Final submission of Computers, Environment and Urban Systems 33 (2009) 325–333

Feature Pruning by Upstream Drainage Area to Support Automated Generalization of the United States National Hydrography Dataset

Lawrence V. Stanislawski [a]

Science Applications International Corporation, Center of Excellence for Geospatial Information Science, United States Geological Survey, 1400 Independence Road, Rolla Missouri, 65401, United States

Abstract

The United States Geological Survey has been researching generalization approaches to enable multiple-scale display and delivery of geographic data. This paper presents automated methods to prune network and polygon features of the United States high-resolution National Hydrography Dataset (NHD) to lower resolutions. Feature-pruning rules, data enrichment, and partitioning are derived from knowledge of surface water, the NHD model, and associated feature specification standards. Relative prominence of network features is estimated from upstream drainage area (UDA). Network and polygon features are pruned by UDA and NHD reach code to achieve a drainage density appropriate for any less detailed map scale. Data partitioning maintains local drainage density variations that characterize the terrain. For demonstration, a 48-subbasin area of 1:24 000-scale NHD was pruned to 1:100 000-scale (100K) and compared to a benchmark, the 100K NHD. The coefficient of line correspondence (CLC) is used to evaluate how well pruned network features match the benchmark network. CLC values of 0.82 and 0.77 result from pruning with and without partitioning, respectively. The number of polygons that remain after pruning is about seven times that of the benchmark, but the area covered by the polygons that remain after pruning is only about 10 percent greater than the area covered by benchmark polygons.

Keywords: automated generalization, hydrographic network, National Hydrography Dataset, directed graph, catchment.

1.0 Introduction

A principal objective of cartographic generalization is reduction of content and detail of geospatial data in a manner that appropriately portrays remaining features at smaller scales. Technology and research have advanced the capacity for cartographic and geospatial database generalization through systems and tools that automate processes using modern database designs, knowledge bases, and artificially intelligent algorithms. Much of the recent (2009) progress is presented or reviewed in the recently published book by the International Cartographic Association [1]. Generalization tools are available that perform specific operations, such as line simplification or smoothing, or polygon collapse or aggregation [2]. Some software systems sequence generalization operators into processes that are suitable for specific data types, such as road networks [3]; however, further development and research are needed to tailor intelligent automated generalization processes that are suitable for primary geospatial data themes operating comprehensively for large regions with diverse conditions, such as the United States.

The U.S. Geological Survey (USGS) vision of *The National Map* is to ensure that "current, complete, consistent, and accurate" geographic base information is readily available through a system of web-based interfaces [4]. With assistance from federal, state, and local data stewards, the

[a] Present address: 1400 Independence Road, Rolla MO, 65401, United States. Tel. +1 573 308 3914, E-mail address: lstan@usgs.gov.

USGS is developing and maintaining eight primary geospatial data themes: transportation, hydrography, boundaries, structures, elevation, land cover, orthographic images, and geographic names [5]. In 2005, the USGS Center of Excellence for Geospatial Information Science (CEGIS) began this generalization project to "research and develop automated methods for generalization to support multiple-scale display and delivery of *The National Map* and other USGS geographic data" [6]. This paper describes ongoing CEGIS research into automated generalization that is focused on the primary hydrography theme of *The National Map*, namely the National Hydrography Dataset (NHD).

The NHD is a comprehensive vector database representing surface-water features of the United States. NHD features have been compiled from several scales and types of USGS digital data and other vector hydrographic data sources. Synoptic coverage is available at two primary levels of detail--the high-resolution (originally compiled from 1:24 000-scale (24K) source data, and 1:63 360-scale in Alaska) and 1:100 000-scale (100K) layers. Additional resolutions are needed to support the USGS automated cartographic mapping and web services. In addition, through ongoing state-level densification efforts, the high-resolution layer is becoming a multi-scale layer of the most accurate hydrographic data for the country that, in places, does not provide a consistent scale representation adequate for cartographic mapping.

The problem is further complicated by the range of surface hydrographic conditions (wet and dry years) and various map compilation standards that existed during the many years that USGS topographic map sheets were compiled, dating back to the 1940's [7].The primary source material for both resolutions of the NHD was the USGS 24K topographic map series. Although standard topographic instructions were available, generalization of features during field collection and subsequently by cartographers to produce the 24K and 100K maps was a difficult task, and the results were not always consistent.

This project develops and tests an automated approach to prune the high-resolution NHD to a consistent level of detail, where needed, and to prune it to other levels of detail that are appropriate for mapping at smaller scales. Pruning is used in this text because it typically refers to the removal of less prominent features, inherently involving a prominence rating, which better describes the approach than similar terms such as thinning, elimination, or selection. The pruning process also must retain topological and data model integrity so that pruned data will function with existing applications. Successful implementation of automated pruning will enhance the USGS NHD program through optimized database maintenance and automation of a fully integrated multiple representation database; both common goals of data generalization and multiple representation databases [1, 8, 9, 10].

With respect to this project, automated generalization has been divided into two primary development tasks—feature pruning and simplification. This article focuses on feature pruning. The next section describes the NHD database to provide perspective on the problem and explain the constraints applied in the selected pruning strategy.

2.0 The NHD Database

The NHD is stored in an ESRI geodatabase model format within an Oracle database. Features in each resolution are separated into five feature classes--NHDArea, NHDFlowline (flowline), NHDWaterbody (waterbody), NHDLine, and NHDPoint—each containing a subset of NHD feature

types represented with the same geometrical shape type. All polygonal features are stored in the waterbody and NHDArea features classes.

The flow line feature class contains features of type artificial path, canal/ditch, coastline, connector, pipeline, and stream/river, which are each represented with a single-part polyline shape type. An artificial path represents a flow path through a polygonal water feature that is connected to other flowline features, and a connector represents a path where surface flow is known to exist, but was not included in the source material (fig. 1). The NHDArea and NHDWaterbody feature classes contain single-part polygon features of various types. All waterbody feature types may have artificial paths passing through them, whereas only three NHDArea types can have artificial paths passing through them. As of January 2008, the high-resolution NHD layer contained more than 27 million features, nearly 20 million of which are flowline features.

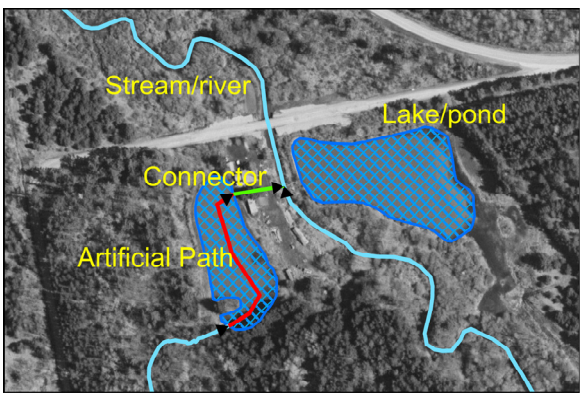

Figure 1: Artificial path, connector, and stream/river features over aerial photo.

The NHD includes a set of surface-water reaches delineated on the vector data. Each reach consists of a significant segment of surface water having similar hydrologic characteristics, such as a stretch of river between two confluences, a lake, or a pond [11]. A unique address, called a reach code, is assigned to each reach. All flowline features receive a reach code address, as well as all lake/pond and reservoir features of the waterbody feature class. Reach addresses and the associated linear referencing system enable the linking of ancillary data to specific features and locations on the NHD; consequently, reach codes are maintained by conflation to new feature representations when acquired [11]. Connected features of compatible feature type can share the same reach code. Likewise, a reach code on the flowline feature class may extend over several confluences because the reach code was conflated from a lower resolution layer.

Flowline features in the NHD are oriented, where possible, in the direction of surface-water flow, and the direction is recorded in a feature attribute. Approximately 94 percent of all high-resolution flowline features have been oriented and assigned flow direction. The structure of the flowline feature class furnishes a drainage network representing water flow over the terrain, which may be referred to as a hydrographic network. Topological connectivity of the flowline network enables the formation of a directed graph [12, 13], which can be used for various analysis functions.

3.0 Methodology: Automated Pruning of the High-Resolution NHD

Pruning, or the initial process of selecting source objects and attributes to be represented in a generalized dataset, is common in generalization strategies [1, 2, 14, 15]. In this paper, pruning the high-resolution NHD consists of eliminating relatively less prominent features, and it is completed in two steps: network pruning and polygon pruning. This paper only discusses processing and analysis of the flowline, waterbody, and NHDArea feature classes, which contain most of the NHD features.

3.1 Network pruning

High-resolution network pruning is completed for a partition, which refers to a watershed or part of a watershed. The first step of the pruning process removes over- or under-passing network features and features without flow direction. These features primarily consist of canal\ditch or pipeline types. The remaining network features in a partition are pruned until a predetermined drainage density is achieved, where drainage density is the ratio of the length of all network features in the partition to the area of the partition. Less prominent network features are identified and pruned based on the relative extent of land surface that flows into the network features.

Before pruning, data enrichment, or preprocessing, assigns a catchment area and upstream drainage area (UDA) estimate to each network feature. Enriching a data layer to prepare it for automated generalization is common practice [3, 16, 17, 18]. The catchment for a network feature is the area of the watershed that drains into the feature. A rapid approach that sums the area of Thiessen polygons derived for evenly spaced points on each flowline is used to estimate a catchment area for each flowline feature (fig. 2) [19]. Subsequently, an augmented directed graph is used to assign UDA estimates to each flowline [20].

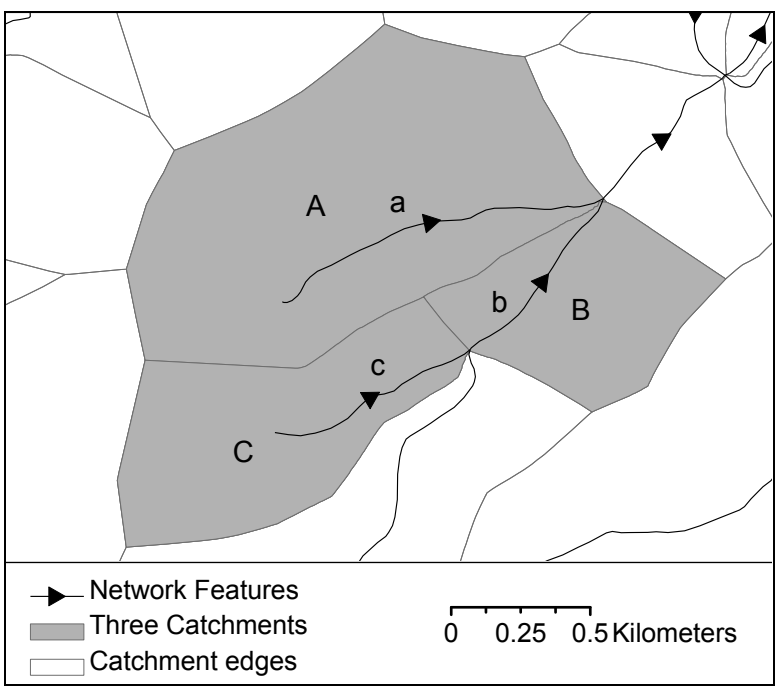

Figure 2: Three catchments labeled A, B, and C derived through Thiessen polygons for network features labeled a, b, and c, respectively.

The pruning process iteratively eliminates network features that drain a minimum UDA, which increases with each iteration until the desired density is achieved for an area. The augmented directed graph generates monotonically increasing values with downstream location on the network, which enable pruning without generating false breaks in the pruned network. At convergences, the augmented graph avoids multiple additions of upstream values, which would improperly magnify the prominence of features in braided areas. An additional constraint applied during network pruning is that complete reaches are extracted to maintain the integrity of the generalized dataset and any links to associated data.

Selecting relative prominence of network features by UDA follows the same logic as the Pfafstetter system for topologically coding river basins and networks [21]. Others have assigned a perceived level of prominence to hydrographic network features for generalization [22]. Perceptual prominence levels assume mapped features are collected in a consistent manner, which cannot be assumed for the high-resolution layer of the NHD. Pruning NHD network features by UDA and reach code is similar to perceptual grouping or "stroke" building [3, 16, 22, 23], but with different emphasis. UDA is the most significant factor for estimating stream flow in the National Flood Frequency Program [24]; hence, UDA-based selection puts greater emphasis on the relative function of a feature rather than on perceived groupings. Assigning prominence by UDA is a feasible and perhaps better alternative for the NHD, although probably more costly to implement. More precise estimates of UDA and flow volumes that are planned for the future should enhance the high-resolution NHD layer [25] and further refine this pruning process.

3.1.1 Partitioning

The NHD is subdivided into region, subregion, and subbasin watershed areas. An evaluation of stream networks mapped at four scales within three NHD subregions identified reliable linear relations between drainage density and map scales ranging from 1:24 000 to 1:2 000 000 [26]. The relation for the Gasconade-Osage subregion is displayed in figure 3. Pruning may be guided by such regression equations; but, pruning to a single density tends to homogenize density and masks natural density variations that help characterize the terrain. A watershed can be partitioned into density classes to alleviate this issue, and each density partition can be separately pruned to an appropriate density. Others have partitioned large datasets into smaller sets for similar reasons [27, 28].

For this study, partitions are generated from catchments estimated for network features. An initial density class is assigned to each catchment based on its size. For instance, the highest density class is assigned to catchments with relatively smaller areas. Edges between catchments of the same class are removed to form clusters with similar densities. Density clusters are constrained to a minimum cluster size, and, therefore, any cluster that is too small receives the class of an adjacent cluster that is the most similar in size and not too small. Edges between clusters with the same class are removed, and the process repeats until all clusters are larger than the minimum size.

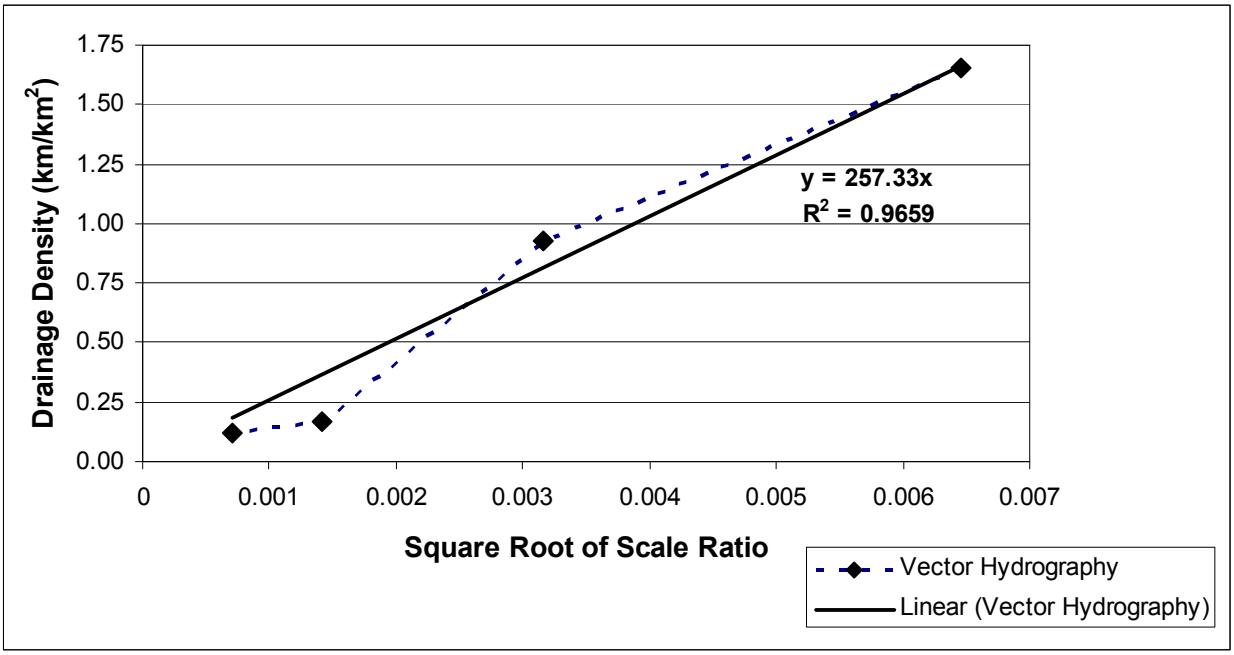

Figure 3: Relation between drainage density and the square root of map scale ratio of source hydrographic features in the Gasconade-Osage subregion watershed of the United States.

3.2.1 Polygon pruning

After network pruning, waterbody and NHDArea polygon features are pruned through feature-type rules based on overlapping conditions and minimum-area criteria derived from the high-resolution NHD standards [29]. In short, a waterbody polygon that contains one or more artificial paths, which have all been pruned, is removed, unless it is a relatively prominent feature. Relatively prominent waterbodies are larger than 6.4516 square centimeters (cm^2) (i.e., 1 in^2) at the generalized map scale. Waterbodies without artificial paths are pruned if smaller than the minimum size at the generalized scale for the associated feature type. NHDArea polygons are pruned with similar rules, but without a condition for relatively prominent features. After waterbody and NHDArea polygons are pruned, a secondary rule removes NHDArea features that must overlap polygons pruned in the previous process. These rules ensure consistency among the remaining network and polygon features.

Network pruning by UDA and associated polygon pruning is an attempt to apply a holistic solution for automated database generalization using knowledge of surface-water drainage and cartography and how it is topologically encoded in the database. Muller recommends the use of "knowledge-based tools to support automated solutions" for generalization [30]. The approach could be classified under the Brassel-Weibel conceptual model [14]. In knowledge-based systems, the knowledge base and inference mechanism are separated [30, 31]. In our case, the knowledge base consists of UDA values and the reach composition of network features in the database and the size, feature type, and network association of polygon features. The inference engine is the partitioning and pruning rules applied through programs.

4.0 Experimentation: Pilot Project

Methods were tested through a pilot project and automated through Arc Macro Language (AML) and Python programs with associated geo-processing functions.

4.1 Test data

Five subregions of high-resolution NHD data near the center of the United States were processed for the pilot project. In this area, the high-resolution NHD was compiled from 1:24 000-scale source data and will be referred to as 24K, for brevity. The Arkansas and Cimarron Rivers are the primary rivers draining these subregions. The study area includes 48 subbasins (fig. 4) covering about 192 670 (square kilometers) km^2.

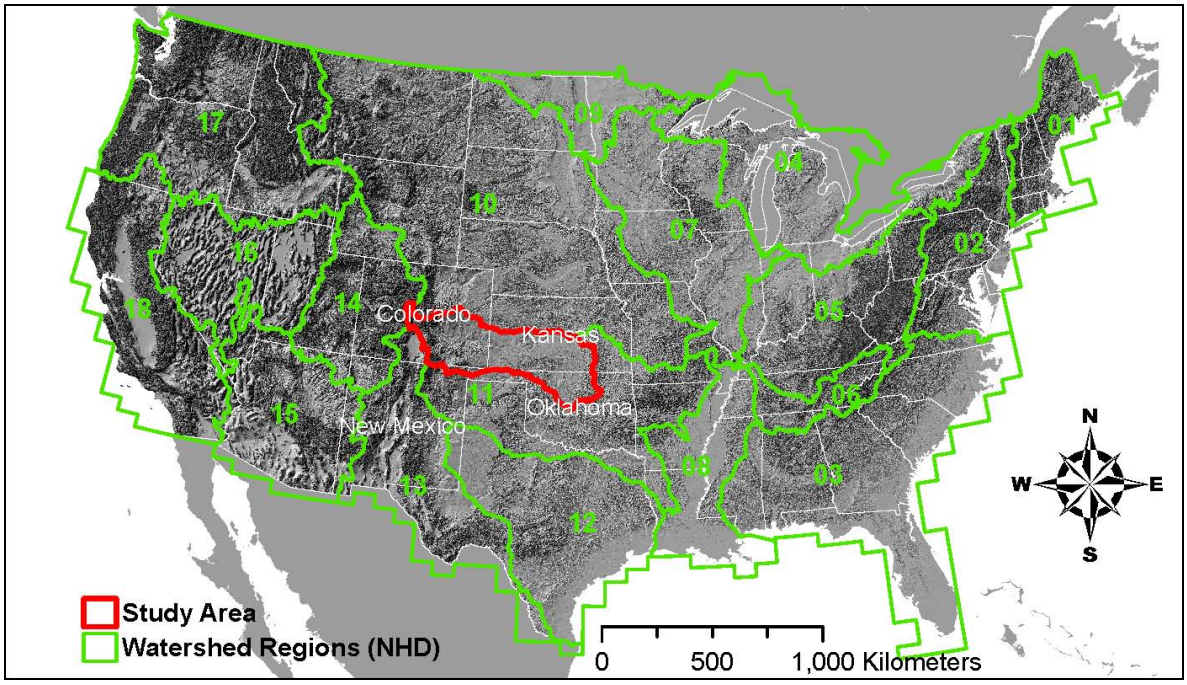

Figure 4: Pilot study area (outline in thick black line) in central United States containing 48 NHD subbasins over hill-shaded terrain model, with NHD regions in thick white lines and numbered, and state boundaries in thin white lines.

After removing all non-flow-directed and over- or under-passing features, a total of 294 607 24K network features remain for subsequent processes. These features total about 215 990 km in length and have a drainage density of 1.121 km/km^2. A single inflow to the 24K network features exists for this study area. An inflowing estimate of 34 856 km^2 from 100K NHDPlus attributes [25] was applied to 24K inflow, and the 24K network features were enriched with catchment area and UDA estimates. The maximum UDA on the 24K features is 209 030 km^2.

4.2 Pruning 1:24 000-scale NHD to 100 000-scale

The enriched 24K NHD features were pruned to 100K. The 100K NHD was used as a benchmark dataset for comparison; therefore, target densities for network pruning were derived from the 100K network features. Target densities from the 100K features are expanded by 7.5 percent to account for extra detail, or granularity, included in 24K network representations. A 100K-to-24K length expansion factor was estimated from 50 matching confluence-to-confluence sections between the

24K and 100K networks. The expansion factor of 1.075 was determined as the average ratio of the 24K to 100K lengths for the 50 matching sections, which are distributed over the study area.

Pruning was tested with and without partitioning. The 100K includes 129 330 km of network features. Without partitioning, the 100K target density derived after applying the 100K-to-24K expansion factor is 0.7216 km/km^2.

For the partitioned case, the study area was partitioned into high and low density partitions from NHDPlus catchments for 100K network features. The NHDPlus catchments originally were derived from an elevation model [25]. Initially, 100K catchments less than 9 km^2 were placed in the high density class, and remaining catchments were assigned low density. A minimum cluster size of 400 km^2 generated two partitions. The resulting high and low density partitions have 16 and 8 clusters, respectively, with each partition covering about one-half of the study area (fig. 5). Applying the expansion factor, target densities determined from the 100K network features falling in each partition are 1.0309 and 0.4106 km/km^2 for the high and low density partitions, respectively. Each 24K network reach was assigned the density class of the partition it mostly falls in. The high- and low-density sets of 24K features were pruned separately to the associated 100K densities.

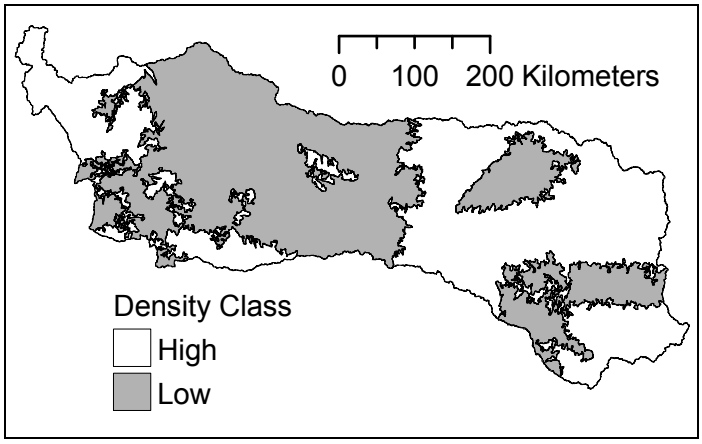

Figure 5: Density partitions for study area determined from 1:100 000-scale catchments.

In addition, the 24K features were pruned to 100K using the same partitions, but target densities were determined from flow-directed 100K network features only. This case helps assess the effect of local high concentrations of 100K network features without flow direction on pruning results.

4.3 Summary statistics and benchmark comparisons

Summary statistics were computed for network and polygon features in the study area for the 24K NHD before and after pruning, and from the 100K NHD benchmark. Quantities that compare the pruned 24K to the 100K benchmark features are described in section 4.3.3.

4.3.1 Network statistics

Network statistics are total network length, average segment length, number of confluence-to-confluence sections, confluence-to-confluence ratio with 24K NHD, density, and percent of study area covered by mapped network lines. Average segment length is the average distance between vertices of all network features, and it equals the network length divided by the number of segments in the network. Confluence-to-confluence ratio with 24K NHD is the ratio of the number of confluence-to-confluence sections in a network to the number of confluence-to-confluence sections

in the 24K network. The percent of study area covered by mapped network lines is computed based on a 0.0203 centimeters (cm; 0.008 inch) line-weight used for the USGS primary series quadrangle maps [32], and it equals $100*w*d*L/A$, where w is the line-weight, d is the scale denominator, L is the length of the network, and A is the study area in km^2.

4.3.2 Polygon statistics

Polygon statistics are number of polygons, number-of-polygons ratio with the 24K NHD, total area of all polygons, percent of study area covered by all polygons, and percent of study area covered by line and polygon features. The number-of-polygons ratio with the 24K NHD is the ratio of the number of polygons in the data set to the number of polygons in the 24K NHD. The percent of the study area covered by line and polygon features equals the percent of the study area covered by polygon features plus the percent of the study area covered by line features, as defined in the previous section.

4.3.3 Network comparison statistics

The coefficient of line correspondence (CLC) estimates how well two sets of lines, representing similar features on the ground, overlap each other. CLC is similar to the coefficient of area correspondence described by Taylor [33]. CLC is the ratio $M/(O+C+M)$, where M is the sum of the lengths of matching benchmark lines, O is the sum of the lengths of benchmark lines that are omitted from the test data set, and C is the sum of the length of test lines that do not have a match in the benchmark data set (commission errors), which is divided by the 100K-to-24K length expansion factor. Reducing test line lengths in C by the expansion factor puts all values on a common scale. The proportions of commission and omission errors equal $C/(O+C+M)$ and $O/(O+C+M)$, respectively. Aside from values for the entire study area, a separate CLC value was computed for each cell of a half-degree, latitude-longitude grid covering the study area, which has 114 cells.

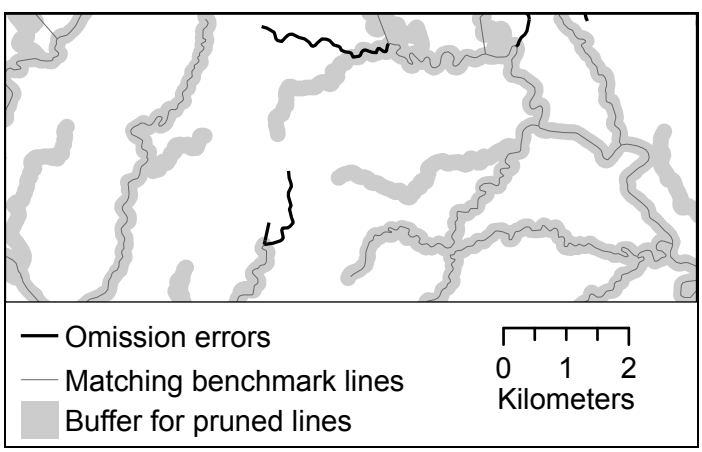

Figure 6: Example of omission errors and matching lines on the benchmark network. Most of the confluence-to-confluence section of an omission error falls outside the buffer for pruned network.

Omission errors are estimated by generating a buffer around the pruned network features and identifying confluence-to-confluence sections of the benchmark network that fall mostly outside the buffer (fig. 6). Matching lines are confluence-to-confluence sections of the benchmark network that fall mostly inside the buffer. Commission errors are estimated by generating a buffer around the benchmark lines and identifying confluence-to-confluence sections of the pruned network that fall mostly outside the benchmark buffer. The buffer size is the combined horizontal positional accuracy estimates for each network. Positional accuracy for each network is estimated as twice the tolerance

for well-defined points from the U.S. National Map Accuracy Standards at the associated scales. The tolerance for well-defined points at 24K and 100K is 1/50th of an inch [34], so the buffer size is 2/50th inch at 24K plus 2/50th inch at 100K, which equals about 126 m on the ground.

5.0 Results

Results of pruning network and polygons features from 24K to 100K with partitioning are illustrated in figure 7. The obvious density change along quadrangle boundaries in the 100K is somewhat maintained by partitioned pruning, but it is not as pronounced as in the 100K.

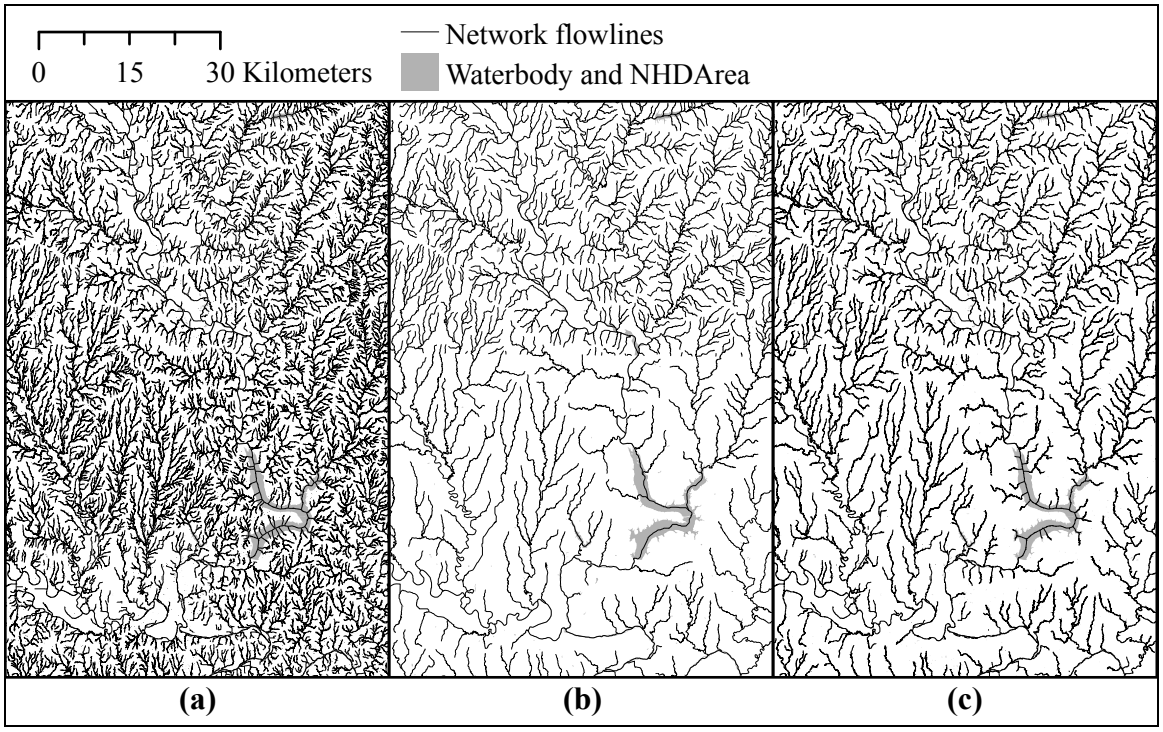

Figure 7: Comparison of (a) original 1:24 000-scale (24K) NHD flowline, waterbody, and area features with (b) the same feature classes from the 1:100 000-scale (100K) NHD benchmark, and (c) the 24K features pruned to 100K with partitioning.

5.1 Network pruning results

Without partitioning, pruning 24K network reaches having a UDA less than 1.8083 km^2 produces the target 100K density of 0.7216 km/km^2. With partitioning, the 24K reaches falling mostly within the high density partition have a density of 1.4980 km/km^2, and those falling mostly within the low density partition have a density of 0.7420 km/km^2. Pruning the 24K high- and low-density reaches having a UDA less than 1.4259 and 3.5332 km/km^2, respectively, produced the 100K densities required for each partition.

Network summary statistics are shown in table 1. Average length of line segments, or the average distance between vertices, in the 24K network is about one fourth the average of the 100K line segments. The general rule of Töpfer's Radical Law [35] suggests the ratio of the number of objects in two maps should equal the square root of the ratio of the map scales [36], but for linear features, and the number of points on linear features, an exponent of two should be applied to the ratio [37], so the Radical Law suggests the ratio of the number of 24K to 100K line features, or the ratio of the

number of points (or line segments) representing the same line features at 24K and 100K should be about 4. The ratio between the number of segments in the pruned and 100K networks is about 4.

Pruning with partitioning provides about 1 percent more confluence-to-confluence sections than pruning without partitioning (table 1). The ratio of the number of confluence-to-confluence sections between pruned and un-pruned 24K network features is about 3.3. The ratio between 100K and un-pruned 24K network features is about 4, which is about the ratio expected from Töpfer's Radical Law for lines, so the pruned networks may have a few more confluence-to-confluence sections than expected.

Table 1: Network summary statistics for the source 24K NHD, 24K NHD pruned to 100K using one density partition, 24K NHD pruned to 100K by two density partitions, and the 100K NHD benchmark. [24K, 1:24 000-scale; NHD, National Hydrography Dataset; 100K, 1:100 000-scale; km, kilometers; m, meters; km^2, square kilometers]

Network data set	Total network length (km)	Average segment length (m)	Number of confluence-to-confluence sections	Confluence-to-confluence ratio with 24K NHD	Density (km/km^2)
24K NHD	215 995	17.2	185 037	1.0	1.1211
24K NHD pruned to 100K by one density	139 023	17.3	55 734	3.3	0.7216
24K NHD pruned to 100K by two densities	139 025	17.3	56 357	3.3	0.7216
100K NHD benchmark	129 328	63.3	45 754	4.0	0.6712

5.1.1 Short tributaries

A larger number of short tributaries appear to be included in the pruned 24K network than in the 100K (fig. 7). Data standards for 100K NHD and 100K USGS topographic maps indicate that, in non-arid regions, stream/river features larger than 1.60 cm should be included in the 100K NHD, but shorter streams may be included if they flow from a two-dimensional (polygon) surface-water feature, or if the stream is more than 0.061 cm wide [38, 39]. The benchmark, 100K NHD network includes 1 704 tributaries that are less than 1.6 km (1.6 cm at 100K) and not flowing from surface-water polygons in the study area, and these tributaries average 1.1 km. But the 24K network features pruned to 100K using partitions includes 4 401 of this type of tributary, averaging about 1 km in length.

5.1.2 Small disconnected sub-networks

Earlier results indicate that small, disconnected sub-networks may be maintained after pruning by UDA [40]. The 100K NHD network includes 229 sub-networks that are less than 1.6 km long totaling about 221 km, or less than 1 km per sub-network. The 24K network pruned to 100K using partitions includes 246 small sub-networks totaling about 193 km, or about 0.79 km per sub-network.

5.1.3 Coefficient of line correspondence

For the entire study area, about 83 percent of the 45 754 confluence-to-confluence sections of the benchmark network match the 24K network pruned to 100K without partitioning. Matching benchmark features total 112 308 km. Matching features in the network pruned without partitioning compose about 83 percent of the 55 734 confluence-to-confluence sections of the pruned network. About 17 020 km of the benchmark features are omitted from the pruned network without partitioning, and about 17 919 km of features in this pruned network are commission errors. These numbers produce a CLC of 0.77 for the study area, with proportions of omission and commission errors at about 0.12 for network pruning without partitioning.

With partitioning, about 87 percent of the confluence-to-confluence sections in the 100K benchmark network match the features in the pruned network. Matching benchmark features total 116 319 km, whereas 13 109 km of the benchmark features are omitted from the pruned network. About 86 percent of the 56 357 confluence-to-confluence section in the pruned network match the benchmark features, and 13 762 km of the pruned features are commission errors, so pruning with partitioning produces a CLC of 0.82 for the study area, with proportions of about 0.09 for omission and commission errors.

CLC values determined for each 0.5-by-0.5 degree cell of the 110-cell grid covering the study area are shown in table 2. Four cells of the grid that do not include at least one confluence-to-confluence section from either network were removed. Partitioning improves the mean CLC value for the 110 grid cells by 0.05, or about 6 percent. A few cells of the 100K benchmark network include a large proportion of features without flow-direction, which are counted as omission errors because only flow-directed features are included in pruned networks. Removing flow-directed benchmark features from the analysis does not affect the mean CLC over all cells, but it increases the minimum CLC value by about 18 percent from 0.40 to 0.47.

Table 2: Summary for 0.5-by-0.5 degree grid (110 cells) of CLC values comparing benchmark features to 24K NHD network features pruned to 100K using one and two density partitions. CLC also is shown comparing only flow-directed benchmark features to 24K NHD network pruned to 100K through the two density partition method.
[CLC, coefficient of line correspondence; benchmark, 1:100 000-scale network; 24K, 1:24 000-scale, NHD, National Hydrography Dataset; 100K, 1:100 000-scale]

Network pruning method	Mean CLC	Standard deviation	Minimum	Maximum
One density	0.75	0.099	0.40	0.93
Two densities	0.80	0.095	0.40	0.93
Two densities, only flow-directed benchmark features	0.80	0.081	0.47	0.93

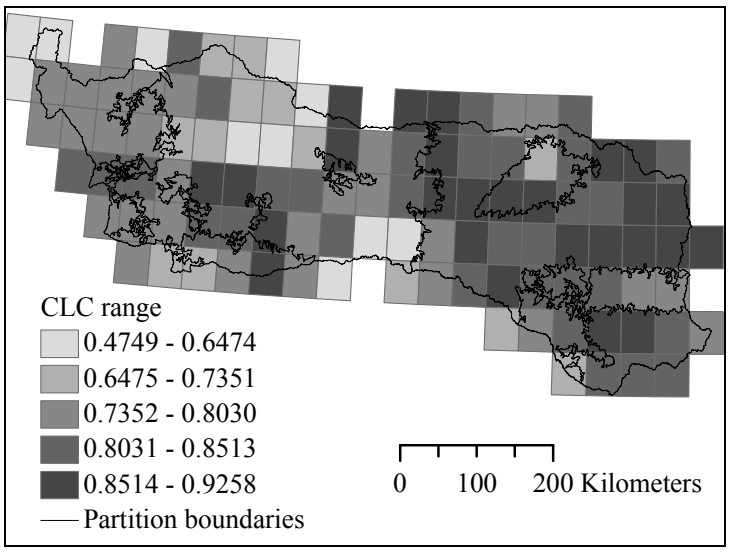

Figure 8: Distribution of the coefficient of line correspondence (CLC) on a 0.5-by-0.5 degree grid, and partition boundaries. CLC values compare the 1:24 000-scale network pruned in partitions to 1:100 000-scale (100K) with the flow-directed features of the 100K NHD benchmark network.

The distribution of CLC values over the study area for partitioned pruning is shown in Figure 8. Relatively smaller CLC values identify vicinities where larger proportions of mismatching features exist. For instance, figure 9 displays a cell in the south central part of the study area where a relatively low CLC (0.54) is caused by a large proportion of omission errors. In figure 9, a group of high-density catchments within the flood plain are clustered in the low-density partition because they are surrounded by larger, low-density catchments; subsequently, this group of high-density features was pruned to the lower density.

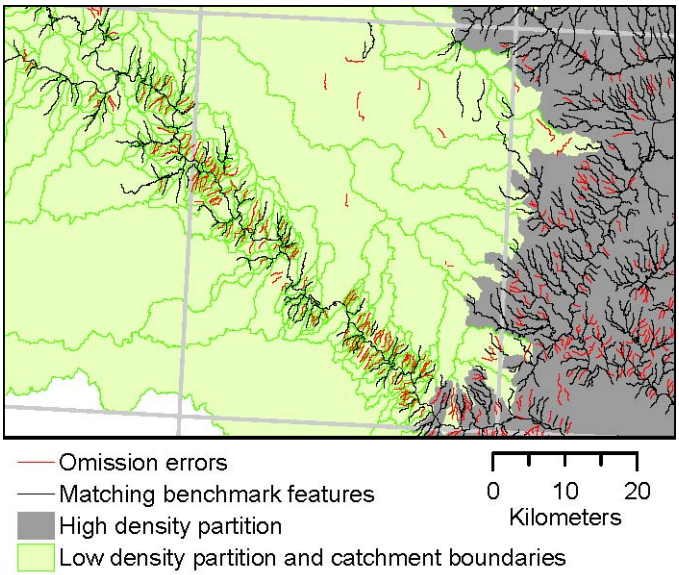

Figure 9: A local high-density area was forced into the low-density partition during clustering. Subsequent pruning of this area to a low density causes a relatively large number of omission errors and a low CLC cell value (0.54).

5.2 Polygon pruning results

A summary of polygon pruning results are presented in table 3. Töpfer's Radical Law for polygons should include an exponent of three [37], suggesting the ratio of the number of polygons in the 24K and 100K should be about 8.5. Ratios in table 3 indicate the benchmark includes much fewer polygons than expected from the Radical Law, and the pruned datasets appear to have more than twice the number of polygons as expected. According to Dutton [37], exponents to Töpfer's Law can vary based on object type and map purpose, and a constant also may be applied. Furthermore, values in table 3 were derived for the benchmark dataset after feature simplification, whereas the pruned datasets were not simplified. Amalgamation should reduce the number of polygons in the pruned dataset, making the ratio comparable.

Table 3: Polygon summary statistics for the source 24K NHD, 24K NHD pruned to 100K using one density partition, 24K NHD pruned to 100K by two density partitions, and the benchmark 100K NHD. [24K, 1:24 000-scale; NHD, National Hydrography Dataset; 100K, 1:100 000-scale; km^2, square kilometers]

Polygon data set	Number of polygons	Number-of-polygons ratio with 24K NHD	Area of all polygons (km^2)	Percent of study area in polygons	Percent of study area covered by lines and polygons
24K NHD	118 435	1.0	1 633.4	0.85	1.39
24K NHD pruned to 100K by one density	37 079	3.2	1 393.1	0.72	2.19
24K NHD pruned to 100K by two densities	36 870	3.2	1 393.9	0.72	2.19
Benchmark 100K NHD	5 868	20.2	1 262.9	0.66	2.02

Estimates of the percent of a map that would be covered by hydrography symbols using the various datasets are shown Table 3. Pruned polygons cover about 10 percent more space than benchmark polygons. Pruned lines cover precisely 7.5 percent more space than benchmark lines because of the applied linear expansion factor. So, a goal for subsequent feature simplification processing could be to reduce polygon space by 10 percent, and line space by 7.5 percent.

6.0 Summary

Network pruning produces slightly more confluence-to-confluence sections than expected when compared to the 100K benchmark or through the Radical Law. More headwater tributaries and small disconnected sub-networks exist in pruned networks than in the benchmark, but many of these may be legitimate given 100K feature specifications. A reasonable pruning enhancement would be to remove extremely short headwater tributaries and tributaries composed of artificial path inside of large polygons, except for the extension of the main path.

The CLC is a good measure of omission and commission errors between networks, and the distributed CLC helps isolate areas where network matching is relatively poor. The CLC estimate assumes each line set has a consistent level of accuracy. CLC may be enhanced through a variable-sized buffer based on feature type or some other attribute associated with positional accuracy. A confidence interval for the mean CLC could strengthen its usefulness.

Partitioning improves the amount of matching features in pruned and benchmark networks by about four percent over non-partitioned pruning. Partitioning could be refined to generate more homogeneous density clusters and better maintain local density variations in pruned networks.

The number of polygons remaining after pruning seems too large in comparisons to benchmark polygons, and based on the Radical Law. However, the map space covered by hydrography polygons is only slightly more than the benchmark, and some of this difference should be removed through subsequent polygon simplification processes. Development of an indicator similar to the CLC could better assess the adequacy of the polygon pruning.

In the pilot project, partitioning and pruning were tailored for the benchmark dataset. Implementation of pruning on the high-resolution NHD requires density partitions for 24K catchments and a method to associate target densities to each partition, which may be estimated through regression or some form of Töpfer's Radical Law.

7.0 Conclusion

This paper presents automated methods to prune flowline and polygon features of the United States high-resolution NHD to support the generalization needs of the USGS. Network and connected polygon features are pruned by UDA and reach codes until a desired drainage density is achieved. Standards-based rules guide remaining pruning operations, and local density variations that characterize terrain and climate conditions are maintained through data partitioning. UDA-based selection associates the prominence of a feature with its relative function, and it performs well on multi-scale representations, which exist in the high-resolution NHD. In comparison, perceptual-based prominence estimates, such as stream order, assume consistent source feature representations and are not reliable for multi-scale data.

Results indicate network pruning achieved very good results, with about 83 percent of the benchmark network features matching those of the pruned network. Partitioned pruning improved results by about 4 percent. CLC values of 0.82 and 0.77 resulted from pruning with and without partitioning, respectively. The number of polygons remaining after pruning is about seven times that of the benchmark, but the amount of area covered by the pruned polygons is only about 10 percent greater than the benchmark. Simplification operations, particularly amalgamation should somewhat compensate for the large difference in polygon numbers.

Future enhancements include revising partitioning, removing inappropriate small network features during pruning, and enhancing CLC estimation, along with developing a similar process for polygon features. The USGS and cooperators will evaluate the approach over the range of geographic conditions in the country through web-distributed tools. Subsequently, the process will be implemented for the NHD, and similar generalization methods will be applied on other vector themes of *The National Map*.

Role of Funding Source
This project is funded through the USGS CEGIS and supported by the USGS National Geospatial Technical Operations Center (NGTOC). The research is being conducted by CEGIS to support *The National Map* Program within NGTOC. Mike Finn is the CEGIS principal investigator.

Acknowledgments
Thanks to E. Lynn Usery and Mike P. Finn of CEGIS, SAIC developer Jess Janssen, former USGS employee Pat Turley, former USGS NHD Program director, Keven Roth, the many contributors to the NHD Stewardship program, the NHDPlus developers, particularly Cindy McKay, and Kathy Laub for editorial assistance.

References Cited

[1] W.A. Mackaness, A. Ruas, L.T. Sarjakoski (Eds.), Generalization of Geographic Information: Cartographic Modelling and Applications, Elsevier for International Cartographic Association, 2007.

[2] N. Regnauld, R.B. McMaster, A synoptic view of generalization operators, 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. 37-66.

[3] G. Touya, A road network selection process based on data enrichment and structure detection, 10th ICA Workshop on Generalization and Multiple Representation: Moscow, August 2-3, 2007.

[4] USGS, The National Map, The Nation's Topographic Map for the 21st Century, United States Geological Survey, online: http://nationalmap.gov/index.html, 2006.

[5] USGS, Implementation Plan for The National Map, (version 1.0), United States Geological Survey, online: http://nationalmap.gov/report/Implementation_Plan_ver_1.0.pdf, 2003.

[6] G. McMahon, S.P. Benjamin, K. Clarke, J.E. Findley, R.N. Fisher, W.L. Graf, L.C. Gundersen, J.W. Jones, T.R. Loveland, K.S. Roth, E.L. Usery, N.J. Wood, Geography for a Changing World – A Science Strategy for the Geographic Research of the U.S. Geological Survey, 2005-2015, U.S. Geological Survey Circular 1281, 2005.

[7] USGS, Map Publication Scales, United States Geological Survey, Book 1, Part B, Chapter 1 of the Geological Survey Topographic Instructions, March 1955.

[8] W.A. Mackaness, Automated cartography in a bush of ghosts, Cartography and Geographic Information Systems 33(4) (2006) 245-256.

[9] O. Chaudry, W.A. Mackaness, Modeling geographic phenomena at multiple levels of detail, AutoCarto 2006: Vancouver, WA, June 26-28, 2006.

[10] S. Mustiêre, J. van Smaalen, Database requirements for generalization and multiple representation, 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. 113-136.

[11] USGS, The National Hydrography Dataset: Concepts and Contents, (February 2000), United States Geological Survey, online: http://nhd.usgs.gov/chapter1/index.html, 2000.

[12] D. McCracken, W. Salmon, A Second Course in Computer Science with Modula-2. New York, New York: John Wiley and Sons, 1987.

[13] U. Manber, Introduction to Algorithms: A Creative Approach. Addison-Wesley Publishing Co., 1989.

[14] R. McMaster, K. Shea, Generalization in Digital Cartography. Washington, D.C.: Association of American Geographers, 1992.

[15] C. A. Brewer, B.P. Buttenfield, Framing guidelines for multi-scale map design using databases at multiple resolutions, Cartography and Geographic Information Systems 34(1) (2007) 3-15.

[16] R.C. Thomson, R. Brooks, Efficient generalization and abstraction of network data using perceptual grouping", Proceedings of the 5th International Conference on GeoComputation: Greenwich, August 23-25, 2000, online: http://www.geocomputation.org/2000/GC029/Gc029.htm.

[17] J.E. Stoter, Generalization within NMA's in the 21st Century, XXII International Cartographic Conference: Coruña, Spain, July 11-16, 2005.

[18] H. Yan, Z. Li, T. Ai, System for automatic generalization of topographic maps, Chinese Geographical Science 16(2) (2006) 165-170.

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

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

[21] K.L. Verdin, A system for topologically coding global drainage basins and stream networks, 1997 ESRI International GIS User Conference Proceedings, online: http://gis.esri.com/library/userconf/proc97/proc97/to350/pap311/p311.htm, 1997.

[22] R.C. Thomson, R. Brooks, Generalization of geographical networks, 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. 255-267.

[23] O. Chaudhry, W.A. Mackaness, Rural and urban road network generalization deriving 1:250,000 from OS MasterMap, XXII International Cartographic Conference: Coruña, Spain, July 11-16, 2005.

[24] USGS, The National Flood Frequency Program, version 3: A Computer Program for Estimating Magnitude of Flood for Ungaged Sites, Unites States Geological Survey, online: http://pubs.usgs.gov/wri/wri024168/#pdf, 2002.

[25] NHDPlus, NHDPlus User Guide (January 21, 2008), United States Environmental Protection Agency and United States Geological Survey, online: ftp://ftp.horizon-systems.com/NHDPlus/documentation/NHDPLUS_UserGuide.pdf, 2008.

[26] L.V. Stanislawski, M. Starbuck, M.P. Finn, E.L. Usery, Generalization for *The National Map* with emphasis on the National Hydrography Dataset, ESRI International User Conference, San Diego, California, July 24-29, 2005, online (abstract and presentation): http://carto-research.er.usgs.gov/generalization/index.html.

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

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

[29] USEPA, USDOI, Standards for National Hydrography Dataset – High Resolution, United States Environmental Protection Agency and United States Department of the Interior, United States Geological Survey, National Mapping Program Technical Instructions, July 1999, online: http://rockyweb.cr.usgs.gov/nmpstds/acrodocs/draft/dlg-f/nhd/NHDH0799.PDF.

[30] L.T. Sarjakoski, Conceptual models of generalization and multiple representation, 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. 11-35.

[31] R. J. Mockler, D.G. Dologite, Knowledge-Based Systems: An Introduction to Expert Systems, Mackmillan Publishing Company, 1992.

[32] USGS, Part 5: Publication Symbols, Standards for 1:24,000- and 1:25,000-Scale Quadrangle Maps, United States Geological Survey, February 2002.

[33] P.J. Taylor, Quantitative Methods in Geography: An Introduction to Spatial Analysis, Chapter 5: Areal association, Houghton Mifflin, Boston, 1977, 396pp.

[34] U.S. Bureau of the Budget, United States National Map Accuracy Standards, June 17, 1947.

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

[36] B. Jiang, L. Harrie, Cartographic selection using self-organizing maps, 6th ICA Workshop on Generalization and Multiple Representation: Paris, France, April 28-30, 2003.

[37] G. Dutton, Scale, sinuosity, and point selection in digital line generalization, Cartography and Geographic Information Systems 26(1) (1999) 33-53.

[38] USEPA and USDOI, Standards for National Hydrography Dataset, United States Environmental Protection Agency and United States Department of the Interior, United States Geological Survey, National Mapping Program Technical Instructions, July 1999. Online: http://rockyweb.cr.usgs.gov/nmpstds/acrodocs/draft/dlg-f/nhd/NHD0799.PDF, 1999.

[39] USGS, Part 3: Feature Specifications and Compilation, Standards for 1:100,000-Scale Quadrangle Maps, United States Geological Survey, January 1994.

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