Intelligent Pruning of the United States National Hydrography Dataset

Lawrence V. Stanislawski[1]

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 National Hydrography Dataset is a comprehensive vector data model representing surface-water features of the United States. The current (October 2008) National Hydrography Dataset model includes three levels of detail, although not all layers are fully populated. Maintenance and integration of a multi-layered database is a costly endeavor, which could be alleviated through an effective automated database generalization process that furnishes less detailed layers from the most detailed layer available, thereby requiring storage and maintenance of only the most detailed layer of data. With this goal in mind, the U.S. Geological Survey has been working on an automated pruning strategy that eliminates less significant features from the highest resolution layer and furnishes data densities appropriate for any map scale smaller than the scale of the source layer. Implementation of this process on a dataset as large as the United States National Hydrography Dataset has several requirements, which include minimum data integrity standards, quality-assurance procedures, and a data partitioning and ordering scheme. This paper describes an approach for implementing automated pruning for the National Hydrography Dataset, which is based on National Hydrography Dataset reach codes, preprocessing estimates of upstream drainage area, and standard's-based criteria. The approach is demonstrated on a five subregion subset of nearly 300 000 hydrographic network features and more than 118 000 polygon features from the high-resolution layer of the National Hydrography Dataset. Pruning results to three levels of detail are summarized for the pilot project.

Keywords: automated generalization, multiple representation database, hydrographic network, National Hydrography Dataset, augmented directed graph, catchment.

1.0 Introduction

Simplified analysis, display, and integration of geospatial data have been research and development goals of cartographic and geospatial data generalization for many years. Technology and research have advanced our 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 (2008) progress is presented or reviewed in the recently published book by the International Cartographic Association [1]. Although there are tools available that perform specific generalization operations such as Environmental Research Systems Institute (ESRI)'s Generalization Toolbox, and some systems may be suitable for specific data types, such as Clarity GIS from 1Spatial with road networks [2], further research is needed to tailor intelligent automated generalization processes that are suitable for primary geospatial data themes having comprehensive national coverage [3].

During recent (2003-present) years, the U.S. Geological Survey (USGS) has been remodeling the way it provides geospatial data to the United States through *The National Map* program [4]. The

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

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]. *The National Map* data will be derived from various sources by a consortium of data stewards. To coordinate these efforts, the 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]. More recently, the U.S. National Research Council (NRC) recommended development of "unique generalization operations that can be automated for the many possible data types and map scales" associated with *The National Map* as a priority research topic for CEGIS in the area of data integration [7]. This paper describes ongoing CEGIS research into automated generalization, focusing 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 of USGS digital data and other vector hydrographic data sources. The NHD is a loosely integrated, seamless, multiple representation database that includes three levels of detail--medium-resolution (1:100 000-scale source data), high-resolution (1:63 360-scale to 1:20 000-scale source data), and local resolution (1:12 000-scale or larger scale source data); local resolution is available only in a few areas as of June 2008. Database maintenance activities do not propagate feature modifications between layers; however, NHD database layers were compiled in order, from medium to local resolution, to allow conflation of reach addresses [8] and geographic names [9] from lower to higher resolution layers. For logistical reasons, maintenance efforts are focused on updating and densifying the high-resolution layer, which includes incorporating features from the local resolution layer to the high-resolution layer. Through current (2008) efforts on the high-resolution layer, the NHD will evolve into a single, multi-resolution layer compiled from 1:63 360 or larger scale source data. Optional smaller-scale resolutions will be derived through automated generalization of the remaining most accurate, multi-resolution layer. Towards that end, a primary goal of this project has been the development and implementation of an automated database generalization process that can render lower resolution NHD layers from the most accurate, high-resolution layer. Successful implementation of this process will enhance the USGS NHD Program through optimized database maintenance and automation of a fully integrated multiple representation database, which are common goals of data generalization and multiple representation databases [1, 10, 11, 12].

Regarding the NHD, automated generalization has been divided into two primary development tasks—feature pruning and simplification. This article focuses on feature pruning. The following sections describe and demonstrate a knowledge-based automated pruning strategy that eliminates less significant hydrographic features from the highest resolution NHD layer, thereby furnishing data densities appropriate for any map scale smaller than the scale of the source layer.

2.0 National Hydrography Dataset

The physical database of the NHD is stored in an ESRI geodatabase model format within an Oracle database that is maintained and distributed by the USGS. Development of the NHD has been a cooperative effort by the USGS, U.S. Environmental Protection Agency (USEPA), U.S. Department of Agriculture Forest Service, and state and commercial organizations. Features in each resolution of the NHD are separated into five feature classes—NHDArea (area), 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 in the NHD are stored in the waterbody and area features classes. Most NHD features are stored in the flowline, waterbody, and area feature classes. Data handling, analysis, and discussion are limited to these feature classes in this paper.

The flowline 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 (figure 1). The NHDArea and NHDWaterbody feature classes contain single-part polygon features having 20 and 6 feature types, respectively. Certain area and waterbody feature types may overlap other features in these two classes. All waterbody feature types may have artificial paths passing through them, whereas only three area feature types (canal/ditch, stream/river, and wash) 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.

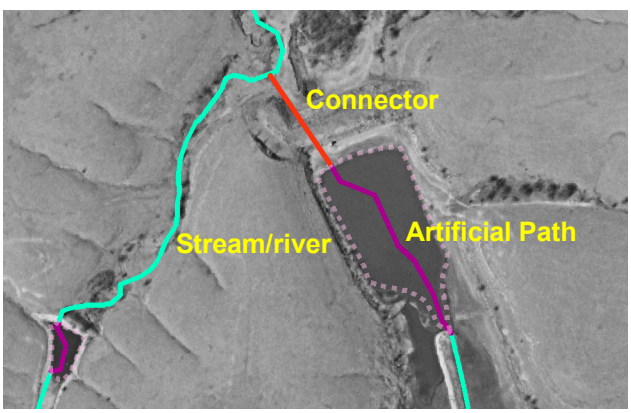


Figure 1: Artificial path, connector, and stream/river features over aerial photo.

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 [8]. 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. On the high-resolution layer, more than 11 million and 6 million reach codes exist on the flowline and waterbody feature classes, respectively. Reach codes are assigned, retired, and conflated between resolutions through a standard system that ensures uniqueness and records a transaction history. Notably, 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. Reach addresses and the associated linear referencing system enable the linking of ancillary data to specific features and locations on the NHD, which explains the need to conflate reach codes to new feature representations when acquired [8].

Flowline features in the NHD are oriented, where possible, in the direction of surface-water flow, and the direction is recorded as “With Digitized” in the associated FlowDir attribute. Approximately 94 percent of all high-resolution flowline features have been oriented and assigned

flow direction. The structure of the flowline feature class within the NHD data model 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 is used to form a directed graph [13, 14] composed primarily of planar components [14]. Subsequently, traversal techniques [14] can be applied on the graph to perform various analysis functions, one of which is accumulating values associated with upstream flowline features. Occasionally, non-planar features that pass over or under other flowline features may exist in the flowline feature class as pipelines or aqueducts.

3.0 Pruning NHD Network and Polygon Features for Automated Generalization

Pruning, or the initial process of selecting source objects and attributes to be represented in a generalized dataset, is common in generalization strategies [1, 3, 15, 16]. 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.

3.1 Network pruning

Network flowline features are pruned until a predetermined drainage density is achieved, where drainage density is the ratio of the length of all features in a drainage network to the area that is drained by the network. The maximum post-pruned drainage density that can be achieved must be less than the density of the source network. Through an evaluation of elevation-derived stream networks and stream networks mapped at four scales within two physiographic regions of the United States, reliable linear relations that estimate an appropriate drainage density for depicting hydrographic networks at map scales ranging from 1:24 000 to 1:2 000 000 were developed through regression [17]. Thus, using the regression equations, the pruning process may be reasonably guided, within the contiguous 48 states, by desired map scale or drainage density.

The network pruning strategy extracts the most prominent network features based on the relative extent of the watershed that flows into the network features. To accomplish this task, catchment area estimates must be acquired for each network feature. The catchment for a flowline 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 [18]. Subsequently, an augmented directed graph approach is used to assign upstream drainage area (UDA) estimates to each flowline [19], which are then used to prune less significant network features. The pruning process iteratively eliminates network features that drain a minimum UDA, which increases with each iteration until the desired density is achieved. The augmented directed graph approach generates monotonically increasing values with downstream location on the network. Monotonically increasing UDA values are needed to properly prune the network without generating false breaks in the pruned network. At convergences, the augmented directed graph approach avoids improperly adding values from upstream divergences multiple times, which would improperly magnify the prominence of features in braided areas. Furthermore, the network pruning process must extract complete reaches to maintain the integrity of the generalized subset and any links to associated data. This restriction is applied during pruning.

Selecting relative prominence of network features by UDA follows the same logic as the Pfafstetter system for topologically coding river basins and networks [20]. Pruning the NHD network features by UDA and reach code is similar to perceptual grouping or “stroke” building [2, 21, 22, 23], but our minimum strokes need not be derived since they already exist as network reach codes. UDA is the most significant factor for estimating stream-flow volumes in the National Flood Frequency

Program [24]. Thus, using UDA as the primary feature selection criteria may be more aptly defined as relative function rather than on a perceptual or contrived ordering scheme.

As described currently (2008), the network pruning process does not fit well within an agent-based framework [25] because the goal achieves a density for the entire database network, or the network within the area of interest. Adding localized constraints, such as subbasin-level density requirements, may be more compatible with the agent approach. The pruning process may be used inside an agent framework as a meso-level algorithm dedicated to a meso-agent "hydrographic network" class; however, it seems more appropriate to classify the automated network pruning strategy as a knowledge-based approach.

3.1.1 Preprocessing for network pruning

Implementing the network pruning process on a dataset as large as the United States has several requirements, which include minimum data integrity standards, quality-assurance, and a data partitioning and ordering scheme. Data integrity issues of network features that affect UDA estimation and pruning include improper feature orientation, improper gaps between features, overlapping features, improper gaps in reach addresses, and branched reaches. NHD data stewards are systematically reviewing the high-resolution data layer for these issues and correcting it as needed. The NHD is subdivided into region, subregion, and subbasin watershed areas. The lower contiguous 48 states comprise 18 regions, 205 subregions, and 2 105 subbasins. Data partitioning extracts all subbasin flowline networks from a set of prestaged subregion geodatabase files and builds a table of intersubbasin network connections. A directed graph representing intersubbasin flow is built from the intersubbasin connections and associated subbasin centroids (figure 2). Using

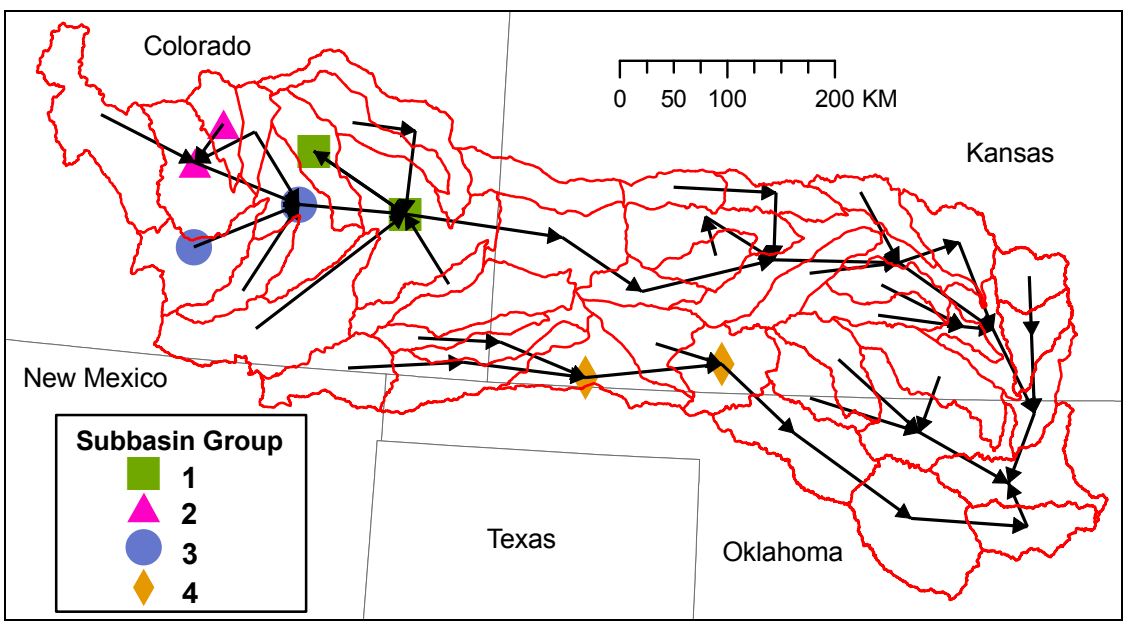


Figure 2: A directed graph showing intersubbasin flow between subbasin centroids for the pilot study area. Subbasins with large common centroid symbols comprise a group having cyclical flow between its subbasins.

the intersubbasin connection graph, subbasin groups having mono-directional flow between them are identified and assigned a processing order. Subsequently, subbasin networks are appended,

where necessary, into subbasin groups that have cyclic flow between subbasins. A more thorough description of network preprocessing is provided by Stanislawski [26].

Network preprocessing provides a table of catchment area and UDA estimates, along with lengths and reach codes for each processed (flow-directed, planar) network feature in the high-resolution NHD, which may be pruned to a desired density by accessing the table. Preprocessing, or enriching a data layer to prepare it for automated generalization is a fairly common practice [27, 28]. In the perceptual grouping or linear 'stroke' building process, a database of network features is enriched with values that facilitate automated network abstraction [2, 21].

3.2 Polygon pruning

Table 1. Minimum map size criteria for NHDWaterbody and NHDArea polygon feature types.

NHDWaterbody Feature Types		NHDArea Feature Types	
Feature type	Minimum area on map (cm^2)	Feature type	Minimum area on map (cm^2)
Playa	.050671	Area to be Submerged	1.266769
Ice Mass	.403225	Area of Complex Channels	14.988357
LakePond	.003167	BayInlet	.000000
Reservoir	.004560	Bridge	.048387
SwampMarsh	.050671	CanalDitch	.008107
Estuary	.000507	DamWeir	.002027
		Flume	.019355
		Foreshore	.025806
		Hazard Zone	.000507
		Lock Chamber	.003167
		Inundation Area	.018241
		Rapids	.000507
		SeaOcean	.000507
		Special Use Zone	.003167
		Spillway	.002027
		StreamRiver	.003167
		Submerged Stream	.012668
		Wash	.201613
		Water IntakeOutflow	.008107
		Levee	.025806

Subsequent to network pruning, waterbody and area 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 having one or more artificial paths passing through it, of which all have been removed through network pruning, 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 any artificial paths are pruned if they are smaller than the minimum size at the generalized map scale for the associated feature type (table 1). Area polygons are pruned using rules similar to the waterbody rules, but with minimum-area criteria determined for NHDArea feature types (table 1); however, no condition is applied in the area pruning rules to maintain relatively prominent area features. After applying the aforementioned rules to prune waterbody and area polygons to a desired generalized map scale, a

secondary rule is applied to remove any area features that must overlap other polygon features that were pruned by the previous process. Area feature types that must overlap other polygon features are area of complex channel, bay inlet, bridge, flume, foreshore, hazard zone, lock chamber, rapids, and submerged stream.

Pruning the high-resolution NHD polygons in this manner ensures consistency among network and polygon features in resulting generalized data layers. Before pruning, high-resolution polygons must be preprocessed to properly populate the WBAreaComID field for artificial path features. The WBAreaComID field identifies the waterbody or area feature through which an artificial path passes.

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. As such, the approach follows Muller's recommendation to "use knowledge-based tools to support automated solutions" for generalization [30]. 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 programs applying the pruning rules. Aside from the aforementioned rules for pruning polygon features, rules governing the network feature selection process are to select reach codes that have a UDA value greater than a tolerance, and to select features composing the selected set of reach codes. A third network rule tests for the desired state, which is the density of the selected set of features less than or equal to a target density; if not, the selection tolerance should be increased and a new set selected, otherwise the results should be summarized.

4.0 Pilot Project

Automated preprocessing, network, and polygon pruning programs were tested through a pilot project. The following sections describe the pilot-project data, preprocessing, and network pruning.

4.1 Data description

High-resolution NHD data from a five subregion area (subregions 1102-1106) near the center of the United States were processed for the pilot project. The Arkansas and Cimarron Rivers are the primary rivers draining these subregions. The western edge of this area abuts the Continental Divide of North America in the Rocky Mountains. The high-resolution NHD within the study area includes 48 subbasins (figure 2) having more than 300 000 and 118 000 network and polygon features, respectively. Waterbody polygons range from less than 0.001 square kilometers (km^2) to more than 65 km^2, with Kaw Lake in Oklahoma being the largest polygon in the dataset. Area polygons range from less than 0.001 km^2 to more than 47 km^2. The inundation area for the John Martin Reservoir on the Arkansas River in Colorado is the largest area polygon in the dataset.

To eliminate some complications in processing network features, all over- or under-passing features were removed from the study area. A total of 294 607 high-resolution network features, with assigned flow direction, remain within the study area and are included in subsequent analyses. The remaining network features are considered planar features because a junction, or node, exists at all intersections of the network features, representing a confluence of the water features on the ground. A single inflowing point to the high-resolution network features in the study area was identified. NHDPlus attributes [32] assigned to the medium-resolution NHD layer indicate that about 34 856 km^2 are draining into the study area from the inflowing, medium-resolution network feature.

4.2 Data preprocessing

Preprocessing and pruning were automated through Arc Macro Language (AML), Python, and Visual Basic programs along with ArcGIS 9.1 and 9.2 geoprocessing functions. Using a single 3 GHz processor, it took about 15 hours, or 18 to 20 minutes per subbasin, to preprocess the 294 607 planar network features having assigned flow direction, which is more than 96 percent of all high-resolution flowline features in the study area. Polygon preprocessing to assign WBAreaComID values for the more than 118 000 features took about 26 hours, or about 33 minutes per subbasin. With these estimates and 2 105 subbasins in the United States, about 75 to 80 days of automated processing on a single machine should be needed to preprocess the high-resolution subbasins in the lower 48 states. Aside from processing time, substantial effort is required to stage data into subregion subsets that can be handled by a single machine, identify inflowing UDA locations and values for each subregion subset, and process subregion subsets in proper order transferring outflowing UDA values to connected subsets.

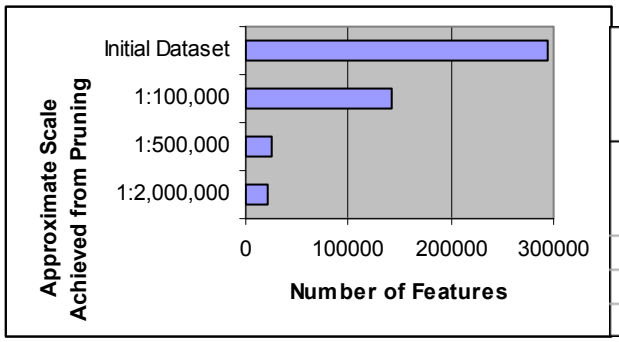
4.2.1 Preprocessing summary

Preprocessing results indicate the maximum UDA estimate of 209 030 km^2 occurs on the outflowing network feature of the high-resolution pilot data. In comparison, the sum of all subbasin areas flowing to the outflowing feature, plus the 34 856 inflowing km^2, is 218 978 km^2. The 9 948 km^2 UDA shortage estimated at the outflow feature is attributed to small sub-networks in the high-resolution network that are not connected to the main network, which subtract drainage area from the primary network. For further review of network preprocessing for this study area see Stanislawski [26]. High-resolution network features were tested for monotonically increasing UDA values with downstream location; no decreasing values were present.

4.3 Network Pruning Results

The high-resolution NHD layer in the five subregion study area was compiled from 1:24 000-scale hydrographic data from USGS digital line graph (DLG) and Tagged Vector Hydro (TVH) files, U.S. Forest Service Cartographic Feature Files (CFF), or other state-collected data. The drainage density of the 294 607 planar high-resolution network features with assigned flow direction in the study area is about 1.123 kilometer (km) per km^2. This set of high-resolution network features is referred to as the source, or initial, network features.

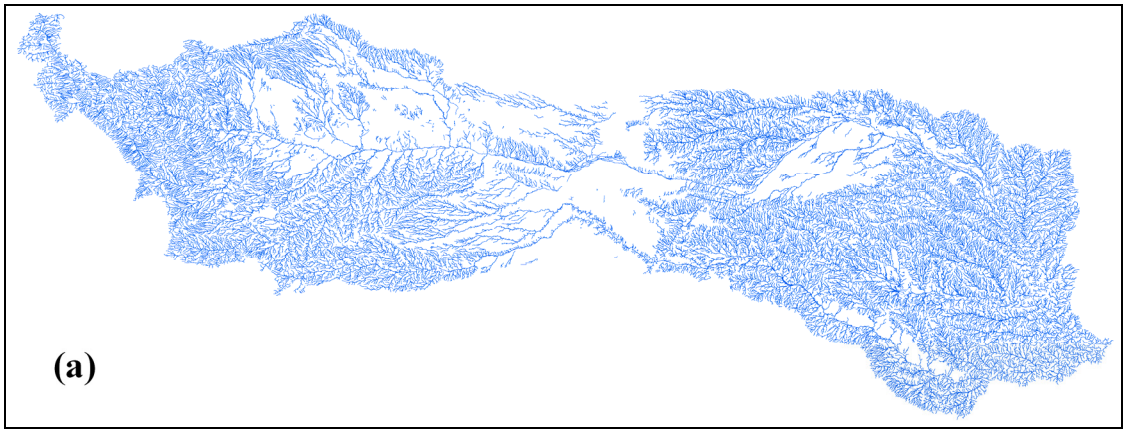
To demonstrate the network pruning process, source network features were pruned to drainage densities appropriate for three map scales. Drainage densities of hydrographic network features appropriate for map scales of 1:100 000, 1:500 000, and 1:2 000 000 were estimated respectively from the medium-resolution NHD layer, river Reach File version RF1 [33], and 1:2 000 000-scale DLG data [34]. Densities estimated from each of these datasets from network features within the five subregions of the study area were 0.630 km per km^2, 0.108 km per km^2, and 0.090 km per km^2, respectively. Features without initialized flow direction were excluded from the 1:100 000-scale density estimate, but all network features were included in the other two datasets because they do not include flow direction information. Subsequently, the source high-resolution network features were pruned to the drainage densities estimated for the three desired map scales. Pruning was completed by iteratively removing the set of features having a UDA estimate less than a tolerance, starting with 0.1 km^2. For each iteration, an interim density of non-pruned features is computed and compared to the desired density. If the interim density is greater than the desired density, then the minimum UDA tolerance is increased, and another iteration is performed.



Map Scale	Drainage Density (sq km)	Minimum UDA (sq km)	Number of Features
High Resolution NHD	1.123	0.00000	294607
1:100,000	0.630	2.60754	142977
1:500,000	0.108	107.92284	26147
1:2,000,000	0.090	158.87178	22313

Figure 3: Pruning summary of initial high-resolution NHD network features within the five subregion study area. Initial dataset was pruned to approximate drainage density appropriate for each associated map scale. Minimum UDA is the final value that was used to prune the initial dataset to achieve the desired drainage density.

Results of pruning the source network features to achieve the densities appropriate for each desired scale are shown in figure 3. Relatively large changes occur between 1:100 000-scale and 1:500 000-scale for minimum UDA to achieve the appropriate drainage density and the number of features retained in the pruned networks; however, drainage densities appropriate for desired scales also have the largest relative differences between 1:100 000 and 1:500 000 scales. The ratio of the number of features in the high-resolution (1:24 000-scale) and 1:100 000-scale is consistent with the general rule of Töpfer's radical law [35]--the ratio of the number of objects in two maps should equal the square root of the ratio of the map scales [36]; however, ratios comparing larger scales to smaller scales suggest the smaller scale (1:500 000 and 1:2 000 000) datasets, relatively, are too sparse. This result can be attributed to the different hydrographic data sources used to estimate appropriate drainage densities for each scale.



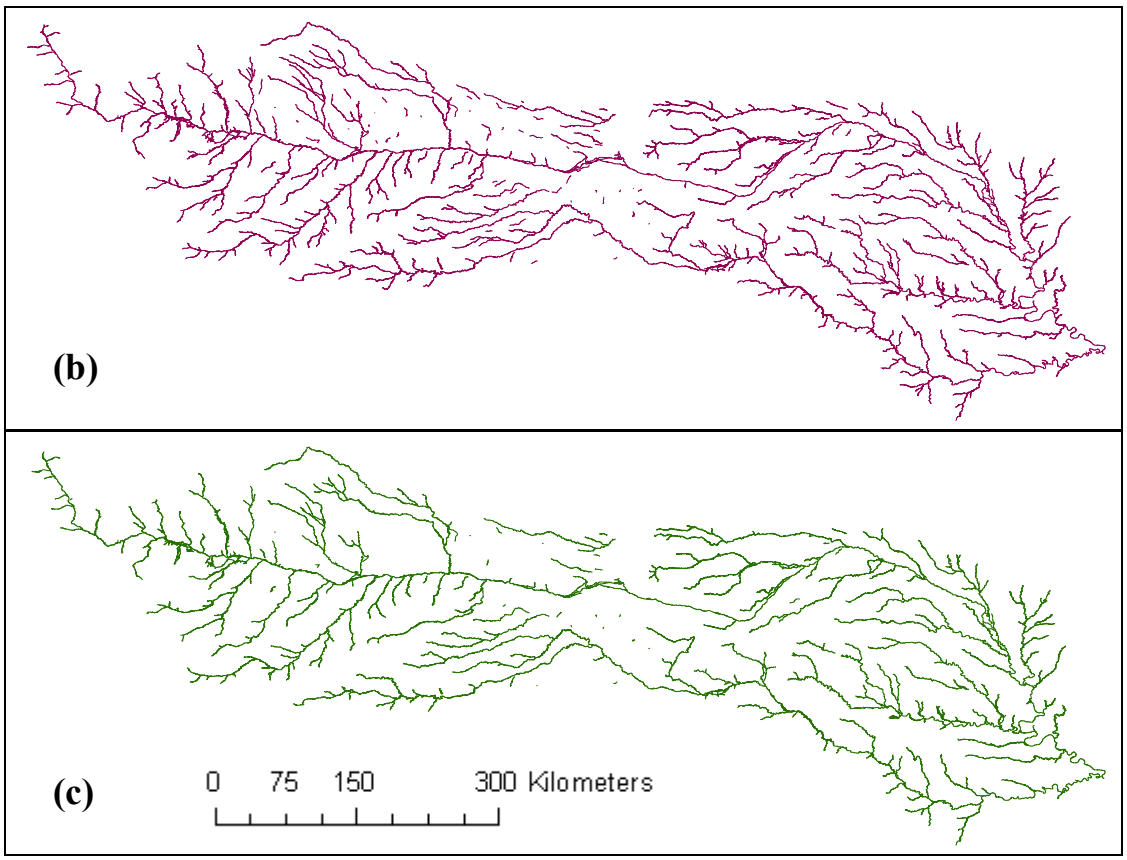


Figure 4: Network pruning results for the five subbasin study area at map scales of : (a) 1:100 000, (b) 1:500 000, and (c) 1:2 000 000.

The source high-resolution network features, after pruning to the three desired map scales, are shown in figure 4. Some small disconnected sub-networks remain in the generalized networks as an artifact of pruning by UDA. Data standards for medium-resolution NHD and 1:100 000-scale USGS topographic maps indicate that perennial stream/river features larger than 1.60 centimeters (cm) should be collected for the 1:100 000-scale NHD layer [37, 38], whereas stream/river features larger than 1.27 cm should be collected for 1:250 000-scale topographic maps [39]. Using the less restrictive, smaller scale criteria of 1.27 cm, sub-networks smaller than 1.27 cm at their respective map scale—6.35 km for 1:500 000 and 25.4 km 1:2 000 000—in total length were identified. The 1:500 000-scale generalized network includes 29 of 77 sub-networks that are less than 6.35 km long, and the 1:2 000 000-scale generalized network includes 34 of 46 sub-networks that are less than 25.4 km long (figure 5). The feature simplification phase of generalization, which focuses on rendering legible graphic products, should eliminate small sub-networks; however, drainage densities that govern network pruning should be magnified in some manner to compensate for the loss of features expected through simplification.

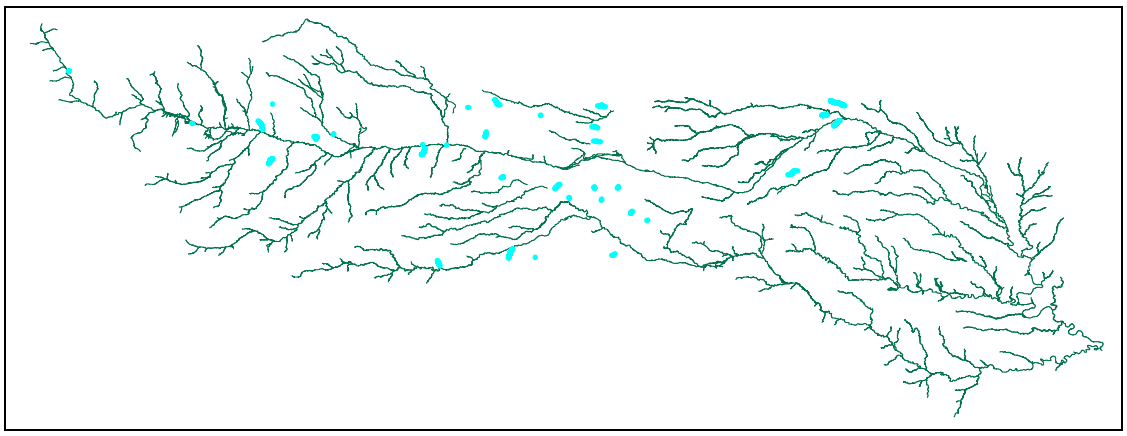


Figure 5: High-resolution NHD of study area generalized to 1:2 000 000-scale with connected groups of features that are less than 25 km long highlighted with cyan.

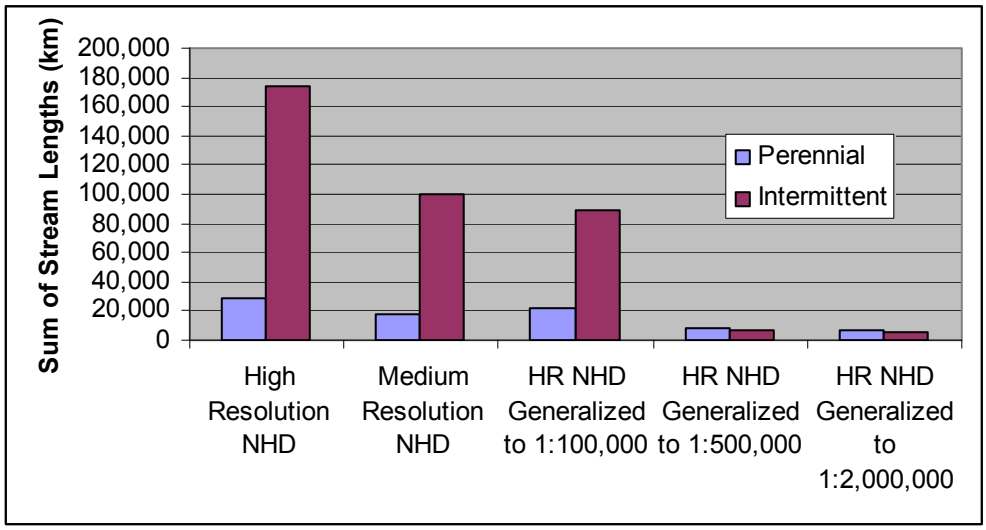


Figure 6: Summary of perennial and intermittent streams in high- and medium-resolution NHD network and in the high-resolution network pruned to achieve densities appropriate for 1:100 000, 1:500 000, and 1:2 000 000 scales. Networks only include features with assigned flow direction.

Network features lengths are summarized in figure 6 by hydrographic category (intermittent, perennial) for the source high-resolution and the medium-resolution planar network features, along with the source high-resolution network pruned to the three generalized map scales. The composition of perennial and intermittent features is nearly the same between the medium-resolution features and the source high-resolution features generalized to the comparable 1:100 000-scale, with nearly four times as many intermittent features as perennial; however, fewer intermittent than perennial features remain after pruning the source network to the scales of 1:500 000 and 1:2000 000.

A review of intermittent streams remaining in the network pruned to 1:2 000 000-scale reveals two anomalies. First, the primary path of some streams has perennial streams flowing into intermittent ones, indicating that either these intermittent streams are losing water through some field condition or inconsistent collections standards were applied. Second, the braided sections within the flood plain of some streams are coded as intermittent except for the primary channel (figure 7). The

braided features are maintained in smaller scale selections because connectivity of the braided features yields UDA estimates nearly the same as the primary channel. Some enhancements in the UDA estimation process may be warranted to better reflect field conditions.

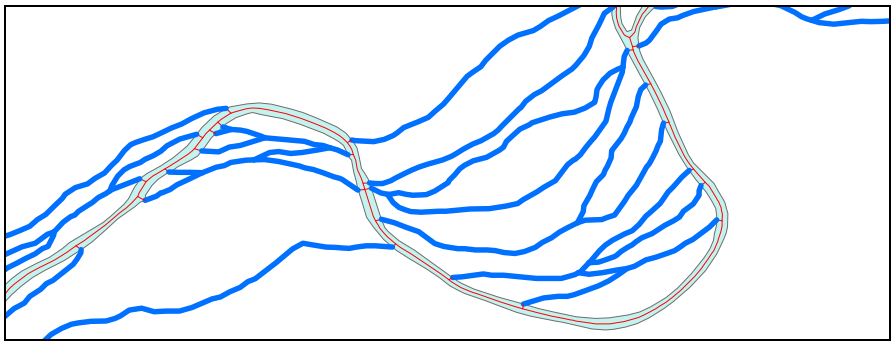


Figure 7: Intermittent streams (blue) in the braided area of this flood plain are maintained in the pruned high-resolution networks because connectivity to the main channel generates UDA estimates that are nearly the same for intermittent braided features and perennial main channel features.

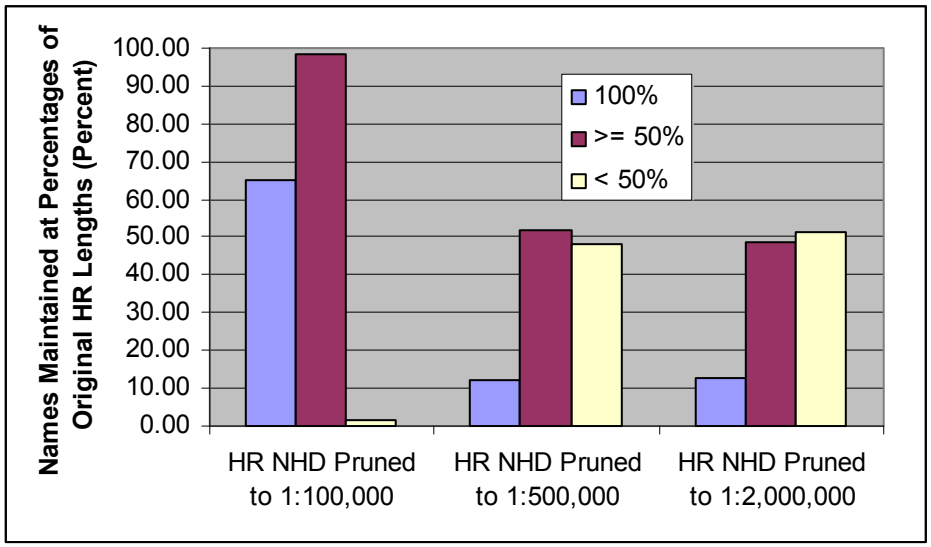


Figure 8: Summary of names maintained on the generalized network features.

Lastly, the lengths of features having common geographic names [9] were compared between the source high-resolution network and the pruned networks. The lengths of named features in the source high-resolution network summarized by name range from 0.147 km to 1535.75 km for a total of 1 835 geographic names. Only 1785, 418, and 315 source names are maintained when the source high-resolution is pruned to 1:100 000-scale, 1:500 000-scale, and 1:2 000 000-scale, respectively. The percentages of the named lengths of source network features that were maintained in each of the pruned networks are summarized into three percentage categories in figure 8. For instance, 99 percent of the named features in the network pruned to 1:100 000-scale maintained greater than or equal to 50 percent of their original length in the source high-resolution network, whereas about 1 percent maintained less than 50 percent of their original length, and about 65 percent of the names in this pruned network maintained 100 percent of their original length. Less

than 15 percent of the names in the lower resolution pruned networks maintained 100 percent of their original lengths.

4.3 Polygon Pruning Results

Applying the polygon pruning rules in conjunction with network features after pruning, the NHDWaterbody and NHDArea feature classes of the high-resolution NHD in the study area were pruned to the 1:100 000, 1:500 000, and 1:2 000 000 scales. The number of features before and after pruning is shown in figure 9. Of 115 802 waterbody polygons in the initial dataset, 30 152, 1 059, and 652 waterbodies remained after pruning to 1:100 000, 1:500 000, and 1:2 000 000-scales, respectively. The 2 633 area features in the initial dataset were pruned to 1 575, 787, and 624 features for the 1:100 000, 1:500 000, and 1:2 000 000-scales, respectively. Comparing the frequencies of features in the pruned datasets to the initial dataset indicates the number of area features remaining after pruning is more than expected from the general rule of Töpfer's law at all three scales, but the number of waterbody polygons remaining after pruning is less than expected from Töpfer's law, and substantially less at the 1:500 000 and 1:2 000 000-scales.

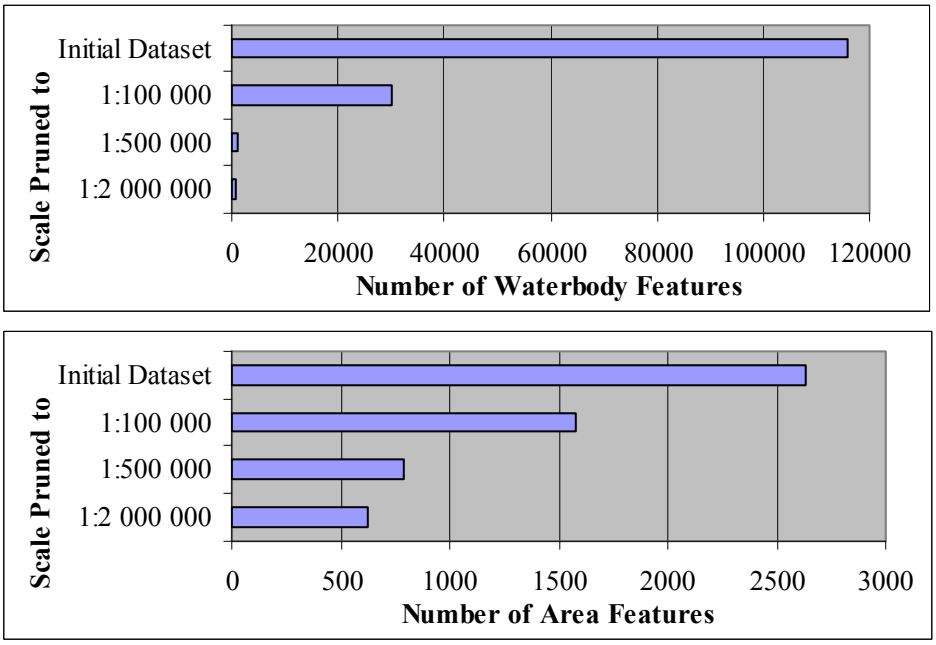


Figure 9. Pruning summary of initial high-resolution NHD waterbody and area polygon features within the five subregion study area. Initial dataset was pruned to 1:100 000, 1:500 000, and 1:2 000 000 scales.

Number of features by feature type before pruning is shown in figure 10. Nearly 98 percent of the waterbody features in the initial dataset are lake/pond type and more than 75 percent of these features are smaller than the associated minimum size criteria for 1:100 000-scale. These results and the Töpfer law comparisons in the previous paragraph suggest the applied standard's-based polygon rules pruned more features from the original dataset than what is needed to appropriately represent waterbodies on graphic products, particularly at the smaller scales. An analysis of the proportions of the mapped area that is covered by polygon features for each scale may furnish useful information for enhancing the pruning rules.

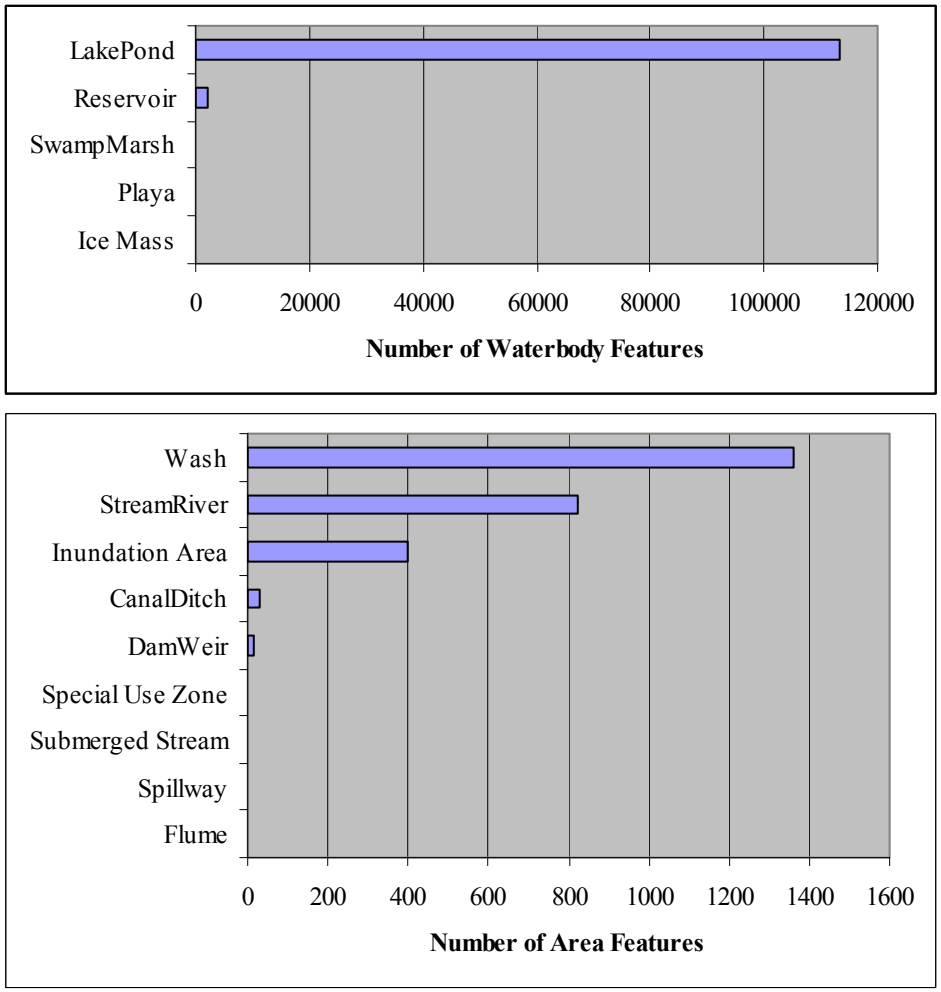


Figure 10. Number of high-resolution NHD waterbody and area polygons by feature type in the study area.

Results of network and polygon pruning for a section of the study area are shown in figure 11. The section of the figure showing the 1:500 000 and 1:2 000 000-scale features (fig 11 c) includes a large polygonal stream feature with a large tributary where all the artificial paths were removed during network pruning. Pruning sub-selects high-resolution features and does not alter the features. This figure illustrates an important requirement for the polygon simplification process; to prune off tributaries or parts of polygonal features where artificial paths are removed, which will furnish a consistent representation of more prominent hydrographic features in lower resolution layers.

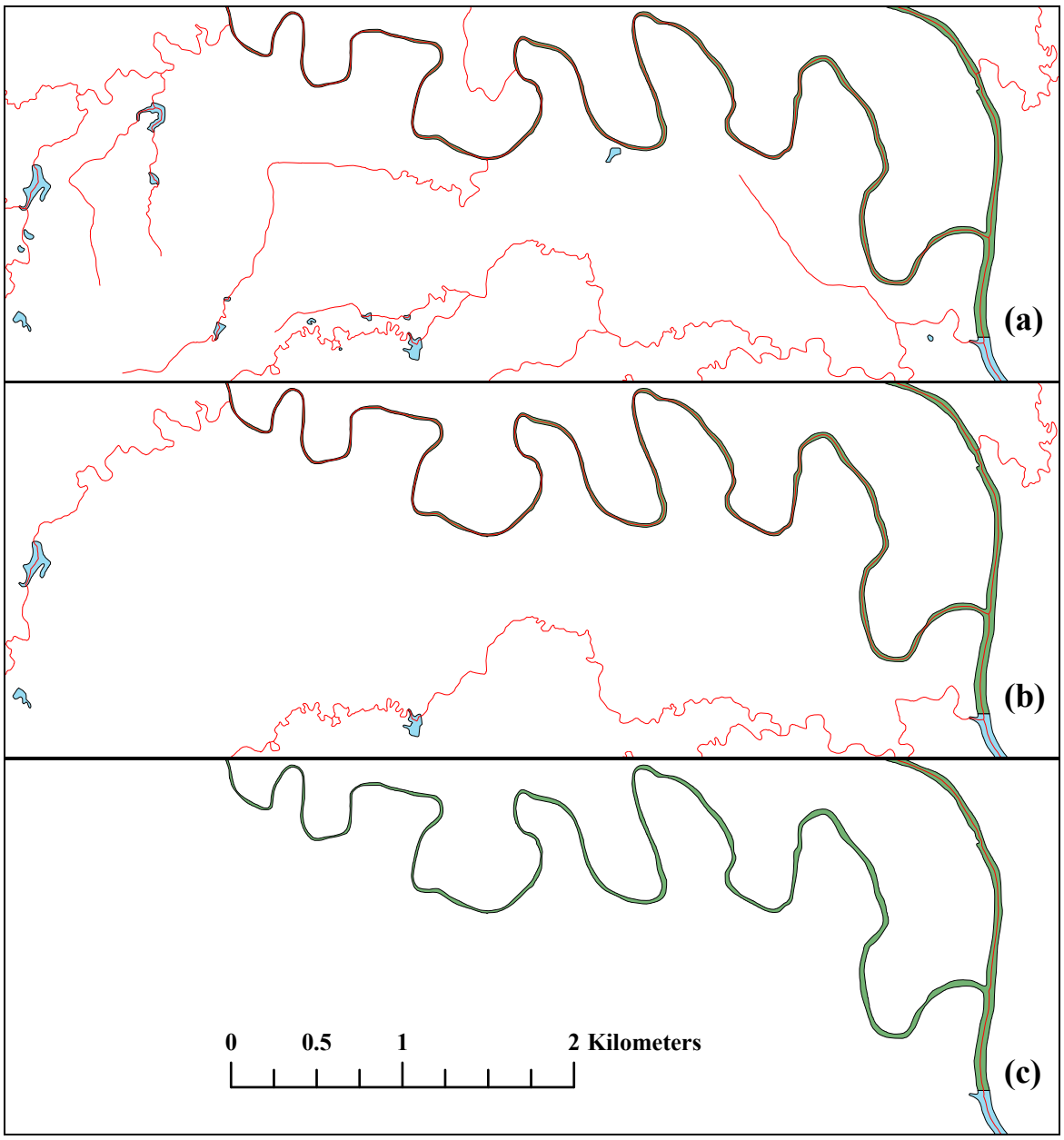


Figure 11. A section of the study area showing (a) source high-resolution NHD network features (red) over waterbody (light blue with black outlines) and area (green with black outlines) features pruned to (b) 1:100 000-scale and (c) 1:500 000 and 1:2 000 000-scales. Pruning to 1:500 000 and 1:2 000 000-scales yields the same features in this area.

5.0 Discussion

An overview of UDA estimates and pruning results for a section of the study area is presented in figure 12. A relatively high density area is present in the southeast quarter of the medium-resolution (1:100 000-scale) NHD network (figure 12 b). The range of surface hydrographic conditions (wet and dry years) and various map compilation standards used by the USGS over the years [40] is likely the cause of this anomaly. Relatively, pruning by UDA and reach code generates more

homogeneity in the density of network features over the area of interest as more features are removed (figure 12 a, c-e). Although pruning can remove data-collection anomalies, it also can mask climate and terrain variations that should be depicted in hydrographic features through density variations. From a cartographic perspective, the NHD pruning process could be enhanced through a system of local relations between map scale and appropriate drainage density that better reflect climate and terrain variations than a single relation for an area of interest or the entire database. Such relations may be implemented for the NHD at the subbasin level. Likewise, Buttenfield [41] indicates a uniform application of Töpfer's law [35] may not be suitable for all sections of a map and that "the geometry of the map symbols must reflect the geographical structure of the landscape, and vary accordingly during map simplification."

Others [2, 22, 23] have applied Töpfer's law to determine an appropriate number of features, or objects, to simplify road networks for various map scales. Simplification by road type furnishes a limited number of feature densities, and, consequently, map scale alternatives. Feature grouping, or stroke building provides hierarchies within road types, which expands the scale options for mapping. Feature lengths likely affect stroke building and ordering, but are not used in the selection process [2, 22, 23]. Conversely, UDA estimates for hydrographic network features are continuous and not likely to be duplicated within a local area. Feature lengths affect preprocessing UDA values and the NHD network selection process. Road network simplification may benefit by using feature lengths in the selection process to achieve scale-dependent network densities.

A braided section of network streams and surrounding waterbody and area polygons were maintained in all pruned datasets and is visible within the (pink) inundation area shown in figure 12 (c-e). This braided section of streams flow into the Great Salt Plains Lake in Oklahoma, and prominently appears in topographic maps and satellite images. NHD data indicate these braids are intermittent, but perennial NHD streams flow into them from the west. Feature simplifications subsequent to pruning can be tailored for remaining network features and aggregates, such as this, to achieve a desired cartographic display.

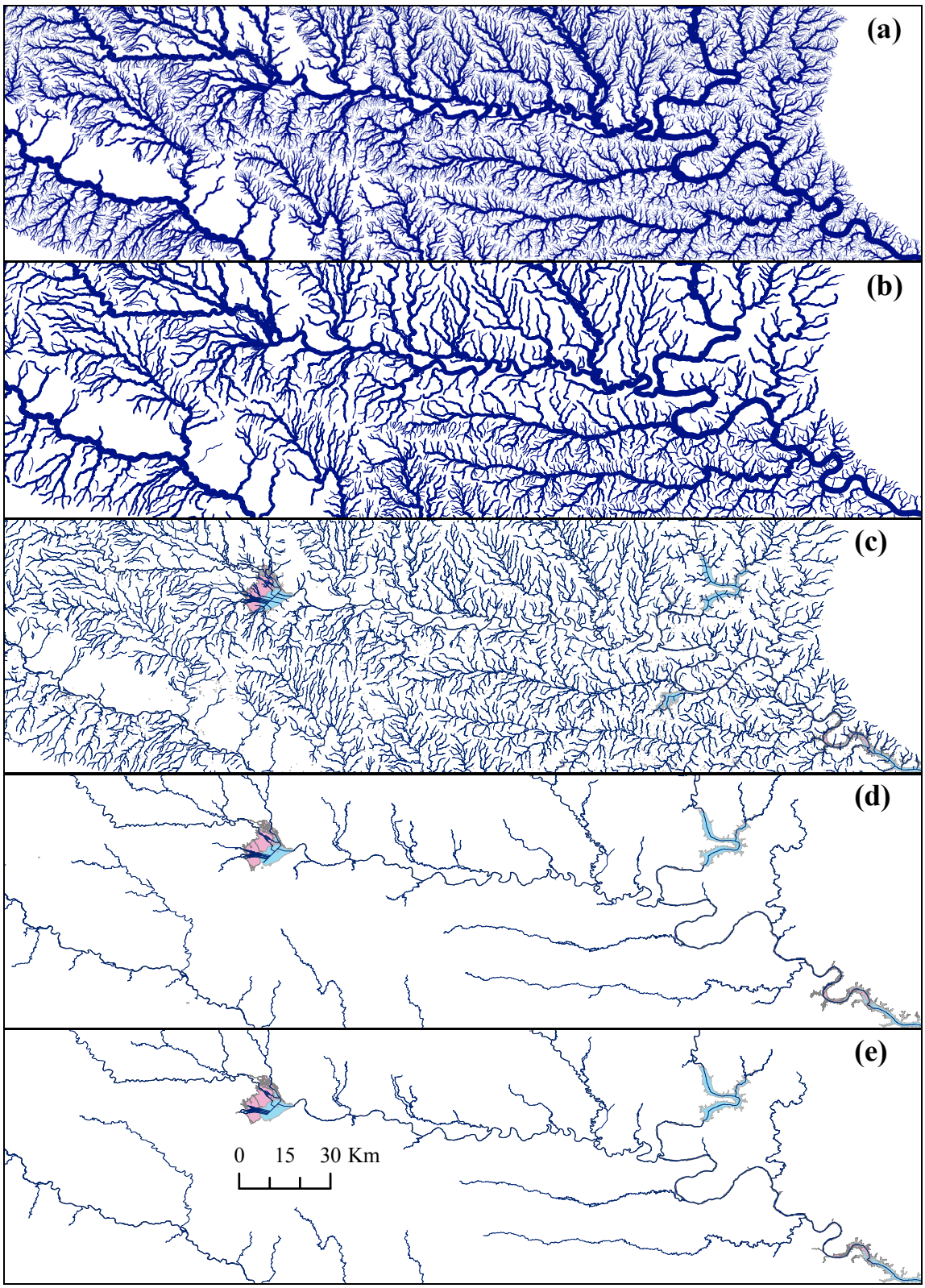


Figure 12: A section of the study area showing (a) source high-resolution NHD network features graduated with UDA estimates, (b) medium-resolution NHD network features graduated with

NHDPlus UDA estimates, and source high-resolution waterbodies (blue with gray outlines), areas (pink with gray outlines, and network features pruned to (c) 1:100 000, (d) 1:500 000, and (e) 1:2 000 000 map scales. Map scale of all figure sections is 1:2 000 000.

Some advantages of the network pruning by UDA and reach code include:

1. the results are unbiased – does not take into account stream names;
2. local high density areas at deltas or other braided areas can be maintained;
3. features in low density, relatively dry, low flow volume areas are maintained because they drain relatively large areas; and
4. some density anomalies because of various data collection standards are eliminated.

Perceived disadvantages of network pruning by UDA and reach code include:

1. the named features may not be maintained to the headwater source;
2. no main path currently is identified in braided areas; and
3. small sub-networks currently (2008) are not removed.

6.0 Conclusion

An automated method for generalizing the United States NHD flowline and polygon features was presented in this paper. Automated preprocessing enriches features in the high-resolution dataset with values required for feature pruning. Subsequently, network features and polygon features connected to the network can be pruned by UDA and NHD reach codes to achieve a drainage density appropriate for any map scale. A set of standard's-based rules were designed to prune polygon features not connected to network features.

Preprocessing and pruning were demonstrated in a pilot study and evaluated. UDA preprocessing may be improved through specialized handling of perennial to intermittent connections in braided flood-plain areas, and network pruning may be improved by maintaining more geographic names to the headwater source and removing small sub-networks not passing through prominent waterbodies. Pruning of network and connected polygon features could be substantially enhanced by enforcing stratified network density estimates that better reflect climate and terrain variations than a single estimate. Comparing results from polygon pruning with Töpfer's law suggests the standard's-based rules removed too many waterbodies, particularly at the smaller scales. Further research is needed to identify possible enhancements for the automated polygon pruning process.

Future efforts will involve preprocessing the NHD for the lower 48 states and pruning the high-resolution network to 1:24 000 and 1:100 000 scales to form the beginnings of a fully integrated multiple representation database. Currently (2008), preprocessing does not include non-planar features or features without an assigned flow direction. Additional development is required to include non-planar features in preprocessing, and flow direction can be assigned through data editing. Further development also may be required to handle large subbasin groups that may be generated around complex coastal drainage networks. Success of this effort will be largely governed by the accuracy of the network features in the database, and the ability of the data stewardship program to expedite revisions. Thiessen polygon catchment area estimates may have an adverse impact on the adequacy of pruning results, particularly in flat or coastal areas, but a program to assign more precise catchment area estimates, such as NHDPlus, could alleviate such impacts.

In conjunction with the aforementioned, feature simplification procedures must be tested and developed. Important tasks to be addressed include removal of small sub-networks not connected to

prominent polygons, handling of braided network and polygon features, and trimming of polygons to remove sections where network artificial paths are pruned.

Although additional processes will be required, successful implementation of this network pruning process can produce several advantages for the USGS NHD Program which include: optimized database maintenance, automation of a fully integrated multiple representation database, improved database integrity, advanced applications for the NHD. The first two advantages are common goals of generalization research and multiple representation databases. Improved database integrity will be achieved through automated review of UDA estimates to detect, and subsequently fix, inappropriate network gaps or features that are improperly oriented or have been assigned an incorrect hydrographic category. Lastly, a multiple representation NHD database will provide better support for existing NHD applications through simplified access, distribution, and display of the integrated layers. Integrated multi-resolution NHD layers with UDA estimates also can assist surface and subsurface hydrologic, geomorphologic, and geophysical terrain investigations.

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 overseeing and participating in the project; however, the design and analysis of this research, along with writing of this article, are contributions of the author.

Acknowledgments
Thanks go to E. Lynn Usery and Mike P. Finn of CEGIS, SAIC developer Jess Janssen, and former USGS employee Pat Turley, the many contributors to the NHD Stewardship program, and the NHDPlus developers, particularly Cindy McKay, and also to 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] 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.

[3] 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.

[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] NRC, A Research Agenda for Geographic Information Science at the United States Geological Survey, National Research Council of the National Academies, Washington, D.C.: The National Academies Press, 2007.

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

[9] USGS, Geographic Names Information System – GNIS, United States Geological Survey, online: http://nhd.usgs.gov/gnis.html, 2007.

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

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

[12] 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.

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

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

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[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] A. Ruas, C. Duchêne, A prototype generalization system based on the multi-agent system paradigm, 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. 269-284.

[26] 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

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

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

[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] 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.

[33] R.C. Horn, L. McKay, S.A. Hanson, History of the U.S. EPA's river reach file: a National Hydrographic Database available for ArcInfo applications, Proceedings of the Fourteenth Annual ESRI User Conference, Palm Springs, California, May 23-27, 1994, online: http://www.epa.gov/waters/doc/historyrf.pdf.

[34] USGS, Streams and Waterbodies of the United States, United States Geological Survey, online: http://nationalatlas.gov/atlasftp.html, 2003.

[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] 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.

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

[39] USGS, 1:250,000-Scale Quadrangle Maps: Supplemental Topographic Instructions, United States Geological Survey, National Mapping Program Technical Instructions, Supplement 84-2-C, 5 November, 1984.

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

[41] B.P. Buttenfield, A rule for describing line feature geometry, in: B.P. Buttenfield, R.B. McMaster (Eds.), Map Generalization: Making Rules for Knowledge Representation, Longman Scientific & Technical, 1991, pp.150-171.