Development of a Knowledge-Based Network Pruning Strategy for Automated Generalisation of the United States National Hydrography Dataset

Lawrence V. Stanislawski
Science Applications International Corporation (SAIC),
Center of Excellence for Geospatial Information Science (CEGIS)
United States Geological Survey (USGS)
Tel: 1 573 308 3914
Email: lstan@usgs.gov

Abstract

The National Hydrography Dataset is a comprehensive vector data model representing surface-water features of the United States. The current (May 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 generalisation 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 network 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 network pruning for the National Hydrography Dataset, which is based on National Hydrography Dataset reach codes and preprocessing estimates of upstream drainage area. The approach is demonstrated on a five subregion subset of nearly 300,000 hydrographic network features from the high-resolution layer of the National Hydrography Dataset. Network pruning results to three levels of detail are summarized for the pilot project.

Keywords: automated generalisation, 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 generalisation for many years. Technology and research have advanced our capacity for cartographic and geospatial database generalisation through systems and tools that automate processes using modern database designs, knowledge bases, and artificially intelligent algorithms. Much of the recent progress is presented or reviewed in the newly published book by the International Cartographic Association (Mackaness and others, 2007). Although there are tools available that perform specific generalisation operations such as ESRI's Generalisation Toolbox, and some systems may be suitable for specific data types, such as Clarity GIS from 1Spatial with road networks (Touya, 2007), further research is needed to tailor intelligent automated generalisation processes that are suitable for primary geospatial data themes having comprehensive national coverage (Regnauld and McMaster, 2007).

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 (USGS, 2006). The 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 (ibid.). *The National Map* data will be derived from various sources by a consortium of data stewards. In coordination of 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 (USGS, 2003a). In 2005, the USGS Center of Excellence for Geospatial Information Science (CEGIS) began this generalisation project to "research and develop automated methods for generalisation to support multiple-scale display and delivery of *The National Map* and other USGS geographic data" (McMahon and others, 2005). More recently, the U.S. National Research Council (NRC) recommended development of "unique generalisation 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 (NRC, 2007). This paper describes ongoing CEGIS research into automated generalisation, 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), but 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 (USGS, 2000) and geographic names (USGS, 2007) from lower to higher resolution layers. For logistical reasons, maintenance efforts are focused on updating and densifying the high-resolution layer, which include 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 generalisation 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 generalisation 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 generalisation and multiple representation databases (Mackaness, 2006; Chaudry and Mackaness, 2006; Mackaness and others, 2007; Mustiêre and van Smaalen, 2007).

Regarding the NHD, automated generalisation has been divided into two primary development tasks—feature pruning and simplification. This article focuses on feature pruning, and feature simplification is not discussed. The following sections describe and demonstrate a knowledge-based automated network pruning strategy that eliminates less significant hydrographic network 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, which 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 other organizations. Features in each resolution of the NHD are separated into five feature classes—NHDArea, NHDFlowline (flowline), NHDLine, NHDWaterbody (waterbody), and NHDPoint—each containing a subset of NHD feature types

represented with the same geometrical shape type. 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 an areal 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). As of January 2008, the high-resolution NHD layer contained more than 27 million features, nearly 20 million of which are flowline features.

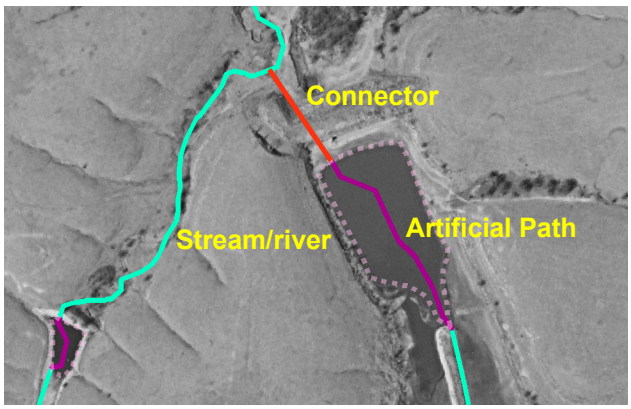


Figure 1: Artificial path, connector, and stream/river features over aerial photo.

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 (USGS, 2000). 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 (USGS, 2000).

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. About 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 (McCracken and Salmon, 1987; Manber, 1989) composed primarily of planar components (Manber, 1989). Subsequently, traversal techniques (Manber, 1989) 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 for Automated Generalisation

Pruning, or the initial process of selecting source objects and attributes to be represented in a generalised dataset, is common in generalisation strategies (McMaster and Shea, 1992; Brewer and Buttenfield, 2007; Mackaness and others, 2007; Regnauld and McMaster, 2007). In this paper, pruning the high-resolution NHD consists of eliminating relatively less prominent flowline network features 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 U.S., 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 (Stanislawski and others, 2005). 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 (Stanislawski and others, 2007). Subsequently, an augmented directed graph approach is used to assign upstream drainage area (UDA) estimates to each flowline (Stanislawski and others, 2006), 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 generalised 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 (Verdin, 1997). Pruning the NHD network features by UDA and reach code is similar to perceptual grouping or “stroke” building (Thomson and Brooks, 2000; Chaudry and Mackaness, 2005; Touya, 2007; Thomson and Brooks, 2007), 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 (USGS, 2002). 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.

The network pruning process, as described currently (2008), does not fit well within an agent-based framework (Ruas and Duchêne, 2007) 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.

Network pruning by UDA is an attempt to apply a holistic solution for automated database generalisation using surface water drainage knowledge 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 generalisation (Sarjakoski, 2007). In knowledge-based systems, the knowledge base and inference mechanism are separated (Mockler and Dologite, 1992; Sarjakoski, 2007). In our case, the knowledge base consists of the UDA values on the network features in the database and the composition of the reaches, and the inference engine is the program applying the pruning rules. Two rules governing the 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. The third 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, alter the tolerance or otherwise summarize results.

3.1 Preprocessing for network pruning

Implementing the network pruning process on a dataset as large as the U.S. 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

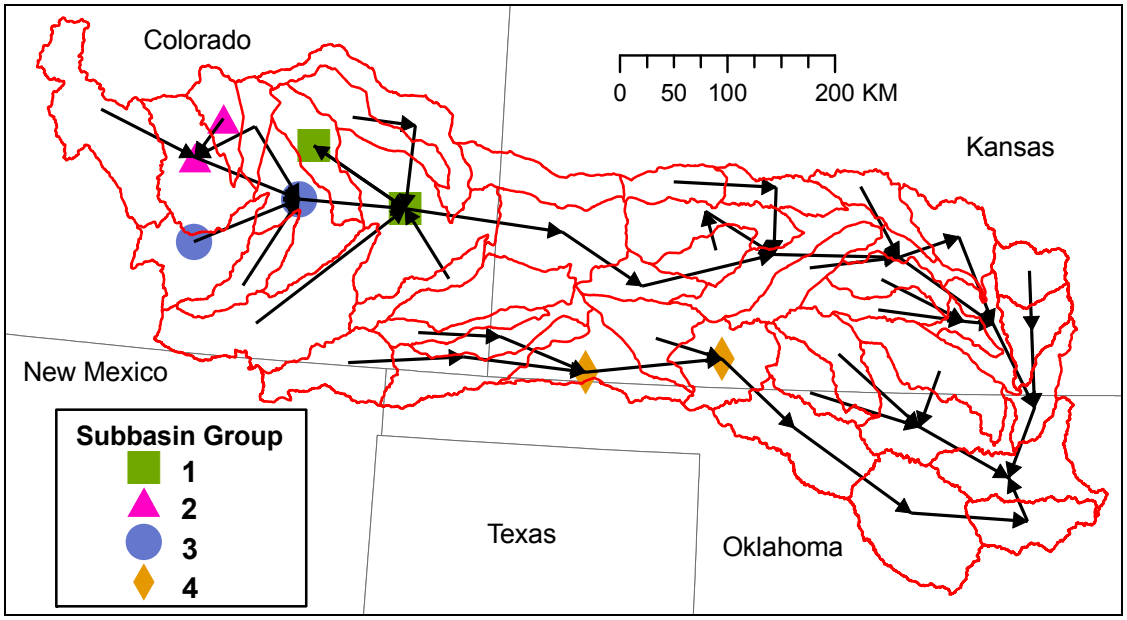


Figure 2: A directed graph showing intersubbasin flow between subbasin centroids for the pilot study area. Subbasins with large common centroid symbols comprise a group having cyclical flow between its subbasins.

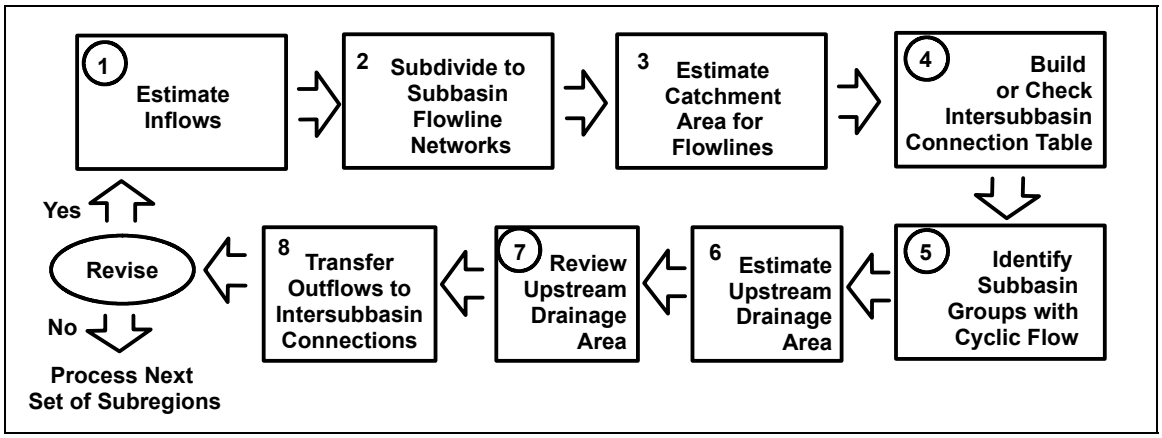


Figure 3: Flowchart of preprocessing steps to generate UDA estimates for each set of subregions. Steps with circled numbers (1, 4, 5, and 7) may not be necessary on secondary passes through the database.

subbasin centroids (figure 2). Using the intersubbasin flow graph, subbasin groups having mono-directional flow between them are identified and assigned a processing order. Subsequently, the subbasin networks are appended, where necessary, into subbasin groups that have cyclic flow between subbasins.

Processing steps required to prepare high-resolution NHD for the automated network pruning process are summarized in the flowchart in figure 3. The steps for each set of adjacent subregions within an area of interest are summarized as: (1) identify UDA estimates that are inflowing to the area of interest and associate them with a high-resolution network feature, (2) subdivide subregion files into subbasin networks, (3) estimate catchment area for all flowline features, (4) build the intersubbasin connection table, (5) identify subbasin groups having cyclic network flow within each group and append associated subbasin networks into group networks, simultaneously build a processing order for resulting subbasin and group networks, (6) estimate UDA for all subbasin and group networks in the proper processing order, transferring outflowing values to downstream networks, (7) review UDA estimates to verify monotonicity or identify anomalous values—such as, relatively large values at dangling to-nodes—indicative of missing connectivity, and (8) transfer the outflowing UDA values to the intersubbasin connection table for use with subsequent sets of subregions. All processing steps, except step (1), have been automated through database queries and batch Python and Arc Macro Language scripts, which also generate queues for interactive data review. Upon completing step (8) for a set of subregions, if it is necessary to fix connectivity issues, then some or all of the processing steps must be rerun, depending on the type of revision.

It is expected that at least two passes through preprocessing will be required to generate adequate UDA estimates for the high-resolution NHD network features. The first pass will identify connectivity problems and supply queue files that data stewards can use for revision. Passes subsequent to revision should provide more accurate estimates; however, completion of the first pass should furnish subregion sets and associated inflow estimates, intersubbasin connections, required subbasin groupings, and the network processing order. Therefore, some processing steps (circled steps in figure 3) may be simplified or eliminated on secondary passes.

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 generalisation is fairly common practice (Stoter, 2005; Yan and others, 2006). In the perceptual grouping or linear 'stroke' building process, a database of network features is enriched with values that facilitate automated network abstraction (Touya, 2007; Thomson and Brooks, 2000).

4.0 Pilot Project

Automated preprocessing and network 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 US 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 study area includes 48 subbasins (figure 2) having more than 300,000 high-resolution hydrographic network features. To eliminate some processing complications, 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 features are considered planar features because a junction, or node, exists at all intersection 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 (NHDPlus, 2008) assigned to the medium-resolution NHD layer indicate that about 34,856 square kilometers (sq km) are draining into the study area from the inflowing, medium-resolution network feature.

4.2 Data preprocessing

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. With these estimates and 2,105 subbasins in the U.S., about 26 to 30 days of automated processing on a single machine should be required 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 sq km 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 sq km, is 218,978 sq km. The 9,948 sq km 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. This point is further substantiated later. High-resolution network features were tested for monotonically increasing UDA values with downstream location; no decreasing values were present.

The high-resolution UDA estimates were compared to medium-resolution estimates on reach codes common to both network layers. UDA estimates are available in NHDPlus attributes that have been compiled for the medium-resolution NHD (NHDPlus, 2008). A total of 39,093 medium-resolution flowline reach codes were common to the high-resolution flowline, which is just less than 84 percent of all the medium-resolution flowline reach codes. The maximum UDA estimate for each common reach code in the high- and medium-resolutions is compared in figure

4. Five cases of large discrepancies between UDA estimates for common reach codes between the two resolutions are identified in figure 4 and evaluated. All networks in figures 5 through 8, which illustrate these cases, have symbols graduated with associated UDA estimates.

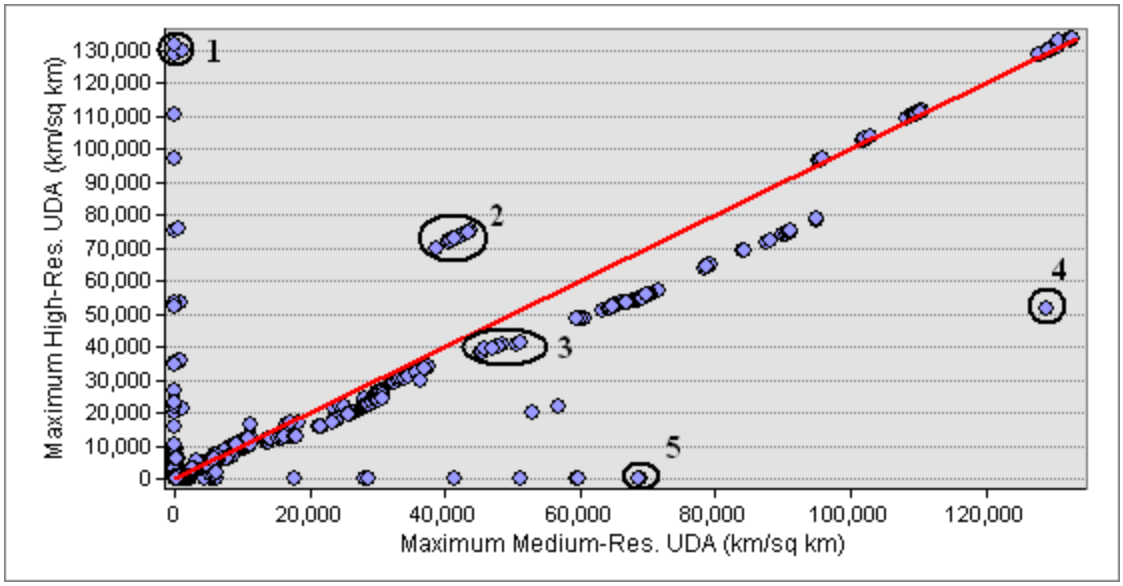


Figure 4: Comparison of maximum upstream drainage area (UDA) estimates for common reach codes in the medium- and high-resolution NHD layers. Each blue circle identifies values associated with one reach code. For reference, a line of equal values is depicted in red. Special cases are circled and labeled.

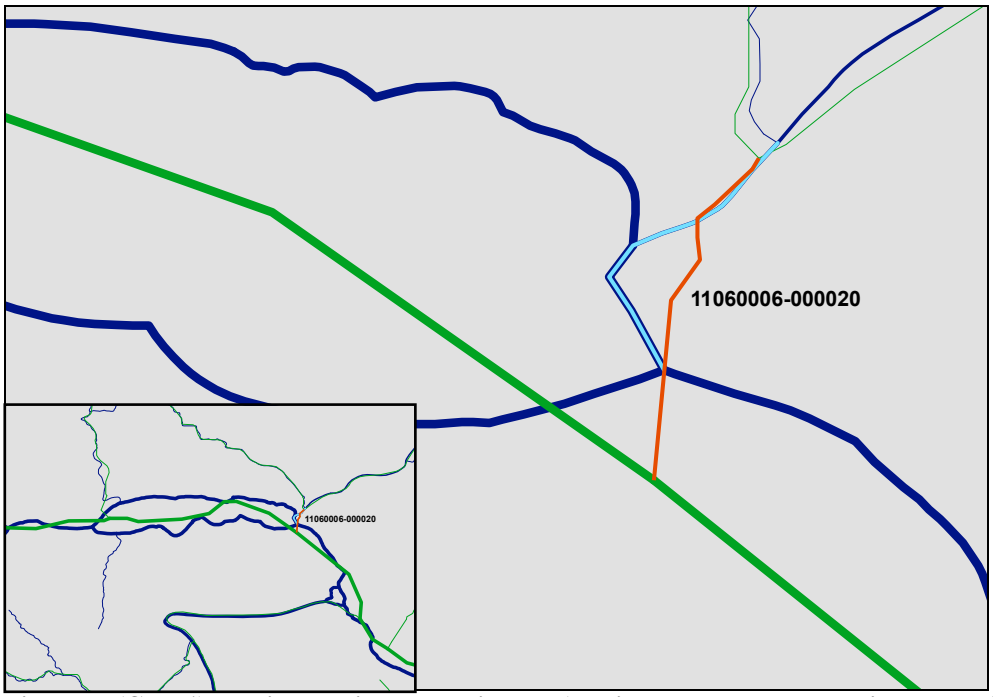


Figure 5: (Case 1) Maximum high-resolution UDA estimate on a reach code is substantially larger than maximum medium-resolution UDA estimate for associated reach code. High-resolution features displayed in blue with selected features in light blue and medium-resolution features shown in green with selected features in red. Common reach code of selected features is reach code 000020 from subbasin 11060006.

Case 1 in figure 4 highlights features with common reach codes that have a large high-resolution UDA estimate and small medium-resolution UDA estimate, which is an example of data points near the y-axis. This situation appears where reach codes occur on multiple network features in the high-resolution layer, with at least one of the features comprising a part of a main channel path; however, in the medium-resolution layer, these reaches do not participate in a main channel path. An example reach with this condition is shown in figure 5.

Case 2 reaches circled in figure 4 occur along the Cimarron River and have high-resolution UDA estimates of about 70,000 sq km, which are about 20,000 sq km larger than associated medium-resolution reaches. This discrepancy is caused by the equivalently large value assigned to the upstream inflow point to the high-resolution network on the south side of the study area (figure 6), where the common feature in the medium-resolution layer only receives 16.6 sq km. This discrepancy impacts all downstream features.

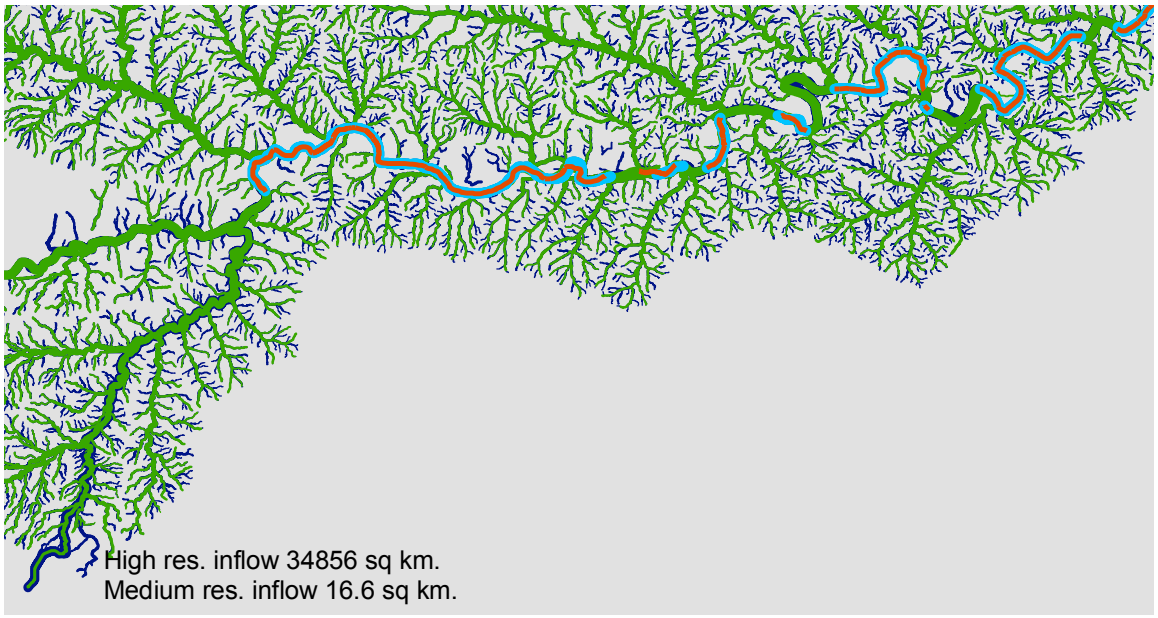


Figure 6: (Case 2) Comparison of maximum UDA estimates on common reach codes between high-resolution (blue with selected features in light blue) and medium-resolution (green with selected features in red). UDA value inflowing to study area from the south is nearly 20,000 sq km lower in medium-resolution layer than in high-resolution layer, which impacts all downstream segments.

For case 3 circled in figure 4, smaller UDA estimates are present on the main channel of the high-resolution layer than on the medium-resolution layer because, in sections of the Arkansas River flood plain, a greater amount of drainage area does not reach the main river channel in the high-resolution layer than in the medium-resolution because more disconnected sub-networks exist in the high-resolution (figure 7). This situation continues downstream on the Arkansas River propagating larger discrepancies between the high-resolution and associated medium-resolution UDA estimates; further explaining why the maximum UDA estimate on the network is less than the sum of associated subbasin areas, which was mentioned at the beginning of this section.

No illustration is provided for the single outlying data point circled for case 4 in figure 4. This discrepancy identifies a braided segment of the medium-resolution layer where the UDA estimate was improperly combined at a convergence, making the UDA estimate about twice as large as the associated high-resolution estimate.

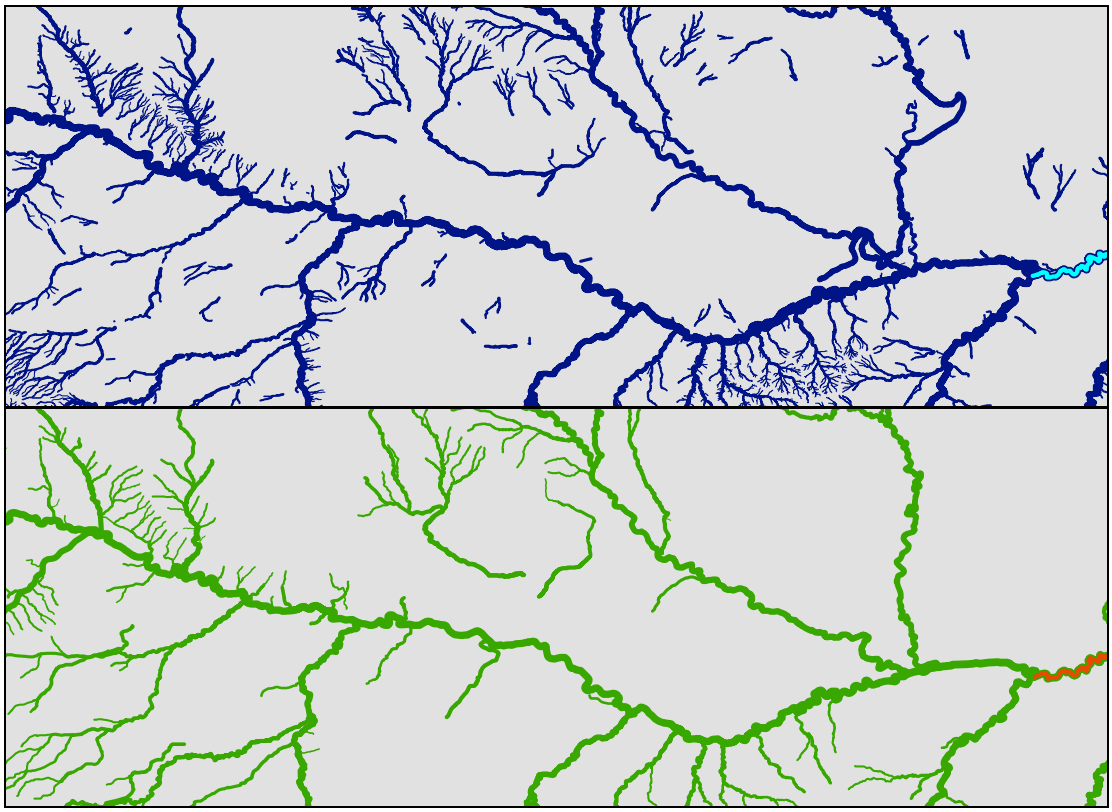


Figure 7: (Case 3) Comparison of maximum UDA estimates on common reach codes between high-resolution (top, selected features in light blue) and medium-resolution (bottom, selected features in reddish). UDA estimates of selected high-resolution features are about 5,000 sq km less than associated features in the medium-resolution. A greater number of disconnected sub-networks in the high-resolution network than in the medium-resolution caused this discrepancy on downstream features.

Figure 8: (Case 5) Comparison of maximum UDA estimates on common reach codes between high-resolution (blue with selected features in light blue) and medium-resolution (green with selected features in red). Pink circle highlights location where high-resolution features are disconnected from main channel, but medium-resolution features are connected to the main channel.

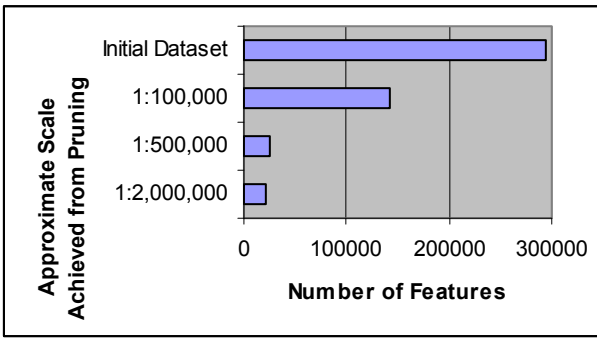
Lastly, in case 5 in figure 4, high-resolution UDA values are close to zero, whereas associated medium-resolution values are much larger. This situation is found where medium-resolution features are diverging from a main channel, but associated features in the high-resolution layer

are not connected to the main channel and, therefore, do not receive a contribution as a divergence. An example of this situation is displayed in figure 8.

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 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 km per square km. This set of high-resolution network features is referred to as the source, or initial, network features.

To demonstrate the network pruning process, the 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 (Horn and others, 1994), and 1:2,000,000-scale DLG data (USGS, 2003b). Densities estimated from each of these datasets from network features within the five subregions of the study area were 0.630 km per sq km, 0.108 km per sq km, and 0.090 km per sq km, 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 sq km. 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 9: 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 is shown in figure 9. 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 (Töpfer and Pillewizer, 1966)--the ratio of the number of objects in two maps should equal the square root of the ratio of the map scales

(Jiang and Harrie, 2003); 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.

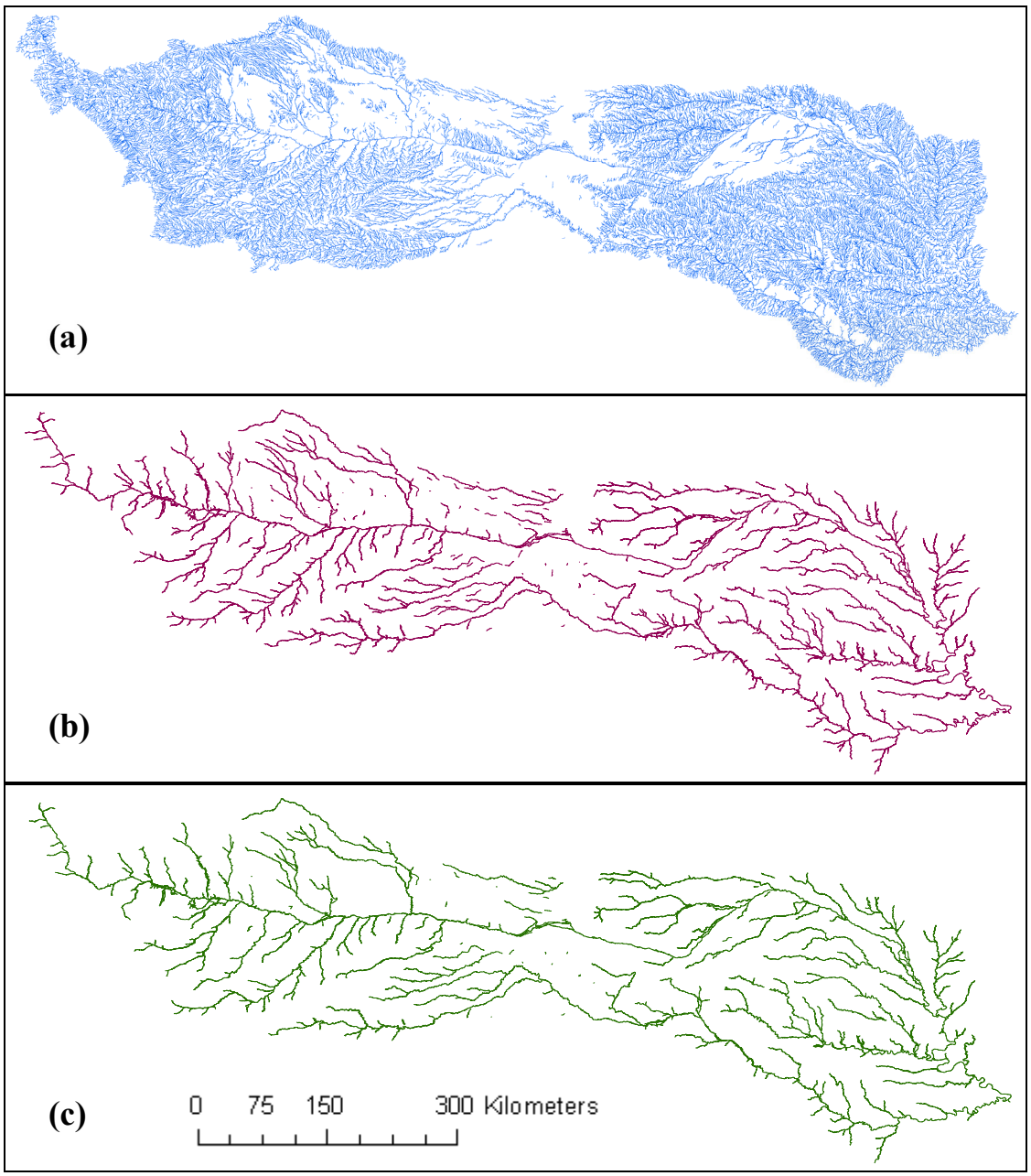


Figure 10: Network pruning results for the five subbasin study area at map scales of : (a) 1:100,000, (b) 1:500,000, and (c) 1:2,000,000.

The source high-resolution network features, after pruning to the three desired map scales, are shown in figure 10. Some small disconnected sub-networks remain in the generalised 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 0.63 inch (1.60

cm) should be collected for the 1:100,000-scale NHD layer (USEPA and USDOI, 1999; USGS, 1994), whereas, stream/river features larger than 0.5 inch (1.27 cm) should be collected for 1:250,000-scale topographic maps (USGS, 1984). Using the less restrictive, smaller scale criteria of 0.5 inch, sub-networks smaller than 0.5 inch 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 generalised network includes 29 of 77 sub-networks that are less than 6.35 km long, and the 1:2,000,000-scale generalised network includes 34 of 46 sub-networks that are less than 25.4 km long (figure 11). Through additional preprocessing, a sub-network number could be assigned to all features and used during the pruning process to eliminate small sub-networks and still achieve the desired drainage density; however, some constraints should maintain small sub-networks passing through prominent waterbody features.

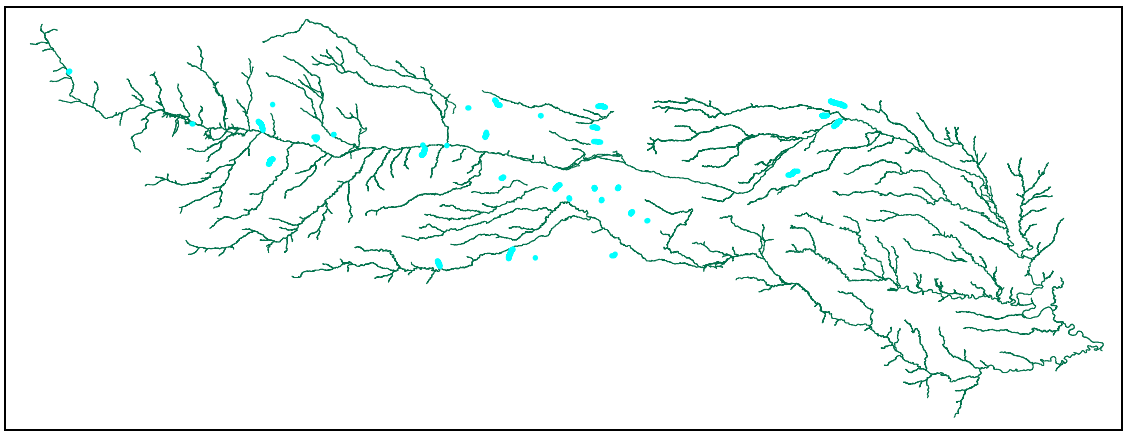


Figure 11: High-resolution NHD of study area generalised to 1:2,000,000-scale with connected groups of features that are less than 25 km long highlighted with cyan.

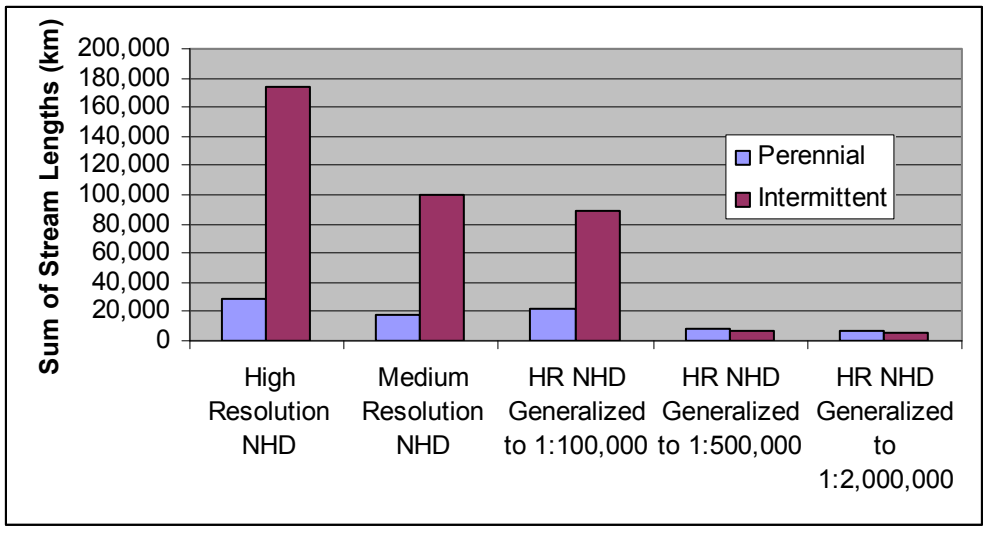


Figure 12: Summary of perennial and intermittent streams in high- and medium-resolution NHD network and in the high-resolution network pruned to achieve densities appropriate for 1:100,000, 1:500,000, and 1:2,000,000 scales. Networks only include features with assigned flow direction.

In figure 12, network features lengths are summarized 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 generalised map scales. The

composition of perennial and intermittent features is nearly the same between the medium-resolution features and the source high-resolution features generalised 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 1:2,000,000.

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, suggesting 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 13). 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.

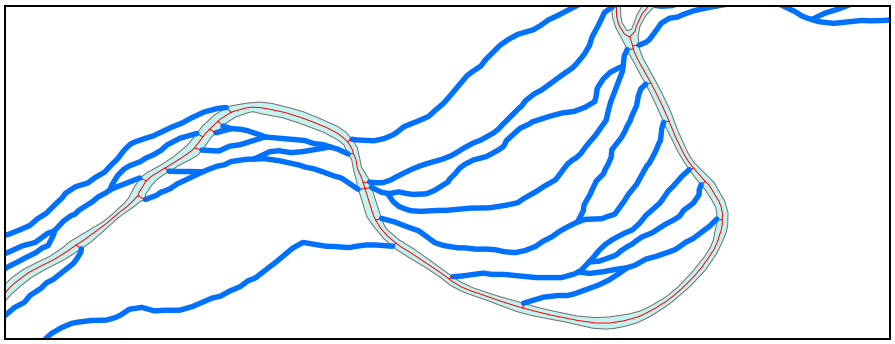


Figure 13: Intermittent streams (blue) in the braided area of this flood plain are maintained in the pruned high-resolution networks because connectivity to the main channel generates UDA estimates that are nearly the same for intermittent braided features and perennial main channel features.

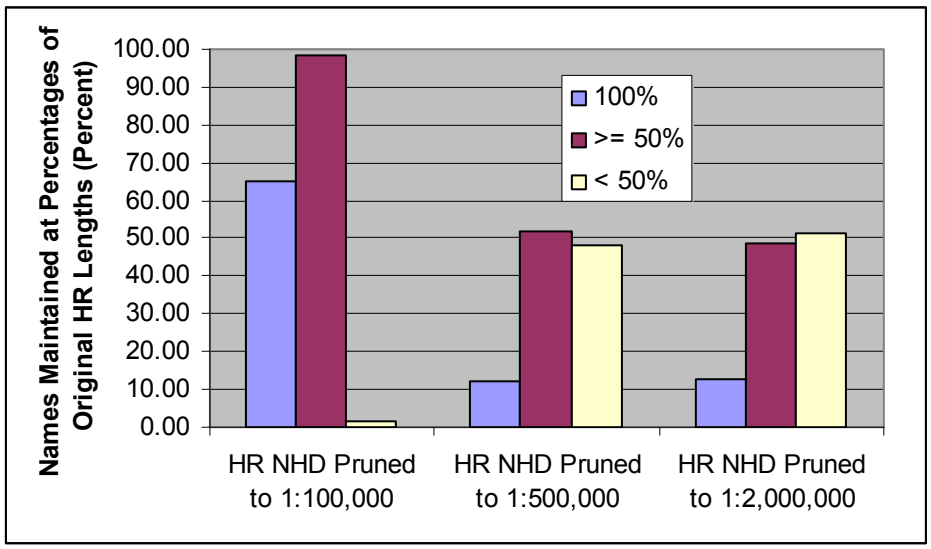


Figure 14: Summary of names maintained on the generalised network features.

Finally, the lengths of features having common geographic names (USGS, 2007) 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 14. 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, while 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.

5.0 Discussion

An overview of UDA estimates and network pruning results for a section of the study area is presented in figure 15. A relatively high density area is present in the southeast quarter of the medium-resolution (1:100,000-scale) NHD network (figure 15 b). The range of surface hydrographic conditions (wet and dry years) and various map compilation standards used by the USGS over the years (USGS, 1955) 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 15 a, c-e). While 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 network 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 (1991) suggests a uniform application of Töpfer's law (Töpfer and Pillewizer, 1966) 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 (Chaudry and Mackaness, 2005; Thomson and Brooks, 2007, Touya, 2007) 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 (Chaudry and Mackaness, 2005; Thomson and Brooks, 2007, Touya, 2007). On the other hand, UDA estimates for hydrographic network features are continuous and not likely to be duplicated within a local area. Feature lengths affect both 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 streams was maintained in all pruned networks and is visible within the (pink) inundation area shown in figure 15 (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 suggest 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.

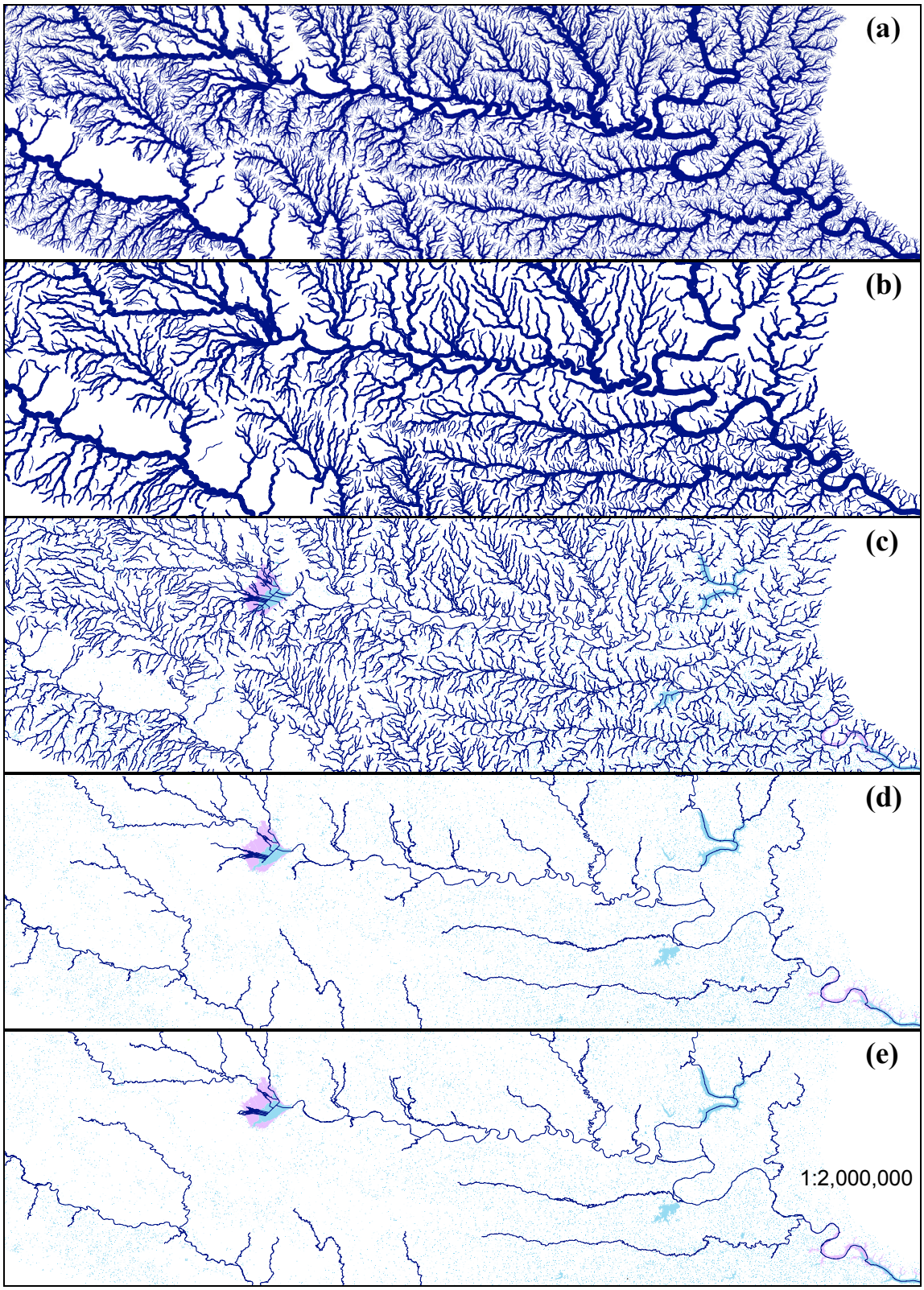


Figure 15: A section of the study area showing (a) source high-resolution NHD network features graduated with UDA estimates, (b) medium-resolution NHD network features graduated with NHDPlus UDA estimates, and source high-resolution waterbodies overlain by source network features pruned to (c) 1:100,000, (d) 1:500,000, and (e) 1:2,000,000 map scales.

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 are not removed.

6.0 Conclusion

An automated method for generalising the United States NHD flowline network was presented in this paper. Automated preprocessing enriches network features with UDA estimates, and, subsequently, the network can be pruned by UDA and NHD reach codes to achieve a drainage density appropriate for any map scale. Preprocessing and network 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. In addition, network pruning could be substantially enhanced by enforcing stratified density estimates that better reflect climate and terrain variations than a single estimate.

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, 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.

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 generalisation 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.

Acknowledgments

This project is funded through the USGS CEGIS and supported by the USGS National Geospatial Technical Operations Center (NGTOC). 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.

References Cited

Brewer, C.A., and Buttenfield, B.P., (2007) "Framing Guidelines for Multi-Scale Map Design Using Databases at Multiple Resolutions", *Cartography and Geographic Information Systems* 34(1): 3-15.

Buttenfield, B.P., (1991) "A Rule for Describing Line Feature Geometry", In: Buttenfield, B.P., and McMaster, R.B., (eds.), *Map Generalization: Making Rules for Knowledge Representation.* Longman Scientific & Technical, pp.150-171.

Chaudhry, O., and Mackaness, W.A., (2005), "Rural and Urban Road Network Generalisation Deriving 1:250,000 from OS MasterMap", *XXII International Cartographic Conference*: Coruña, Spain, July 11-16.

Chaudry, O., and Mackaness, W.A., (2006) "Modeling Geographic Phenomena at Multiple Levels of Detail", *AutoCarto 2006*: Vancouver, WA, June 26-28.

Horn, R.C., McKay, L., and, Hanson, S.A., (1994) "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. Online: http://www.epa.gov/waters/doc/historyrf.pdf

Jiang, B., and Harrie, L., (2003) "Cartographic Selection Using Self-Organizing Maps", *6th ICA Workshop on Generalisation and Multiple Representation*: Paris, France, April 28-30.

Mackaness, W.A., (2006) "Automated Cartography in a Bush of Ghosts", *Cartography and Geographic Information Systems* 33(4):245-256.

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

Manber, U., (1989) *Introduction to Algorithms: A Creative Approach*. Addison-Wesley Publishing Co., 478 pages.

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

McMahon, G., Benjamin, S.P., Clarke, K., Findley, J.E., Fisher, R.N., Graf, W.L., Gundersen, L.C., Jones, J.W., Loveland, T.R., Roth, K.S., Usery, E.L., and Wood, N.J., (2005) "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*, 76 pages.

McMaster, R., and Shea, K., (1992) *Generalization in Digital Cartography.* Washington, D.C.: Association of American Geographers, 134 pages.

Mockler, R.J., and Dologite, D.G., (1992) *Knowledge-Based Systems: An Introduction to Expert Systems*. Mackmillan Publishing Company, 793 pages.

Mustiêre, S., and van Smaalen, J., (2007) "Database Requirements for Generalisation and Multiple Representation", In: Mackaness, W.A., Ruas, A., and Sarjakoski, L.T., (eds.), *Generalisation of Geographic Information: Cartographic Modelling and Applications*. Elsevier for International Cartographic Association, pp. 113-136.

NHDPlus, (2008) *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

NRC, (2007) *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, 143 pages.

Regnauld, N., and McMaster, R.B, (2007) "A Synoptic View of Generalisation Operators", In: Mackaness, W.A., Ruas, A., and Sarjakoski, L.T., (eds.), *Generalisation of Geographic Information: Cartographic Modelling and Applications*. Elsevier for International Cartographic Association, pp. 37-66.

Ruas, A., and Duchêne, C., (2007) "A Prototype Generalisation System Based on the Multi-Agent System Paradigm", In: Mackaness, W.A., Ruas, A., and Sarjakoski, L.T., (eds.), *Generalisation of Geographic Information: Cartographic Modelling and Applications*. Elsevier for International Cartographic Association, pp. 269-284.

Sarjakoski, L.T., (2007) "Conceptual Models of Generalisation and Multiple Representation", In: Mackaness, W.A., Ruas, A., and Sarjakoski, L.T., (eds.), *Generalisation of Geographic Information: Cartographic Modelling and Applications*. Elsevier for International Cartographic Association, pp. 11-35.

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

Stanislawski, L.V., Finn, M.P., Starbuck, M., Usery, E.L., and Turley, P., (2006) "Estimation of Accumulated Upstream Drainage Values in Braided Streams Using Augmented Directed Graphs", *AutoCarto 2006*: Vancouver, Washington, June 26-28.

Stanislawski, L.V., Finn, M.P., Barnes, M., and Usery, E.L., (2007) "Assessment of a Rapid Approach for Estimating Catchment Areas for Surface Drainage Lines", *ACSM-IPLSA-MSPS 2007*: St. Louis, Missouri, March 9-12.

Stoter, J.E., (2005) "Generalization within NMA's in the 21st Century", *XXII International Cartographic Conference*: Coruña, Spain, July 11-16.

Töpfer, F., and Pillewizer, W., (1966) "The Principles of Selection: A Means of Cartographic Generalization", *The Cartographic Journal* 3(1): 10-16.

Touya, G., (2007) "A Road Network Selection Process Based on Data Enrichment and Structure Detection", *10th ICA Workshop on Generalisation and Multiple Representation*: Moscow, August 2-3.

Thomson, R.C., and Brooks, R., (2000) "Efficient Generalisation and Abstraction of Network Data Using Perceptual Grouping", *Proceedings of the 5th International Conference on GeoComputation*: Greenwich, August 23-25. Online: http://www.geocomputation.org/2000/GC029/Gc029.htm

Thomson, R.C., and Brooks, R., (2007) "Generalisation of Geographical Networks", In: Mackaness, W.A., Ruas, A., and Sarjakoski, L.T., (eds.), *Generalisation of Geographic Information: Cartographic Modelling and Applications*. Elsevier for International Cartographic Association, pp. 255-267.

USEPA and USDOI, (1999) *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

USGS, (1955) "Map Publication Scales", United States Geological Survey, *Book 1, Part B, Chapter 1 of the Geological Survey Topographic Instructions,* March 1955, 13 pages.

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

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

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

USGS, (2002) *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

USGS, (2003a) *Implementation Plan for The National Map*, (version 1.0), United States Geological Survey. Online: http://nationalmap.gov/report/Implementation_Plan_ver_1.0.pdf

USGS, (2003b) *Streams and Waterbodies of the United States,* United States Geological Survey. Online: http://nationalatlas.gov/atlasftp.html

USGS, (2006) *The National Map, The Nation's Topographic Map for the 21st Century,* United States Geological Survey. Online: http://nationalmap.gov/index.html

USGS, (2007) *Geographic Names Information System – GNIS*, United States Geological Survey. Online: http://nhd.usgs.gov/gnis.html

Verdin, K. L., (1997) "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

Yan H., Li Z., and Ai T., (2006), "System for Automatic Generalization of Topographic Maps", *Chinese Geographical Science* 16(2):165-170.