Stratified Database Pruning to Support Local Density Variations in Automated Generalization of the United States National Hydrography Dataset

Lawrence V. Stanislawski, Science Applications International Corporation
U.S. Geological Survey, Center of Excellence for Geospatial Information Science
1400 Independence Road, Rolla, MO 65401, USA
lstan@usgs.gov

Barbara P. Buttenfield, Department of Geography
UCB 260, University of Colorado—Boulder, Boulder, CO 80309, USA
babs@colorado.edu

Michael P. Finn, U.S. Geological Survey
Center of Excellence for Geospatial Information Science
Denver Federal Center, Box 25046, Mail Stop 510, Denver, CO 80225, USA
mfinn@usgs.gov

Keven Roth, GIS Consultant, Former USGS NHD Coordinator
Victor, MT 59875, USA
kevenroth@yahoo.com

Abstract
Introduction: The U.S. Geological Survey's (USGS) Center of Excellence in Geospatial Information Science (CEGIS) is conducting generalization research in cooperation with the University of Colorado—Boulder and Pennsylvania State University to support display and delivery of *The National Map* and other USGS geospatial data at multiple scales. This paper focuses on generalization of the National Hydrography Dataset (NHD).
Objectives: Objectives of this research are to develop methods to sub-select, or prune, features from the multi-scale high-resolution (HR) NHD layer to automate generating a multiple representation database (MRDB) and simplify, or further generalize, remaining features for cartographic display. Methods should maintain hydrographic network connectivity and local density variations that typify physiographic or climate variations.
Methodology: A four-subbasin region of HR NHD data in Iowa, having obvious natural network density variations, was pruned to four smaller scales—1:100,000 (100K); 1:500,000 (500K); 1:2,000,000 (2M); and 1:10,000,000 (10M). Methods employ a stratified database pruning approach that partitions HR NHD data based on localized network densities and prunes features to densities appropriate for smaller map scales based on reach codes and upstream drainage area (UDA) estimates. Database enrichment,

partitioning, pruning, and subsequent generalization operations are completed through automated processes. Validation processes, such as the coefficient of line correspondence (CLC), that automate comparisons of pruned data with existing lower resolution benchmark data also are being developed and employed.
Results: For the tested region, stratified density pruning produced reasonably good results that maintain local density variations where needed and mimic data conditions existing in other sources of available hydrographic data. The CLC is 0.88 when HR NHD pruned to 100K is compared to 100K NHD. Automated stratified pruning to 500K or smaller scales produces some network discontinuities near partition boundaries, but stratification may not be needed for these smaller scales.
Conclusions: Results indicate stratified pruning can be guided by the Radical Law, but with tailoring that handles large and small scale data conditions. Future research should eliminate the creation of isolated sub-networks during stratification and further compare generalization operations performed after data pruning.

Introduction

The process of generalizing detailed geospatial data to render less detailed data and cartographic products can be enhanced substantially with a MRDB. One method to build a MRDB is to assign a level of importance, or prominence, to each feature in a theme's base layer containing the most detailed feature representations. Subsequently, a hierarchical set of layers can be sub-selected from the base layer and stored in the database. A MRDB built in this manner can enhance cartographic generalization if the level of prominence for each sub-selected layer is tailored appropriately for a specific range of map scales. Ensuing generalization operations and symbology changes associated with base-map generation can be focused efficiently on layers needed for each desired scale range.

The NHD is a comprehensive vector database of surface-water features for the United States. The NHD includes an address system, known as reach codes, for linking to other environmental databases, such as stream flow, water quality, fish populations, and habitat classifications. The NHD database includes three levels of detail—medium, high, and local resolution; however, only the medium-resolution (MR) and high-resolution (HR) layers are populated fully and attributed with reach address and ancillary hydrographic information. Features in the MR layer are compiled from 1:100,000-scale (100K) source hydrographic data. The local-resolution layer includes 1:4,800-scale data for the state of Vermont only.

Initially, the HR NHD layer was compiled from 1:63,360-, 1:24,000-, and 1:20,000-scale data for Alaska, the conterminous 48 states, and Puerto Rico and the Virgin Islands, respectively. During recent years (2007-2009) maintenance has focused on correcting features and collecting more detailed features for the HR layer. In some cases, 1:4,800-scale source data are being included in the HR layer; consequently, the HR NHD layer is a

multi-resolution, or multi-scale, layer, generally with the most current and accurate hydrographic data available. Development of this multi-scale layer has been conducted, anticipating that database pruning will satisfy the needs to generate uniformly-scaled data.

Objectives This paper demonstrates a stratified database pruning approach for automatically generating a MRDB for the NHD of the United States. Database pruning operations must be applied to the high-resolution layer to support multi-scale delivery of NHD data and subsequent cartographic generalization operations.

NHD pruning extracts network features that produce a drainage density appropriate for a target map scale, with two goals. First, the pruning process must preserve local variations in drainage density. Pruning a region to a single drainage density tends to normalize, or homogenize, the density of network features in the region. It also subdues local density variations that typify physiographic or climate variations, which are important traits for hydrographic data. Hence, pruning network and connected polygon features is enhanced by enforcing stratified network density estimates that better reflect climate and terrain variations than a single estimate.

The second goal relates to protecting network topology. Stream channel connectivity is a characteristic that is geographically important for modeling and analysis, and cartographically important for preserving visual logic on smaller scale maps of generalized hydrography. Pruning procedures must protect connectivity between stream reaches, and among channels connected to hydrographic areas and water bodies for every subset in the database hierarchy; therefore, the action of pruning any given stream reach involves eliminating all upstream reaches, which is accomplished by using logically-enriched attribute tables; consequently, pruned data preserve the network overall.

Methodology

The USGS has been working on an automated pruning strategy that eliminates less prominent features from the HR NHD layer, and furnishes data densities appropriate for smaller map scales (Stanislawski 2009). The level of prominence used in the pruning process is based on NHD reach codes, UDA estimates, and criteria derived from NHD data-collection standards. HR flowline features are enriched with catchment area and UDA estimates before pruning.

Study Area Stratified pruning is demonstrated on a four-subbasin area in the central Iowa, a region in the midwest United States, which covers about 20,172 square kilometers (km^2). In this part of the country, climate is humid and hydrography diverse. The study area includes the drainage area for the Raccoon River and the Middle Des Moines River upstream from the confluence with the Raccoon River. The four subbasins straddle two

physiographic regions where a glacial lake borders a till plain. The stream network clearly captures the glacial history of the area. The hydrography shows a clear distinction between glaciated areas and the area known as the dissected till plains. More recently glaciated areas in the glacial lake region show fewer stream channels and relatively poor drainage. More drainage ditches exist in this area indicating a lack of well-developed drainage. The stream network is still developing in this area and eroding the terrain. The dissected till plains, which occupy a large part of Iowa, were developed during an earlier glacial period; more time has passed, resulting in a well developed, more dense drainage network.

Partitioning To perform stratified pruning, the study area was subdivided into two classes or partitions, based on drainage density. Partitioning has been used previously to separately handle areas with different data densities during generalization (Bobzien et al. 2008, Chaudhry & Mackaness 2008, Stanislawski 2009). Catchments were automatically estimated for each HR NHD flowline feature, where a catchment is the area on the ground that drains into the associated flowline (Stanislawski et al. 2007). Catchments for the HR flowlines were automatically clustered based on catchment drainage density, where drainage density for a catchment equals the length of the associated flowline feature divided by the catchment area. A histogram of catchment densities for the 48 contiguous states was subdivided into five density classes (fig. 1), and associated density classes were assigned to each catchment in the study area. Study area catchments were then clustered by removing edges between catchments with the same density class. Any cluster smaller than 10 km^2 was assigned the density class of the next largest adjacent cluster greater than 10 km^2. This clustering process was repeated until all clusters were greater than 10 km^2.

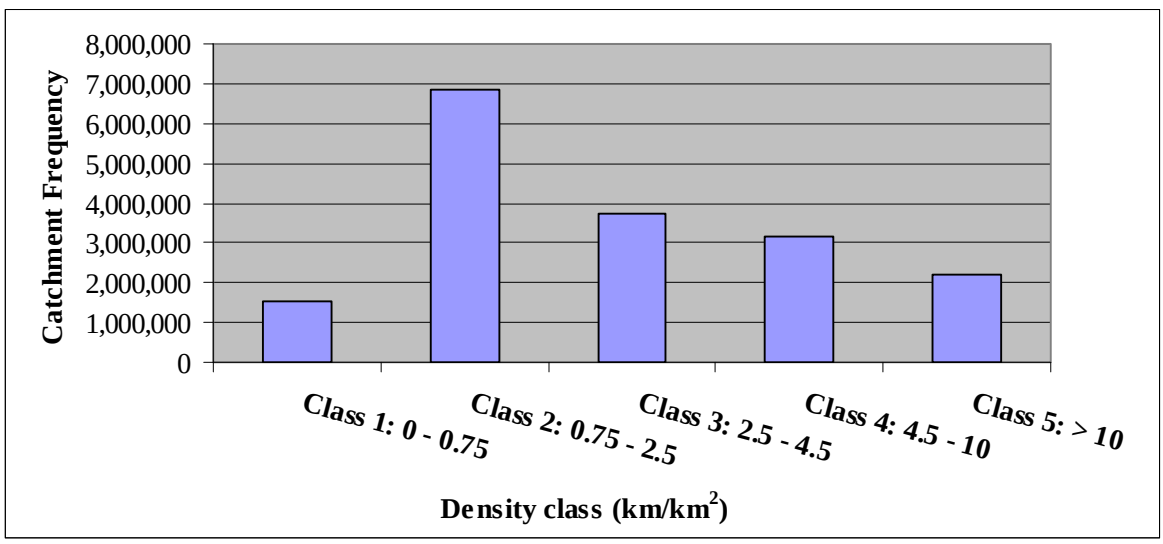

Figure 1. Frequency histogram of catchment density classes for Thiessen-polygon derived catchments in the 48 contiguous United States.

Feature Pruning Relations characterizing hydrographic network drainage density within the study area from the representative fraction for map scale were estimated from five hydrographic data sources—HR NHD, MR NHD, river reach file version RF1 (Horn et al. 1994), digital line graph (DLG) from the National Atlas (U.S. Geological Survey 2003), and vector digital data from the National Atlas (U.S. Geological Survey 2006)—that respectively have the following scales: 1:24,000 (24K), 100K, 1:500,000 (500K), 1:2,000,000 (2M), and 1:10,000,000 (10M). Before estimating target drainage densities from the lower resolution datasets, lengths of the lower resolution network lines were expanded by a length expansion factor to account for the extra detail, or granularity, included in the 24K lines. Thirty, ten, ten, and three confluence-to-confluence sections that match 24K network lines were respectively identified for the 100K, 500K, 2M, and 10M to estimate a length expansion factor for each lower resolution dataset. For each scale, a ratio of the 24K length to the length at that scale was computed for each matching confluence-to-confluence section. The length expansion factor for a scale is the average ratio of all matching confluence-to-confluence sections at that scale. A target drainage density was estimated for each of the four smaller scales by multiplying the associated expansion factor by the sum of the lengths of linear hydrographic features from the associated data source located within the study area, and dividing this product by the sum of the HR catchment areas within the study area. Subsequently, HR NHD features in the study area were pruned to the estimated target drainage densities for the four smaller scales.

A coefficient of line correspondence (CLC) with the 100K (MR) NHD layer as the benchmark validates the pruning process for the 100K level of detail. The CLC is computed between two linear datasets and estimated as the ratio of the length of matching features to the length of matching features, plus the length of omission and commission errors (Stanislawski 2009). In addition, a CLC value comparing the HR NHD pruned to 100K and the 100K NHD was estimated for each 0.15-degree cell within the study area.

In addition, numbers of confluence-to-confluence segments, associated lengths, and stream orders are compared between linear features in pruned and validation datasets at associated scales. Similar comparisons are reported for polygon features.

Results

Length expansion factors of 1.08, 1.16, 1.25, and 1.50 were estimated for the 100K, 500K, 2M, and 10M scales, respectively. Partitioning of the HR catchments produced two density partitions having 0.46 and 1.41 km/km^2 of flow-directed HR network features in the high and low density partitions, respectively (fig. 2).

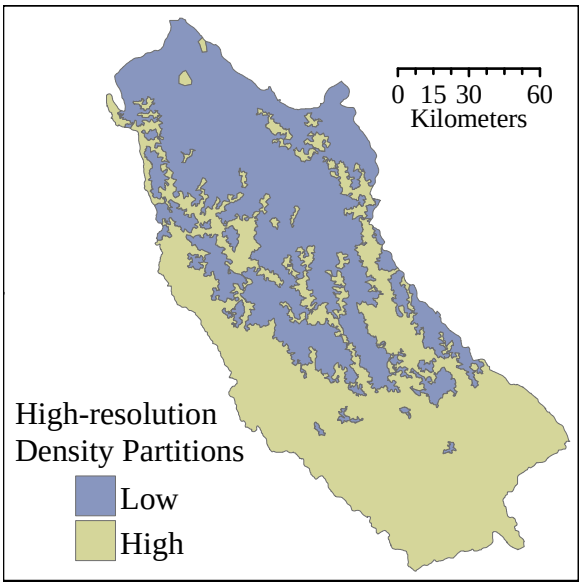

Figure 2. High-resolution (HR) density partitions generated from HR catchments. Network feature densities are 0.46 and 1.41 km/km^2 in the low and high density classes, respectively.

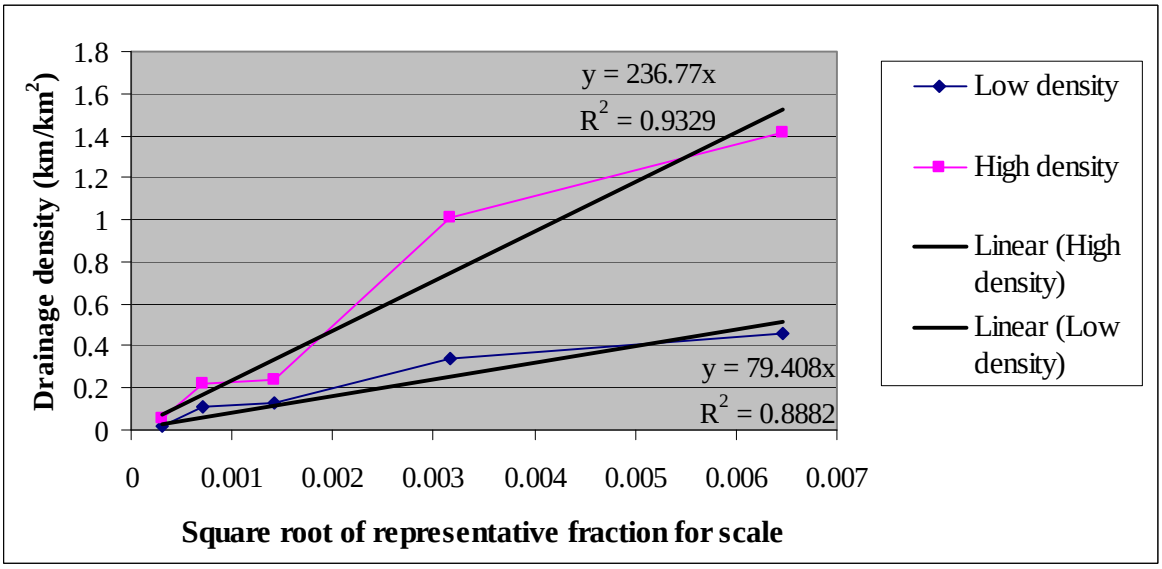

Figure 3. Linear relations estimating drainage density from the square root of the representative fraction for scale for hydrographic network features from source scales of 1:24,000, 1:100,000, 1:500,000, 1:2,000,000, and 1:10,000,000. Regression relations are shown separately for the low and high density partitions within the four- subbasin Iowa study area.

Length expanded drainage densities for each of the four lower resolution datasets range from 0.02 to 0.34 km/km^2 in the low density partition and from 0.05 to 1.01 km/km^2 in the high density partition. Regression equations characterizing the linear relations between drainage density and the square root of the representative fraction for map scale are shown for each density partition in figure 3. Drainage density estimates from the linear relation for both density partitions range in error from 12 to 50 percent for the three smaller scales and from 8 to 27 percent for the two larger scales. These errors may be too big for more sensitive, larger scale hydrographic analyses. Non-linear, such as logarithmic, relations may furnish more suitable fits for the more sensitive, larger scale data.

For comparison, the proportions of the total HR network length that should be retained for the smaller scales based on the general form of the Radical Law (Töpfer and Pillewizer 1966) were computed, along with associated drainage densities and regression relations with map scale (fig. 4). The general form of the Radical Law indicates that the proportion of the larger scale data to retain is the square root of the ratio of the source and target scales (Töpfer and Pillewizer 1966). Regression relations derived for Radical Law densities are nearly the same as those derived from available hydrographic data densities. This result

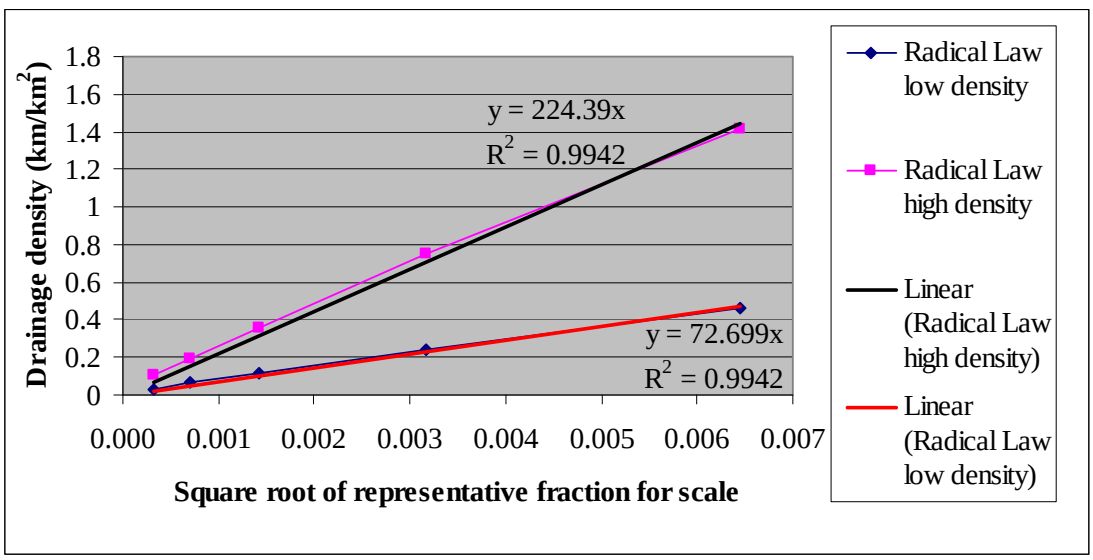

Figure 4. Drainage densities expected from the general Radical Law (Töpfer and Pillewizer 1966) given starting densities from the 1:24,000-scale NHD flowlines in the high and low density partitions of the four-subbasin Iowa study area. Regression relations estimating Radical Law-expected drainage density from the square root of the representative fraction for scale at 1:24,000, 1:100,000, 1:500,000, 1:2,000,000, and 1:10,000,000. Regression relations are shown separately for low and high density partitions.

indicates smaller scale NHD network densities for MRDB generation can be reasonably estimated from the Radical Law if sufficient density partitions can be established for the HR NHD.

Feature pruning the HR NHD with partitioning to the 100K scale produced 6230 confluence-to-confluence sections of an average length of about 2.30 km, which is nearly the same as the MR NHD, which has 6220 confluence-to-confluence sections averaging 2.25 km in length. The overall CLC comparing the HR NHD pruned to 100K with the 100K NHD is 0.88, with a 0.08 omission proportion and 0.04 commission proportion. More omission errors are caused by undirected flowline features included in the 100K NHD.

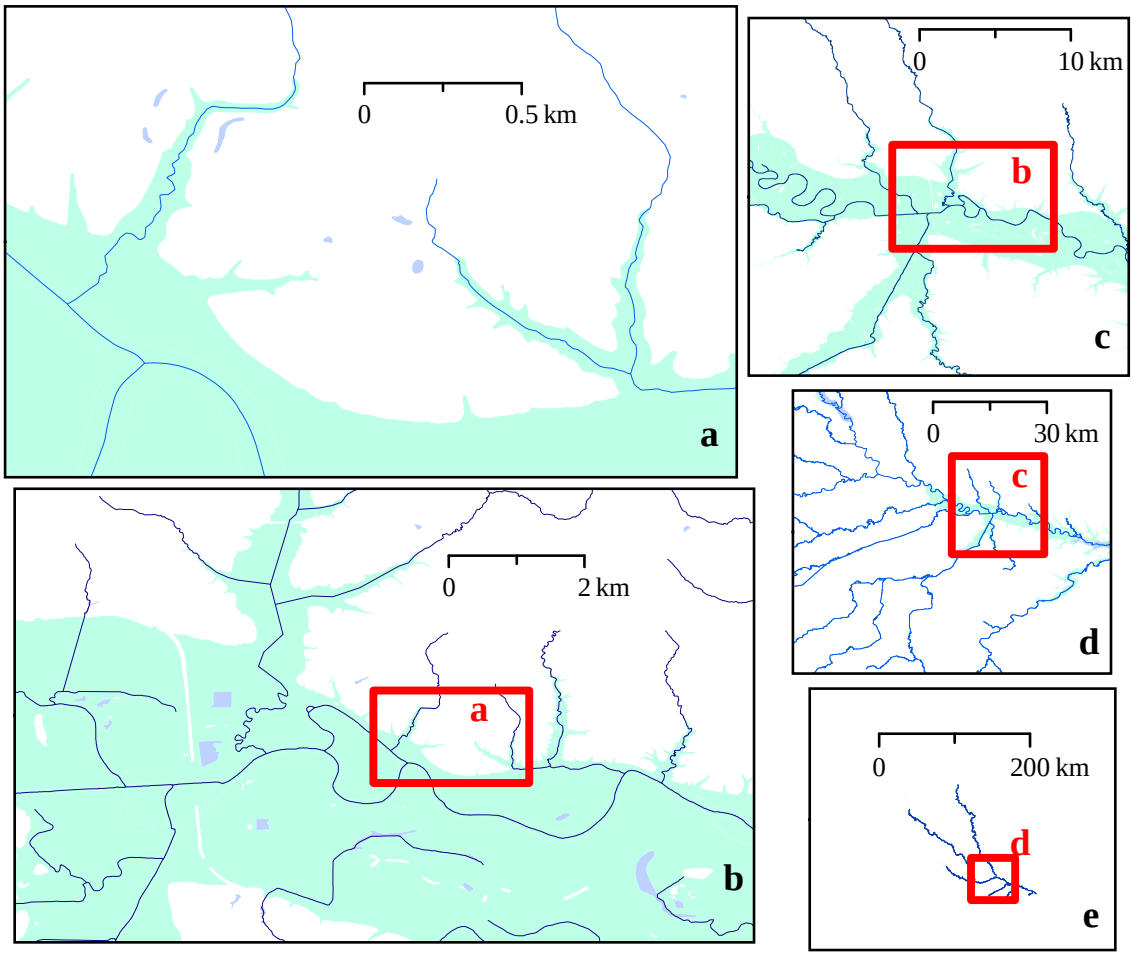

Figure 5. Parts of the four-subbasin study area in Iowa showing results of feature pruning the (a) high-resolution NHD to (b) 1:100,000, (c) 1:500,000, (d) 1:2,000,000, and (e) 1:10,000,000 scales. Red boxes show map extent of next larger scale.

Some small isolated sub-networks were improperly generated when the HR NHD was pruned to 500K, 2M, and 10M scales with partitioning. Topological network inconsistencies between scales can affect hydrographic network studies and should be avoided, if possible. Therefore, pruning without partitioning was also tested for these scales. Pruning the HR NHD without partitioning to the three smaller scales maintains network connectivity in all cases. The number of confluence-to-confluence features maintained in the 2M and 10M data after pruning is about twice the amount expected based on Radical Law (Töpfer and Pillewizer 1966). But, many of the extra sections maintained in the smaller-scaled pruned data consist of small braids that should be eliminated through subsequent feature simplification or generalization operations. Parts of the pruned datasets are illustrated in figure 5.

Generalization

As an experiment, we generalized the 100K-pruned NHD data to the four target scales (100K, 500K, 2M, and 1:7,500,000 (7.5M)). Generalization in this case involved selection, elimination and simplification; no aggregation or displacement was applied (fig. 6). The process was automated using ArcGIS ModelBuilder.

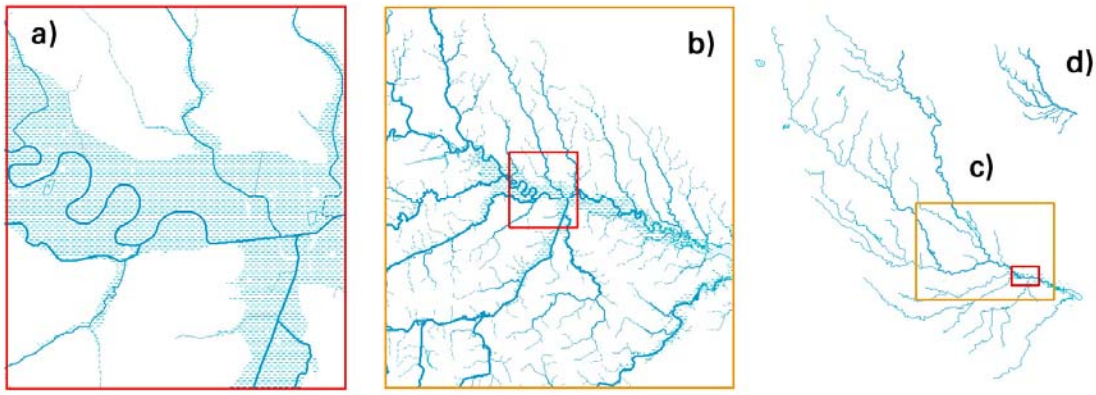

Figure 6. 24k NHD data pruned to 100K and generalized to the four target scales: a) 100K; b) 500K; c) 2M; and d) 7.5M. The red box shows the 100K footprint; the orange box shows the 500K footprint. Methods for generalization are described in the text. Stream channels are symbolized in a quantitative progression of upstream drainage area; dashed channels are intermittent. Lakes and ponds are solid blue; inundation areas are shown with a blue texture.

One objective of the experiment was to determine if stratified pruning can provide a viable source for subsequent generalization. Stoter et al. (2004) distinguish “incremental” generalization (generalizing data in stages from a large-scale source to a smaller destination scale) from “star” generalization (creating the destination scale data in a single jump, such as generalizing from 24K down to 500K with a single processing pass). In the short scope of this paper, only incremental generalization from a 100K source is reported on.
A second objective was to explore how the two density class differences described above could be preserved during generalization, and at what target scale the density class differences become insignificant. To our knowledge, empirical results of differentially generalizing hydrographic data based on local density has not been reported widely in the literature, with the exception of Brewer et al. (2009) and Stanislawski (2009). This work suggests that at scales below 500K the differences between density classes are no longer significant, insofar as data generalization is concerned. It’s difficult to abstract a guiding principle from a single data set, of course, but this result is reported to inform others working on automatic selection of tolerance parameters.

A third objective of the experiment was to compare reduction of detail on the basis of stratified density partitioning, as described earlier, against the same approach but with a more thorough manual assignment of density classes that assigns complete sub-networks to a density class. The second, manual approach always preserves topology in the generalized stream network. As previously mentioned, network topology is sometimes broken at density class boundaries by the simpler automated partitioning process. However, stratified partitions can be established automatically for a large region (potentially, the entire contiguous United States), and thereby reduce labor and production costs. The goal is to establish metrics to compare the two outcomes, beginning with a single data source.

The workflow for each target scale varied somewhat. The target scale of 500K involved the most complicated workflow, selecting flowlines on density class, selecting on UDA thresholds (described below) and then reducing detail using Wang and Muller’s (1998) Bend Simplify algorithm. NHD waterbodies include, for example, lakes, ponds, swamps, and reservoirs. These were selected on size and then simplified. NHD areas include inundation areas, and hydrographic channels such as streams, ditches, rapids, and levees which are customary to map using polygons at large and intermediate scales, and as linear features at small scales. After eliminating submerged streams, the NHD area polygon boundaries were simplified.

The two density classes produced through partitioning were used. Threshold values for UDA were established using an exponential model based on a principle similar to the Radical Law (Töpfer and Pillewizer, 1966), which states that the number of features retained at smaller scales should be a function of the area covered by those features. The

Radical Law was intended for number of map symbols and size of map area, but easily can be extended to ground area or other metrics; for example, how much stream channel length to retain (fig 4).

The initial pruning used data from a source scale of 24K for a target scale of 100K, and the Radical Law dictates this scale change should retain 48% (that is, the square root of (24/100) of the detail. However, the 100K pruning retains about 61% of the original features (22,243 in the 24K source database and 13,493 in the target database) and 71% of stream length (14,332 km retained in the 100K pruning), which is greater than the density expected through the Radical Law because the 100K pruning was tailored to the density of the MR NHD. Using the 100K pruning as source for a 500K target scale, the Radical law dictates retaining 45% of the 100K detail. A trial generalization to 500K retaining only 45% (6,071) of the 100K features produced a network that was much too sparse, both analytically and cartographically. Refining the 500K generalization to 64% of stream channel length produced a stream network with 9,279 km stream channels, preserving many additional upper tributaries (fig. 6). For analytical purposes, these headwaters carry relatively large amounts of channel length, and this factor could be significant in many types of hydrologic models.

It may be that a third coefficient could be added to the original Radical Law computation to account for data use in cartographic display or in numerical modeling. It's not realistic, of course, to derive such a coefficient based on a single experiment, and this idea remains a direction for future research.

Conclusions

This paper demonstrates a stratified pruning process that preserves network connectivity and yields localized density variations in hydrographic data to reflect different hydrographic conditions for two physiographic regions in a four-subbasin area in Iowa. CLC comparing the HR NHD pruned to 100K with the 100K NHD is 0.88. Results suggest pruning criteria for density strata can be estimated through some form of the Radical Law, which should handle large and small scale data conditions.

The goal of this research is to develop a valid approach to prune the multi-scale HR layer of the NHD to a range of scales, effectively converting the NHD database to an integrated MRDB that supports cartographic generalization, mapping, and essential hydrographic analysis needs of the user community. As demonstrated, stratified pruning in conjunction with tailored generalization and symbolization operations can fulfill these requirements, but implementation on the dynamic, multi-resolution NHD layer will be a challenge.

Acknowledgments

Dr. Buttenfield's work forms a portion of USGS CEGIS Cooperative Agreement 4121HS0029 with the University of Colorado. Symbolization styles for simplified data courtesy of Professor Cindy Brewer, Penn State University.

References

Bobzien, M, Burghardt, D, Petzold, I, Neun, M &Weibel, R 2008, 'Multi-representation databases with explicitly modeled horizontal, vertical, and update relations', *Cartography and Geographic Information Systems*, vol. 35, no. 1, pp. 3-16.

Brewer, CA, Buttenfield, BP & Usery, EL 2009, 'Evaluating generalizations of hydrography in differing terrains for *The National Map* of the United States', *Proceedings, 24th International Cartographic Congress*, Santiago, Chile, November 2009 (forthcoming).

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

Horn, RC, McKay, L & Hanson, SA 1994, 'History of the U.S. EPA's river reach file: a National Hydrographic Database available for ArcInfo applications', in *Proceedings of the Fourteenth Annual ESRI User Conference*, Environmental Systems Research Institute, Redlands, CA.

Stanislawski, LV 2009, 'Feature pruning by upstream drainage area to support automated generalization of the United States National Hydrography Dataset', *Computers, Environment and Urban Systems*, vol. 33, no. 5, pp. 325-333.

Stanislawski, LV, Finn, MP, Barnes, M, & Usery, EL 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, JE.,Kraak, MJ & Knippers, RA 2004, 'Generalization of framework data: A research agenda', *Proceedings of the ICA Workshop on Generalisation and Multiple Representation*, Leicester, UK. 14 pp.

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

U.S. Geological Survey 2006, North American Atlas – Hydrography, Accessed July 23, 2009, at URL: http://nationalatlas.gov/atlasftp-na.html.

U.S. Geological Survey 2003, Streams and Waterbodies of the United States, Accessed June 14, 2005, at URL: http://nationalatlas.gov/atlasftp.html.

Wang, Z & Muller, JC 1998, 'Line generalization based on analysis of shape characteristics', *Cartography and Geographical Information Systems*, vol. 25, no. 1, pp. 3 - 15.