CO-361

A RASTER ALTERNATIVE FOR PARTITIONING LINE DENSITIES TO SUPPORT AUTOMATED CARTOGRAPHIC GENERALIZATION

STANISLAWSKI L.(1), BUTTENFIELD B.(2)

(1) United States Geological Survey, ROLLA, UNITED STATES ; (2) University of Colorado-Boulder, BOULDER, UNITED STATES

Abstract

An automated data-partitioning algorithm that uses a raster-based approach is described and demonstrated in this paper. The algorithm delineates an area of interest (AOI) around a set of line features, and then subdivides the AOI into line density partitions based on predefined density classes. In this case, the primary objective of partitioning is to enable stratified feature pruning, which supports automated cartographic and database generalization. The approach is demonstrated on hydrography and transportation vector data and compared, visually and metrically, to alternative partitioning approaches, comparing vector-based to raster-based and manual to automated approaches. Comparisons validate results and measure the relative efficiency of the process. Results indicate the raster algorithm is as effective as alternative approaches at partitioning hydrographic lines for stratified pruning. In addition, the algorithm generated partitions for transportation data that are consistently defined over multiple adjacent data subdivisions. Although additional testing is required, the approach appears more efficient than an alternative automated clustering method, and it is well-suited for implementation through parallel operations.

1. Background and Objectives

Although various degrees of manual intervention continue to be required for interactive editing, inspection, or parameterization, automated cartographic generalization is becoming a reality. Procedures for automated cartographic generalization that produce smaller scale maps or derive intermediate levels of detail from higher resolution representations are being implemented by several National Mapping Agencies (NMAs). For instance, the National Land Survey of Finland is using automated generalization to produce 1:100 000-scale (100K) level of detail from 1:5 000 to 1:10 000-scale representations (Pätynen and Ristioja 2009). Turkey has developed and implemented KartoGen software to automate generation of 1:50 000- and 100K maps from 1:25 000-scale base data (Simav et al. 2010). Likewise, several other NMAs are developing and implementing a variety of cartographic generalization software systems that approach full automation wherever possible (Stoter et al. 2009, Stoter et al. 2010, Touya et al. 2010, Renard et al. 2010).

Automated generalization and associated database enrichment often involve intensive computer processing algorithms that require logistical data management for use on large databases—such as those housing the primary geospatial data themes of The National Map for the United States Geological Survey (USGS 2006). Data partitioning subdivides large datasets into smaller units that can be handled by processing algorithms. Aside from subdividing data into manageable sections, several researchers suggest that data partitions for automated generalization should form context-based geographic spaces that allow the proper application of site-specific generalization operations (Bobzien et al. 2008, Chaudhry and Mackaness 2008a, Stanislawski et al. 2009, Touya 2010, Touya et al. 2010). Through similar reasoning in earlier work, Burghardt and Neun (2006) propose using collaborative filtering to determine the best generalization sequence to apply to features from various geographic landscape or landuse characteristics.

Partitioning to facilitate automated generalization may be performed on individual or multiple data themes, and partitioning methods depend on the generalization strategy (Bobzien et al. 2008, Chaudhry and Mackaness 2008a). For instance, Chaudhry and Mackaness (2008b) clustered buildings around road nodes to estimate settlement boundaries for developing partonomic relations in a multiple representation database and for generalization processing. Stanislawski (2009) partitioned hydrography by flowline catchment areas to form density partitions that regulate density variations when pruning a hydrographic network.

This paper presents a raster-based alternative for partitioning linear features by line density. The approach is demonstrated on two vector themes of network data: hydrography and transportation. The objective of partitioning is to facilitate stratified pruning for cartographic and database generalization. Stratified pruning enables the maintenance of density variations that reflect local geographic content (Buttenfield et al. 2010, Stanislawski 2009, Stanislawski et al. 2009, Stanislawski et al. 2010).

The raster-based partitioning algorithm is described in the next section. An analysis of the approach tested on a selection of hydrographic networks from the United States National Hydrography Dataset (NHD) follows, along with a summary of tests on the transportation network around St. Louis, Missouri. Summary statements and future plans are discussed in the final section.

2. Approach and Methods

Several methods of spatial density estimation are available for geographic analyses. Silverman (1986) describes various approaches, including kernel density estimation (KDE). Raster-based KDE can be applied to point or line features in a variety of problem domains. Recent uses of KDE on point features are described by Borruso (2003), Downs (2010), and Lüscher and Weibel (2010). Density can be estimated by other methods as well. In a non-raster approach, Xiang et al. (2008) used Delaunay triangulation with feature vertices to estimate density for map generalization purposes.

User-specified parameters—such as cell size, search radius, and neighborhood attenuation function—control raster-based operations. Consequently, the parameters for raster-based algorithms must be specifically tailored for each use. In this study, several tailored raster operations are designed in a sequence to automatically delineate line density partitions that can support subsequent generalization operations. The goal of the algorithm is to generate a smooth density surface that can be subdivided into two to six density partitions, with any patch within a partition spanning a minimum of 15 square kilometers (km2). The sequence of raster operations and associated parameters were tailored through trial and error to efficiently partition sections of the vector hydrography and transportation themes of The National Map.

The raster-based partitioning algorithm is implemented in the ESRI ArcGIS environment through a Python geoprocessing tool. KDE is not used in the algorithm. The program completes three tasks for a target set of line features: delineate the area of interest (AOI), delineate generalized line density partitions within the AOI, and estimate average line density of each partition.

The AOI is the area that encloses the target line features, which is similar to, but a little larger than a convex hull. The role of the AOI is to provide a compact envelope within which densities may be meaningfully estimated. The AOI is delineated through raster and vector processes as follows (figure 1):

- a) Rasterize target lines using a 300 meter (m) cell size;
- b) Convert raster lines to polygons and build a 2 800 m buffer around the polygons, and then buffer this area inward 2 600 m to generate a boundary polygon around the target lines that extends about 200 to 300 m beyond the extent of the target lines.
- c) Remove interior holes (gaps) that may exist within the boundary polygon due to large spaces within the AOI that do not contain any line features.

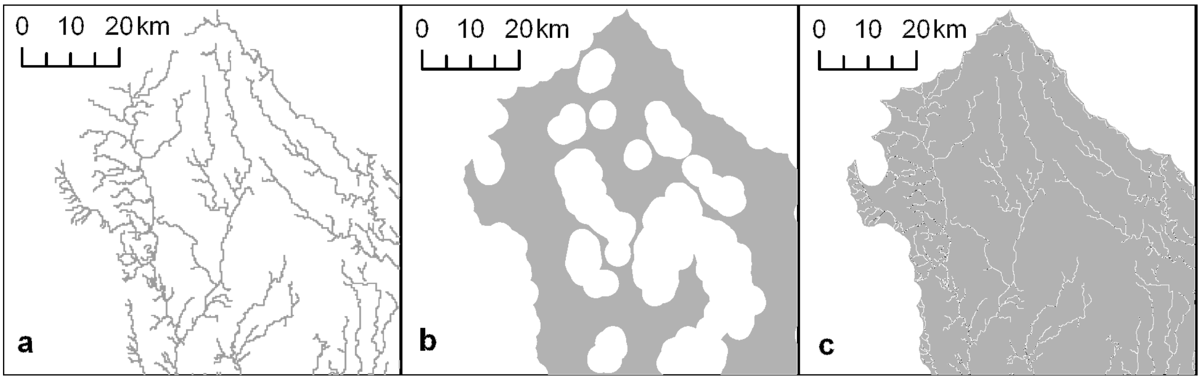

Figure 1. Steps in delineation of area of interest (AOI). Section of rasterized target lines (a), section of AOI (gray area) with holes from buffer process (b), and AOI (gray area) with holes removed (c) (target vector lines overlain).

Afterward, target lines are converted to a 300-m resolution line density grid that covers the AOI. Line density for each cell is estimated, in kilometers per square kilometer (km/km2), as the sum of the length of all lines within a 1 500-m radius of the cell center divided by the area of a circle with a 1 500-m radius. The line density grid is smoothed through a focal-mean process that assigns to the target cell the average density of a 4-by-4 cell window centered on the target cell. Next, the smoothed line density grid is reclassified into predefined density categories that are tailored to follow conspicuous density variations (or breaks) for the target line theme (eg., hydrography) in the AOI. Manual and automated analysis techniques assist the selection of density class breaks. The reclassified density grid is further generalized to remove cell clusters of the same density class that are smaller than 15 km2, and to extend (typically greater valued) density class clusters that end near the boundary of the AOI to the AOI boundary. The latter is necessary because density estimates fade (decrease) near the edge of the AOI because adjacent data are not represented in the target lines.

The grid of density classes within the AOI is converted to polygons and intersected with the target line data to transfer the density classes to the lines and compute average line density values for each class.

3. Tests on High-Resolution NHD Flowlines

The line-density partitioning algorithm was tested on two datasets of high-resolution (HR) NHD flowlines, compiled at 1:24 000 (24K). The NHD is a comprehensive set of vector data (points, lines, and polygons) representing surface-water features within the United States (USGS 2000), and it is distributed in an ESRI geodatabase format. The NHD flowline feature class includes linear features oriented in the direction of surface-water flow (where possible), which represent streams, rivers, canals, ditches, and pipelines, along with artificial paths that complete the surface drainage network where it is broken or interrupted by polygonal features.

3.1 Piceance-Yellow subbasin: a dry landscape

The first test dataset was the Piceance-Yellow subbasin in the Rocky Mountains, which contains 3 768 HR flowlines (figure 2). Stanislawski and Buttenfield (2010) manually partitioned this subbasin for stratified pruning purposes (figure 2a). Densities were averaged for the three manual partitions, and these formed the basis for tailoring density class breaks for the following two approaches. Three density classes with ranges, in km/km2, from 0 to less than 1.0, from 1 to less than 1.5, and from 1.5 to less than 2.75 defined the partitions for the raster algorithm (figure 2b). A third, vector-based approach (figure 2c) derived partitions by clustering Thiessen-polygon-derived catchments for the target lines (Stanislawski 2009).

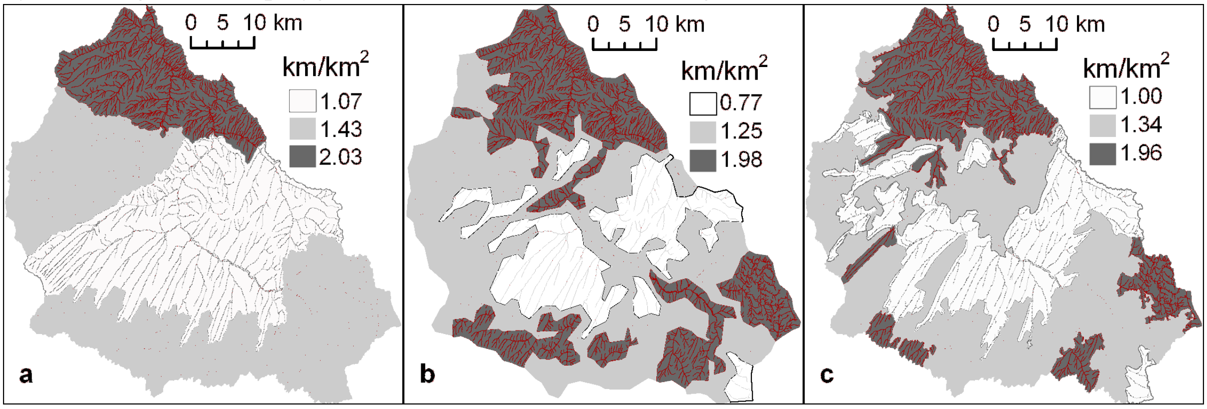

Figure 2. Line density partitions estimated for the high-resolution NHD flowline features in the Piceance-Yellow subbasin in Colorado through (a) manual delineation, (b) raster-based algorithm, and (c) vector-based catchment clustering algorithm. Average line density for each partition in kilometers per square kilometer (km/km2) is shaded by density class.

As the figures show, density partitions produced by each method have similar average density values, and similar proportions of coverage. They differ in terms of partition compactness and contiguity. Visual inspection indicates that the raster and vector approaches generate partitions that reflect a finer granularity of line density distinctions exhibited in the data than does the manual approach, but the manual partitions more aptly follow primary sub-watershed boundaries. Also, it appears that the vector and manual approaches miss some localized high-density pockets in the flowline data.

Partitions generated by the three methods were compared using the coefficient of areal correspondence (CAC) (Taylor 1977). CAC is a ratio comparing two polygon datasets to establish what proportion of respective polygons match, relative to polygons that have been omitted from one or the other dataset. Three CAC values were computed for the subbasin, comparing all possible pairings of manual, raster, and vector based partitions. In each case, the CAC for the entire AOI was determined from the union of two sets of partitions, and computed as the area of the union having the same density class from the two approaches, divided by the sum of all area in the union.

CAC comparisons between the manual and raster partitions and between the manual and vector partitions produce values of 0.59 and 0.64, respectively. This suggests vector partitions match the manual partitions about 5 percent better than do the raster partitions. Comparing the raster partitions to the vector partitions results with a 0.67 CAC, indicating about 67 percent of the subbasin was assigned the same density partition through the raster and vector approaches. The raster approach adequately delineates the AOI, but it includes a slightly larger area than the subbasin boundary, which also slightly reduces CAC values (by about 1 percent).

Another advantage of the raster-based approach is computational speed. Partitioning the 3 768 line features of the Piceance-Yellow Colorado subbasin using the raster approach took about 2 minutes [Windows XP with 2.66 gigahertz and 3.25 gigabytes of random access memory (RAM)], which is about 40 percent faster than the vector approach.

3.2 Four subbasins in Iowa: a glaciated landscape

The second test NHD dataset consists of 21 926 HR flowline features from four adjacent subbasins in Iowa. The four subbasins straddle two physiographic regions where a glacial lake borders a till plain. The density of the hydrography shows a natural distinction between glaciated areas and more densely dissected till plain areas. Two sets of partitions were generated through the raster and vector approaches using three density classes ranging, in km/km2, from 0 to less than 0.75, from 0.75 to less than 2.50, and from 2.50 to less than 4.50. Density class breaks were derived in an earlier study of this area (Stanislawski et al. 2009).

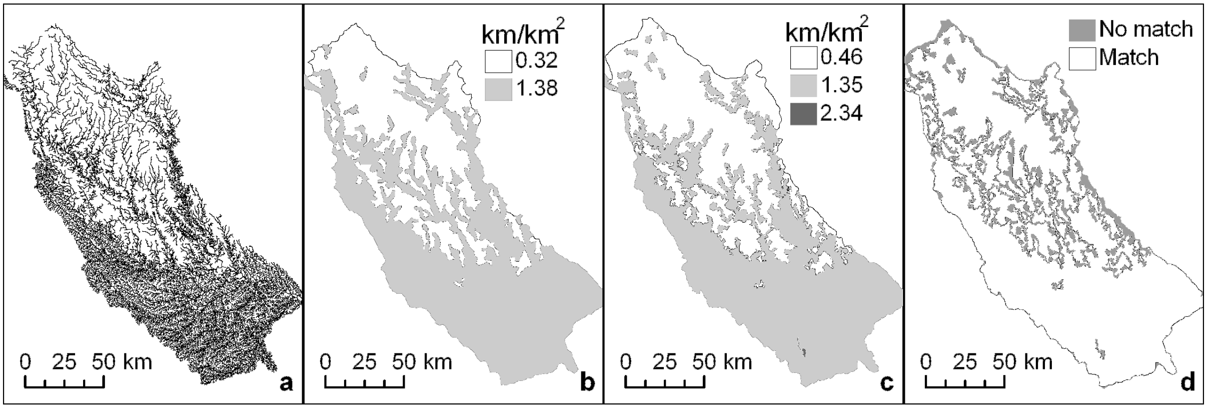

Figure 3. Line density partitions estimated for high-resolution NHD flowline features (a) from four subbasins in Iowa that straddle the boundary between glacial lake and glacial till subsurface material. Partitions generated through the raster-based (b) and the vector-based catchment clustering (c) algorithms were overlaid and matching and mismatching areas were identified (d). Average line density for each partition in kilometers per square kilometer (km/km2) is shaded by density class (b and c).

Partitions generated from the raster and vector approaches for the flowlines in the four Iowa subbasins are shown in figure 3, along with matching and mismatching areas identified between the two approaches. The CAC summarizing the comparison between the two sets of partitions is 0.86, indicating that the two approaches yield very similar partitions. This may be due to the geographically distinct terrain variations formed from glacial processes in the region, along with well-tailored density class breaks that should be apparent in any density partitioning strategy. It is obvious from figure 3d that mismatching areas are found near partition boundaries, which is likely due to the inherently different ways the boundaries are established between the two approaches. Also, the vector approach delineates a small high-density partition, which has an average density below the associated class break. This anomaly does not exist in the raster approach, indicating the raster approach may be more reliable. As in the Colorado subbasin, the boundary for the AOI from the raster approach is nearly identical to the subbasin boundaries.

Results indicate stratified pruning of flowlines in these four subbasins could be accomplished using either raster or vector generated partitions with minor differences; however, the raster approach took about 5 minutes to process, which is nearly five times faster than the vector approach. Processing times estimated for the vector approach include time to generate catchment areas, which are not needed for the raster approach but are required for subsequent generalization of NHD flowlines. Consequently, the two approaches appear equally capable of supporting generalization of hydrography, but the raster approach can more efficiently support applications that do not require catchments.

4. Test on Road Network Data

It is important to establish that results reported above are not biased in some way by the choice of data theme. The following test explores the robustness of the raster partitioning algorithm with respect to another data domain. The algorithm was tested on a section of a road network around St. Louis, Missouri, which was extracted from the USGS Best Practices Dataset (BPD). The BPD includes the transportation, structures, and governmental units data themes for The National Map. The transportation theme was compiled from 2008 U.S. Census Bureau Topographically Integrated Geographic Encoding and References (TIGER) dataset (U.S. Census Bureau 2008) and is adequate for 24K mapping.

4.1 Partitioning roads of the Best Practices Dataset

A section of roads with more than 275 000 features was subdivided manually into four parts containing 38 330, 54 928, 62 089, and 119 756 road features, respectively, for the west, central, south, and northeast parts of the St. Louis area (figure 4). Using the raster approach, each part was partitioned separately into three density classes, which approximate rural, suburban, and urban road densities. Line densities for the rural, suburban, and urban classes were partitioned automatically to range from 0 to less than 2, 2 to less than 4, and greater than 4 km/km2, respectively. Average line densities were compared between the four sets of partitions, and partition boundaries were inspected visually to identify how well the boundaries match along edges between the four parts of the St. Louis area.

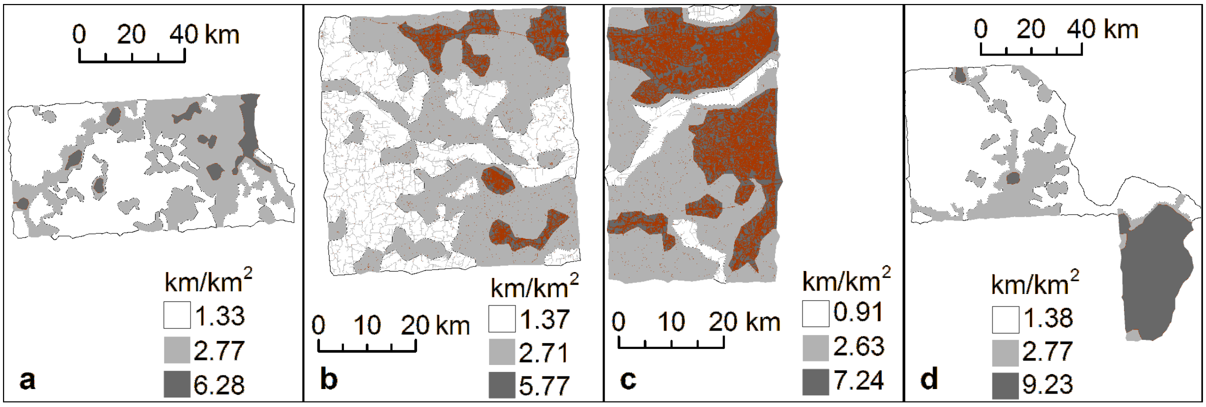

Figure 4. Line density partitions generated through the raster-based approach for four adjacent sections of roads covering the St. Louis, Missouri area. Roads were extracted from the USGS Best Practices Dataset. Average line densities are shown for each partition in kilometers per square kilometer (km/km2) and shaded by density class for the south (a), west (b, roads included), central (c, roads included), and northeast (d) sections of the St. Louis area.

Partitioning results for the road data in the four parts of the St. Louis area indicate that generated AOI boundaries adequately follow the edge of the road data, as demonstrated for the west and central sections in figure 4 (b and c). Roads are not displayed for the other two sections in figure 3 because the line densities at the smaller displayed scales would obscure the partitions. The density class breaks appear to be adequately tailored for the rural, suburban, and urban classes because small pockets of urban roads are delineated for the small towns and cities in the rural areas (figure 4b). These class breakdowns may be sufficient for delineating the rural to urban road density partitions for the BPD in the rest of the country. This remains an area for further research.

Overall, partition boundaries match well along the edges of the four data parts (figure 5a); however, partition boundaries between the rural and suburban categories do not match well along one edge where the west and central data parts meet (figure 5a). Although this is a minor issue, it may be alleviated by partitioning road data to include a minimum number of road lines—such as about 50 000—for each partitioning process. (The west data section includes 38 330 features only). In addition, sufficient overlap could be included between data subdivisions to help alleviate this issue. Further testing is necessary for this refinement.

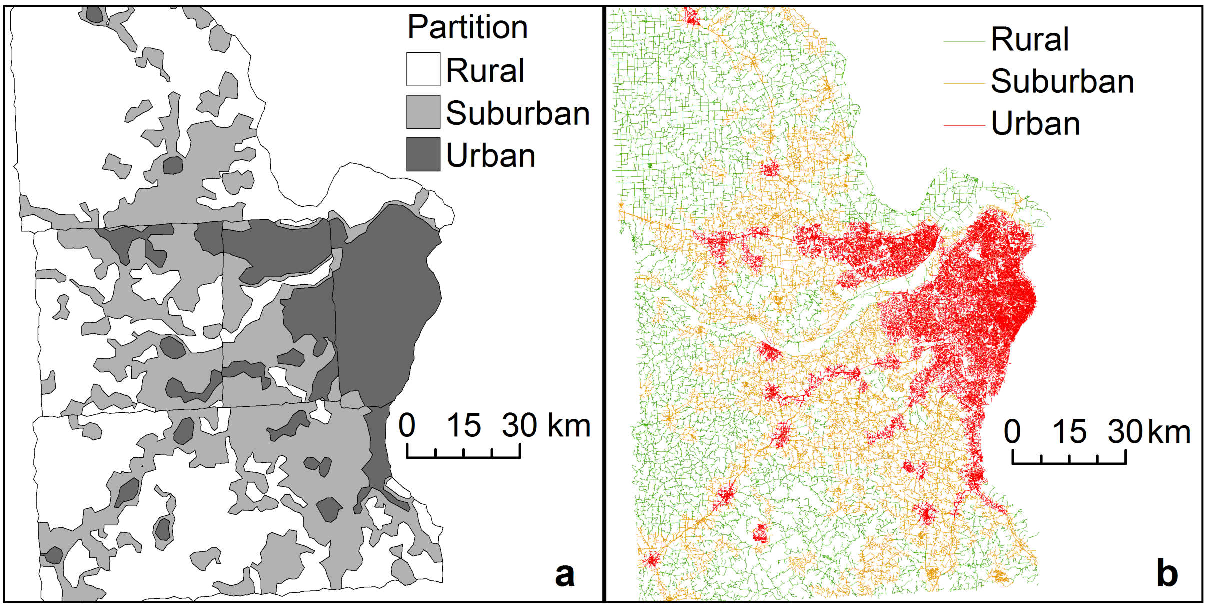

Figure 5. Line density partitions generated through the raster approach for four adjacent sections of road data from the Best Practices Data around St. Louis, Missouri (a). Partition density classes tailored for rural, suburban, and urban road densities transferred to the road lines (b).

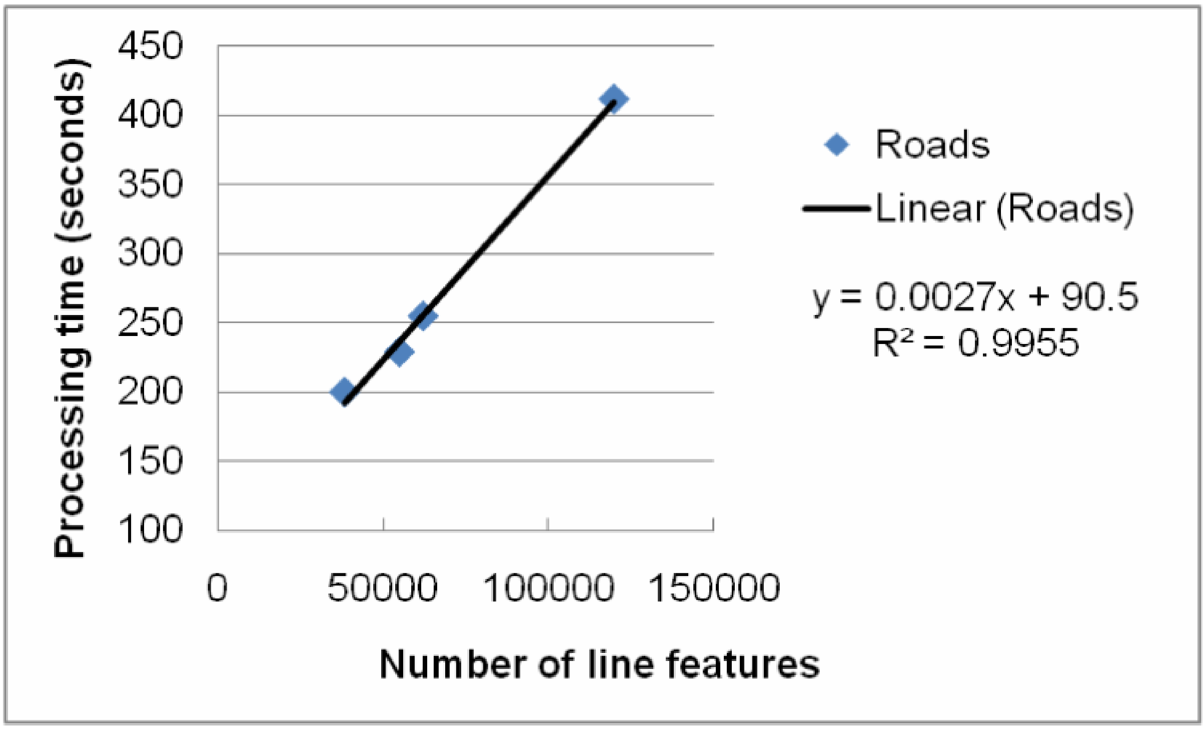

Figure 6. Processing speed of raster partitioning algorithm compared to number of line features processed. Results are shown for four sections of road data extracted for the St. Louis, Missouri area from the Best Practices Dataset.

The speed of raster partitioning ranged from about 3.5 to 7 minutes, depending on the number of road features processed (figure 6). Processing speed is linearly related to the number of features processed, as depicted by the regression line in figure 6. Consequently, the raster partitioning algorithm appears well-suited for parallel operations.

5. Discussion and Summary

An automated raster-based data partitioning algorithm that delineates the AOI surrounding a set of line features and subdivides the data into line density partitions based on predefined density classes has been described and demonstrated. The approach was demonstrated on hydrography and transportation data commonly used for cartography and geospatial analysis in the United States. Results from the raster partitioning approach were compared to a vector and a manual approach for two sets of hydrographic data, and evaluated for consistency and computing efficiency through tests on hydrographic and transportation data. Results indicate the raster approach delineates localized density differences better than either the manual or vector-clustering approaches, while preserving a reasonably high correspondence (60 percent or better) with density classes selected by those methods. In this sense, it is at least as effective as the vector and manual approaches for delineating partitions that support stratified pruning for automated generalization. Clearly, the automated approaches will provide better consistency and less misclassification than the manual approach. In terms of computation, the raster process is considerably faster than the alternatives, particularly for supporting applications that do not require surface-water drainage catchments; however, further analysis is needed to verify this conclusion.

Parameters for the raster partitioning process—such as cell size, radius for density smoothing, and buffer distance for delineation of the AOI—were tailored for hydrography and transportation data compiled for use at 24K. Parameters may need to be adjusted for other types and scales of data. Furthermore, the raster algorithm requires predefined density ranges that identify natural or artificial density breaks in line features that should be adequately maintained after generalization in the first case, but removed in the second case. However, the process is intended as a data-exploration tool that can be iteratively used to delineate, review, and refine density class boundaries for line features in an efficient manner, which can subsequently enhance generalization. Further testing and refinements are necessary to build appropriate data subdivisions that can be handled by the raster partitioning process for large databases, such as those spanning physiographic regions, or the entire contiguous United States.

Although additional testing and refinements are needed, the raster-based line-density partitioning process appears suitable to support automated generalization of hydrography and roads. It is more efficient than the vector clustering method, and could be made even more so through easily implemented parallel operations. Aside from automated generalization, variations of the algorithm may have other useful applications.

Acknowledgements

Barbara Buttenfield's research is supported by USGS-CEGIS grant # 04121HS029.

References

Bobzien, M., Burghardt, D., Petzold, I., Neun, M., and 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.

Borruso, G. 2003. Network density and the delimitation of urban areas. Transactions in GIS, vol. 7, no. 2, pp. 177-191.

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

Buttenfield, B.P., Stanislawski, L.V., and Brewer, C.A. 2010. Multiscale representations of water: Tailoring generalization sequences to specific physiographic regimes. Short Abstract Proceedings of the 6th International Conference on Geographic Information Science (GIScience'2010).

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

Chaudhry, O., and Mackaness, W.A. 2008b. Automatic identification of urban settlement boundaries for multiple representation databases. Computers, Environment and Urban Systems, vol. 32, no. 2, pp. 95-109.

Downs, J.A. 2010. Time-geographic density estimation for moving point objects. Geographic Information Science, Lecture Notes in Computer Science, vol. 6292/2010, pp. 16-26.

Lüscher, P., and Weibel, R. 2010. Semantics matters: Cognitively plausible delineation of city centres from point of interest data. 13th ICA Workshop on generalization and multiple representation. September 12-13, 2010, Zurich, Switzerland.

Pätynen, V., and Ristioja, J. 2009. Vector based generalization application in the production of small scale databases of the NLS of Finland, 24th International Cartography Conference, November 15-21, 2009, Santiago, Chile.

Renard, J. Gaffuri, J., and Duchêne, C. 2010. Capitalization problem in research—example of a new platform for generalization: CartAGen. 13th ICA Workshop on generalization and multiple representation. September 12-13, 2010, Zurich, Switzerland.

Silverman, B.W. 1986. Density Estimation for Statistics and Data Analysis. New York, Chapman and Hall.

Simav, Ö., Aslan, S., Çetinkaya, B., and Çobankaya, O.N. 2010. Implementation of comprehensive modeling techniques on Kartogen generalization software. 13th ICA Workshop on generalization and multiple representation. September 12-13, 2010, Zurich, Switzerland.

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

Stanislawski, L.V., Buttenfield, B.P, Finn, M.P., and Roth, K. 2009. Stratified database pruning to support local density variations in automated generalization of the United States National Hydrography Dataset, 24th International Cartography Conference, November 15-21, 2009, Santiago, Chile.

Stanislawski, L.V., and Buttenfield, B.P. 2010. Hydrographic features generalization in dry mountainous terrain. 2010 AutoCarto Proceedings, November 14-18, Orlando, Florida.

Stanislawski, L.V., Buttenfield, B.P., and Samaranayake, V.A. 2010. Automated metric assessment of hydrographic feature generalization through bootstrapping. 13th ICA Workshop on Generalization and Multiple Representations, September 12-13, 2010, Zurich, Switzerland.

Stoter, J., Burghardt, D., Duchêne, C., Baella, B., Bakker, N., Blok, C., Pla, M., Regnauld, N., Touya, G., and Shmid, S. 2009. Methodology for evaluating automated map generalization in commercial software, Computers, Environment and Urban Systems, vol. 33, no. 5. pp. 311-324.

Stoter, J., Baella, B., Blok, C., Burghardt, D., Dávila, F., Duchêne, C., Pla, M., Regnauld, N., and Touya, G. 2010. EuroSDR research on state-of-the-art of automated generalization in commercial software: main findings and conclusions. 13th ICA Workshop on generalization and multiple representation. September, 12-13, 2010, Zurich, Switzerland.

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

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

Touya, G., Duchene, C., and Ruas, A. 2010. Collaborative generalization: Formalization of generalization knowledge to orchestrate different cartographic generalization processes. GIScience 2010, September 14-17, 2010, Zurich, Switzerland.

U.S. Census Bureau. 2008. TIGER/Line Shapefiles 2008 Technical Documentation., U.S. Department of Commerce, online: http://www.census.gov/geo/www/tiger/tgrshp2008/TGRSHP08.pdf, November 2008.

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

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

Xiang, Z., Tinghua, A., and Stoter, J. 2008. The evaluation of spatial distribution density in map generalization, The International Archives of the Photogrammetry, Remote Sensing, and Spatial Information Sciences, vol. 37, part B2, Beijing, China, 2008.