CLUSTERING APPROACHES FOR HYDROGRAPHIC GENERALIZATION

Alper SEN, Turkay GOKGOZ

Geomatics Engineering Department, Civil Engineering Faculty, Yildiz Technical University, Davutpasa Street, 34220, Esenler, Istanbul, Turkey

{alpersen, gokgoz*}@yildiz.edu.tr*

Abstract

In many GIS applications users need to visualize and inspect data in different scales. This case requires that the different representations are stored at different levels of detail. One possibility is to store maps from the representation levels in a multiple representation database. Generalization methods are used to maintain and update the spatial databases.

Hydrographic data at different resolutions or scales are needed in various spatial studies. This paper presents a generalization operation, selection/elimination, for river networks based on clustering methods, namely k-means and Self Organizing Maps (SOM).

This is a model generalization operation, that considers geometric, topological and semantic river attributes as input. Clustering methods group all rivers into different categories according to similarities of the attributes, and then select the rivers for the reduced map scale based on these categories.

The clustering methods are compared using the data of United States Geological Survey (USGS) National Hydrography Dataset (NHD) at scales of 1:24000 and 1:100000. Clusters are eliminated based on the coefficient of line correspondence (CLC) is used to evaluate how well the features of clusters match the target 1:100000-scale. If the features of a cluster do not match any features of the target 1:100000-scale (CLC value is zero) or CLC value is different from zero, yet the selection of the cluster decreases the total CLC value of the derived network, then the cluster is eliminated.

The derived networks are visually compared to USGS 1:24000-scale and 1:100000-scale NHD. Töpfer's "Radical Law" is also used for a quantitative comparison. The case study applied to the network illustrates that both clustering approaches for hydrographic generalization have pleasing visual impact. However, SOM-based approach can be used as an effective method for the selection of rivers via data visualization and exploration for multi dimensional geospatial data.

Keywords: hydrographic generalization, cluster, self organizing maps, k-means, river network

INTRODUCTION

By the development of the technology and science, substantial investments towards Spatial Data Infrastructure (SDI) projects are occurred at local, national, regional/multi-national and global levels. Although various definitions exist, the term Spatial Data Infrastructure (SDI) generally describes a framework of components that collectively create an environment in which spatial data can be accessed and widely used; often it is the co-ordinating infrastructure underpinning the data assets of a country. The earliest and perhaps most prominent examples of SDIs are the US National Spatial Data Infrastructure (NSDI) in USA, The Authoritative Topographic-Cartographic Information System (ATKIS) in Germany and the Ordnance Survey Digital National Framework in Great Britain. The central task of National Mapping Agencies has been to establish and standardize digital geographic databases from which to produce maps. Working Committee of the Surveying Authorities of the States of the Federal Republic of Germany (AdV) defines mapping agencies in a common and harmonized data model, called the AAA data model (AFIS-ALKIS-ATKIS Data Model) can generally be used for standardization of the spatial infrastructure.

The maintenance and updating of these databases has become an urgent problem for which there remains no uniform solution. Since problems in updating digital geographic databases have become a major impediment to the effective use of geographic data in production environments, multiple representation databases have become one of the most important researches. The different representations are stored at

different levels of detail. Its flexibility lies in its ability to derive different types of maps from the representation levels of a multiple representation database, using generalization methods. In this respect, automated spatial data generalization techniques are very important, besides the methods such as data modeling, data management and data distribution (Kilpälainen, 1997; Sarjakoski, 2007).

Generalization is a process used for reducing the volume of data of a spatial data set while preserving important structures (Sester, 2008). Map generalization operations concerned with the abstraction of the database come under the heading of "model generalization", whilst the set of operations concerned with the optimal visualization of the selected data are grouped under "cartographic generalization". Model generalization is relevant to activities other than the visual. In particular it has relevance to data mining (Mackaness, 2007).

Model and cartographic generalization is used for deriving Digital Cartographic Model (DCM) from Digital Landscape Model (DLM). Both DLM and DCM are spatial databases in GIS. DLM is a generic term for a comprehensive description of the landscape, usually in the form of a topographic basic scale map. Databases derived from the primary DLM, through model generalization, are special-purpose secondary models of reality. Both primary and secondary DLMs may be used to create Digital Cartographic Model (DCM), through the process of cartographic generalization.

Model generalization is mainly a filtering process to obtain a subset of an original database for data analysis. Cartographic generalization is used for graphic display and it aims to improve the visual effectiveness and readability of a map. GIS need to perform model and cartographic generalization in order to satisfy both analysis and display purposes, respectively. (Joao, 1998).

Selection and elimination operation has primary importance in map generalization. Competition for map space is a fundamental principle of map design. Unnecessary features are eliminated and important features are retained. The completeness of a map is affected by the elimination of features due to generalization procedures. Robinson et al. (1995), defines that the selection is a generalizing, but it is not part of cartographic generalization included simplification, classification and symbolization. Töpfer and Pillewizer (1966) suggest a mathematical formula, the Principle of Selection also known as Radical Law, is one of the well known selection method. From the study of atlases of different countries, the Principle of Selection proves to be a great potential value for the derivation of smaller scale maps from larger scale source material. The Principle of Selection is expressed by an equation that relates the number of occurrences of a particular feature at a source map scale and at a derived map scale. The principle can be expressed in its simplest form as

$$n_f = n_a \sqrt{M_a / M_f} \qquad (1)$$

where n_f is the number of objects that can be shown at the derived scale, n_a is the number of objects shown on the source material, M_a and M_f are the scale denominators of the source and the derived map, respectively. The formula yields the number of symbols to be displayed, but it does not reveal which of the symbols should be chosen. Töpfer's law is the only quantitative rule in the selection and elimination of the features.

Stream order is a common technique to assign a hierarchy to the components of a river network is used for selection (Horton, 1945; Strahler, 1952; Shreve, 1966). Kadmon (1972) uses the Radical Law and weighted values for selecting the cities on Israel maps. Stenhouse (1979) applies the Radical Law to towns in England. Fitzsimons (1985) examines the selection of base information for thematic maps. Wolf (1988) suggests a weighted network data for river generalization. Richardson (1994) presents a method to select rivers by creating a fine hierarchy and filtering. Mackaness (1995) focuses on reduction of roads derived from axial maps. Ruas (1998, 1999) uses the building density as a threshold value for reducing the buildings in a block. The good continuation grouping principle derived from the Gestalt theory can serve as the basis for analyzing road and river networks into a set of linear elements called as strokes. Thomson and Richardson (1999) apply the perceptual grouping principle of good continuation to build strokes in road network selection. Thomson and Brooks (2000) also use the good continuation principle to build strokes in road and river networks in order to perform the selection. Jiang and Claramunt (2004), and Jiang and Harrie (2004)

introduce the centrality measures (i.e. degree, closeness, and betweenness) based on the connectivity graph for characterizing the structural properties of an urban street network, and for the selection of important streets using SOM algorithm. Ai et al. (2006) present a method to select the river network by the watershed area threshold based on the watershed hierarchical partitioning. Cetinkaya (2006) uses Critical Path Method for selecting the rivers according to the weights. Touya (2007) focuses on model generalization and uses the principle of good continuation to enrich the database with river strokes. In Touya's study, in order to determine which strokes are to be selected, the main criterion used is a hierarchical organization of the strokes. Stanislawski (2009) uses a pruning algorithm considering the upstream drainage area. Stanislawski and Savino (2011) compare two pruning approaches which are stratified pruning, and the length and density pruning for hydrographic networks. Gulgen and Gokgoz (2011) suggest a block-based selection method for road network generalization.

In this study, an approach for a model generalization of a river network at scale of 1:24000 is proposed in order to obtain the river network at scale of 1:100000. In this approach, geometric (length and sinuosity), topological (degree centrality, betweeness centrality and closeness centrality) and semantic (type, stream level, lake connection and flow accumulation) attributes are considered as inputs, and clustering methods, k-means which looks for a partition based on a given number of clusters (*k*) and an unsupervised learning method of artificial neural networks, self-organizing maps (SOM) are used. SOM is different from the more popular approach k-means by reason of using topological relationships between clusters and geographic visualization environment. Some studies discusses about the relationship between GIScience and the SOM method. Sester (2005) suggests SOM for the typification of buildings. Agarwal and Skupin (2008) edit a book about SOM applications in GIS.

Clustering methods group all rivers into different categories according to similarities of various attributes, and then selects rivers at reduced map scale based on these categories. The basic idea behind clustering is the attempt to organize objects into groupings based on certain shared characteristics. The clustering methods are compared using the data of United States Geological Survey (USGS) National Hydrography Dataset (NHD) at scales of 1:24000 and 1:100000. Clusters are eliminated based on the coefficient of line correspondence (CLC) is used to evaluate how well the features of the clusters match the target 1:100000-scale. If the features of a cluster do not match any features of the target 1:100000-scale (CLC value is zero) or CLC value is different from zero, yet the selection of the cluster decreases the total CLC value of the derived network, then the cluster is eliminated.

Finally, the derived networks are visually compared to USGS 1:24000-scale and 1:100000-scale NHD. Töpfer's "Radical Law" is also used for a quantitative comparison. The case study applied to the network illustrates that both clustering approaches for hydrographic generalization have pleasing visual impact. However, SOM-based approach can be used as an effective method for the selection of rivers via data visualization and exploration for multi dimensional geospatial data.

METHODOLOGY

Cluster Analyses

The complexity of the problem of generalization is that the knowledge is difficult to formalize, where objects, attributes, tools and objectives are of a multiple variety. There are complex relations between geometric, topological and semantic attributes of the features in the level of details and multidimensional spatial data space requires spatial data mining techniques to analyze.

Unlike classification and prediction, which analyze class-labeled data object, clustering analyzes data objects without consulting a known class label and it can be used for the selection process of model generalization. In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such selection/elimination labels.

The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters.

Since clustering is the most frequent interpretation and implementation of the SOM method, it is useful to compare it to one of the most popular approach, k-means clustering which look for a partition based on a given number of clusters (*k*) (Han and Kamber, 2001). There is a disadvantage of using clustering approach for the selection, since clustering methods used are not deterministic.

Self Organizing Maps (SOM) Method

SOM is part of a large group of techniques known as artificial neural network (ANN). One quickly realizes that, apart from seeing the SOM only in the context of other ANN methods, depending on its purpose and training parameters one could also interpret it primarily as a clustering or dimensionality reduction technique. SOM adjusts weights of neuron vectors for minimizing the distance and maximizing the similarity to input vectors; thus the similar inputs are associated with closely positioned neurons. In SOM, the input vectors do not correspond to classes known a priori. Output nodes compete for the input vectors on the basis of certain similarity functions and the weights of winning nodes are adjusted according to the weights of respective input nodes. The basic idea behind clustering is the attempt to organize objects into groupings based on certain shared characteristics (Agarwal and Skupin, 2008).

Input data as consisting of n-dimensional vectors:

$$x = [\varepsilon_1, ..., \varepsilon_n]^T \in R^n \quad (2)$$

Each of k neurons has an associated reference vector:

$$m_1 = [\mu_1, ..., \mu_n]^T \in R^n \quad (3)$$

During training, one x is compared with all m_i to find the reference vector m_c that satisfies a minimum distance or maximum similarity criterion. Through a number of measures are possible, in this study Euclidean distance is used.

$$\|x - m_c\| = \min_i \{\|x - m_c\|\} \quad (4)$$

The best-matching unit and neurons within its neighbourhood are modified by;

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)] \quad (5)$$

Neighborhood function defines a distance-weighed model for adjusting neuron vectors. The Gaussian model:

$$h_{ci}(t) = \alpha(t).e^{-d_{ci}^2 / 2\sigma_i^2(t)} \quad (6)$$

d_{ci} : Distance between the best matching unit (BMU) and the respective neuron

In the Gaussian model, that neighborhood's size appears as kernel width (σ) and is not a fixed parameter. Similarly, the initial learning rate (α_0) is an input parameter, which is then gradually decreased as t progresses. SOM training stops when a predetermined number of training cycles (t_{max}) are completed.

In this study, Matlab SOM tool is used for clustering the river network. 5-by-5 SOM in a two dimensional grid is used for clustering the 272-by-9 inputs (272 features and 9 attributes). The output vectors are trained based on the input vectors. During the training process, a Gaussian neighborhood function is chosen. Empirically chosen parameter settings are listed in Table 1.

Table 1. Parameter settings for the SOM training

Parameter	Value
Size	5x5
Dimensionality	2
Inputs	272x9
Topology	Hexagonal
Neighbourhood	Gaussian
Distance	Euclidean
Initial learning rate	0.5
Training cycles	20000

The left side of Fig.1 shows how many of the training data are associated with each of the neurons (hexagons). The topology is a 5-by-5 hexagonal, so there are 25 neurons. Totally 14 clusters are occurred in 25 neurons according to their similarities after the SOM training as shown by hexagons in different sizes. The larger hexagon is associated with more input vectors. The maximum number of hits associated with any neuron is 45. Thus there are 45 input vectors in that cluster. The right side of Fig.1 shows the U-matrix, on which distances between input vectors can easily be identified, developed by Ultsch and Siemon (1989). In this figure, the blue hexagons represent the neurons. The red lines connect neighboring neurons. The colors in the regions containing the red lines indicate the distances between neurons. The darker colors represent larger distances, and the lighter colors represent smaller distances. Any boundaries cannot be recognized in the U-matrix except the darker one at the upper right corner that neuron represents the longest central stem (main stem), and very different from the other vectors which similar to each other.

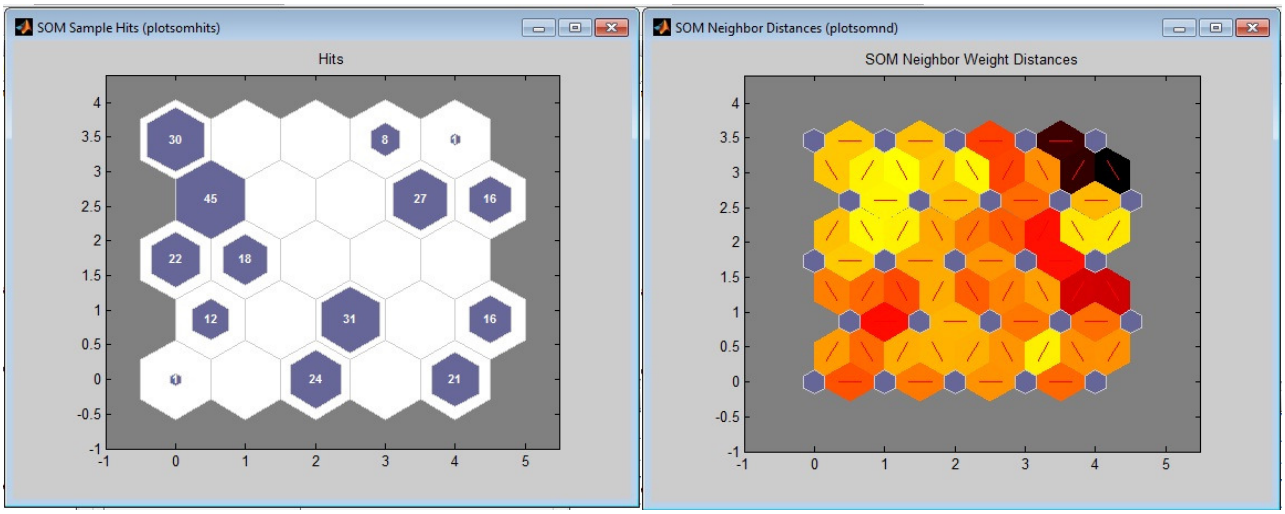

Fig. 1. 5-by-5 SOM network: Clustered samples in neurons (left) and neighbor weight distances / U-matrix (right)

k-means Method

Given a database of *n* objects and *k*, the number of clusters to form, a partitioning algorithm organizes the object into *k* partitions *(k≤n),* where each partitions represents a cluster. The clusters are formed to optimize an objective partitioning criterion, often called a similarity function, such as distance, so that the objects within a cluster are “similar”, whereas the objects of different clusters are “dissimilar” in terms of the database attributes.

The k-means algorithm for partitioning based on the mean value of the objects in the cluster. The number of clusters *k* and a database containing *n* objects are the inputs. A set of *k* clusters that minimizes the squared-error criterion is the output.

In this method: (1) arbitrarily *k* objects as the initial cluster centers are chosen; (2) each object is assigned to the cluster to which the object is the most similar based on the mean value of the objects in the cluster (3) the cluster means are updated (4) the iterations are continued until no change (Han and Kamber, 2001).

In this study, Matlab is also used for k-means clustering which partitions 272 features-by-9 dimensions into 14 and 25 clusters using the squared Euclidean distance. Each centroid is the mean of points in the cluster. The numbers of clusters (*k*), 14 and 25 are determined based on the SOM results and the total sum of point-to-centroid distances, respectively. In order to decrease the total sum of point-to-centroid distances *k* number is changed from 14 to 25.

THE CASE STUDY

The National Hydrography Dataset (NHD) is a vector geospatial data layer of the National Map, being developed by the United States Geological Survey (USGS) and created from many data sources and Web Map Services, representing the surface water hydrography of the United States. It is available nationwide as medium resolution at 1:100000-scale, and as high resolution at 1:24000-scale or better.

The study area is the subbasin 102901070403 in the 1:24000 The National Hydrography Dataset, forms the watershed for the Little Pomme de Terre River, in the Midwest United States, in Missouri. The basin area is 159.56 km^2.

Input Data

The geometric, topological and semantic attributes of the rivers are derived from the geodatabase files, namely NHDH1029 (1:24000-scale) and NDHM1029 (1:100000-scale) and National Elevation Dataset 1/3 arc-second (approx. 10m) DEM and used as the inputs of the clustering methods, k-means and SOM.

During selection it is important that a river is processed as a whole, and segments are not separately eliminated, which may disconnect the graph (Stanislawski and Savino, 2011). Before deriving the input data, river segments are combined according to the stream levels. Since the limitations of the elimination of some river segments between the confluences in order to protect the main stems (Sen and Gokgoz, 2011), note that river segments are combined according to the stream levels. After combining the segments of rivers, there are totally 272 features included tributaries and stems in the river network.

A stream level is assigned to each reach in the drainage network. Reaches are delineated between confluences in the drainage network, and each reach is assigned a unique and permanent address called a reach code. The stream level value is a numeric code that identifies a hierarchy of main paths of water flow through the network. Level values are established for the purpose of computationally traversing the drainage network through flow relations identified between the reaches. The level coding system appears to be the upstream routine of the Horton and Strahler stream ordering system (NHD standards) (Fig. 2). The derived attributes are normalized and all become in the range of [0,1].

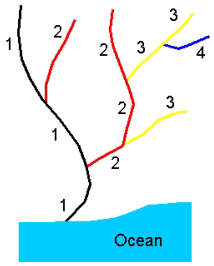

Fig. 2. Stream levels

Geometric attributes:

*Length (ratio-scaled)***:** The lengths of tributaries and stems.

*Sinuosity (ratio-scaled)***:** The ratio of the Euclidean distance between the end points to the stream length.

Topological attributes:

Pajek software is used to calculate the topologic attributes via the connectivity graph of the river network (Nooy, Mrvar and Batagelj, 2005). A graph is a set of points (vertices) in a mathematical space, which are interconnected by a set of lines (links). A graph-theoretic representation of river networks uses streams as vertices and stream junctions as links of a connectivity graph to understand the topology of river networks. Centerlines of the lake polygons and the river lines are not in the same feature type. Therefore, centerline of the lake polygon represents different vertex and it is not included into the clustering. Fig. 3 illustrates the

Little Pomme de Terre River network at 1:24000-scale and its connectivity graph, and the representative vertex of the lake polygon centerline shown in the circle.

Degree centrality (ratio-scaled) of a vertex is defined as the number of links incident upon a vertex (i.e., the number of ties that a vertex has).

$$C_D(V_i) = \frac{\sum_{k=1}^{n} r(V_i, V_k)}{n-1} \tag{7}$$

Closeness centrality (ratio-scaled) of a vertex is based on the inverse of the distance of each vertex to every other vertex in the network.

$$C_C(V_i) = \frac{n-1}{\sum_{k=1}^{n} d(V_i, V_k)} \tag{8}$$

Betweenness centrality (ratio-scaled) of a vertex counts the number of shortest paths between *i* and *k* that vertex *j* resides on. Vertices that occur on many shortest paths between other vertices have higher betweenness than those that do not.

$$C_B(V_i) = \sum_{j=1}^{n} \sum_{k=1}^{j-1} \frac{p_{ikj}}{p_{ij}} \tag{9}$$

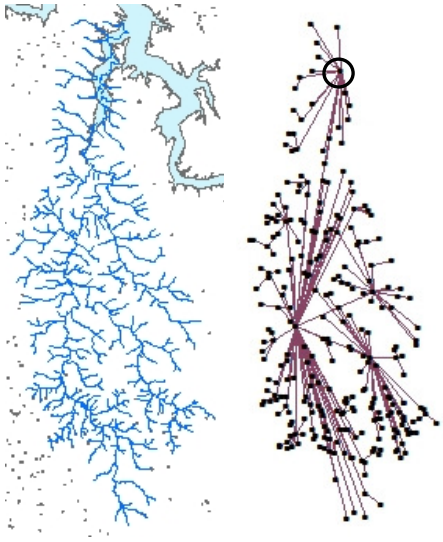

Fig.3. Little Pomme de Terre River network at 1:24000-scale (left) and its connectivity graph (right)

Semantic attributes:

*Feature code (binary)***:** Features are classified by type codes. These feature types are such as intermittent and perennial stream/river.

Stream level (ordinal): A numeric code identifies each main path of water flow through a drainage network based on Horton and Strahler order. The lowest value of stream level represents the highest hierarchy in the NHD database.

*Lake connection (binary)***:** If a stream/river intersects a lake/pond at 1:24000-scale then the value gets one, else the value is zero.

*Average flow accumulation (ratio-scaled)***:** ArcHydro tools functions are used for the hydrologic computations. The flow accumulation grid that contains the accumulated number of cells upstream of a cell is computed for each cell in the input grid. Then the average flow accumulations of the linear rivers located on the grid cells are determined. In order to compute the average flow accumulation, “DEM Reconditioning” function which

modifies the DEM by imposing linear river network (burning/fencing) (AGREE method by Hellweger, 1997), “Fill Sinks” function which modifies the elevation value to eliminate the problems about cells surrounded with higher elevation and resumes the water flow, and “Flow Directions” using D8 method are used.

Selection and Elimination of Clusters

The coefficient of line correspondence (CLC) estimates how well two sets of lines, representing similar features on the ground, overlap each other. Clusters are eliminated based on the coefficient of line correspondence (CLC) is used to evaluate how well the features of the clusters match the target 1:100000-scale. If the features of a cluster do not match any features of the target 1:100000-scale (CLC value is zero) or CLC value is different from zero, yet the selection of the cluster decreases the total CLC value of the derived network, then the cluster is eliminated.

Totally 9 clusters are eliminated from 14 SOM clusters, whilst 9 and 16 clusters are eliminated from 14 and 25 k-means clusters, respectively. Selected clusters with CLC values are given in Table 2.

CLC value is calculated by the given equation (10).

$$M / (O + C + M) \quad (10)$$

where *M* is the sum of the lengths of matching target lines, *O* is the sum of the lengths of target lines that are omitted from the generalized data set, and *C* is the sum of the length of lines in the generalized data set that do not have a match in the target data set (commission errors), which is divided by the 1:100000 – 1:24000 scales length expansion factor. Reducing line lengths in *C* by the expansion factor puts all values on a common scale (Stanislawski, 2009). The expansion factor of 1.03 is determined as the average ratio of the 1:24000-scale to 1:100000-scale lengths for the 15 matching rivers, which are distributed over the study area. Clusters by SOM and two variations of k-means method, and selection and elimination of clusters according to CLC values are given in Table 2 and Table 3.

Table 2. Selected clusters with CLC values

SOM		k-means (k=14)		k-means (k=25)	
Cluster no	**CLC**	**Cluster no**	**CLC**	**Cluster no**	**CLC**
10	0.09	3	0.09	3	0.05
11	0.24	8	0.07	4	0.06
19	0.18	10	0.09	6	0.02
24	0.05	12	0.37	7	0.11
25	0.22	13	0.18	13	0.12
				18	0.04
				20	0.08
				21	0.07
				22	0.22

CLC value of 0.02 is quite low in Table 2. However, the selection of these clusters increases the total CLC value of the derived network (M>C).

Table 3. Eliminated clusters with CLC values

SOM		k-means (k=14)		k-means (k=25)	
Cluster no	**CLC**	**Cluster no**	**CLC**	**Cluster no**	**CLC**
3	0.06	2	0.01	2	0.01
6	0.01	7	0.06	9	0.07
12	0.07	11	0.06	10	0.02
1, 5, 8, 16,		1, 4, 5, 6,		12	0.06
20, 21	0	9, 14	0	14	0.01
				1, 5, 8, 11,	
				15, 16, 17,	
				19, 23, 24,	
				25	0

In Table 3, some clusters, whose CLC values are different from zero, are eliminated, because of decreasing the total CLC value of the derived networks (M<C). The total CLC values of 0.65, 0.63 and 0.73 result from SOM, k-means whose cluster numbers are 14 and 25, respectively.

RESULTS

Results are given in several tables and figures below. The visual and quantitative comparisons are done among the derived networks using clustering methods for selection/elimination, USGS 1:24000-scale and 1:100000-scale NHD maps. Töpfer's "Radical Law" which links the scale of the map to the amount of details it should contain is also used for quantitative comparison.

Since the limitations of the elimination of some rivers with respect to protection of main stems, note that stream segments are combined according to the stream levels based on upstream routine of the Horton and Strahler stream ordering system. Therefore, the number of combined river segments (tributaries and stems) is not equal to the number of river segments in the NHD. Combining the river segments task increases the quality of clustering (Sen and Gokgoz, 2011).

According to the results, SOM and k-means clustering are very similar when the number of clusters is determined based on the SOM results. The number of selected hydrographic features by SOM and k-means (k=14) are closer to the Töpfer's number (Table 4). The number of features of the target 1:100000-scale NHD is far from the Radical Law number because of the high elimination rate of the short rivers and some of them are in high hierarchical order. In order to compare the original NHDs in a qualitative manner, the river networks are overlapped in 1:24000 and 1:100000 scales (Fig. 4). Some short rivers in high hierarchical order are eliminated at target 1:100000-scale as shown in Fig. 4 close-up.

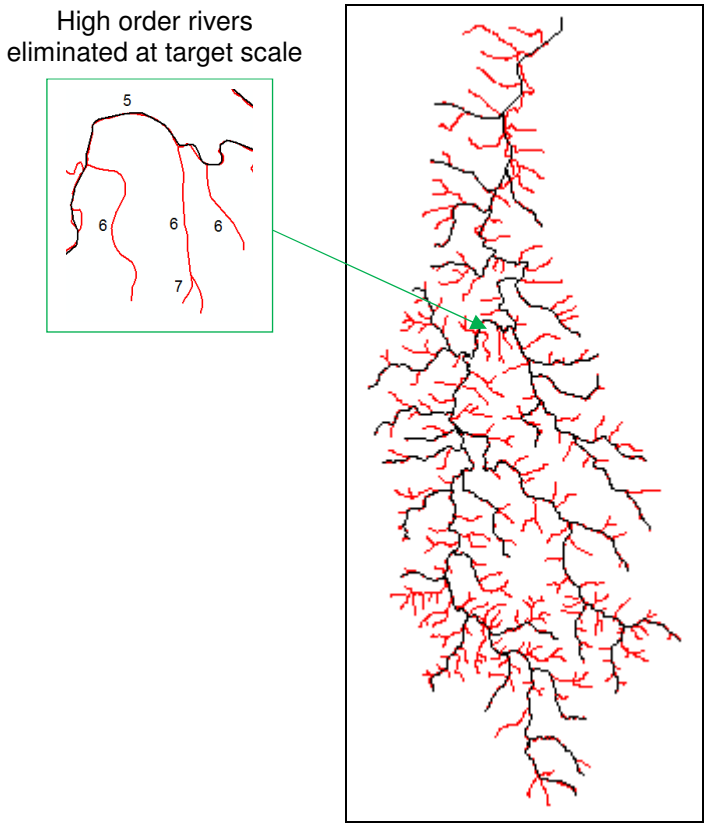

Fig. 4. 1:24000 NHD (red), 1:100000 NHD (black)

In Table 4, total network length and drainage density values are given as statistical information. Selection percentages of the total network length and features are also given in brackets. The drainage density is average length of rivers within the basin per unit of area. The drainage density is expressed as:

$$D_d = \sum L / A \qquad (11)$$

where $\sum L$ is the total length of the rivers and A is the area, both in units of the same system (Horton, 1945). Drainage densities of derived networks using SOM and k-means (k=14) methods are almost same. However, the network derived from 25 k-means clusters is close to the target drainage density.

Since the minimum river length is 30 meters according to the minimal dimensions of data displayed on a screen (Spiess, 1995), the length of 0.05 km can be accepted. Selection by k-means (k=25) method is better than SOM considering the minimum lengths.

Table 4. Total network length, minimum river length, drainage density and the number of features on the original and derived maps, the number of features at 1:100000-scale by the Radical Law, and the selection percentages of the total network length and features in brackets

Scale	Number	∑ Length (km)	Min. Length (km)	Density (km/km²)
1:24000	272	275.047	0.02	1.72
1:100000	30 (11%)	122.388 (44%)	1.58	0.77
SOM 1:100000	74 (27%)	140.768 (51%)	0.05	0.88
k-means 1:100000 (k=14)	77 (28%)	144.442 (52%)	0.05	0.9
k-means 1:100000 (k=25)	35 (13%)	111.45 (40%)	0.21	0.7
Töpfer 1:100000	133 (49%)	-	-	-

As it is given in Table 5, SOM and k-means (k=14) methods retain all the rivers in the highest hierarchical level, namely 5. However, k-means (k=25) retains half of them. The percentage of selected features decreases in the lower hierarchy. All of the features are eliminated in the low stream levels of 8 and 9.

Table 5. Numbers of features in the hierarchy classes of the original and derived networks, and selection percentages of the features in brackets

Stream level	1:24000	1:100000	SOM 1:100000	k-means (k=14)	k-means (k=25)
5	18	5 (28%)	18 (100%)	18 (100%)	9 (50%)
6	69	17 (25%)	32 (46%)	35 (51%)	18 (26%)
7	126	8 (6%)	24 (19%)	24 (19%)	8 (6%)
8	53	-	-	-	-
9	6	-	-	-	-

The clustering methods consider the geometric, topological and semantic attributes and try to group all rivers into different categories according to the similarities of attributes. Fig. 5 and 6 are given for visual comparison between derived networks using clustering methods for selection/elimination and NHD networks. Hydrographic generalization using clustering approaches has pleasing visual impact. The proposed k-means method using 25 clusters to select the rivers matches the target scale of 1:100000 better.

In Fig. 5.a and 5.b, derived network using k-means method with 14 clusters and original networks are overlapped and compared. The derived networks of k-means (k=14) and SOM are almost same. The blue circles show the rivers which are retained different from the derived network using SOM. However, it is not recognized a specific characteristic of these three rivers.

Derived network using k-means method with 25 clusters and original networks are overlapped and compared in Fig. 5.c and 5.d. The blue frame in Fig. 5.c shows the elimination of perennial rivers in the highest hierarchy, but SOM and the other variation of k-means retained them. The perennial rivers in 7^{th} stream level are retained only if they intersect with lakes and betweeness centrality is different from zero, as shown in the green framed close-up. However, the intermittent rivers in 7^{th} stream level are eliminated. The intermittent rivers in 6^{th} stream level are retained only if they intersect with lakes as shown in red framed close-up.

Finally, derived network using SOM method and the original NHD networks as shown in red are overlapped for comparison in Fig. 6. The perennial rivers in 7^{th} stream level are retained only if they intersect with lakes as shown in the green framed close-up. However, the intermittent rivers in 7^{th} stream level are eliminated. The intermittent rivers in 6^{th} stream level are retained only if they intersect with lakes as shown in red framed close-up. Also, the perennial rivers in 6^{th} stream level which intersect with lakes and the high accumulated ones are retained as shown in the blue framed close-up.

As a result, two clustering methods can be used in hydrographic generalization. Two methods organize objects into groupings putting forward the different characteristic. SOM-based approach is an effective method for the selection of rivers via data visualization and exploration for multi dimensional geospatial data and works better for the selection of the rivers considering Töpfer's "Radikal Law". However the derived network using k-means (k=25) matches the target scale of 1:100000 better.

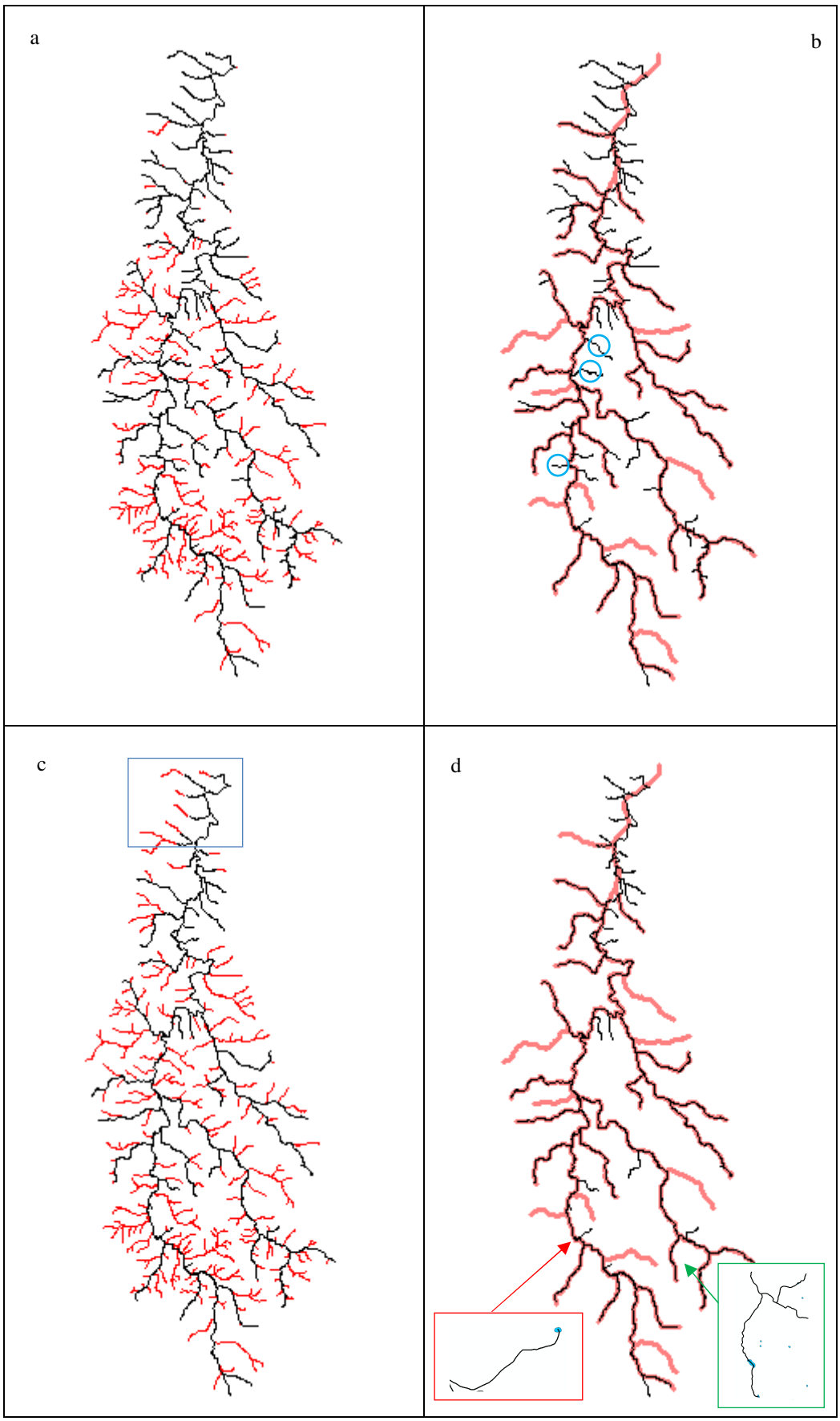

Fig. 5. a) 1: 100000 k-means (k=14) clustering (black), 1:24000 NHD (red), b) 1: 100000 k-means (k=14) clustering (black), 1:100000 NHD (red), c) 1: 100000 k-means (k=25) clustering (black), 1:24000 NHD (red), d) 1: 100000 k-means (k=25) clustering (black), 1:100000 NHD (red)

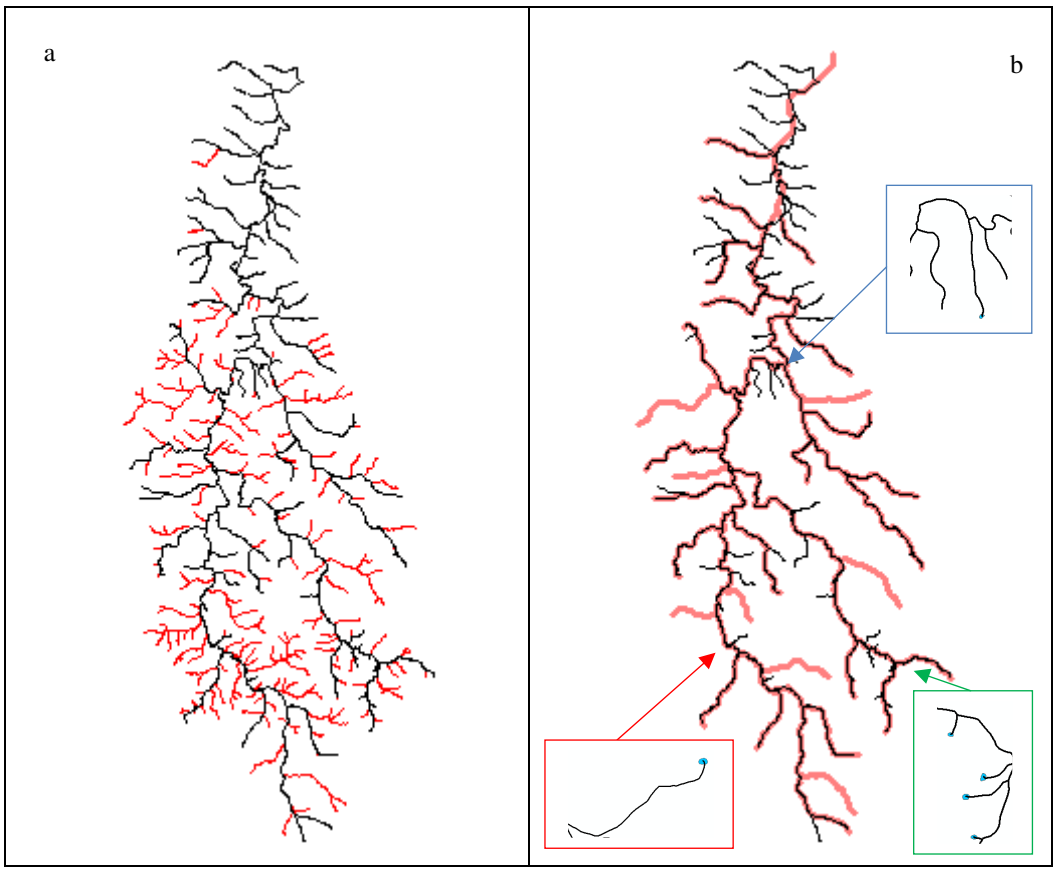

Fig. 6. a) 1:100000 SOM clustering (black), 1:24000 NHD (red) b) 1:100000 SOM clustering (black), 1:100000 NHD (red)

CONCLUSIONS AND FUTURE STUDY

The clustering approaches and CLC values provide an automated model generalization. There is a disadvantage of using the clustering approach for selection. Clustering methods used are not deterministic, which means that different runs of the algorithm lead to different result. This is due to the random selection at the beginning of the process. However, the results are still reflecting the density and distribution of the original situation.

The number of selected hydrographic features by SOM generalization is closer to the Töpfer's number than k-means (k=25) clustering and target 1:100000 NHD. However the derived network by k-means (k=25) matches the target scale of 1:100000 better considering the total CLC value. Moreover, the shortest river in this network is more legible.

The case study applied to the network illustrates that the SOM-based approach can be used as an effective method for the selection of rivers via data visualization and exploration for multi dimensional geospatial data. It is not effective to define the number of clusters by examining the total sum of point-to-centroid distances in k-means clustering.

In the future study, different and more effective attributes will be researched and applied to the net in order to determine the clusters more definite. Some optimization algorithms such as Genetic algorithms will be used as a determiner of the parameters in the SOM architecture.

Support vector machines (SVM) which is a concept in statistics and computer science for a set of related supervised learning methods that analyze data and recognize patterns will be used for training the input data to select the rivers in model generalization.

ACKNOWLEDGEMENTS

The authors would like to thank Prof. Dr.-Ing. habil. Monika Sester, the director of the Cartography and Geoinformatic Institute, Leibniz University Hannover, for the valuable comments and TÜBİTAK (The Scientific and Technological Research Council of Turkey) for their supports.

REFERENCES

Agarwal, P., Skupin, A. (2008) Self-Organising Maps : Applications in Geographic Information Science, Wiley.

Ai, T., Liu, Y. and Chen, J., (2006) The hierarchical watershed partitioning and data simplification of river network. In: Riedl, Andreas; Kainz, Wolfgang; Elmes, Gregory A. (Eds.), Progress in Spatial Data Handling, Springer, Part 11, 617-632.

Cetinkaya, B. (2006) Topografik haritaların üretiminde eş yükseklik eğrileri, akan su ve su iletim hatları coğrafi verilerin otomasyon süreçleri. Dissertation, Istanbul Technical University, Fen Bilimleri Enstitüsü, Istanbul (in Turkish).

Fitzsimons, D. E. (1985) Base data on thematic maps. The American Cartographer, 12(1): 57-61

Gulgen, F. and Gokgoz, T. (2011) A block-based selection method for road network generalization. International Journal of Digital Earth, 4(2), 133-153.

Han, J. and Kamber, M. (2001) Data mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2001, USA.

Hellweger, 1997, http://www.ce.utexas.edu/prof/maidment/gishydro/ferdi/research/agree/agree.html

Horton, R.E. (1945) Erosional development of streams and their drainage basins; hydrophysical approach to quantitative morphology. Bulletin of the Geological Society of America, 56(3), 275-370.

Jiang, B. and Claramunt, C. (2004) A structural approach to the model generalization of an urban street network. GeoInformatica, 8(2), 157-172.

Jiang, B. and Harrie, L. (2004) Selection of streets from a network using self-organizing maps. Transactions in GIS, 8(3), 35-350.

Joao, E.M. (1998) Causes and Consequences of Map Generalisation. Taylor & Francis Ltd, London, 35-37.

Kadmon, N. (1972). Automated selection of settlements in map generalization. The Cartographic Journal, 9(2), 93-98.

Kilpälainen, T. (1997) Multiple representation and generalization of geo-databases for topographic maps. Dissertation, Helsinki University of Technology, Publications of the Finnish Geodetic Institute.

Mackaness, W. A. (1995) Analysis of urban road networks to support cartographic generalization. Cartography and Geographic Information Systems, 22(4), 306-316.

Mackaness, W. A. (2007) Understanding geographic space. In: Mackaness, W. A., Ruas, A. and Sarjakoski, L. T., (Eds.), Generalization of Geographic Information: Cartographic Modelling and Applications. Elsevier, Amsterdam, 1-10.

Nooy, Mrvar and Batagelj, (2005), Exploratory social network analysis with Pajek review. Cambridge University Press.

Richardson, D. (1994) Generalisation of spatial and thematic data using inheritance and classification and aggregation hierarchies. In: Advances in GIS Research 2, Taylor and Francis, 957-972.

Robinson, A. H., Morrison, J. L., Muehrcke P. C., Kimerling A. J. and Guptill, S. C. (1995) Elements of Cartography, Sixth Edition, Wiley, USA.

Ruas, A. (1998) O-O Constraints modelling to automate urban generalisation process. Proceedings of SDH'98, Vancouver, 225-235.

Ruas, A. (1999) The role of meso level for urban generalisation. Workshop on Progress in Automated Map Generalisation, ICA, Ottowa.

Sarjakoski, L. T. (2007) Conceptual models of generalisation and multiple representation. In: Mackaness, W. A., Ruas, A. and Sarjakoski, L. T., (Eds.), Generalization of Geographic Information: Cartographic Modelling and Applications, Elsevier, Amsterdam, 11-35.

Sen, A. and Gokgoz, T. (2011) Hydrographic selection / elimination in automated generalization by using artificial neural networks. (Poster), in Proceeding of the 25th International Cartographic Conference, 3-8 July 2011, Paris, France.

Sester, M. (2005) Optimization approaches for generalization and data abstraction, International Journal of Geographical Information Science,19(8), 871 – 897.

Sester, M. (2008) Self-Organizing Maps for density-preserving reduction of objects in cartographic generalization. In: Agarwal, P. and Skupin, A., (Eds.), Self-Organizing Maps Applications in Geographic Information Science, John Wiley & Sons, England, 107-120.

Shreve, R.L. (1966) Statistical law of stream numbers. Journal of Geology, 74, 17-37.

Spiess, E. (1995) The need for generalization in a GIS environment. In: Gıs and Generalization, Methodology and Practise, Eds. Müller, J.C., Lagrange, J.P. and Weibel, R., Tatlor and Francis London.

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, 33, 325-333.

Stanislawski, L.V. and Savino S. (2011) Pruning of hydrographic networks: A comparison of two approaches. 14th ICA/ISPRS Workshop on Generalisation and Multiple Representation, 2011, Paris.

Stenhouse, H. (1979) Selection of towns on derived maps. The Cartographic Journal, 16(1), 30-39.

Strahler, A.N. (1952) Dynamic basis of geomorphology. Bulletin of Geological Society of America, 63(7), 923-938.

Thomson, R.C. and Richardson, D.E. (1999) The 'good continuity' principle of perceptual organisation applied to the generalisation of road networks. Proceeding of the 19th International Cartographic Conference, Ottawa, Canada, 1215-1225.

Thomson, R.C. and Brooks, R. (2000) Efficient generalization and abstraction of network data using perceptual grouping. Proceedings of the 5th International Conference on GeoComputation: Greenwich, August 23-25.

Touya, G. (2007) River network selection based on structure and pattern recognition. Proceedings of the 23rd International Cartographic Conference, 4-9 August 2007, Moscow, Russia.

Töpfer, F. and Pillewiser, W. (1966) The principles of selection. Cartographic Journal, 3(1), 10-16

Ultsch, A. and Siemon, H. (1989) Technical report 329, University of Dortmund, Dortmund, Germany.

USGS National Hydrography Dataset, Feb. 2000.

Wolf, G.W. (1988) Weighted surface networks and their application to cartographic generalization. In: Barth, W. (ed) Visualization Technology and Algorithm. Springer-Verlag, Berlin, 199–212