An evaluation of unsupervised and supervised learning algorithms for clustering landscape types in the United States

Jochen Wendel[1], Barbara P. Buttenfield[2], Lawrence V. Stanislawski [3]

[1] European Institute for Energy Research (EIFER), Karlsruhe Institute of Technology, Karlsruhe Germany
Email: jochen.wendel@kit.edu

[2] Department of Geography, University of Colorado, Boulder Colorado USA
Email: babs@colorado.edu

[3] Center for Excellence in Geospatial Information Science (CEGIS),
United States Geological Survey (USGS), Rolla, Missouri USA
Email: lstan@usgs.gov

Landscape regions may be defined using terrain roughness, precipitation, and geologic structure. Knowledge of landscape type can inform cartographic generalization of hydrographic features, because landscape characteristics provide an important geographic context within which to understand variations in channel geometry, flow pattern and network configuration. The U.S. Geological Survey is exploring methods to automate generalization of features in the National Hydrography Dataset (NHD), which represents surface water in the coterminous USA and spans multiple landscape regions. The goal is to associate specific sequences of processing operations and specific ranges for processing parameters, obviating manual selection of a processing strategy for every NHD watershed unit. Several methods have been applied to delineate physiographic regions, beginning with Fenneman and Johnson (1945) and moving to an existing landscape classification by maximum likelihood methods based on seven input variables including average elevation, elevation standard deviation, bedrock density, drainage density, area of inland surface water, runoff, and average slope. The existing solution shows problems in areas of high aridity and areas of high landscape diversity. There is some indication that additional input variables and more sophisticated classification methods can refine the existing classification, improving methods for generalizing hydrography.

This research compares unsupervised and supervised learning algorithms to establish an optimal number of classes and to refine the existing classification. Seven grids with 5 km spatial resolution were classified using unsupervised methods (Hierarchical Clustering, k-Means, Self-Organizing Maps) and supervised methods (k-Nearest Neighbor [k-NN], Random Forest, Support Vector Machines), and repeated for 7-12 classes. Evaluation metrics for unsupervised methods included the Davies-Bouldin index, the Silhouette index, and the Dunn index; and cross-validation evaluated supervised classification methods. Multiple evaluation criteria were compared progressively across the range of number of classes. The paper will report on the comparative analysis and its impact on the selection of landscape regions.

Keywords: clustering, classification, landscape regions, automatic generalization

Word count: 298