ISEIS

Journal of Environmental Informatics 11(2) 90-102 (2008)

Journal of Environmental Informatics

www.iseis.org/jei

A Comparison of Nebraska Reservoir Classes Estimated from Watershed-Based Classification Models and Ecoregions

H. N. N. Bulley[1,*], D. B. Marx[2], J. W. Merchant[3], J. C. Holz[4] and A. A. Holz[4]

[1]*Department of Geography and Geology, DSC 260, University of Nebraska-Omaha, Omaha, NE 68182, USA*
[2]*Department of Statistics, 342 Hardin Hall, University of Nebraska-Lincoln, Lincoln, NE 68583, USA*
[3]*Center for Advanced Land Management Information Technologies, School of Natural Resources, 306 Hardin Hall, University of Nebraska-Lincoln, Lincoln, NE 68583-0973, USA*
[4]*School of Natural Resources, 507 Hardin Hall, University of Nebraska - Lincoln, Lincoln, NE 68583-0995, USA*

Received 29 August 2006; revised 4 May 2007; accepted 1 January 2008; published online 30 May 2008

ABSTRACT. Regulatory agencies have been investigating a number of alternatives for classifying lakes into hydogeologically and ecologically similar assemblages that will facilitate establishment of attainable water quality standards. Concerns over the ability of traditional statistical classifiers to effectively classify environmental data have led to increasing interest in machine (predictive) learning classification tools such as decision trees. This paper compares the performance (classification strength) of a classification tree-based watershed classification model of Nebraska reservoirs to a discriminant analysis (DA)-based watershed classification system and reservoir classes derived from Omernik's Level IV Ecoregions. The performance of classification tree and DA-based watershed classification methods were also compared with respect to their cross-validation prediction errors. The results suggest that both watershed-based classification approaches (classification tree and DA) were more effective than Omernik's Level IV ecoregions in accounting for observed variations in water quality characteristics of Nebraska reservoirs. Moreover, this study demonstrates the utility of a classification tree algorithm, either as a supplement or alternative to DA, in handling the complexities of watershed variables and classifying Nebraska reservoirs for the purpose of water quality management. The classification tree also provides water resource managers with a useful interpretive classification interface.

Keywords: classification tree, discriminant analysis, ecoregions, reservoirs, water quality, watershed

1. Introduction

Concerted efforts to improve the quality of U.S. surface waters began over 40 years ago. As the U.S. Environmental Protection Agency (EPA) continues to develop criteria for lake water quality assessment, there is a need to account for hydrogeologic and ecological differences among lakes since these differences determine the inherent capacities of lakes to meet such criteria. A number of investigators have suggested that regulatory agencies should group lakes into hydogeologically and ecologically similar classes to improve management and decision-making processes (Conquest et al., 1994; Hawkins et al., 2000). The rationale is to establish water quality standards for lakes in different classes according to a set of benchmark conditions unique for each class.

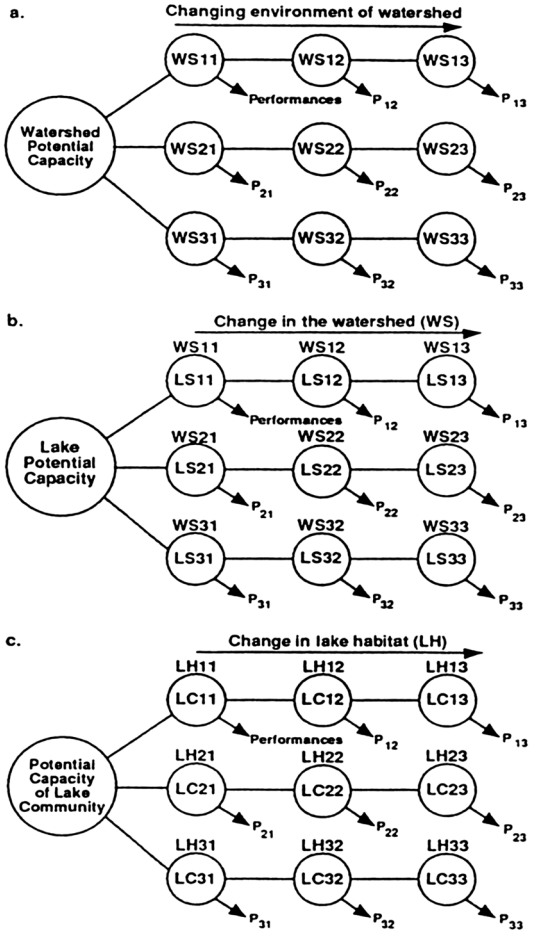
Bulley et al. (2007) recently demonstrated a GIS-based approach to classifying lake watersheds using classification trees. Focusing on Nebraska reservoirs, a procedure was developed to determine the optimal number of classes required to capture the underlying variability in the watersheds of Nebraska reservoirs (9 classes), and to identify the key watershed characteristics that contributed to the classification (e.g., in Nebraska, soil organic matter) (Bulley, 2004; Bulley et al., 2007). According to Bulley et al. (2007), "the goal was to model pre-settlement (potential) conditions for the purpose of establishing reference water quality conditions, therefore anthropogenic impacts were assumed to be minimal. This assumption was necessary in order to address the long term water management goal of developing baseline information (reference conditions) against which to assess human impacts on the reservoirs". A classification tree predictive model was used to describe the reservoir class structure.

This paper further examines the utility of the classification tree approach to watershed classification (see Bulley, 2004; Bulley et al., 2007) as a tool for lake water quality management. To achieve this goal, we compared the aforementioned classification tree approach with two other commonly used approaches to resource classification: Omernik's Level IV ecoregions (Omernik, 1987; EPA, 2002) and a discriminant analysis (DA)-based classification method (Momen and Zehr, 1998). The DA method was chosen because it has widely been used in ecological classifications despite issues with

* Corresponding author. Tel.: +1 402 554 3107; fax: +1 402 554 3518.
E-mail address: hbulley@mail.unomaha.edu (H. N. N. Bulley).

ISSN: 1726-2135 print/1684-8799 online
 doi:10.3808/jei.200800114

underlying assumptions that are often not met. Ecoregions, on the other hand, have frequently been employed as a framework for environmental and water resource assessment and management despite concerns that higher order ecoregions are often not congruent with watersheds (Omernik and Bailey, 1997; EPA, 2002; Omernik, 2003).



Notes: Each system develops according to its potential capacity and changes in its environment. The environment of the watershed is the biogeoclimatic region. The watershed forms the environment of the lake. The lake habitat is the environment of the community, the community and its habitat together forming the lake as a whole. Each state or stage in development of the watershed (a), the lake (b), and the community (c) are designated as Wij, LSij, and LHij, respectively, where i designates developmental environment and j designates developmental state. For example, W11, W12, W13 are three consecutive stages of development of the watershed in biogeociimatic environmental context (a). If the watershed develops in this way, then the lake, influenced by watershed changes, will develop through states LS11, LS12, and LS13 (b). Development of the lake entails concordant development of the lake habitat (LH11, LH12, and LH13) and the co-development, the watershed, the lake, and the community exhibit observable performances (p).

Figure 1. Conceptual hierarchical interaction among co-developing systems of lake watershed environment (a), habitat (b), and water quality (c) based on the potential capacity of each system (Modified from Lomnicky, 1995).

Sampled water quality data were used to assess the utility of these classification schemes for lake water management. This is because watersheds have a high degree of influence on biophysical and chemical characteristics of streams and reservoirs (see Figure 1). Furthermore, the GIS-based approach to classifying lake watersheds using classification trees proposed by Bulley et al. (2007) was intended to address the U.S. EPA's efforts to establish base-line water quality information or nutrient criteria (EPA, 2000), against which the impact of land use could be assessed via tools like *Total Maximum Daily Load* (EPA, 2003a; EPA, 2003b). Since the U.S. EPA is concerned about managing land use in lake watersheds, it seems inconsistent to employ water quality data, directly impacted by land use activities, as predictor variables in determining lake classes. However, we believe sampled water quality variables can be used for testing the utility of different lake classification schemes for water management (EPA, 2000; Robertson and Saad, 2003).

2. Background

2.1. Ecoregions

Ecoregions are defined as areas that comprise similar ecosystems demarcated principally on the basis of landforms, climate, potential vegetation and soils (Loveland and Merchant, 2004; Omernik, 1987; Omernik and Bailey, 1997; EPA, 2002). Ecoregions have frequently been used as a spatial framework for aquatic ecosystem management and assessment (Omernik 1987; Omernik and Bailey, 1997; EPA, 2002; Rohm et al., 2002; Omernik, 2003). Although ecoregions have been defined in a number of different ways, the EPA has most often employed those developed by Omernik (EPA, 2002).

Omernik's ecoregions are defined hierarchically at four scales, with Level I being the most general depiction and Level IV ecoregions being the most detailed. These products have been used extensively, although sometimes inappropriately, as a framework for sampling, and to assist in assessment of fisheries, wildlife communities, lake acidification and in other applications. Omernik and Bailey (1997) note that "ecoregions are generally useful for structuring the research, assessment, and management of all environmental resources, but may not be the best framework for any one particular resource." In the case of water resources, it is important to note that the boundaries of ecoregions often do not coincide with watershed boundaries, the most important units of analyses in evaluating surface water quality. Additionally, previous research has shown that, although ecoregions such as those defined by Omernik, are useful for general ecosystem management and analyses, they do not adequately account for the inherent variations in stream and lake water quality (e.g. Van Sickle and Hughes, 2000; Severn et al., 2001; Winter, 1999; Jenerette et al., 2002; Detenbeck et al., 2003 and 2004; Robertson and Saad, 2003).

2.2. Discriminant Analysis

Supervised, i.e. *apriori*, classification is most useful when one has clear classification objectives (i.e., target clas-

ses are known). Discriminant analysis (DA) is multivariate statistical approach to classification that has often been used in analyzing environmental datasets even though these datasets are sometimes problematic. DA uses empirical hypothesis testing approaches to determine which linear combination of input variables discriminate between two or more naturally occurring groups (Dunteman, 1984; Ripley, 1996; Legendre and Legendre; 1998, McGarigal et al., 2000). The discriminant function (δ) of the linear model is computed as a series of linear combinations of input vectors (x) that seek to maximize the separation between training classes as:

$$y = \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + ...\delta_p x_p = \delta' x \quad (1)$$

The classification problem then reduces to identifying the appropriate function (δ) in equation 1. Discussions of the use of discriminant analysis as well as applications for environmental analyses are provided by Legendre and Legendre (1998), McGarigal et al. (2000), and Huberty and Olejnik (2006).

According to De'ath and Fabricius (2000), environmental data are usually complex (e.g. have unequal variances) and values for certain variables may be missing. Such data are often characterized by multimodal distributions, and the relationships among variables are commonly non-linear and involve high-order interactions that may render traditional statistical techniques ineffective for data exploration, pattern recognition and modeling. This depiction is true of lake biophysical and chemical water quality parameters (e.g. phosphorus) and watershed characteristics (e.g. watershed area and soil organic matter).

Breiman et al. (1984), James and McCulloch (1990) and Quinlan (1993) discussed the limitations of DA. In summary, effective use of DA must meet certain distributional assumptions. These include the assumption that all explanatory variables follow a multivariate normal distribution for each class of response variable, and the variance-covariance matrices for each class are equal. Although the assumption of normality is critical to DA, this method is usually applied irrespective of whether the assumption is true for every explanatory variable employed in the analysis. Since the DA classification method is mostly suitable for dichotomous predictor variables, categorical variables need to be transformed into a series of dummy variables and this can lead to problems of dimensionality. Moreover, the DA method may be limited in dealing with cases of missing explanatory variables and hence observations with missing variables are usually dropped from the analyses. This can lead to unintended bias due to elimination of variables that might otherwise be critical to developing an appropriate classification rule.

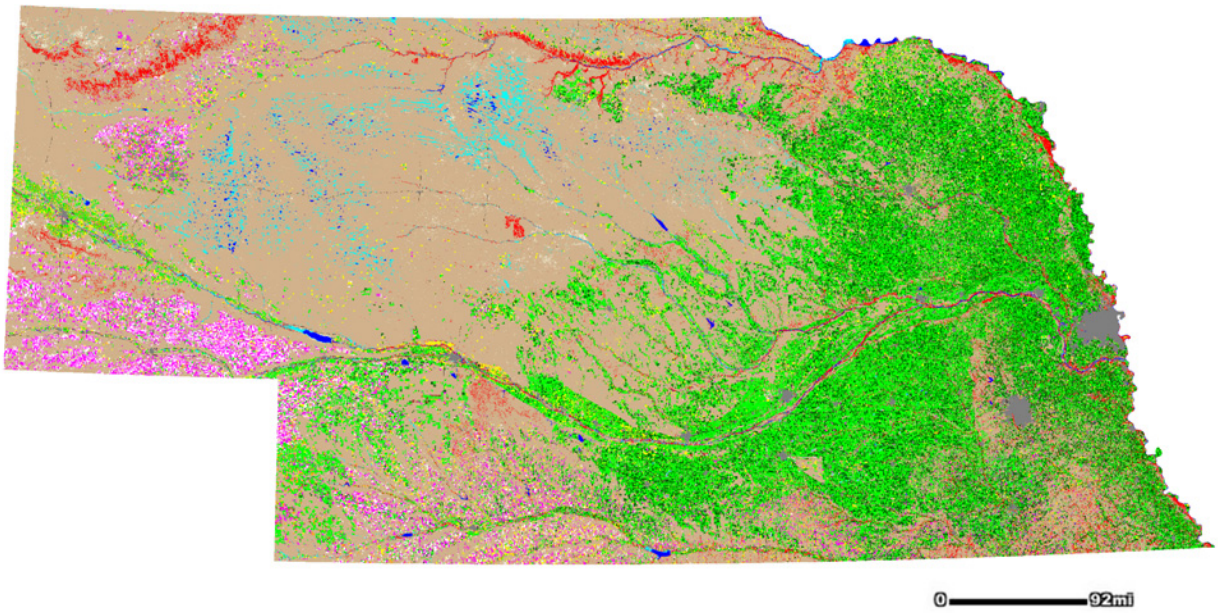
Some alternatives to the use of DA in resolving classifycation problems include the multinomial logistic regression, mixture discrimant analyses and probit models. However, these alternatives have some limitations similar to DA since they are only suitable for categorical data, and may produce biased results when the dataset contains missing variables (Ripley, 1996; Hastie et al., 2001; Feldesman, 2002).

2.3. Decision Tree Classifiers

Concerns over the ability of multivariate statistical classifiers, such as discriminant analysis (DA), to effectively classify environmental datasets have led to increasing interest in machine learning classification tools such as neural networks, genetic algorithms (or genetic programming), and decision tree classifiers (e.g., German et al., 1999; Vayssieres et al., 2000; Park et al., 2003; Karels et al., 2004; Joy and Death, 2005; Luoto and Hjort, 2005; Baker et al., 2006; Ghaffari et al., 2006; Baker et al., 2006). Machine learning involves the application of inductive algorithms to resolve classification problems. After an extensive review of machine (predictive) learning methods, Friedman (2006) indicated that decision trees are the most frequently used predictive learning techniques. This is partly due to their invariance to monotone transformations of the predictor variables; resistance to irrelevant predictor variables; they do not require a lot of training as with artificial neural networks; and the classification rules of decision trees are simple and interpretable (e.g., Gahegan and West, 1998; Olden and Jackson, 2002; Goel et al., 2003). Our previous work focused on evaluating classification trees as potential modeling tool for watershed-based reservoir classification (see Bulley et al., 2007). Hence, this paper is parti ularly concerned with classification trees, a form decision trees method.

Decision tree classifiers are usually implemented as rule-based classifiers (Hunt et al., 1966; Breiman et al., 1984; Quinlan, 1986 and 1993; Verbyla, 1987; Ripley, 1996; Mitchell, 1997; De'ath and Fabricius, 2000; Rokach and Maimon, 2005). A simple form of rule-based classifier is a hierarchical construction (tree) with various levels (leaves). A more rigorous form of decision trees employs the recursive partitioning non-parametric statistical method, which can account for non-linear relationships, higher order interactions and missing values in a dataset (Breiman et al., 1984; Verbyla, 1987; De' ath and Fabricius, 2000; Feldesman, 2002). There are two types of decision tree models: regression trees are appropriate when the dependent variable is numeric, whereas classification trees are more relevant for instances with categorical dependent variables, e.g. lake classes.

In general, decision tree approaches offer several advantages over multivariate statistical approaches, such as DA, when dealing with environmental datasets. For example, classification tree methods are not limited by prior knowledge of dataset distributions, since modeling of these distributions is not required. Thus, in contrast to Bayesian approximators (such as DA) and maximum likelihood classifiers, decision tree algorithms can easily handle multimodal distributions and they have no restrictions on sample size. Moreover, decision tree algorithms have been shown to outperform multivariate statistical approaches in accounting for variations in environmental datasets for classification tasks (e.g., Verbyla, 1987; Emmons et al., 1999; German et al., 1999; De' ath and Fabricius, 2000; Vayssieres et al., 2000; Feldesman, 2002;



Notes: Green and light brown areas represent agricultural land use and prairies respectively

Figure 2. Map of Nebraska land cover (Source: CALMIT).

Robertson and Saad, 2003; Karels et al., 2004; Luoto and Hjort, 2005).

Previous authors (Breiman et al., 1984; Quinlan, 1986 and 1993; Ripley, 1996; Mitchell, 1997; De'ath and Fabricius, 2000) have provided detailed discussions of decision tree procedures. Here, only the classification tree method is reviewed because it was used to implement one of the watershed-based classifications of interest in this paper. Classification tree methods discriminate the attribute space of a dataset into K disjoint groups, K_r (r = 1, 2, …, k), based on decision rules that are parallel or orthogonal to the attribute axis. The classification tree identifies the best possible path (and attributes) to partition the feature space and traces a path down the tree from the root node (dataset) to leaves (classes). Each node of the tree represents a set of rules that progressively refines the classification in a top-down hierarchical approach. Classification trees can represent higher levels of complexity or deep trees (where the class segregation is difficult) and more simplistic rule sets (short trees) when appropriate.

The classification tree process involves a binary recursive partitioning of the data into successive nodes. The process is binary because the parent nodes are always split into exactly two subsequent nodes and recursive because the process can be repeated by treating each subsequent node as a parent until there are no more splits, i.e. terminal nodes (e.g. reservoir classes) (Breiman et al., 1984; Quinlan, 1993). The basic components of the classification tree building process include: a set of questions; splitting criteria; and rules for assigning a class to a terminal node. Attributes that do not seem to contribute to defining ultimate terminal nodes are usually excluded in the final tree structure, leaving only those attributes that influence the overall classification process (Breiman et al., 1984; Quinlan, 1993).

The foregoing sections provide an overview of the relative strengths and limitations of ecoregions and DA as classification tools. As mentioned earlier, the primary objective of this paper is to compare the performance of a classification tree-based watershed classification model of Nebraska reservoirs, to a discriminant analysis (DA)-based watershed classification system (Momen and Zehr, 1998) and to reservoir watershed classes based on Omernik's ecoregions (Omernik, 1987; EPA, 2002). The two watershed-based reservoir classification approaches were hypothesized to outperform ecoregions in defining *apriori* classes of Nebraska reservoirs. Our study focuses on reservoir watersheds located in the agriculturally-dominated landscape of Nebraska, a region that is representative of mid-latitude states with substantial agricultural-based economies in the United States.

3. Methods

The classification tree-based watershed classification developed by Bulley et al. (2007) was compared to Omernick's Level IV ecoregions (Omernik, 1987; EPA, 2002) and discriminant analysis (DA)-based watershed classification methods (Momen and Zehr, 1998). The comparison was a two step process: first, the watershed-based classifications and Omernik's ecoregions were assessed with respect to their abilities to account for observed variations in water quality parameters of Nebraska reservoirs; second, the classification tree and DA approaches to reservoir classification were compared with regards to their respective prediction errors. Comparing differ-

Table 1. Environmental Datasets Used in Nebraska Reservoir Classification

Dataset	Abbreviation	Units	Source
Climate data (annual means)			
Maximum temperature	Temp_max	^oC	DAYMET (Thornton et al., 1997)
Minimum temperature	Temp_min	^oC	DAYMET (Thornton et al., 1997)
Total precipitation	Ppt_tot	mm	DAYMET (Thornton et al., 1997)
Precipitation intensity	Ppt_intns	mm	DAYMET (Thornton et al., 1997)
Humidity	Humidity	mmHg	DAYMET (Thornton et al., 1997)
Growing degree days	GDD(base 10^{o}C)	degrees	DAYMET (Thornton et al., 1997)
Terrain data:			
Lake Area	LA	ha	Updated Nebraska lakes map
Watershed area	WA	ha	EDNA DEM-derived watersheds
Lake area : watershed area	LA:WA	unitless	
Mean watershed slope	Slope	degrees	EDNA DEM (edna.usgs.gov)
Mean watershed elevation	Relief	degrees	EDNA DEM (edna.usgs.gov)
Watershed relief	Elevation	m	EDNA DEM (edna.usgs.gov)
Total drainage length	Drn_T	m	EDNA streams (edna.usgs.gov)
Drainage density	Drn_D	mm^{-2}	EDNA streams (edna.usgs.gov)
Soils biophysical data:			
Erodibility	Kfact	unitless	STATSGO (Soil Survey Staff, 1993)
Clay content	Clay	% weight	STATSGO (Soil Survey Staff, 1993)
Permeability	Perm	$inhr^{-1}$	STATSGO (Soil Survey Staff, 1993)
Infiltration rate	Infilt	$inhr^{-1}$	STATSGO (Soil Survey Staff, 1993)
Organic matter content	OM	% weight	STATSGO (Soil Survey Staff, 1993)
Soils chemistry data:			
Salinity	Sal	$mmhoss^{-1}$	STATSGO (Soil Survey Staff, 1993)
Soil reaction	pH	unitless	STATSGO (Soil Survey Staff, 1993)
Cation exchange capacity	CEC	unitless	STATSGO (Soil Survey Staff, 1993)
Soil carbonate	$CaCO_3$	% $CaCO_3$	STATSGO (Soil Survey Staff, 1993)

ent classification methods can be problematic since there are different ways to set up each classifier. Hence, only default forms of classification tree (See5® software) and DA (implemented in SAS software) were used (i.e., without accuracy enhancements such as prior probabilities for DA and boosting for classification trees).

It is important to note at this point that sampled lake water quality data were used to assess the utility of these classification approaches for lake water management. As mentioned earlier, it is inconsistent to use land cover or water quality data that are directly impacted by land use activities, as predictor variables in determining lake classes in order to manage land use in the lake watersheds. However, the water quality variables can be used in testing the utility of lake classification schemes for water quality management (EPA, 200b; Robertson and Saad, 2003).

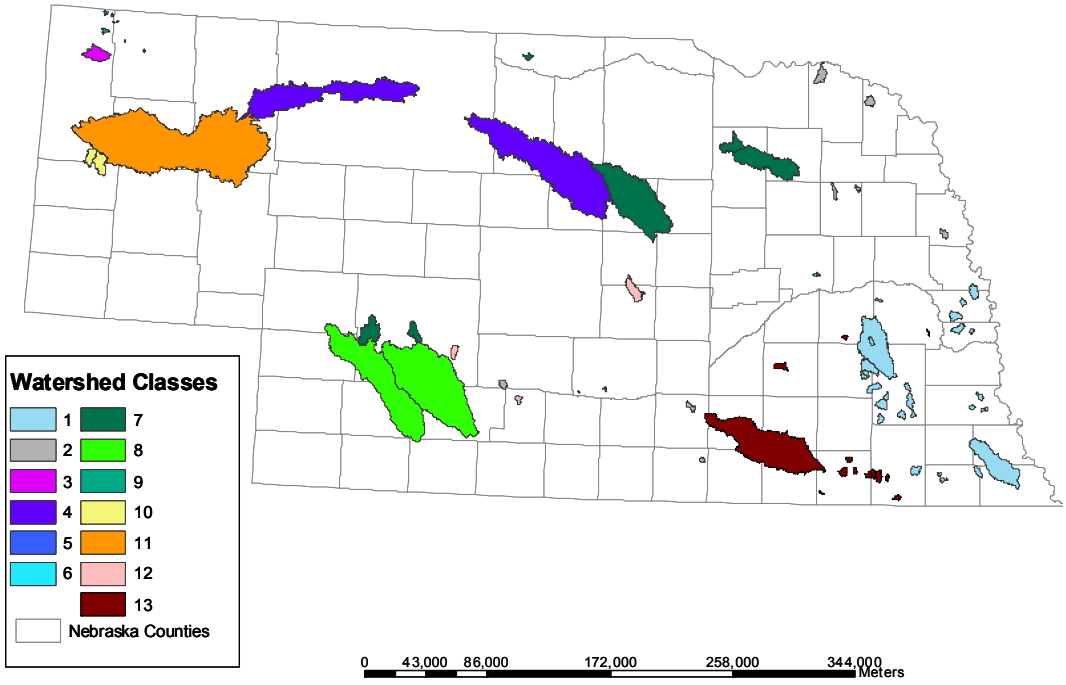
3.1. Study Area

This research focuses on Nebraska, an agriculturally-dominated state that covers a broad range of climatic, physiographic, land use and water quality conditions (Figure 2). The Sand Hills (grass covered sand dunes mostly devoted to grazing) in western Nebraska occupy about 30 percent of the state. Elevations range from 256 meters in the east to 1,654 meters in the west, and climate follows a gradient of rainfall and temperature regimes from east to west. The average annual precipitation ranges from 36 cm in the northwest to 86 cm in southeast Nebraska (Johnsgard, 2001; Kuzelka et al., 1993). Nebraska has about 13,500 lakes including natural lakes, reservoirs, and sand pits. The primary cause of water quality impairment in Nebraska is the transport of soil sediments, agrochemicals and animal wastes through surface runoff from farmlands into streams and lakes.

3.2. Dataset Development

Water quality data for 78 sampled reservoirs were derived from a database of Nebraska lake water quality developed through field sampling conducted between 1988 and 2003 in the months of May-August. The water quality variables employed in this study are chlorophyll-a, Secchi depth, total phosphorus, total nitrogen and alkalinity of lake waters (EPA, 2001; Holz, 2002). These variables have been identified as candidate reference water quality parameters by the U.S. EPA for use in developing lake nutrient criteria (EPA, 2000). Additionally, two potential agrochemical herbicide pollutants (Atrazine and Alachlor) were included in the analysis because the outcome of this study may also have implications on how the reservoir classification methods could assist in managing non-point source pollution of lake water quality from agrochemical effluents via stream runoff. For each lake, the annual mean value was determined for the water quality variables.

Environmental characteristics for each reservoir watershed were extracted from a variety of sources (Table 1, Figures 3 and 4) (Bulley et al., 2007). These characteristics in-



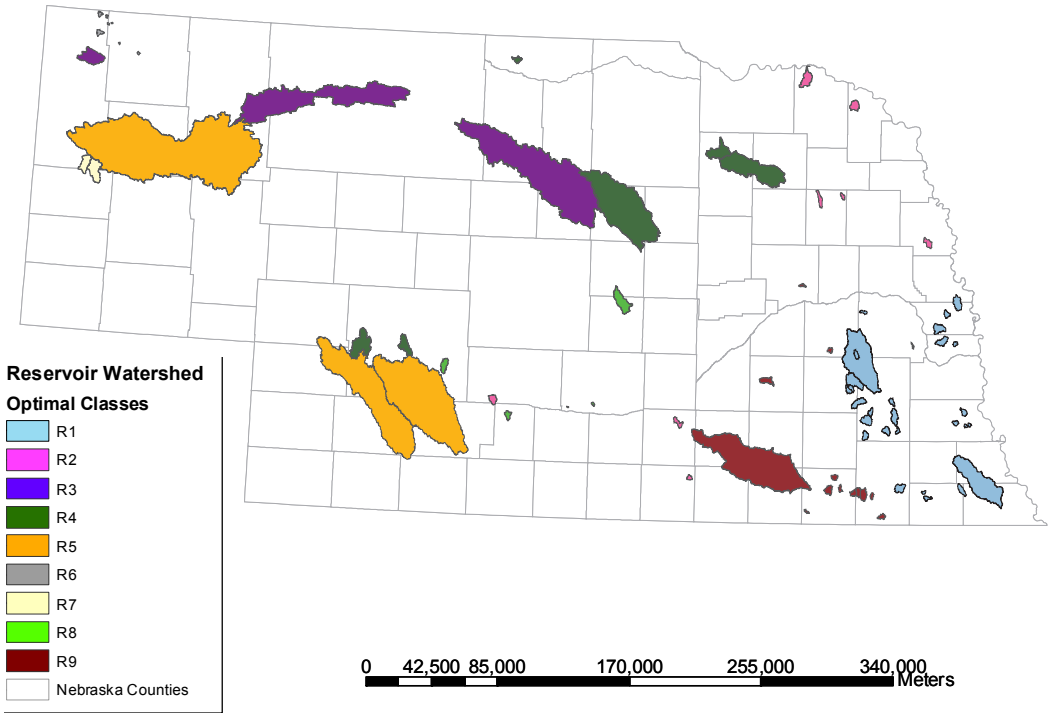
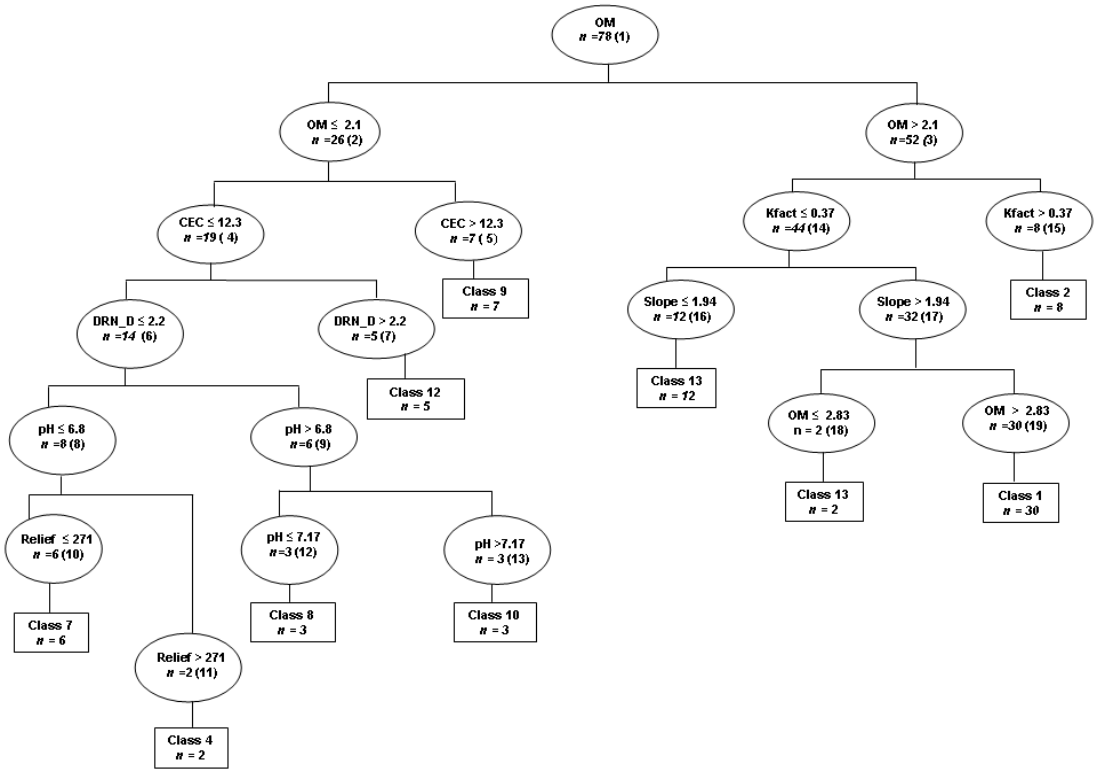


Figure 3. Map of revised reservoir watershed classes: (a) 13 preliminary classes, and (b) 9 optimal classes (Source: Bulley et al., 2007).

clude watershed area, watershed slope and relief, soil erodibility, soil infiltration rate, soil organic matter, soil reaction (pH), soil cation exchange capacity, soil carbonate, soil clay content, soil water holding capacity, soil permeability, and climate (precipitation, temperature and humidity). An important consideration in selecting data was the potential to obtain data

Table 2. Nebraska Reservoir Classes Derived from Watershed-Based Classification Tree Method

9 classes	13 classes	No. of reservoirs	Characteristics of Class
R1	1	30	Located in southeastern Nebraska. Most of the reservoir watersheds in this group are small on average; characterized by high organic matter content and relatively low erodibility. Adjacent to R9, but these watersheds have higher erosion potential (steeper slopes) than the R9 watersheds.
R2	2	8	Located in northeastern Nebraska and average reservoir size is in the lower 25th percentile of the data. Watersheds are generally small and characterized by low relief, high soil erodibility and organic matter content.
R3	3 & 4	3	Located in northwestern and north central Nebraska. This group is characterized by both large and medium watersheds, relatively low soil organic matter content and high relief.
R4	7	6	Watersheds aligned diagonally between central and southwestern Nebraska. Watershed conditions are similar to those of R8 and R6 watersheds, except that the R4 watersheds have lower relief and pH than the R8 and R6 watersheds, respectively.
R5	8 & 11	3	Watersheds aligned between southwest and northwestern Nebraska. Watersheds in this group are characterized by high relief, and alkaline soils with low soil organic matter content.
R6	9	7	Located in northwestern Nebraska and characterized by high buffering capacity. This is indicative of the soil and vegetation of the Niobrara shrub land.
R7	10	2	Located in northwestern Nebraska and adjacent to R5 watersheds. However, R7 watersheds are relatively smaller and characterized by lower relief and highly alkaline soils as compared to R5 watersheds
R8	5 &12	5	Located in low relief areas along the Platte river valley in central Nebraska and characterized by small sized watersheds, low soil organic matter content.
R9	6 & 13	14	Located in southeastern part of Nebraska and adjacent to R1 watersheds. Watersheds in this group are characterized by relatively lower erosion potential and soil organic matter content than R1 watersheds.



Notes: Terminal nodes (classes) are represented by rectangular boxes, while oval boxes represent non-terminal nodes that required additional splitting. The number in parenthesis indicates the node number.

Figure 4. Classification tree model for Nebraska reservoir classes (Source: Bulley et al., 2007).

nationally in GIS-compatible format (Bulley et al., 2007). ArcGIS software was used to append the annual mean water quality data to the 9 and 13 reservoir classes as defined by Bulley et al. (2007) (Table 2).

Ecoregions of Nebraska were extracted from a dataset of Omernik Level IV ecoregions of the conterminous United States (Omernik, 1987; EPA, 2002; http://www.epa.gov/wed/pages/ecoregions.htm). The United States ecoregions dataset was clipped to a GIS coverage of Nebraska using ArcGIS software. A GIS "point" coverage of the sampled reservoir locations was overlaid on Omernik's Level IV ecoregions of Nebraska in order to identify those ecoregions that included sampled reservoirs. The annual mean values of growing season index period data for each water quality variable were computed and summarized for the ecoregions.

The U.S. Geological Survey Elevation Derivatives for National Applications (EDNA) dataset was used for delineating the reservoir watersheds. The foundation for EDNA is a seamless 30-meter resolution digital elevation model (DEM) for the conterminous United States (Verdin and Verdin, 1999; Gesch et al., 2002; USGS, 2003). Therefore, the comparison between watershed-based classifications and ecoregions derived reservoir classes has potential national applications.

3.3. DA-based Reservoir Classification

Discriminant analysis (DA) was performed on watershed characteristics of 78 sampled reservoirs that were previously used in a classification tree-based watershed classification (Bulley, 2004; Bulley et al., 2007). This was done in order to compare the effectiveness of DA, ecoregions and classification tree approaches in grouping Nebraska reservoirs for water quality management. Since DA is a parametric method, its distributional assumptions may limit the validity of the prediction error, since we are especially interested in comparing DA to the classification tree non-parametric method.

Resampling approaches (jackknifing, bootstrapping and cross-validation) offer non-parametric means to perform statistical significance tests of DA (Stone, 1974; Breiman et al., 1984; Efron and Tibshirani, 1993; Ronchetti et al., 1997; Efron, 2003). The jackknife approach involves resampling without replacement, while bootstrapping involves resampling with replacement. Cross-validation is fundamentally different from jackknife and bootstrapping in that the latter are used to compute estimates of bias and variances whereas cross-validation is used for model selection. For this reason the DA approach in this study employed the cross-validation resampling option.

The DA was implemented in SAS® software using a "Discrim" procedure with a cross-validation option (SAS Institute, 2000). The output of the cluster analysis of watershed characteristics datasets based on 13 and 9 classes respectively was employed (see Figure 3) (Bulley et al., 2007). This was done to explore the effectiveness with which the DA handles the more complicated 13-class dataset (involving 13 classes and single object classes) as compared to the less complicated 9-class dataset (no single object classes). Cross-validation prediction errors were determined for both 13 and 9 class datasets respectively. The predicted reservoir class memberships for 13 and 9 classes, derived from the DA-based watershed classification, were extracted. ArcGIS was then used to append the class membership information to a watershed characteristics dataset that included predicted reservoir classes for classification tree-based watershed classification (13 and 9 classes) and ecoregions. This dataset was then used to compare DA to the other two classification schemes.

3.4. Comparison of Classification Methods

The watershed-based classification tree and DA classifycations methods were compared to ecoregions regarding their abilities to account for variations in observed water quality parameters of Nebraska reservoirs. This was done using the concept of classification strength, which measures how strongly different landscape classification approaches can separate water quality water conditions (Van Sickle and Hughes, 2000). A modified version of classification strength (CS) was estimated as the extent to which average within-class water quality variations exceeded the average variations between reservoir classes. The CS is defined as a function of within-class heterogeneity and between-class separation (modified from Van Sickle and Hughes, 2000) as:

$$CS = \frac{\varpi}{\beta} \tag{2}$$

where β is variability in water quality conditions between classes; and ϖ is the variability in water quality conditions within classes.

The variance in mean annual water quality is given as:

$$s^2 = \sum \frac{(x_i - X)^2}{n-1},\ i = 1, 2 \ldots n \text{ reservoirs} \tag{3}$$

where x is annual mean value of water quality (e.g. chlorophyll-a) for each reservoir; X is sample mean; and n is the number of reservoirs in each class.

The CS was computed for each water quality parameter and the results were summarized into three categories as follows:

1) Biophysical water quality (chlorophyll-a and Secchi depth);

2) Chemical nutrient water quality (total phosphorus, total nitrogen and alkalinity);

3) Agrochemical herbicide effluents (Atrazine and Alachlor).

Since the aim of this study was to identify the Nebraska reservoir classes that could be used to establish water quality and nutrient criteria, it was expected that a decrease in CS value would represent an increase in interclass heterogeneity or increase in within-class homogeneity. Consequently, the classification method with the lowest CS value in each of the 3 water quality categories was considered to be most effective. Finally, the performance of classification tree and DA me-

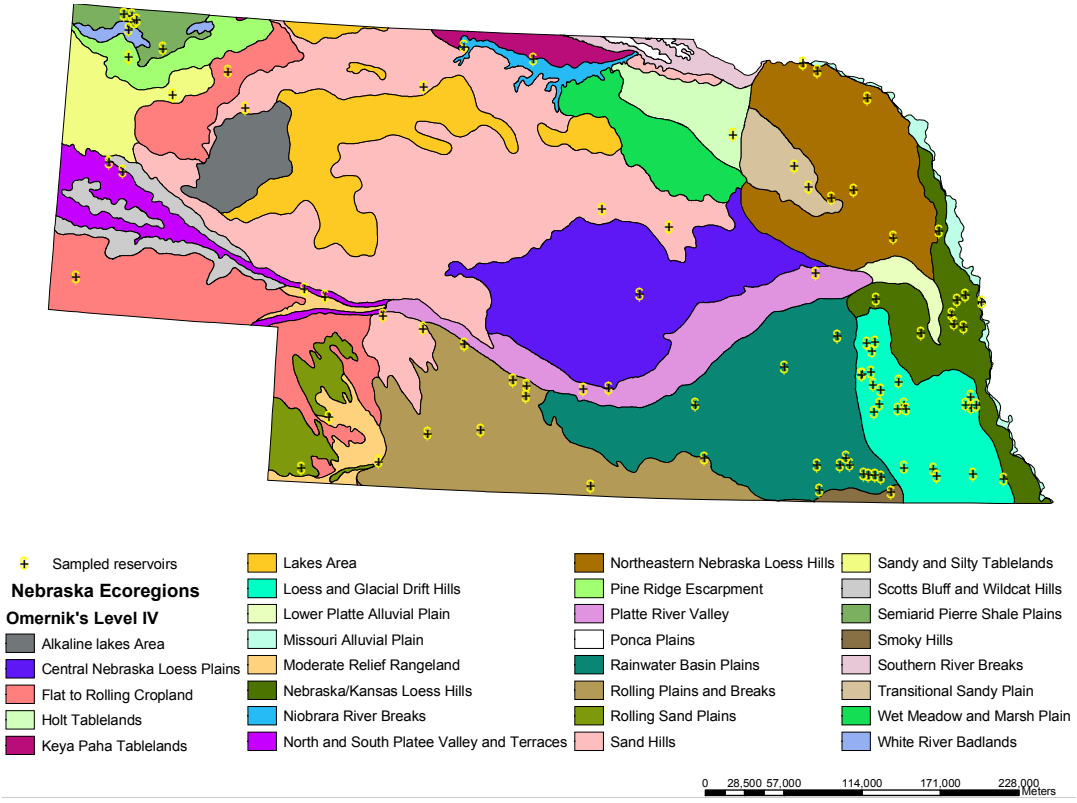


Figure 5. Sampled reservoirs sites overlaid on Omernik's Level IV Ecoregions of Nebraska.

thods were compared based on their cross-validation prediction errors.

4. Results and Discussions

A map of the sampled reservoirs overlaid on Omernik's Level IV ecoregions of Nebraska is shown in Figure 5. Although 20 out of the 27 Nebraska Level IV ecoregions contained sampled reservoirs, only 9 of these ecoregions had sufficient water quality data or contained more than one reservoir. Consequently, 9 Omernik Level IV ecoregions of Nebraska were used in evaluating the effectiveness of the two watershed-based classification methods.

4.1. Classification Comparison

Classification strength (CS) was used to assess the utility of the three classification methods as potential tools for lake water quality management. The CS of ecoregions, discriminant analysis (DA) and classification tree-based reservoir classifications are shown in table 3. These results were summarized with respect to three water quality categories; namely, biophysical, chemical nutrients and agrochemical herbicide effluents. For each category, the classification method with lowest CS value was considered to be most effective. Overall, both watershed-based classification approaches (classification trees and DA) were more effective than ecoregions in accounting for the variations in water quality characteristics of Nebraska reservoirs. This is not surprising since reservoirs are highly impacted by their watersheds, and ecoregions are generally not congruent with watershed boundaries.

The DA method was most effective in separating biophysical water quality parameters. Also, the classification tree approach was most effective in accounting for variations in both nutrients and herbicide water quality parameters. Although ecoregions seem to have lower CS values than both watershed-based classification methods with respect to total nitrogen and total phosphorus (Table 3), the relatively high CS value for alkalinity lessens any potential usefulness of ecoregions as water quality assessment tool. This is particularly important because the alkalinity of lake waters determines their natural buffering capacity; thus alkalinity helps to regulate pH changes and photosynthetic uptake of plant nutrients like phosphorus and nitrogen (Wetzel and Likens, 2000).

The above results are in agreement with previous findings that ecoregions do not adequately account for variations in lake water quality parameters (Van Sickle and Hughes, 2000; Severn et al., 2001; Jenerette et al., 2002; Detenbeck et al., 2003 and 2004). For example, Jenerette et al. (2002) tested the hypothesis that Omernik's ecoregions will allow for discriminating lakes of different water quality and suggested that the spatial distribution of lake ecosystems is more complicated than the biophysical characteristics represented by

Table 3. Comparison of Classification Strength of Reservoir Classification Methods (CS = ϖ / β *)

	DA Watershed Classes	Ecoregions	See5® Watershed Classes
Number of classes	8	9	8
Water Biophysical Parameters:			
Secchi Depth	1.520	1.053	1.538
Chlorophyll-a	3.090	5.992	4.893
Average	2.305	3.523	3.215
Water Chemistry Parameters:			
Alkalinity	2.075	6.584	2.146
Total Nitrogen**	1.047	0.874	1.015
Total Phosphorus	2.760	0.4904	2.532
Average	1.961	2.649	1.897
Agrochemical Herbicide Effluents:			
Atrazine	1.4214	1.301	1.249
Alachlor	1.877	1.644	1.371
Average	1.649	1.472	1.310

* ϖ is within class variation; and, β is between class variations; ** Adjusted mean value of total nitrogen was used in this analysis;
Note: DA was implemented using SAS® "Discrim" procedure (SAS Inc., 2000); Classification tree was implemented using See5® software (RuleQuest Research).

ecoregions. The use of classification strength in assessing the effectiveness of classification methods is dependent on sampled water quality data. Hence, the classification strength comparison is to some extent affected by limitations of sampling lake water quality parameters. These limitations include the need for extensive and frequent sampling of lakes in a given region which can be costly in terms of manpower and equipment.

Subsequent comparison of classification tree and DA approaches to reservoir watershed classification was based on their respective cross-validation prediction errors (Table 4). Cross-validation prediction error or accuracy evaluations often tend to be conservative compared to other accuracy evaluation methods, and that may explain the relatively high error values in Table 4. For the 13-class dataset, the percent cross-validation prediction error of classification tree and DA are 26.33 and 40.59 respectively. This suggests that the classify-cation tree did a better job than DA in segregating the 13-class dataset which consisted of single-member classes. On the hand, these single-member classes were reassigned to the closest class in the 9-class dataset (Table 2 and Figure 3b). Consequently, the percent cross-validation prediction error of classification tree and DA are 16.84 and 10.29 respectively for the 9-class dataset. That, the DA performed better than the classification tree with regards to the 9-class dataset supports previous suggestions that DA is still useful when we have complete datasets devoid of any complexities such as missing values or in this case, single-member classes. However, the considerable change in prediction error from the 40.59 (13-class dataset) to 10.29 (9-class dataset) is indicative of how perturbations or complexities (such as single-member classes) in a dataset can reduce the predictive effectiveness of the DA method (Breiman et al., 1984; Quinlan, 1993; De'ath and Fabricius, 2000; Vayssieres et al., 2000; Feldesman, 2002; Karels et al., 2004; Luoto and Hjort, 2005).

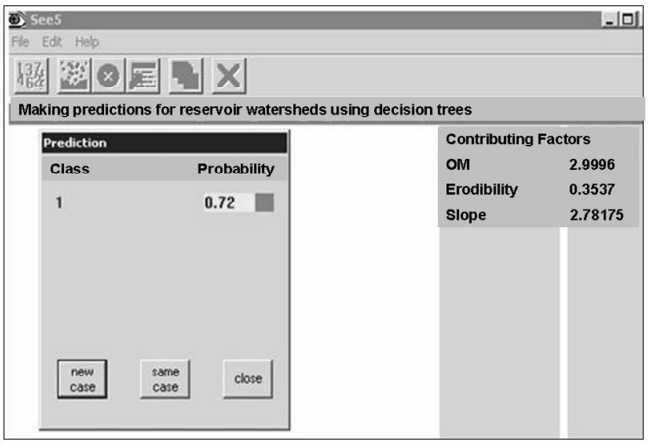
Table 4. Comparison of the Performance (Prediction Strength) for Watershed-Based Reservoir Classification Methods

	Prediction Strength (percent cross-validation error)	
Number of classes	13-classes	09-classes
SAS® DA	40.59	10.29
SEE5® Classification tree	26.33	16.84

These findings seem to be in agreement with previous assertions that classification trees are an especially useful alternative (or supplement) to DA when dealing with environmental datasets that are characterized by complexities such as missing data, and multimodal distribution (Breiman et al., 1984; Quinlan, 1993; German et al., 1999; De'ath and Fabricius, 2000; Vayssieres et al., 2000; Feldesman, 2002; Karels et al., 2004; Luoto and Hjort, 2005). In particular, the results of comparison between DA and classification tree methods were in agreement with previous analyses (Emmons et al., 1999; German et al., 1999; De' ath and Fabricius, 2000; Vayssieres et al., 2000; Feldesman, 2002; Karels et al., 2004; Luoto and Hjort, 2005). For example, Emmons et al. (1999) found that the decision tree method resulted in lower-rates of misclassification and more interpretable classes of Northern Wisconsin lakes than DA-derived classes.

We have so far found no studies where classification trees have been used in watershed-based analyses, for which the unit of analysis is the lake, reservoir or even stream watershed. Our results suggest that the watershed is a more appropriate unit of analyses, than ecoregions, for reservoir assessment in agriculturally dominated ecosystems. Furthermore, we have demonstrated the utility of classification tree algo-

rithm, either as a supplement or alternative to DA, in classifying reservoirs for the purpose of water quality management. It is also noteworthy that the SEE5® classification tree software employed in this study can be used to generate an interpretative user-interface and this has potential applications for water resource managers (Figure 6).



Notes: Example showing use of the interface to assign Yankee Hill Reservoir (near Lincoln, Nebraska) to class 1 with 72% probability.

Figure 6. Interpretative classification interface.

5. Conclusions

A classification tree-based reservoir watershed classification method was compared to Omernik's Level IV ecoregions and discriminant analysis (DA-based watershed classification methods). This comparison was done to assess the utility of the three classification approaches as potential tools for lake water quality management. Initially, the abilities of the watershed-based classifications and ecoregions to account for variations in water quality parameters of Nebraska reservoirs were evaluated with respect to their classification strength. Secondly, the predictive effectiveness of classification tree and DA-based reservoir watershed classification methods were assessed based on cross-validation prediction errors.

Sampled Nebraska reservoirs (78) were grouped into various classes using the classification tree and DA-based methods and ecoregions. Annual mean summaries of water quality parameters (chlorophyll-a, Secchi depth, alkalinity, total phosphorus, total nitrogen, Atrazine and Alachlor) were generated and appended to classification tree, DA, and ecoregions derived reservoir classes respectively. A classification strength metric (which measures how strongly different landscape classification approaches could separate water quality conditions) was used to evaluate the effectiveness of watershed-based reservoir classifications and ecoregions-derived reservoir classes. The results suggest that both watershed-based classification approaches (classification tree and DA) were more effective than Omernik's Level IV ecoregions in accounting for the variations in water quality characteristics of Nebraska reservoirs. This outcome was in agreement with previous findings that despite their usefulness in other ecological applications, ecoregions may not adequately account for variations in lake water quality parameters.

Also, the classification tree and DA-based watershed classification methods were compared with respect to their cross-validation prediction errors. The results suggest that the classification tree method was more effective in handling the complexities of watershed characteristics dataset and Nebraska reservoir classes. These results are in line with previous assertions that classification trees are especially useful alternative (or supplement) to DA when dealing with environmental datasets that are characterized by complexities such as nonlinear relationships, multimodal distributions and missing data. An important feature of the See5® classification tree software is the interpretive classification interface that can be used to predict the classes to which new cases (reservoirs) belong. This interface is particularly useful to water resource managers interested in identifying the class membership of a particular lake, in order to explore management options for the reservoir being considered.

It is noteworthy that classification trees do not allow for the inclusion of prior knowledge of known relationships between watershed characteristics and reservoir water quality to improve the classification results, e.g. weighting of watershed characteristics using lake area (Minka and Picard, 1997). However, this limitation can be overcome by exploring expert systems to incorporate prior knowledge of watershed characteristics and water quality parameters in a post-classification process to refine the results of the classification tree (Lauritzen and Spiegelhalter, 1988; Neapolitan, 1990).

Relationships between watershed characteristics and water quality parameters have been derived using parametric statistical methods such as correlation and linear regression analysis. Regression trees are non-parametric methods and can be used to derive simple but ecologically interpretable associations between watershed characteristics and water quality parameters. This is because the regression trees algorithm uses both numeric and categorical explanatory variables in assessing relationships or associations among the variables of interest. For example, Robertson and Saad (2003) used regretssion trees to determine the most statistically significant environmental characteristics affecting the water quality parameters of streams in the upper Midwestern United States. Hence, additional work is needed to explore how to incorporate the relationships between watershed characteristics and water quality in either pre-processing the input variables of classification tree modeling (to enhance the splitting process) or into post-classification expert systems (to refine the classification tree modeling results).

Despite the aforementioned issues, the study results provide some useful insights into the utility of the watershed-based classification tree and DA versus ecoregions in grouping reservoirs for water quality management. Both watershed-based classification approaches (classification tree and DA) outperformed Omernik's Level IV ecoregions and hence were more effective in accounting for the variations in water qua-

lity characteristics of Nebraska reservoirs. These results concur with previous suggestions to explore classifications frameworks other than ecoregions as tools for water quality assessment and management.

Of the watershed-based classification methods that were examined in this study, the DA seems to perform better when the dataset is not complex (e.g., no missing data or single-member classes). However, the predictive strength of DA decreases sharply in the face of data perturbations (in this case, inclusion of single-member classes). This is coupled with the need to meet all distributional assumptions of DA in order to make the discriminant functions meaningful in segregating the dataset. Hence, DA may not be the best tool for watershed based reservoir classification for water quality management. As a result, it is concluded that the classification tree method was better than DA at handling environmental data of reservoir watersheds as well as complexities the in reservoir class datasets.

Even though all three methods performed reasonably well, the results of this study suggest that the classification tree method may be the best classification tool overall, i.e. a watershed-based classification tree method for grouping Nebraska reservoirs for water quality management. Finally, the SEE5® classification interface is user-friendly and has potential applications for water resource managers.

Acknowledgments. This research was partially supported by the US EPA Science to Achieve Results (STAR) Program (Grant # R828635). K. Verdin and N. Bliss, USGS EROS Data Center, provided critical assistance in using EDNA and STATSGO datasets respectively and their efforts are gratefully acknowledged. This paper is a contribution of the University of Nebraska Agricultural Research Division, Lincoln, NE 68583. Journal Series No.14564.

References

Baker, C., Lawrence, R., Montagne, C. and Patten, D. (2006). Mapping wetlands and riparian areas using Landsat ETM+ Imagery and Decision-Tree-Based models, *Wetlands*, 26(2), 465-474, doi:10.1672/0277-5212(2006)26[465:MWARAU]2.0.CO;2.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees*, Wadsworth, Belmont, California.

Bulley, H.N.N. (2004). *A Watershed-Based Classification System for Lakes in Agriculturally Dominated Ecosystems: A Case Study of Nebraska Reservoirs*, Ph.D. Dissertation, University of Nebraska.

Bulley, H.N.N., Merchant, J.W., Marx, D.B., Holz, J.C. and Holz, A.A. (2007). A GIS-Based Approach to Watershed Classification for Nebraska Reservoirs, *J. Am. Water Resour. Assoc.*, 43(3), doi:10.1111/j.1752-1688.2007.00048.x.

CALMIT (Center for Advanced Land Management Information Technologies), University of Nebraska-Lincoln; http://www.calmit.unl.edu/gap/landcover.shtml

Conquest, L.L., Ralph, S.C. and Naiman, R.J. (1994). Implementation of large-scale stream monitoring efforts: Sampling design and data analysis issues, in L. Loeb and A. Spacie (Eds.), *Biological Monitoring of Aquatic Systems*, Lewis Publishers, Boca Raton, Florida, pp. 69-90.

De' ath, G. and Fabricius, K.E. (2000). Classification and regression trees: a simple yet powerful technique for ecological data analysis, *Ecol.*, 8(11), 3178-3192, doi:10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2.

Detenbeck, N.E., Elonen, C.M., Taylor, D.L., Anderson, L.E., Jicha, T.M. and Batterman, S.L. (2004). Region, landscape, and scale effects on Lake Superior tributary water quality, *J. Am. Water Resour. Assoc.*, 40(3), 705-720, doi:10.1111/j.1752-1688.2004.tb04454.x.

Detenbeck, N.E., Elonen, C.M., Taylor, D.L., Anderson, L.E., Jicha, T.M. and Batterman, S.L. (2003). Effects of hydrogeomorphic region, catchment storage and mature forest baseflow and snowmelt stream water quality in second-order Lake Superior Basin tributaries, *Freshwater Biol.*, 48(5), 912-927, doi:10.1046/j.1365-2427.2003.01056.x.

Dunteman, G.H. (1984). *Introduction to Multivariate Analysis*, Sage Publications, Beverly Hills, California.

Efron, B. (2003). Second thoughts on the bootstrap, *Stat. Sci.*, 18(2), 135-140, doi:10.1214/ss/1063994968.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman-Hall, New York.

Emmons, E.E., Jennings, M.J. and Edwards, C. (1999). An alternative classification method for northern Wisconsin lakes, *Can. J. Fisheries Aquatic Sci.*, 56(4), 661-669, doi:10.1139/cjfas-56-4-661.

EPA (US Environmental Protection Agency) (2003a). Withdrawal of revisions to the water quality planning and management regulation and revisions to the national pollutant discharge elimination system program in support of revisions to the water quality planning and management regulation, final rule, *Federal Register*, 68(53), 13607-13614.

EPA (US Environmental Protection Agency) (2003). Total Maximum Daily Loads. http://www.epa.gov/owow/tmdl/intro.html.

EPA (US Environmental Protection Agency) (2002). *Levels III and IV Ecoregions of the Continental United States (revision of Omernik, 1987)*, EPA National Health and Environmental Effects Laboratory, Western Ecology Division, Corvallis, Oregon.

EPA (US Environmental Protection Agency) (2000). *Nutrient Criteria Technical Guidance Manual for Lakes and Reservoirs*, Report No. EPA-822-B00-001, Washington DC.

Feldesman, M.R. (2002). Classification trees as an alternative to linear discriminant analysis, *Am. J. Phys. Anthropol.*, 119, 257-275, doi:10.1002/ajpa.10102.

Friedman, J.H. (2006). Recent advances in predictive machine learning, *J. Classification*, 23(2), 175-197, doi:10.1007/s00357-006-0012-4.

Gahegan, M. and West, G. (1998). The Classification of Complex Geographic Datasets: An Operational Comparison of Artificial Neural Network and Decision Tree Classifiers. http://www.geocomputation.org/1998/61/gc_61.htm.

German, G.W.H., West, G.A.W. and Gahegan, M.G. (1999). Statistical and AI techniques in GIS classification: A comparison, in *Proc. of the 11th Annual Colloquium of the Spatial Information Research Centre*, University of Otago, Dunedin, New Zealand.

Gesch, D., Oimoen, M., Greenlee, S., Nelson, C., Steuck, M. and Tyler, D. (2002). The national elevation dataset, *Photogramm. Eng. Remote Sensing*, 68(1), 5-11.

Ghaffari, A., Priestnall, G. and Clarke, M.L. (2006). Artificial neural networks and decision tree classifier performance on medium resolution ASTER data to detect gully networks in southern Italy, in *Proc. SPIE Vol. 6064, 60641Q*.

Goel, P.K., Prasher, S.O., Patel, R.M., Landry, J.A., Bonnell, R.B. and Viaub, A.A. (2003). Classification of hyperspectral data by decision trees and artificial neural networks to identify weed stress and nitrogen status of corn, *Comput. Electron. Agric.*, 39(2), 67-93, doi:10.1016/S0168-1699(03)00020-6.

Hastie, T., Tibshirani, R. and Freiedman, J.H. (2001). *The Elements of Statistical Learning*, Springer-Verlag, New York.

Hawkins, C.P., Norris, R.H., Gerritsen, J., Hughes, R.M., Jackson, S.K., Johnson, R.K. and Stevenson, R.J. (2000). Evaluation of landscape classifications for the prediction of freshwater biota: synthesis and recommendations, *J. North Am. Benthol. Soc.*, 19(3), 541-556, doi:10.2307/1468113.

Holz, J.C. (2002). Lake and reservoir classification in agriculturally dominated ecosystems, in *EPA 2002 Aquatic Ecosystem Classification Workshop*, Denver, CO.

Huberty, C.J. and Olejnik, S. (2006). *Applied Manova and Discriminant Analysis*, 2nd Edition, John Wiley & Sons, New York.

Hunt, E.B., Marin, J. and Stone, P.J. (1966). *Experiments in Induction*, Academic Press, New York.

James, F.C. and McCulloch, C.E. (1990). Multivariate Analysis in Ecology and Systematics: Panacea or Pandora's Box? *Ann. Rev. Ecol. Syst.*, 21, 129-166, doi:10.1146/annurev.es.21.110190.001021.

Jenerette, G.D., Lee, J., Waller, D. and Carlson, R.E. (2002). Multivariate analysis of the Ecoregion delineation for aquatic ecosystems, *Environ. Manage.*, 29(1), 67-75, doi:10.1007/s00267-001-0041-z.

Johnsgard, P.E. (2001). *The Nature of Nebraska: Ecology and Biodiversity*, University of Nebraska Press, Lincoln, Nebraska.

Joy, M.K. and Death, R.G. (2005). Modeling of freshwater fish and macro-crustacean assemblages for biological assessment in New Zealand, in S. Lek, M. Scardi, P.F.M. Verdonshot, J.P Descy, and Y.S. Park (Eds), *Modeling Community Structure in Freshwater Ecosystems*, Springer, Berlin, pp. 76-89.

Karels, T.J., Bryant, A.A. and Hik, D.S. (2004). Comparison of discriminant function and classification tree analyses for age classification of marmots, *OIKOS*, 105(3), 575-587, doi:10.1111/j.0030-1299.2004.12732.x.

Kuzelka, R.D., Flowerdale, C.A., Manley, R.N. and Rundquist, B.C. (1993). *Flat Water: A History of Nebraska and its Water*, Conservation and Survey Division, IANR, University of Nebraska-Lincoln, Resource Report No. 12.

Lauritzen, S.L. and Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their application to expert systems, *J. Royal Stat. Soc.*, 50(2), 157-224.

Legendre, P. and Legendre, L. (1998). *Numerical Ecology*, 2nd Edition, Elsevier Science, BV, Amsterdam.

Lomnicky, G.A. (1995). Lake classification in the glacially influenced landscape of the North Cascade Mountains, Washington, USA. PhD Dissertation. Oregon State University, Orego.

Loveland, T.R. and Merchant, J.W. (2004). Ecoregions and ecoregionalization: Geographical and ecological perspectives, *Environ. Manage.*, 34, S1-S13, doi:10.1007/s00267-003-5181-x.

Luoto, M. and Hjort, J. (2005). Evaluation of current statistical approaches for predictive geomorphological mapping, *Geomorphology*, 67(3-4), 299-315, doi:10.1016/j.geomorph.2004.10.006.

McGarigal, K., Cushman, S. and Stafford, S. (2000). *Multivariate Statistics for Wildlife and Ecology Research*, Springer-Verlag, New York.

Minka, T.P. and Picard, R.W. (1997). Interactive learning using a "society of models", *Pattern Recog.*, 30(4), 565-581.

Mitchell, T.M. (1997). *Machine Learning*, McGraw-Hill, New York.

Momen, B. and Zehr, J.P. (1998). Watershed classification using discriminant analyses of lake water-chemistry and terrestrial characteristics, *Ecol. Appl.*, 8(2), 497-07, doi:10.2307/2641089.

Neapolitan, R.E. (1990). *Probabilistic Reasoning in Expert systems: Theory and Algorithms*, John Wiley & Sons, New York.

Olden, J.D. and Jackson, D.A. (2002). A comparison of statistical approaches for modelling fish species distributions, *Freshwater Biol.*, 47(10), 1976-1995, doi:10.1046/j.1365-2427.2002.00945.x.

Omernik, J.M. (2003). The misuse of hydrologic unit maps for extrapolation, reporting and ecosystem management, *J. Am.Water Resour. Assoc.*, 39(3), 563-573.

Omernik, J.M. and Bailey, R.G. (1997). Distinguishing between watersheds and ecoregions, *J. Am. Water Resour. Assoc.*, 33(5), 935-949, doi:10.1111/j.1752-1688.1997.tb04115.x.

Omernik, J.M. (1987). Ecoregions of the Conterminous United States, *Ann. Assoc. Am. Geogr.*, 77, 118-125.

Park, Y.S., Verdonschot, P.F.M., Chon, T.S. and Lek, S. (2003). Patterning and predicting aquatic macroinvertebrate diversities using artificial neural network, *Water Res.*, 37(8), 17749-1758, doi:10.1016/S0043-1354(02)00557-2.

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, CA.

Quinlan, J.R. (1986). Induction of decision trees, *Mach. Learn.*, 1(1), 81-106.

Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*, Cambridge University Press.

Rokach, L. and Maimon, O. (2005). Top-down induction of decision trees classifiers-a survey, *IEEE Trans. Syst., Man, Cybern.*, 35(4), 476-487, doi:10.1109/TSMCC.2004.843247.

Robertson, D.M. and Saad, D.A. (2003). Environmental water-quality zones for streams: A regional classification scheme, *Environ. Manage.*, 31(5), 581-602, doi:10.1007/s00267-002-2955-5.

Rohm, C.M., Omernik, J.M., Woods, A.J. and Stoddard, J.L. (2002). Regional characteristics of nutrient concentrations in streams and their application to nutrient criteria development, *J. Am. Water Resour. Assoc.*, 38(1), 213-239, doi:10.1111/j.1752-1688.2002.tb01547.x.

Ronchetti, E., Field, C. and Blanchard, W. (1997). Robust linear model selection by cross-validation, *J. Am. Stat. Assoc.*, 92(439), 1017-1023, doi:10.2307/2965566.

RuleQuest Research. See5®: An Informal Tutorial. http:// rulequest.com/see5-win.html.

SAS Institute Inc. (2000). *SAS© Version 8 Users Manual*, SAS Institute Inc., Cary, NC.

Severn, A.A., Holz, J.C., Barrow, T.A., Hoagland, K.D., Bulley, H. and Merchant, J.W. (2001). Lake classification in the Sand Hills Region of Nebraska, Poster presentation, *North America Lake Management Society 21st International Symposium*, Madison, WI.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions, *J. Royal Stat. Soc.*, Series B, 36, 111-147.

Van Sickle, J. and Hughes, R.M. (2000). Classification strengths of ecoregions, catchments and geographic clusters for aquatic vertebrates in Oregon, *J. North Am. Benthol. Soc.*, 19(3), 370-384, doi:10.2307/1468101.

USGS (US Geological Survey) (2003). Elevation Derivatives for National Applications (EDNA). http://edna.usgs.gov.

Vayssieres, M.P., Plant, R.E. and Allen-Diaz, B.H. (2000). Classification trees: An alternative non-parametric approach for predicting species distributions, *J. Vegetation Sci.*, 11(5), 679-694, doi:10.2307/3236575.

Verbyla, D.L. (1987). Classification trees: A new discrimination tool, *Can. J. For. Res.*, 17, 1150-1152.

Verdin, K.L. and Verdin, J.P. (1999). A topological system for delineation and codification of the Earth's river basins, *J. Hydrol.*, 218(1-2), 1-12, doi:10.1016/S0022-1694(99)00011-6.

Wetzel, R.G. and Likens, G.E. (2000). *Limnological Analysis*, 3rd Edition, Springler Verlag, New York.

Winter, T.C. (1999). The relation of streams, lakes, and wetlands to groundwater flow systems, *Hydrogeol. J.*, 7(1), 28-45, doi:10.1007/s100400050178.