12-2004

A WATERSHED-BASED CLASSIFICATION SYSTEM FOR LAKES IN AGRICULTURALLY-DOMINATED ECOSYSTEMS: A CASE STUDY OF NEBRASKA RESERVOIRS

Henry N. N. Bulley
University of Nebraska-Lincoln, hbulley@clarku.edu

A WATERSHED-BASED CLASSIFICATION SYSTEM FOR LAKES IN AGRICULTURALLY-DOMINATED ECOSYSTEMS: A CASE STUDY OF NEBRASKA RESERVOIRS

By

Henry N. N. Bulley

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Geography

Under the Supervision of Professor James W. Merchant

Lincoln, Nebraska

December, 2004

CHAPTER 1. INTRODUCTION

1.0. General background

In recent decades substantial progress has been made in improving the quality of surface waters in the United States (Hawkins *et al.,* 2000; EPA, 2000; EPA, 2001); nevertheless, much work remains to be done in assessing the state of impairment of lake waters. Impairment implies that the existing water quality of a lake, as measured by selected criteria (e.g., nitrogen, phosphorus, chlorophyll-a, Secchi depth), exceeds a threshold value or standard that presumably reflects optimal attainable lake water quality conditions (or "reference" conditions) (Hughes, 1995; EPA, 2000; EPA, 2001). Such impaired waters are not suitable for designated uses such as drinking, irrigation, recreation or fishery (Carpenter *et al.*, 1998). The management of lake water quality requires an effective means to establish which lakes are most impaired (and, hence, may require restoration) and which lakes are least impaired.

It is estimated that about 43 percent of the 16.4 million hectares comprising the United States' lake area have been adequately assessed for water quality (EPA, 2000). Of the lakes that have been assessed, 45 percent are "impaired" and 9 percent of the impaired lakes are listed as threatened. Nutrients exported from agricultural lands contribute about 50 percent of water quality problems in impaired lakes (Figure 1.1) (EPA, 2000). Water quality standards are particularly difficult to establish for lakes located in areas highly modified by humans, such as agricultural landscapes of the Midwest. In these areas (a) few, if any, lakes may represent pre-settlement "reference" conditions, and (b) many lakes are human constructed (e.g., reservoirs). The principal objective of this research is to develop and evaluate an approach for establishing lake water quality standards using watershed-based classification of lakes.

Lakes are inland water bodies that serve as sources of drinking water, flood control, and outdoor recreation in addition to providing habitat for many wildlife species. The different types of lakes include natural lakes, reservoirs, and sand pits (or gravel pits) (Whittier *et al.,* 2001). Natural lakes were created as a result of geologic processes like glacial movement, while reservoirs in Nebraska were created by communities for flood control, drinking water, irrigation, hydroelectric power, and recreation. Sand pits are generally by-products of road construction activities where the sand or gravel was removed to provide aggregate materials. Natural lakes and sand pits are fed primarily by lower order stream and groundwater respectively, so both natural lakes and sand pits usually have very small or negligible watersheds. On the other, the primary source of water for reservoirs is high order streams. The response of reservoirs to climatic conditions is intricately linked to the lakes' morphology and watershed characteristics. As such, a watershed approach to lake classification seems more applicable to reservoirs than natural lakes and sand pits. Consequently, the focus of this research is on the watersheds of Nebraska reservoirs. There are about 6796 reservoirs in Nebraska. While each lake is unique, it is impossible to manage all of these lakes individually. Moreover, the term lakes and reservoirs will be used interchangeably in the following paragraphs.

The U.S. Environmental Protection Agency (EPA) is charged with establishing national standards and criteria for assessing lake water quality. However, it is increasingly evident that a single set of national water quality standards that does not take into account the hydrogeologic and ecological differences among lakes will not be viable, since lakes have different inherent capacities to meet such standards (EPA, 2000; EPA, 2001). For example, in Nebraska, the EPA suggested criteria for the management of lake phosphorus (30 μgL^{-1}) has likely never been met in Nebraska lakes even under natural (pre-settlement) conditions. A more realistic standard might be about 60μgL^{-1} (John Holz, *pers. comm.*). This inconsistency is partly due to the fact that Nebraska lakes are typically assessed in the same manner as lakes in nearby regions, such as the Ozarks of Missouri, which have very different hydrogeologic settings and relatively undisturbed environmental conditions.

A more tenable approach would be to define different standards for groups ("classes") of lakes determined to be similar to one another in terms of their potential to attain a certain level of water quality. Standards could then be established independently for lakes in different classes according to a set of "reference" target conditions unique for each class. Lake classification is used to group lakes into ecologically relevant or environmentally similar classes, enhance our understanding of complex systems, and to improve management and decision-making processes (Conquest *et al.,* 1994; Hawkins *et al.,* 2000). To be effective, a lake classification system designed to assess potential lake conditions must be based on environmental variables that underlie, determine, and explain the patterns of change in physical, chemical or biological water quality performances over seasonal or annual cycles (Warren, 1979). It is therefore important to differentiate between the natural or potential capacity of a lake to meet a certain level of water quality from actual water quality conditions that exist at a specific time of sampling.

The watershed classification approach that is proposed here is based on the premise that in the absence of human interference, lake ecosystems evolve in response to physical, chemical, and biological processes in their watersheds. It reflects an emerging emphasis on the watershed framework for water resource management (e.g., Warren, 1979; Satterlund and Adams, 1992; EPA, 1993; EPA, 1997; National Research Council, 1999; Mehan 2002; Bohn and Kershner, 2002). The lake watershed provides an important spatial framework to develop a classification system because it is the source of runoff water, sediments and nutrients for lakes. In general, lake watersheds integrate the effects of all the natural and anthropogenic processes on water quality.

A watershed is a topographically defined area that collects all surface runoff and groundwater and discharges them into the lake up to the furthest downstream point (Ponce, 1989; Satterland and Adams, 1992). The term watershed has been used synonymously with drainage basin or catchments (Viessman *et al.,* 1977; Ponce, 1989; Satterland and Adams, 1992). Watersheds influence lake water quantity (e.g., peak flows and seasonal low flows) and quality (e.g., sedimentation rate and nutrient enrichment or eutrophication) (Welch, 1978; Warren 1979; Wetzel, 1983; Frissell *et al.,* 1986; Ponce, 1989; Satterlund and Adams, 1992; Omernik, 2003). This makes watershed boundaries the most appropriate spatial and topographic units for lake classification, assessment, and management.

1.1. Previous Research

Previous attempts to classify lakes have been based either on actual, measurable biochemical conditions of lakes, or on biogeographic characteristics of ecological regions or zones (Vollenweider 1968; Carlson 1977; Schindler 1971; Jensen and Van der Maarel, 1980; Omernik, 1987; Omernik *et al.*, 1991; Lomnicky, 1995; Niles *et al.,* 1996; Heiskary, 2000; Winter, 2001, EPA, 2002a; Jenerette *et al.* 2002; Moss *et al.,* 2003; Detenbeck *et al.,* 2004). Schindler (1971), Carlson (1977), and Heiskary (2000) for example, classified lakes based on indices of lake performance that required extensive or repeated sampling of lake water quality parameters, e.g. nitrogen, phosphorus, chlorophyll-a and Secchi depth. On the other hand, Omernik *et al.* (1991), Maxwell *et al.* (1995), Hargrove and Luxmoore (1998), Winter (2001), McMahon *et al.* (2001), EPA (2002a) and Moss *et al.,*(2003) have developed landscape classification systems that may represent potential conditions of lakes and other water bodies, based on the characteristics of biogeographic or hydrogeologic regions, i.e., ecoregions and hydrologic landscapes.

Existing watershed-based classification systems for lakes and other water bodies have most often used actual water quality conditions in combination with topographic, soils, land use, and other data (Heywood *et al.,* 1980; Paulsen *et al.,* 1998; Momen and Zehr, 1998; Emmons *et al.,* 1999; Detenbeck *et al.,* 2000; Hawkins *et al.*, 2000; Johnson *et al.*, 2001; Lu and Lo, 2002; Bryd and Kelly, 2003; DeNicola *et al.,* 2004). Momen and Zehr (1998), for example, used discriminant function analysis (DFA) of lake water chemistry and land use data in a watershed-based lake classification. Emmons *et al.* (1999) compared DFA with a non-parametric statistical method, i.e. a decision tree model, in classifying northern Wisconsin lakes based on actual lake water quality data. They found that the decision tree method resulted in lower-rates of misclassification and more interpretable lake classes than those derived from DFA. Also, decision tree models can account for non-linear relationships, variable interactions and missing values in a given dataset (Breiman *et al.,* 1984; Verbyla, 1987; De'ath and Fabricius, 2000). Even though decision tree is a promising new tool for lake classifications it has not been applied extensively.

Other watershed-based classification systems and watershed assessments have been developed using the smallest (or fourth level) division of U.S. Geological Survey hydrologic units, i.e. hydrologic cataloging units (e.g., Smith *et al.* 1997, Griffiths *et al.,* 1999; EPA, 2002b; Bryd and Kelly, 2003, Papahicolau *et al.*, 2003). However, these hydrologic units are not topographic watersheds and limitations of their use as surrogates for watersheds have been documented (e.g., Verdin and Verdin, 1999; Gesch *et al.*, 2002; Omernik, 2003).

According to Grigg (1965), "Classifications should be designed for a specific purpose since they rarely serve two purposes equally well". The purpose for classifying lakes in this research is in part to help the U.S. Environmental Protection Agency to establish reasonable attainable water quality standards ("targets") for groups of lakes that are considered to share similar potential capacity to meet these standards. Classification frameworks such as those cited above, while quite effective for a number of applications, do exhibit several major shortcomings for setting lake water quality standards. For example:

1. Lake classification based on observed water quality does not provide adequate insights into the potential of lakes to meet water quality standards for the following reasons:

- Human activities, such as land use, impact water quality.
- Water quality data represent observed water quality conditions, not the potential to meet a water quality standard.

- Extensive and frequent sampling of lakes in a given region is required, and lake sampling campaigns can be costly, in terms of personnel and equipment.
- Sampled lakes may not adequately represent the lakes in a given region.
- Lake water quality parameters are sometimes so variable that one lake may change classes over seasonal or annual cycles.

2. Omernik's ecoregions are inappropriate because they were based on subjective criteria of perceived patterns of land surface form, climate, vegetation, soils and land use. Hence, these ecoregions can not be easily replicated. Also, ecoregion boundaries do not coincide with watershed boundaries, and the inclusion of land use data reflects the impact of human activities.

3. Attempts to delineate ecoregions via quantitative and objective methods (e.g. Hargrove and Luxmoore, 1998; Zhou *et al.,* 2003) are not appropriate because:

- These ecoregions include aspects of human influence, such as land use.
- The unit of analysis, e.g. 1 kilometer pixel of satellite data, does not take into consideration the terrain effect of watershed boundaries.

4. Existing watershed-based classifications are not appropriate because:

- They include aspects of human influence, such as measured water quality condition and land use.
- They are sometimes based on hydrologic cataloging units which do not conform to the natural hydrologic boundaries of the terrain.
- They are usually based on parametric statistical approaches such as discriminant function analysis (DFA) and regression analysis, which presume the use of

normally distributed data, although most watershed data are multimodal and not normally distributed.

- Lake classes as well as some watershed data are categorical, and these types of data usually require transformations, when using traditional statistical approaches.

In summary, most lake classifications are based on observed, extant water quality data or on environmental variables that are often impacted by human activities and, thus, usually cannot be used directly for determining lake classes and subsequently setting lake reference conditions; data collection is expensive and time consuming. Regionalization schemes, on the other hand, generally use subjective criteria for delineating boundaries (e.g., ecoregions) which do not coincide with watersheds. In both cases, there is an apparent arbitrary and often subjective choice of the number of classes.

This research focuses on the development of a watershed-based lake classification system that is based on: topographic boundaries that represent the lake watersheds; watershed characteristics that underlie, determine and explain the patterns of change in physical, chemical or biological water quality performances of lakes; and non-parametric statistical approaches that can account for the multimodal and categorical nature of watershed variables and lake classes.

1.2. Objectives and research questions

The primary objective of this dissertation is to develop a watershed-based approach to classify reservoir watersheds and to evaluate the effectiveness of the classification method to account for variations in water quality data that are pertinent to reservoir water quality management. The utility of Geographic Information Systems (GIS) and decision tree algorithms in developing a watershed-based approach to reservoir classification is also evaluated. This work is based on Nebraska reservoirs because most of the lakes in the state are constructed and located in agriculturally-dominated landscapes. Nebraska has a broad diversity of environments and landscapes, and is representative of many mid-latitude regions of the United States.

The specific research objectives are to:

1. *Determine the optimal number and characteristics of classes of Nebraska reservoirs based on their watershed characteristics. The research question that was addressed with respect to this objective is, "what watershed characteristics are required to classify reservoirs based on their potential to attain certain water quality standards?"*

An important component of managing reservoir water quality effectively is to segregate the reservoirs into similar "groups" or "classes", in terms of their potential to achieve certain water quality standards. However, information on the number of classes of Nebraska reservoirs is not available. This lack of knowledge limits our understanding of the biophysical characteristics of Nebraska reservoir classes and prevents accurate estimation of potential reservoir water quality. Such information is useful for many applications including predictive modeling of potential water quality impairment of reservoirs based on their class membership.

A vital step in developing a classification is to determine the optimal number of classes to be used. This requires partitioning a dataset such that the entities in one group are more similar to each other than to those in other groups (i.e., clustering). Similarity refers to the distance between two data points (reservoirs), where the distance decreases for more similar reservoirs (Gordon, 1999). The fundamental issue in any clustering approach is to determine which number of clusters best describes the class structure (or optimal number of classes) of the dataset (i.e., cluster validation). A cluster validation approach was used to identify the optimal number of classes that exist among Nebraska reservoirs. Additionally, a decision tree model was applied to describe the structure of watershed classes.

2. *Evaluate the watershed-based decision tree classification model to predict the class membership of Nebraska reservoirs. The research questions that were addressed with respect to this objective are (a) which decision tree rules are optimal for assigning a reservoir watershed to a class? (b) how does the level of discrimination achieved by the decision tree approach compare to other water resource classification systems, i.e. DFA-based reservoir watershed classes and ecoregion-derived reservoir classes?*

Once the numbers of underlying lake groups as well as essential watershed characteristics have been identified, a rule-based decision tree classification model can be used to classify the reservoirs based on their watershed characteristics. There are two types of decision tree models, i.e. classification trees and regression trees. 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 class (Breiman *et al.,* 1984; Ripley, 1996; De'ath and Fabricius, 2000). As such, a classification tree was used in this study.

According to De'ath and Fabricius (2000), the classification tree model can be used for data description (i.e., represent the systematic structure of the data) and for prediction (i.e., accurately predict the class membership of new observations). A classification tree-based predictive model of reservoir watershed classes is developed and the performance of classification tree-based reservoir watershed classification method is compared to DFA-based watershed classification system (Momen and Zehr 1998), and Omernik's Level IV ecoregions (Omernik 1987; EPA, 2002a). This comparison was done to assess the effectiveness of watershed-based classifications and ecoregions in accounting for variations in water quality parameters of Nebraska reservoirs; and also to determine the prediction accuracy of classification tree-based and DFA-based reservoir watershed classification methods.

1.3. Study area

This research focuses on Nebraska, a state representative of many mid-latitude areas having agriculturally-dominated economies (Figure 1.2). The eastern boundary of the state is defined by the Missouri river and the line of 105° W constitutes the westernmost boundary. Nebraska encompasses a broad range of climatic, physiographic, land use and water quality conditions. Elevations range from about 256 meters in the east to 1654 meters in the west. About 30 percent of the state is dominated by the Sand Hills, grass covered sand dunes predominately devoted to grazing. The climate is characterized by a gradient of rainfall and temperature regimes along an east-west axis. Average annual precipitation varies from 36 cm in the northwest to 86 cm in the southeast; temperatures vary between -20 to 30 C° (Johnsgard, 2001; Kuzelka *et al.,* 1993).

In semiarid agriculturally-dominated environments such as Nebraska, water quality impairment stems primarily from the transport of soil sediments, agrochemicals and animal wastes via runoff from croplands and livestock operations into streams and lakes. There are about 13,500 lakes in Nebraska including natural lakes, reservoirs, and sand pits. The condition of Nebraska's lake waters is largely unknown, although it is suspected that many are impaired to some degree. Over the past two decades, the

Nebraska Department of Environmental Quality (NDEQ) and the School of Natural Resources (University of Nebraska–Lincoln) have sampled about 225 Nebraska lakes and have developed a database that describes their chemical and biophysical water quality characteristics (Holz, 2002). These data provide a valuable resource for studies of lake water quality. Only reservoirs with total surface area greater than 4 hectares (10 acres) were considered in this study in order to conform with the U.S. EPA requirements for developing nutrient water quality criteria for lakes in the United States (EPA, 2001). Furthermore, geospatial datasets that were used in characterizing Nebraska reservoir watersheds are available for the entire U.S.; thus, the research approach has potential national applications.

1.4. Structure of the dissertation

This dissertation is comprised of six chapters. This introductory chapter is followed by chapter 2, a review of relevant literature pertaining to lake water quality assessment and lake classification. Chapter 3 discusses the first part of the research: the development of an updated digital map of Nebraska lakes in order to identify reservoirs in the state and to delineate watershed boundaries for selected Nebraska reservoirs. This chapter also includes a summary of preliminary statistical analyses of the watershed datasets. Chapter 4 deals with the implementation of a watershed-based classification approach for Nebraska reservoirs. The classification process includes an assessment of the optimal number of watershed classes for Nebraska reservoirs. This assessment was based on k-means clustering algorithm and a unique cluster validation technique. The cluster validation approach uses relative criteria that employs indices (in this case, Calinski-Harabasz statistic) extracted from the clustering results to identify the optimal number of classes. Finally, a classification tree model was used to describe the structure of Nebraska reservoir watershed classes and also determine the final structure of the reservoir classes (number of classes and classification rules). Chapter 5 describes comparisons of the performance of classification tree-based reservoir watershed classification method with DFA-based reservoir watershed classification system (Momen and Zehr, 1998) and Omernik's Level IV ecoregions derived reservoir classes (Omernik 1987; EPA, 2002). Chapters 6 wraps up the dissertation report with a summary of major research results, conclusions and recommendations for future research.

References cited

Bohn, B.A. and J.L. Kershner 2002. *Establishing aquatic restoration priorities using a watershed approach.* **Journal of Environmental Management**. 64: 355–363.

Breiman, L., J. H. Friedman, R.A. Olshen and C. J. Stone. 1984. **Classification and Regression Trees**. Wadsworth, Inc. Belmont, California. 358p.

Bryd, K.B. and Kelly N.M. 2003. *Development of a classification system for linking watershed land-use change and wetland vegetation response in the Elkhorn slough watershed, Monterey County, California*. **Proceedings of the 3rd Biennial Coastal GeoTools Conference.** Charleston, South Carolina. January 6-9, 2003.

Carpenter, S.R, N.F. Caraco, D.A. Correll, R.W. Howarth, A.N. Sharpley and V.H. Smith. 1998. *Nonpoint pollution of surface waters with phosphorus and nitrogen.* **Issues in Ecology.** 3:1-11

Carlson, R. E. 1977. *Trophic state index for lakes.* **Limnology and Oceanography**. 22:361-369.

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

De' ath, G and K.E. Fabricius. 2000. *Classification and regression trees: a simple yet powerful technique for ecological data analysis.* **Ecology.** 8(11):3178-3192.

DeNicola, D. M. E. de Eyto, A Wemaere and K. Irvine. 2004. *Using epilithic algal communities to assess trophic status in Irish lakes.* **Journal of Phycology.** 40 (3): 481 – 495.

Detenbeck, N.E., C.M. Elonen, D.L. Taylor, L.E. Anderson, T.M. Jicha, and S.L. Batterman. 2004. *Region, landscape, and scale effects on Lake Superior tributary water quality.* **Journal of the American Water Resources Association.** 40 (3): 705 – 720.

Detenbeck, N.E., S.L. Batterman, V.J. Brady, J.C. Brazner, V.M. Snarski, D.L. Taylor and J.A. Thompson. 2000. *A test of watershed classification systems for ecological risk assessment.* **Environmental Toxicology and Chemistry.** 19:1174-1181.

Emmons, E.E., M.J. Jennings and C. Edwards. 1999. *An alternative classification* *method for northern Wisconsin lakes.* **Canadian Journal of Fisheries and Aquatic Sciences.** 56 (4):661-669.

EPA (U.S. Environmental Protection Agency). 1993. **The Watershed Protection Approach**. EPA 840-s-93-001. Washington, D.C.

EPA (U.S. Environmental Protection Agency). 1997. **The Index of Watershed Indicators**. EPA841-R7-010. Washington, D.C.

EPA (U.S. Environmental Protection Agency). 2000. **A Summary of the National Water Quality Inventory: 2000 Report to Congress.** EPA841-S-00-001. Washington, D.C.

EPA (U.S. Environmental Protection Agency). 2001. **Nutrient Criteria Technical Guidance Manual for Lakes and Reservoirs.** Report No. EPA-822-B00-001. Washington, D.C.

EPA (U.S. Environmental Protection Agency). 2002a. 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 (U.S. Environmental Protection Agency). 2002b. States **Unified Watershed Assessments.** http://www.epa.gov/owow/uwa/states/.

Frissell, C.A., W.J. Liss, C.E. Warren ad M.D. Hurley. 1986. *A hierarchical framework for stream habitat classification: viewing streams in a watershed context.* **Environmental Management**. 10:199-214.

Gesch, D., M. Oimoen, S. Greenlee, C. Nelson, M. Steuck and D. Tyler. 2002. *The National Elevation Dataset.* **Photogrammetric Engineering and Remote Sensing**. 68(1): 5-11.

Gordon, A. 1999. **Classification,** 2nd Edition. Chapman and Hall/CRC. London. 256p.

Griffiths, G.E., J.M. Omernik and A.J. Woods. 1999. *Ecoregions, watersheds, basins and HUCS: how state and federal agencies frame water quality.* **Journal of Soil and Water Conservation**. 54 (4):666-677.

Grigg, D. 1965. *The logic of regional systems*. **Annals of the American Association of Geographers**. 55: 465-491.

Hargrove, W.W. and R.J. Luxmoore. 1998. *A New High-Resolution National Map of Vegetation Ecoregions Produced Empirically Using Multivariate Spatial Clustering.* **ESRI ARC/INFO User Conference.** http://gis.esri.com/library/userconf/proc98/PROCEED/TO350/PAP333/P333.HTM

Hawkins, C.P., R.H. Norris, J. Gerritsen, R.M. Hughes, S.K. Jackson, R.K. Johnson and R. J. Stevenson. 2000. *Evaluation of landscape classifications for the prediction of freshwater biota: synthesis and recommendations.* **Journal of the North American Benthological Society**. 19(3): 541-556.

Heiskary, S. A. 2000. *Ecoregional Classification of Minnesota Lakes.* Pages B4-B5 *in* EPA (U.S. Environmental Protection Agency). 2000. **Nutrient Criteria Technical Guidance Manual for Lakes and Reservoirs**. Report No. EPA-822-B00-001. Washington, D.C.

Heywood, R.B., H.J.G. Dartnall and J. Piddle. 1980. *Characteristics and classification of lakes of Signy Island, Outh Orkney Islands, Antarctica.* **Frcshwater Biology.** 10:47–60.

Holz, J.C. 2002. **Lake And Reservoir Classification In Agriculturally Dominated Ecosystems.** EPA 2002 Aquatic Ecosystem Classification Workshop, Denver, CO, September, 2002, oral presentation, invited.

Hughes, R.M. 1995. *Defining acceptable condition.* Pages 31 – 41 *in* W.S. Davis and T.P. Simon (eds.). **Biological and Nutrient Criteria. Tools for Water Resource Planning and Decision making**. Lewis Publishers, Boca Raton, Florida.

Jenerette, G.D., J. Lee, D. Waller and R.E. Carlson. 2002. *Multivariate analysis of the Ecoregion delineation for aquatic ecosystems.* Environmental Management. 29 (1): 67- 75.

Jensen, S. and E. Van Der Maarel. 1980. *Numerical approaches to lake classification with special reference to macrophyte communities.* **Vegetatio.** 42: 42:117-128.

Johnson, G.D., W.L. Meyers and G.P. Patil. 2001. *Predictability and of surface pollution loading in Pennsylvania using watershed-based landscape measurements.* **Journal of the American Water Resources Association**. 37(4):821–835.

Kuzelka, R.D., C.A. Flowerdale, R.N. Manley and B.C. Rundquist. 1993. **Flat water: A history of Nebraska and its water.** Resource Report No. 12. Conservation and Survey Division, IANR, University of Nebraska-Lincoln. 291p.

Lomnicky, G.A. 1995. **Lake Classification in the Glacially Influenced Landscape of the North Cascade Mountains, Washington, USA.** PhD. Dissertation. Oregon State University, Oregon.

Lu, R.S. and S.L. Lo. 2002. *Diagnosing reservoir water quality using self-organizing maps and fuzzy theory.* **Water Research.** 36 (9): 2265 – 2274.

Maxwell, J.R., C.J. Edwards, M.E. Jensen, S.J. Paustian, H. Parrott and D.M. Hill. 1995. **A hierarchical framework of aquatic ecological units of North America (Nearctic Zone).** Technical Report NC-176:1-76. United States Department of Agriculture, Forest Service. Washington D.C, USA.

McMahon, G., S.M. Gregonis, S.W. Walton, J.M. Omernik, T.D. Thorson, J.A. Freeouf, A.H. Rorick and J.E. Keys. 2001. *Developing spatial frameworks of common ecological regions of the conterminous United States.* **Environmental Management**. 28 (3): 293 -316.

Mehan, G.T. 2002. **Committing EPA's Water Program to Advancing the Watershed Approach.** EPA memo to Regional Water Division Directors. December 3, 2002. http://www.epa.gov/owow/watershed/memo.html. Accessed February 27, 2004.

Momen, B. and J.P. Zehr. 1998. *Watershed classification using discriminant analyses of lakewater-chemistry and terrestrial characteristics.* **Ecological Applications**. 8 (2):497-507.

Moss, B D. Stephen, C. Alvarez, E. Becares, W. Van de Bund, S.E. Collings, E. Van Donk, E. De Eyto, T . Feldmann, C. Fernandez-Alaez, M. Fernandez-Alaez, R.J.M Franken, F. Garcia-Criado, E.M. Gross, M. Gyllstrom, L.A, Hansson, K. Irvine, A. Jarvalt, J.P. Jensen, E. Jeppesen, T. Kairesalo, R. Kornijow, T. Krause, H. Kunnap, A. Laas, E. Lille, B. Lorens, H. Luup, M.R. Miracle, P. Noges, T. Noges, M. Nykanen, I. Ott, W. Peczula, E.T.H.M. Peeters, G. Phillips, S. Romo, V. Russell, J. Salujoe, M. Scheffer, K. Siewertsen,H. Smal, C. Tesch, H. Timm, L. Tuvikene, I. Tonno, T. Virro, E. Vicente and D. Wilson. 2003. *The determination of ecological status in shallow lakes - a tested system (ECOFRAME) for implementation of the European Water Framework Directive.* **Aquatic Conservation-Marine and Freshwater Ecosystems**. 13: 6 – 507.

National Research Council. 1999. **New Strategies for America's Watersheds.**

Niles, R.K., D.L. King and R. Ring. 1996. *Lake classification systems - part I.* **The Michigan Riparian.** http://www.mslwa.org/lkclassif1.html.

Omernik, J.M., 2003. The misuse of hydrologic unit maps for extrapolation, reporting and ecosystem management. **Journal of the American Water Resources Association**. 39(3):563–573.

Omernik, J.M., 1987. *Ecoregions of the Conterminous United States.* **Annals of the Association of American Geographers.** 77:118-125.

Omernik, J.M. C.M. Rohm, R.A. Lillie, and N. Mesner. 1991. *Usefulness of natural* *regions in lake management: analysis of variation among lakes in Northwestern Wisconsin, USA.* **Environmental Management.** 15:281-293.

Papahicolau, A.N., A. Bdour, N. Evangeloulos and N. Tallebeydokhti. 2003. *Watershed and instream impacts on the fish population in the South Fork of the Clearwater River, Idaho.* **Journal of the American Water Resources Association.** 39(1):191–203.

Paulsen, S.G., R.M. Hughes, and D.P. Larsen. 1998. *Critical elements in describing and understanding our nation's aquatic resources.* **Journal of the American Water Resources Association.** 34 (5): 995 – 1005.

Ponce, V.M. 1989. Engineering **Hydrology: Principles And Practices**. Prentice-Hall, Inc. New Jersey. 627p.

Ripley, B.D. 1996. **Pattern Recognition and Neural Networks.** Cambridge University Press. 403p

Satterlund, D.R. and P.W. Adams. 1992. **Wildland Watershed Management.** 2nd Ed. J. Wiley and Sons, New York, N.Y. 436p.

Schindler, D.W. 1971. *A hypothesis to explain the differences and similarities among lakes in experimental lakes area, Northwestern Ontario.* **Journal of fisheries Research, Canada**. 28: 295-301.

Smith, R.A., G.E. Schwarz and R.B. Alexander. 1997. Regional interpretation of water quality data. Water Resources Research. 33:2781-2798.

Verbyla, D.L. 1987. *Classification trees: a new discrimination tool.* **Canadian Journal of Forestry Research.** 17:1150–1152.

Verdin, K.L. and J.P. Verdin. 1999. *A topological system for delineation and codification of the Earth's river basins.* **Journal of Hydrology**. 218:1 – 12.

Viessman, W., J.W. Knapp, G.L. Lewis and T.E. Harbuagh. 1977. **Introduction to Hydrology**. Harper and Row, Publishers. New York, N.Y. 704 p.

Vollenweider, R.A. 1968. **Scientific fundamentals of the eutrophication of lakes and flowing waters, with particular reference to nitrogen and phosphorus as factors in eutrophication**. Tech. Rpt. DAS/SCI/68.27. Organization for Economic Cooperation and Development (OECD), Directorate for Scientific Affairs, Paris, France. 192p.

Warren, C. E. 1979. **Toward classification and rationale for watershed management**

and stream protection. EPA - 600 / 3-79-059. United States Environmental Protection Agency. Corvallis, Oregon. 143p.

Welch, D.M. 1978. **Land/Water Classification. A Review of Water Classifications and Proposals for Water Integration into Ecological Land Classification.** Ecological Land Classification Series, No.5. Environment Canada. Ottawa. 54p.

Wetzel, R.G. 1983. **Limnology.** Saunders College Publishing, Philadelphia, PA. 767p.

Whittier, T.R., D.P. Larson, S.A. Peterson and T.M. Kincaid. 2001. *A comparison of impoundments and natural drainage lakes in the Northeast USA.* **Hydrobiologia.**

Winter, T.C. 2001. *The concept of hydrologic landscapes.* **Journal of the American Water Resources Association**. 37(2):335–349.

Zhou, Y., S. Narumalani, W.J. Waltman, S.W. Waltman and M.A. Palecki. 2003. *A GIS-based spatial pattern ecoregion mapping and characterization.* **International Journal of Geographical Information Science**. 17(5):445-462.

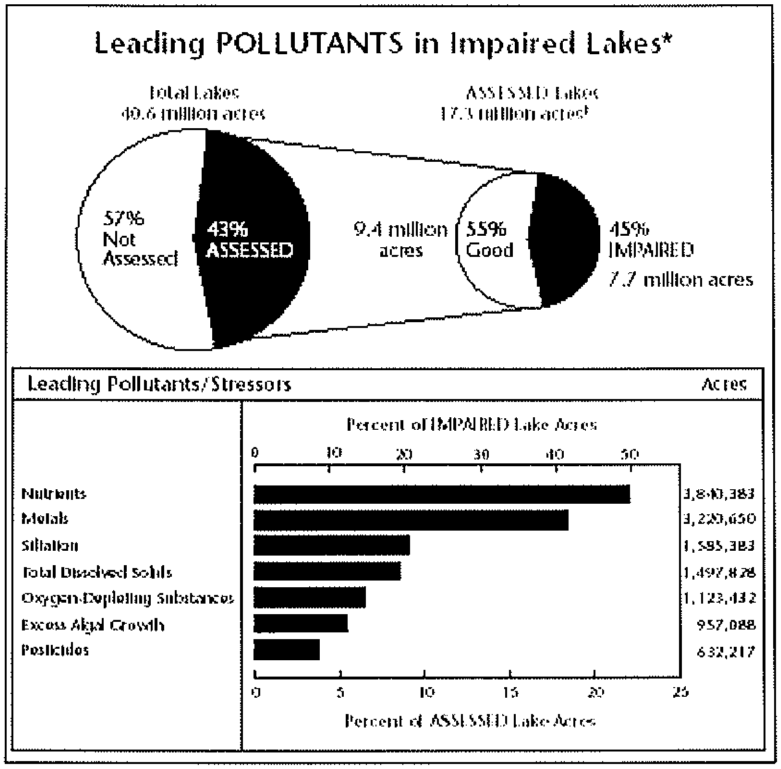


Figure 1.1. National Water Quality Inventory. Agricultural nutrients are the most common pollutants affecting assessed lakes; are found in 22% of assessed lakes; Contribute to 50% of reported water quality problems in impaired lakes.
Source: EPA (2000)

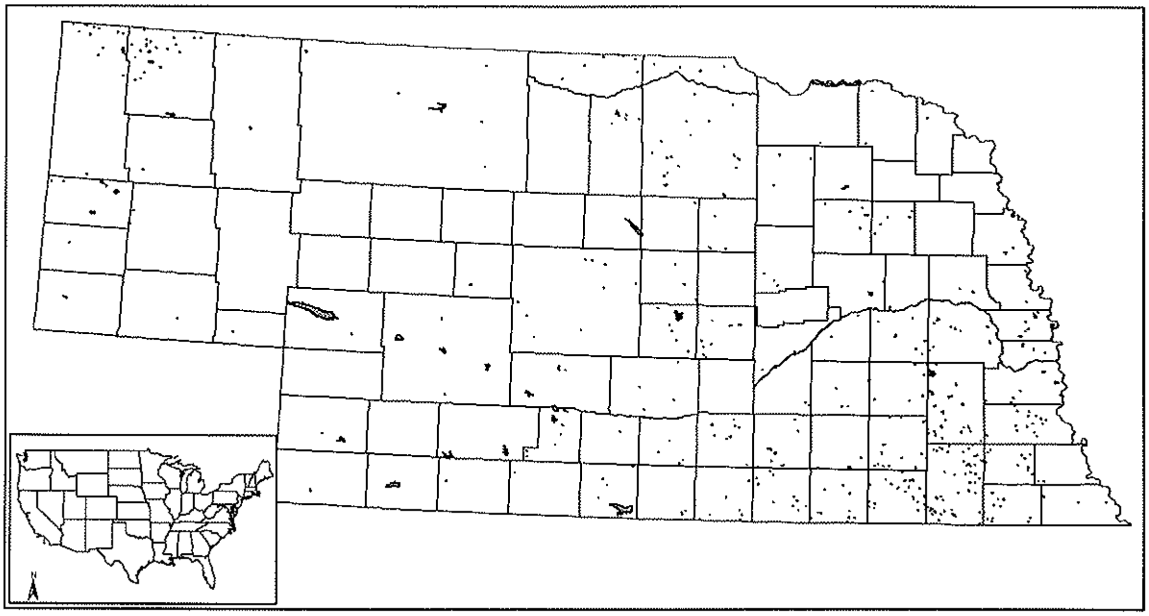


Figure 1.2. Map of Nebraska reservoirs (reservoir size at least 4 hectares).

CHAPTER 2. BACKGROUND REVIEW

2.0. Introduction

Lakes are inland water bodies that serve as sources of drinking water, flood control, outdoor recreation and provide habitat for many wildlife species. The different types of lakes include natural lakes, reservoirs and sand pits (or gravel pits) (Whittier *et al.,* 2001). Natural lakes develop as a result of geologic processes like glacial movement, while reservoirs are created artificially to meet diverse land use objectives including flood control, irrigation, recreation and drinking water supply. Sand pits are generally by-products of road construction activities where the sand or gravel was mined to provide building materials, leaving behind huge craters.

According to Thornton *et al.* (1990) and Cooke *et al.* (1993), lake ecosystems analyses generally ignore the differences between lake types because the fundamental hydrological and watershed processes that govern the chemistry and biology of natural and man-made lakes were thought to be similar. However, there is now increased emphasis on treating the different lake types as unique due to the differences in their origin, water residence time, and water source (Thornton *et al.,* 1990; Cooke *et al.,* 1993; EPA, 2000; Whittier *et al.,* 2001). Whittier *et al.* (2001) found significant differences between small impoundments and natural lakes. For example, the primary water source for natural lakes and sand pits is lower order streams and groundwater respectively while reservoirs are mostly fed by higher order streams. Furthermore, the response of a reservoir to climatic conditions is intricately linked to its morphology and watershed characteristics.

Since stream-fed lakes tend to reflect their hydrogeologic setting (watershed characteristics) the environmental conditions in the lake watershed can be good indicators of lake water quality. Therefore, the following review will be focused on lake and watershed characteristics of reservoirs, with the goal of articulating a rationale for a lake classification system based on potential lake water quality. This will include an overview of the lake aging (eutrophication) process, different approaches to lake classification and a discussion of some of the factors that affect lake water quality.

2.1. Lake eutrophication process

Lake eutrophication (or lake aging) is a slow process by which a lake progresses from its creation or youth through sedimentation and nutrient enrichment to extinction (Figure 2.1). This process usually occurs over a period of centuries, but anthropogenic influences like agricultural land use can hasten the process to take place over a few decades.

Lake and watershed characteristics interact with geomorphologic or gradational processes that eventually convert the lake into a lacustrine plain (an ancient lake bed), a site typical of extinct lakes (Mortimer, 1942; Larson, 1970; Carpenter *et al.* 1998). The primary causes on the lake aging are the deposition and accumulation of soil sediments and organic materials (Mortimer, 1942; Wetzel, 1965; Larson, 1970; Wetzel 1983; Carpenter *et al.* 1998). Three key stages in the lake eutrophication process under natural conditions were chronicled by Mortimer (1942). The initial or primary phase is characterized by a slow increase in lake productivity, followed by a second stage evidenced by a sharp rise in productivity. Continued influx of nutrients and sediments will accelerate decomposition of organic materials at the lake bottom, leading to precipitation of insoluble iron compounds (e.g., ferrous sulfide). Consequently, essential nutrients for primary production are bound with iron compounds and become unavailable for photosynthesis. The third phase is the point when the lake becomes incapable of utilizing the influx of nutrients and these nutrients are bounded to sediments and deposited at the lake bottom. Over centuries, this lake loses its capacity to hold water and becomes a lacustrine plain.

The process described above can be hastened to occur over a few decades, especially in agricultural ecosystems. Agricultural activities increasingly expose soil particles to erosion and contribute excessive amount of nutrients from fertilizer applications in farmlands. The secondary stage in the eutrophication process follows closely the influx of non-point sources of nutrients and sediments. These nutrient influxes are usually intermittent and closely associated with seasonal and annual cycles of agricultural activity, such as planting and plowing, or climatic activities like heavy rainfall. The secondary eutrophication phase is usually made evident by cyanobacteria blooms, which are the primary causes of fish kills, bad drinking water taste and offensive odors that emanate from affected waters.

Strong anthropogenic influences on lakes and their watersheds tend to change the natural eutrophication process as function of the land use and land cover types within the lake watershed (Battaglin and Goolsby, 1996; Carpenter *et al.,* 1998). The net gain in phosphorus and nitrogen through intensive fertilizer application results in a nutrient surplus on croplands, and this is the underlying cause of non-point pollution in agriculturally-dominated ecosystems (Carpenter *et al.* 1998). Along with increased anthropogenic influences on lakes and their watersheds, comes the need to manage and/or restore lake water quality. The challenge to lake water quality management in these environments lies in identifying the potential capacity of these lakes to attain certain water quality level, in order to mitigate the acceleration of the lake-aging process. While each lake is unique, it is impossible to manage all lakes individually. Lake classification is used to group lakes into ecologically similar classes (Conquest *et al.*, 1994). Different approaches to lake classification are discussed in the following sections in order to articulate the rationale behind the watershed-based lake classification system.

2.2. Lake classification approaches

Lake classification is generally designed to enhance our understanding of the complex environment and improve lake management and decision-making processes. According to Hawkins *et al.* (2000), an effective lake classification should (a) enhance our understanding of the effects of spatial and temporal environmental stressors, i.e., predicting stressors likely to cause impairment; and (b) help to establish attainable water quality conditions, e.g., establish a network of reference lake sites for setting expected conditions at potentially impaired sites. However, lake classifications need to be designed for a specific purpose since classifications rarely serve two purposes equally well (Grigg, 1965). Thus, the review of lake classifications addressed in this study will be done with respect to those that facilitate efforts to establish attainable reference water quality conditions. There have been several efforts to classify lakes for various management goals including nutrient criteria development for fisheries, drinking water, and recreational use. These classifications can be summarized as either based on actual lake water quality conditions or based on potential lake water quality conditions.

2.2.1. Classification of actual water quality conditions

Four systems for classifying actual lake conditions are in common use. One system, based on the trophic state (nutrient richness or primary productivity) of lakes, identifies lakes as oligotrophic to eutrophic and is exemplified by the Carlson's trophic state index (TSI) (Carlson, 1977). A second system is based on the timing and extent of mixing in lakes as well as lake area and depth (Niles *et al.,* 1996). A third system is based on the fishery resources in lakes, i.e., fish type and productivity (Niles *et al.,* 1996). The fourth system is based on multivariate statistics such as discriminant function analyses (DFA) of actual lake water quality data (e.g., Heywood *et al.,* 1980; Willen *et al.,* 1990; Momen and Zehr, 1998; Paulsen *et al.,* 1998; An and Kim, 2003; DeNicola *et al.,* 2004). Although the second and third classification systems are important for fisheries management, the TSI and DFA classification systems are widely used in assessing lakes ecosystem functioning for water quality management.

The aforementioned lake classification approaches all depend on extensive survey of biological and/or physical water quality parameters derived from water samples that reflect water conditions at the time of observation or sampling. The number of samples and the spatial distribution of the sampled lakes will clearly affect the classification outcome. Often, the sampling records are from different times and different locations (Heiskary and Wilson, 1989). This problem perhaps could be partly resolved by using remotely sensed data (e.g., Lathrop, 1992; Dekker and Peters, 1993; Olmanson *et al.*, 2001, Yang *et al.*, 2001; Nelson *et al.*, 2003). However, there are some limitations to the use of multispectral remote sensing data in assessing of lake water quality, such coarse spectral resolution (Dekker and Peters, 1993). Hyperspectral sensors like "Hyperion" onboard the EO-1 satellite collect scenes in coordination with the Landsat 7 Enhanced Thematic Mapper (ETM+) (Earth-Observing-1, 2002). Koponen *et al.,* (2002) also demonstrated the integration of satellite derived and airborne hyperspectral data in lake water quality classification. These examples show that remote sensing data may be useful in complementing field water sampling data for water quality analysis and in identifying lake reference conditions. However, the use of remote sensing data in augmenting lake water quality data is beyond the scope of this research.

Lake reference conditions are quantitative descriptions of lake conditions used as a standard of comparison. Although reference conditions are intended to portray pristine environmental conditions, it is generally recognized that they realistically portray least impacted or most sustainable conditions (Hughes, 1995; EPA, 2000; EPA, 2001). There are three main approaches to characterizing lake reference conditions; namely, (i) direct observation of reference sites or entire lakes in a class, (ii) paleolimnological reconstruction of past conditions, and (iii) model-based prediction (EPA, 2000). The direct observation of the population of lakes in a given class can be used to develop histograms from which different quantile values of reference water quality can be derived. Four important actual lake water parameters have been identified as candidate variables for setting lake reference conditions, i.e., total phosphorus (TP), total nitrogen (TN), Secchi depth and chlorophyll-a (EPA, 2000). For example, when TP concentrations for all minimally impaired lakes are plotted on a distribution curve, the 75th percentile threshold will represent an acceptable reference condition for TP (Figure 2.2a). When there are no identifiable minimally impaired reference sites, the 25th percentile of the TP concentration for all lakes in a given class will represent a fairly acceptable reference condition (Figure 2.2b) (EPA, 2000). Therefore, the assessment of actual lake water quality conditions can play an important role in identifying and monitoring changes in lake conditions, once the potential lake classes have been identified.

2.2.2. Classification of potential water quality conditions

Potential water quality is usually estimated based on landscape characteristics that reflect the potential hydrogeologic and ecological conditions that are expected to exist in a particular area (Omernik, 1987; Omernik and Bailey, 1997). Examples of landscape classifications include Omernik's ecoregions, Kuchler's potential natural vegetation, U.S. Department of Agriculture major land resource area (MLRA) and Winter's hydrologic landscape units. Landscape classifications for water quality management should ideally be based on the inherent characteristics of a region instead of those characteristics that are subject to anthropogenic influences (EPA, 2000).

Ecoregions are areas with presumed relative homogeneity in terrestrial and aquatic ecosystems and are currently being used to develop lake nutrient criteria by states across the nation (EPA, 1996; EPA, 2000; EPA, 2001; Heiskary, 2000). The goal has generally been to represent the spatial heterogeneity inherent in most landscapes via stratifications (or regionalization) based on presumed similarity in ecosystem function within a given strata or region (Omernik *et al.,* 1991; Omernik and Bailey, 1997; EPA, 2000; EPA, 2001; Heiskary, 2000). However, the relevance of ecoregions as the basis for lake classification is not clear because recent studies do not uniformly agree on this issue (e.g., Omernik *et al.,* 1991; Hughes *et al.,* 1994; Hawkins *et al.,* 2000; Van Sickle and

Hughes, 2000; Johnson *et al.,* 2001; Jenerette *et al.,* 2002; Rohm *et al.*, 2002; Detenbeck *et al.,* 2003 and 2004).

For example, Van Sickle and Hughes (2000) tested the utility of landscape classification approaches in Oregon, and found that the ecoregions were somewhat useful in explaining the variations in lake water quality conditions. Rohm *et al.*, (2002) also found that the spatial patterns in nutrient concentrations from 928 sites across the United States corresponded with Omernik level III ecoregions. However, Jenerette *et al.* (2002) compared ecoregions with other classification models and suggested that ecoregions may not account for the variation in lake water quality data. This assertion was later confirmed by Detenbeck *et al.,* (2003) and (2004) that the use of ecoregions for setting water quality criteria may lead to misrepresentation of reference conditions in the Lake Superior region. These findings are in agreement with the observation by Omernik and Bailey (1997) that ecoregions may not be the best framework for a particular resource problem, despite their broad use in structuring research activities and management of natural resources and environments. In fact, Omernik and Bailey (1997) and Omernik (2003) cautioned against the apparent misuse and comparisons of ecoregions.

Despite efforts to promote an ecoregions approach to nutrient criteria development, the U. S. EPA expressed a willingness to consider other landscape-based lake classification approaches to developing nutrient criteria provided they are scientifically defensible (EPA, 2000; EPA, 2001). It is therefore possible to develop other landscape classification approaches that provide a hydrologically consistent framework for the lake classification with the aim of setting nutrient water quality criteria and standards for similar groups of lakes.

The concept of "hydrologic landscapes" was proposed as a way of establishing an appropriate framework for water resource assessments, monitoring and management (Winter, 1999; Winter, 2001). According to Winter (1999) an effective framework for water resource analysis must consider the complete hydrologic system that integrates ground water, surface water and climatic variations in a given region. These hydrologic landscapes are multiples or variations of fundamental hydrologic landscape unit (FHLU) (Winter, 2001). The FHLUs are therefore defined by land-surface form, geologic texture, and climatic setting of the landscape and this can be achieved by integrating geographic information systems (GIS) with multivariate statistics (Winter, 2001; Wolock *et al.*, 2000). However, the hydrologic landscape concept has not been embraced by the water resource management community.

Others researchers (e.g. Hargrove and Luxmoore, 1998; Hatch *et al.*, 2001; Zhou *et al.*, 2003) have developed spatial clustering and agglomerative methods for landscape stratification. However, landscape classifications derived from these approaches are not appropriate for assessing lakes water quality impairment potential. This is because the ecoregions generated by Hargrove and Luxmoore (1998) are based on spatial clustering of one-kilometer pixel satellite data, and they do not take into consideration the terrain effect of watershed boundaries. Also, Zhou *et al.* (2003) used STATSGO polygons as the primary classification units to generate agro-ecoregions of Nebraska. While these may be appropriate for other land resource management issues like crop monitoring or range management, they are not appropriate for lake water quality management.

Hatch *et al.* (2001) also developed agro-ecoregions for agricultural watersheds in Minnesota. They used GIS to combine various landscape data and compared the agro-ecoregions with hydrologic cataloging units. The limitation to applying these agro-ecoregions to lake water quality management is the arbitrary nature of selecting the number of agro-ecoregions. Also, the agro-ecoregions, like Omernik's ecoregions, are based primarily on existing land use conditions. Again, it is important to note that the agro-ecoregions developed by Hatch *et al.* (2001) were intended to be used for major watersheds or river basins. Therefore, they are not likely to provide a useful framework for lake classification.

Since lakes tend to reflect their hydrogeologic setting and watershed characteristics, it seems reasonable to expect that the environmental conditions as well as the nature of change in these conditions in a lake's watershed could provide a more representative framework within which to characterize the lake potential water quality. Omernik and Bailey (1997) and Omernik (2003) have argued that the use of watersheds in water resources assessments is complicated by the general lack of agreement on the appropriate spatial scale (i.e., basin, watershed, or hydrologic cataloging units) as well as difficulties in delineating watershed boundaries. However, the availability of Elevation Derivatives for National Applications (EDNA) datasets from the U.S. Geological Survey (USGS) provide an optimal basis for delineating watershed boundaries, because the digital elevation models (DEM) obtained from EDNA datasets are comprehensive and seamless for the conterminous United States (Verdin and Verdin, 1999; Gesch *et al.*, 2002). The EDNA project, previously known as the National Elevation Dataset-Hydrologic derivatives (NED-H), was aimed at a systematic derivation of standard hydrologic derivatives (Verdin, 2000; Kost and Kelly, 2001). Also, recent developments in computer modeling have improved the accuracy assessment and reliability of watershed boundary delineation algorithms (Garbrecht and Martz, 2003). These developments make the watershed approach to lake classification a more tenable option.

2.2.3. Watershed-based classification rationale

The lake watershed provides an important spatial framework to develop a lake classification system because it is the source of runoff water, sediments and nutrients for lakes. A watershed is a topographically defined area of the earth's surface that collects runoff water and discharges it at the furthest downstream point (Ponce, 1989; Satterlund and Adams, 1992). Watersheds influence lake water quantity (e.g., peak flows and seasonal low flows) and quality (e.g., rate of sedimentation and nutrient enrichment or eutrophication) (Welch, 1978; Warren 1979; Wetzel, 1983; Frissell *et al.,* 1986; Ponce, 1989; Satterlund and Adams, 1992; Bohn and Kershner, 2002; Omernik, 2003).

Lakes, watersheds, and climatic processes are intimately linked, co-developing systems. The aging processes of lake systems are constrained or enhanced by watershed processes, while climate also affects the evolution of these watersheds (e.g., vegetation, drainage pattern, and soil organic matter content). These co-developing system processes regulate the path and net movement of water in the watersheds and consequently, the accumulation of sediments and nutrients in lakes. The development of a lake system is conceptualized as being determined by the development of the watershed within which the lake is located, as well as the potential capacity of the lake to counteract adverse influences from the watershed. The potential lake capacity refers to all possible developmental directions of a lake when exposed watershed processes, e.g., changes induced by anthropogenic activities (Warren, 1979).

As such, any changes (optimal, irregular or catastrophic) in the environment of lakes (i.e. watersheds) will invariably be reflected in the response or performance of lakes (Warren, 1979). Therefore, the observed performance of lakes (e.g., chlorophyll concentration, dissolved solids, and transparency) is a manifestation of the realized capacity, i.e. one of all possible performances that could have occurred under different developmental paths in the environment of the lake system. Since these observed performances could shift with seasonal and annual cycles of changes in climate and watershed conditions, it may be improper to classify the lakes based primarily on such actual water quality conditions like chlorophyll concentration and dissolved solids. Also, there is a huge financial cost in developing an appropriate sampling framework, locating all the lakes, and repeated measurements to account for seasonal and annual changes in actual water quality conditions.

For the preceding reasons, a lake classification system must be designed to assess potential lake water quality capacity. This classification must be based on watershed variables that underlie, determine and explain the patterns of change in physical, chemical or biological water quality performances over seasonal or annual cycles (Warren, 1979). Also an effective classification tool should be able to distinguish the various levels at which the environmental variables influence the segregation of lake classes. This issue is addressed in a review of hierarchical classification approaches, as described below.

2.2.4. Hierarchical classification

Hierarchical classification is based on the concept that ecosystems are affected by natural (and anthropogenic) processes that operate at a variety of spatial scales ranging from regional to local level (Frissel *et al.,* 1986; Allen and Hoekstra, 1992; Lomnicky, 1995; Maxwell *et al.,* 1995; Davies *et al.,* 2000; Edmunson and Mazumder, 2002). Furthermore, hierarchical principles make it possible to observe and analyze ecological complexity without confusing upper level environmental controls with lower level lake water quality possibilities (Allen and Hoekstra, 1992). For example, Lomnicky (1995) developed a watershed-based, three-level hierarchical classification for lakes in the northern Cascade Mountains of Washington. He found that the primary components of the hierarchical classification were lake position relative mountain crest, vegetation zone and basin origin. These key components were attributed to the glacially influenced landscape of the Pacific Northwest U. S. and the predominance of natural lakes in this region.

In another study of sub-arctic Alaskan lakes, Edmunson and Mazumder (2002) also examined the influences of climatic setting, morphology, transparency and typology on thermal characteristics of the lakes including water temperature, mixing depth, and heat content. They found that climatic setting, lake morphology, and lake typology showed a hierarchical regulation of growing season characteristics, lake water temperature, heat retention and stratification. Bohn and Kershner (2002) also developed a watershed-based hierarchical analytic template to improve understanding of the impact of non-point pollution sources on stream water quality. It is possible then that watershed-based lake classifications could serve as the foundation for a hierarchical lake classification that integrates the functional and spatial attributes of the landscape (e.g., erosion potential) in characterizing potential lake water quality conditions.

A simple form of hierarchical classification is the rule-based decision tree. The tree is comprised of a sequence of simple questions, the answer to each of which traces a path down the tree. The classification or prediction made by the model is determined when a final point is reached. The prediction may be qualitative (e.g., least vulnerable lakes) or quantitative (e.g., temperature class). A more rigorous form of decision trees is 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 Fabricus, 2000).

Lake water quality datasets often have missing data as well as inconsistencies in spatial and temporal sampling frequency. Also, landscape level data may only be available at different scales and for different time periods. These dataset problems make the use of decision tree (e.g., classification tree) an appropriate choice for dealing with lake classification. For example, Emmons *et al.* (1999) compared the use of classification tree method to discriminant function analysis (DFA) in classifying northern Wisconsin lakes based on actual water quality data. They found that the classification tree method resulted in lower-rates of misclassification and more interpretable lake classes than those classes derived by DFA. The classification tree method is therefore useful in defining potential lake classes by integrating lake morphology, watershed characteristics, and climate datasets. Having discussed the different approaches to lake classification, it is now appropriate to review factors that affect lake water quality in order to understand the data that may be needed to characterize and group lake watersheds.

2.3. Factors that affect lake water quality

Factors that affect lake productivity and water quality are usually interrelated and often complex (Figure 2.3). Some of these factors include surface area, lake landscape position relative to stream order, altitude or elevation, watershed area, mean watershed slope, soil erodibility and infiltration rate, as well as precipitation amount, intensity and frequency, air temperature, and light energy. The response of stream-fed lakes to climatic conditions, are intricately linked to the lakes morphology and watershed characteristics. Hence, the following review will be focused on lake and watershed characteristics of reservoirs.

Landscape position (or lake order) - the influence of lake order on water quality has been documented for some lakes in the mid-western United States (Kratz *et al.* 1997; Reira *et al.*, 2000; Magnuson and Kratz, 2000). For example, Reira *et al.* (2000) reported significant relationships between lake order and chlorophyll-a, total nitrogen, dissolved silica, Secchi depth, pH, calcium, and conductivity. They found that pH, specific conductance and calcium of lake water increased significantly with an increase in lake order. Total nitrogen and chlorophyll-a were also found to increase with increasing lake order. On the other hand, total phosphorus did not show any significant increase with increases in lake order. Generally, lake order can provide insights into the geomorphic constraint of the landscape on the physical, chemical and biological water quality characteristics of lakes.

Lake depth - is a primary determinant of heat retention and the extent of thermal stratification, which in turn can impact nutrient cycling and dissolved oxygen levels of lakes (Gorham 1961; Schindler 1971; Wetzel, 1983). Thermal stratification is the process by which lakes develop a layer of dense cooler water that underlies a surface layer of less dense, warmer water. The lower layer of water is termed the hypolimnion, the upper warmer layer is the epilimnion and the transitional layer, which acts as a barrier between the two layers, is the metalimnion. The metalimnion is usually identified by a temperature change of $1^{o}C$ per meter (Wetzel and Likens, 2000). The mixing depth of the lake or reservoir represents the thickness of the epilimnion. However, maximum lake depth controls the extent of mixing of the epilimnion and hypolimnion, which releases nutrients (especially phosphorus) and dissolved oxygen from the hypolimnion. Such mixing usually result in increased primary production and concentration of chlorophyll-a in the epilimnion. The maximum depth of a lake also affects the volume ratio of epilimnion to hypolimnion and consequently the primary productivity of lakes. Deeper lakes have epilimnion to hypolimnion ratio of less than one and tend to be less productive and so can assimilate higher nutrient loads than shallow lakes. For example, Lampert and Sommers (1997) indicated that deep lakes have lower chlorophyll-a concentration in the epilimnion than shallow lakes.

Lake surface area - is another primary determinant of heat retention and the extent of thermal stratification, which in turn can impact the nutrient cycling and dissolved oxygen levels of lakes (Gorham 1961; Schindler 1971; Wetzel, 1983; Wetzel and Likens, 2000). Smaller deeper lakes are more likely to stratify than larger and shallow lakes, because the mixing potential of larger lakes is increased by the contact between water surface and air circulating above the water. Surface area also affects the amount of direct precipitation into the lake (Wetzel, 1983). Larger lakes receive more nutrients (especially nitrogen in the form of nitrates) from precipitation than smaller lakes. The hydraulic retention time of large lakes is greater than that of smaller lakes. Retention time controls the difference between phosphorus and nitrogen concentration in lakes and the surrounding watershed (Lampert and Sommers, 1997). According to Canfield *et al.* (1989), the surface area of a lake can be used as a proxy for other factors that affect the internal nutrient cycling and water quality, e.g., mean lake depth, depth of mixed area, thickness of ice and snow cover, and shoreline development. This is why surface area is a key morphological characteristic in the assessment of lake water quality.

Lake altitude or elevation - is inversely related to the primary productivity (chlorophyll-a) of lakes (Canfield *et al.,* 1989). Lakes in hilly regions are generally less productive (i.e., lower concentration of chlorophyll-a) than lakes at lower altitudes. This is because air temperature and solar radiation decrease with increasing altitude, which affects the rate of photosynthesis (primary productivity) in lakes. The effect of altitude on lake productivity co-varies with the latitude of the lake or reservoir, because latitude integrates the effects of day length, length of the growing season, angle of incident solar radiation, and temperature on the photosynthetic processes in water bodies. Generally, lakes at lower latitudes and altitudes are more productive than lakes located in higher latitudes and altitudes (Brylinsky and Mann, 1973).

Mean watershed slope - water that reaches the soil or land surface via precipitation moves down slope in the general direction of the point of minimum gravitational force. The slope of a watershed also affects the contact time between soil and water. For example, given similar soil permeability and infiltration rate, water moves faster through soils on steeper slopes than on flat areas. The inclination (aspect) of the slope is also important in determining the physical and chemical properties streams and soil water that enter the lakes. Lakes fed by streams from west facing slopes are more likely to be rich in organic matter and have higher temperatures than lakes fed by streams from east facing slopes. This is because the west facing slopes receive warmer afternoon solar radiation than the east facing slopes. The temperature regime of lakes is critical to primary productivity, which contributes to increased organic matter content. Hence the slope and elevation together influence the microclimate of a lake.

Watershed area - the nutrients supplied to lakes (e.g., phosphorus and nitrogen) via precipitation and surface runoff are directly proportional to the area of lakes watershed area and inversely proportional to the volume or surface area of lakes (Schindler, 1971; Satterlund and Adams, 1992; Lomnicky, 1995; Lampert and Sommers, 1997). When the watershed area is small with respect to lake area, then the nutrient loading (nitrogen and phosphorus) to the lake or reservoir will be low and vice-versa. Also, where the inflow water volume is higher than outflow volume from the lake, nutrients and organisms can be flushed out before they exceed critical levels that trigger algal blooms (Canfield *et al.*, 1989). This effect is usually represented by the ratio of lake area to watershed area, a measure of lake flushing rate and hence the potential for nutrient enrichment from runoff sources (Figure 2.4). Drainage lakes receive nutrients primarily through surface water (i.e. soil erosion by the surface runoff) and atmospheric deposition of nutrients (e.g., nitrogen and phosphorus) from precipitation. Schindler (1971) hypothesized that the biological productivity of experimental lakes in Ontario, with no cultural or anthropogenic nutrient inputs, was directly proportional to ratio of watershed area to lake volume. Recent studies indicate that nutrient loading per unit lake volume is a function of the ratio of watershed area to lake volume (e.g., Lomnicky, 1995; Lampert and

Sommers, 1997). Therefore lakes with low watershed area to lake volume ratio will have relatively low nutrient loading, while lakes with high watershed area to lake volume ratio will have high nutrient loading.

The nutrient content of drainage water from a watershed is modified as the water travels through the terrestrial, stream and wetland (or littoral) areas before reaching the lake (Wetzel, 1983). The area that contributes runoff to the lakes may vary with season due to the hydrologic response of the watershed. The hydrologic response (i.e. generation of stream flow) of the watershed is explained by the "variable source area concept", which states that, "a portion of the watershed actively generates runoff in response rainfall or snowmelt" (Hewlett, 1961). This watershed response varies in recognizable pattern with season (Satterlund and Adams, 1992). In times of excessive rainfall or snowmelt, portions of the watershed that seldom contribute runoff become active contributors of runoff, which could reach a lake as either surface water or ground water. However, the nature of inflow into the lake and the lake's morphometric characteristics will ultimately determine its response to a rainfall or snowmelt event.

Furthermore, the delivery ratio of sediments from watersheds to lakes decreases with an increase in watershed size (Satterlund and Adams, 1992). The decrease in sediment delivery to lakes can be attributed the dampening of velocity as stream runoff is routed via various portions of a large watershed to lakes. Since sediments are the primary sources of lake nutrients, especially phosphorus, the reduction in sediment delivery may counteract the possible increase in nutrient loading due to an increase in the watershed area

.

Geology - since lakes are intimately linked to their watersheds by movement of materials from land to water, lake chemistry is to some extent influenced by the surface geology of the watershed. For example, under conditions of limited or absent cultural inputs, weathering of sedimentary rock materials and subsequent transport by runoff to a lake will determine the concentration of phosphorus in the lake (Golterman, 1973). This is because sedimentary rocks generally have highest concentration of phosphates, followed by metamorphic and igneous rock (Golterman, 1973; Canfield *et al.,* 1989). The geologic age of weathered rocks (soils) in a watershed also affects the salinity of lake water. For example, lakes in watersheds containing young glacial soils exhibit higher salinity than lakes in watersheds that contain older weathered soils (Jones and Bachmann, 1978).

Soil erodibility (K-factor) - lakes in watersheds where the soil K-factor is high are likely to have higher sediment and nutrient loads due to erosion than lakes in watersheds where the soil is less susceptible to erosion (Satterlund and Adams, 1992). Sediments act as conveyors of attached nutrients and chemicals like organic nitrogen and phosphorus.

The velocity of stream runoff affects the energy available to dislodge soil particles. However, runoff velocity is influenced by the rainfall intensity and slope of the watershed (Satterlund and Adams, 1992). Obviously, an intense rainfall over a hilly area is likely to result in more erosion than the same intensity rainfall over a flat area. In general, a greater percentage of eroded sediment is delivered to streams and lakes with a smaller watershed area, steeper slopes and fine-textured soils. On the other hand, the K-factor of large watersheds is poorly correlated with the transparency of lake water (Secchi depth) and lake nutrients (nitrogen and phosphorus) (Satterlund and Adams, 1992).

Soil infiltration rate - affects the amount of precipitation that enters the lake via stream runoff and ground water flow. This eventually affects the concentration of dissolved solutes (e.g., calcium and sodium) in lake waters. The intensity of precipitation received in a watershed will affect the contact time between water and the soil particles. An intense and short duration rainfall event will lead to more surface runoff compared to the same amount of rainfall over an extended duration. Moreover, fine textured sandy soils are likely to have higher infiltration rates than clayey soils.

Soil permeability - also affects the contact time between the soil particles and water passing through the soil. The longer the water stays in the soil column the greater will be the concentration of dissolved solutes in the water that eventually reaches the lake as seepage water or base flow recharge. Alkalinity of low flow streams and lakes is generally correlated to soil permeability (Woolock *et al.*, 1989). Water reaching streams and lakes via low permeability soil are likely to have high concentration of dissolved solutes such as calcium, magnesium and sodium. The slope of the watershed, especially along the path of runoff, modifies the effects of the soil permeability on solute concentration in streams and lakes.

Soil organic material - is often deposited as a mixture of peat. The general kinds of peat, according to origin are: sedimentary peat (derived from floating aquatic plants, as well as remains and fecal material of aquatic animals); moss peat (derived from mosses, including *Sphagnum*); herbaceous peat (derived from herbaceous plants); and, woody peat (derived from woody plants). In areas where anaerobic decomposition occurs in soils, biological nitrification and denitrification can affect the nitrogen flux from watersheds into lakes. High levels of organic material input into lake systems affect photosynthetic activity (chlorophyll-a) in the lakes in several ways.

For example, increased bacterial oxygen consumption at the bottom of a lake may change the chemical balance at the sediment-water interface, which influences the nutrient diffusion rates from the sediments. This situation can lead to increased phosphorus concentration and phytoplankton bloom (Canfield *et al.,* 1989). However, when the amount of carbon in organic material input from the watershed exceeds that of phosphorus, the bacteria will compete with phytoplankton for phosphorus, thereby reducing the concentration of chlorophyll-a in the lake (Canfield *et al.,* 1989).

Soil pH - is a major factor in determining the acidity or alkalinity of lake water, since lake pH is strongly correlated to soil pH (Wetzel, 1983; Lampert and Sommers, 1997). Soil pH affects the solubility of metallic ions (e.g., aluminum Al^{3+}) and dissociation of ammonium ions in soil and lake waters. The solubility of metallic ions including aluminum (Al^{3+}), iron (Fe^{2+}), copper (Cu^{2+}), zinc (Zn^{2+}), and lead (Pb^{2+}), increases with decreasing pH of soil and lake waters. Since aluminum is a major component of siliceous rocks (most common element in the earth's crust), it is often abundant in most lake watersheds and is more likely to be transported by stream runoff into lakes (Lampert and Sommers, 1997). Hence, substantial decreases in lake pH will lead to increases in dissolved aluminum ion concentrations to toxic levels. A problem associated with high pH of soil and lake waters, is the conversion of harmless ammonium ions to toxic ammonia. Ammonium ions dominate lake water at pH of less than 8. When the pH increases beyond 10.5 (critical point), almost all the ammonium ions are converted into ammonia (Lampert and Sommers, 1997). Lakes waters with high pH are likely to experience abrupt disruption of aquatic life (e.g., fish kills) when pH exceeds the critical point. Therefore, soil pH controls the solubility of aluminum ions in lake waters and the conversion of ammonium ions into toxic ammonia.

Soil cation exchange capacity (CEC) - controls the weathering process of soils and consequently the amount of calcium (Ca^{2+}) and magnesium (Mg^{2+}) ions that reaches streams and lakes (Wetzel, 1983). CEC also determines the extent to which acidic waters draining the soil surface can be neutralized (Wetzel, 1983). Soils with high CEC will have enough basic ions like calcium (Ca^{2+}) and magnesium (Mg^{2+}) to neutralize the effects of acidic ions, e.g. sodium (Na^{2+}) and aluminum (Al^{3+}) (Wetzel, 1983). Streams and lakes in watersheds that have soils with higher CEC are therefore more likely to be alkaline waters.

The general climate of a region also influences the CEC of soils. Under arid conditions, atmospheric deposition of salts gradually increases the concentration of sodium ions in soil solutions, which result in the gradual replacement of Ca^{2+} and Mg^{2+}exchange sites with Na^{2+}ions. When the Na^{2+} ions are flushed into streams or lakes during rainstorm or snowmelt, high Na^{2+} ion concentrations (sodic water) endangers aquatic organisms (Wetzel, 1983).

Soil salinity - is one of the major factors that control the salinity of surface waters (Gibbs 1970; Wetzel, 1983). Lakes that receive runoff inputs from moderate to strongly saline soils tend to contain large amounts of cations (especially Ca^{2+} and Mg^{2+}) and thus become more alkaline. On the other hand, lakes that receive runoff from non-saline to slightly saline soils contain lesser amount of cations than anions (e.g., Cl^{-}) and become more acidic (Gibbs, 1970; Wetzel, 1983).

Lake salinity is also controlled by water source, surface area, atmospheric deposition of salts directly into lakes, and the dynamics between evaporation-induced fractional crystallization and precipitation. The chemical composition of open lakes (i.e. lakes with outlets) is controlled almost entirely by the dissolved ion constituents of runoff in the lake watershed (Wetzel, 1983). On the other hand, the salinity of closed lakes (i.e., lakes without outlets) is controlled not only by the inputs of dissolved ions in runoff, but also by the fate of the dissolved ions in evaporation (Hutchison, 1957; Wetzel, 1983). In semi-arid regions like parts of Nebraska, some lakes could dry out during drought periods thereby exposing nearby lakes to nutrient input via wind action (Wetzel, 1983). Furthermore, an intense rainfall following a prolonged drought could lead to a sudden influx of nutrients to lakes and cause harmful effects such as toxic algal blooms.

The preceding discussion provides some insight into the lake eutrophication process, limitations of previous approaches to lake classification for water quality management and a rationale for watershed-based lake classification. The latter approach emphasizes the need to employ watershed characteristics that underlie, determine and explain the patterns of change in physical, chemical or biological water quality of lakes (e.g., watershed area, watershed slope, soil organic matter, soil pH, and soil erodibility). A review of the effects some of these watershed characteristics on lake water quality was done with respect to available data that were used in the dataset development process described in Chapter 3.

References cited

Allen, T.F.H. and T. W. Hoestra. 1992. **Toward a Unified Ecology**. Columbia University Press. New York, New York. 384 pp.

An, K.G. and D.S. Kim. 2003. *Response of reservoir water quality to nutrient inputs from streams and in-lake fish farms.* **Water Air and Soil Pollution.** 149 (1-4): 27- 49.

Battaglin, W.A. and D.A. Goolsby. 1996. *Using GIS and regression to estimate annual concentrations in outflow from reservoirs in the Midwestern USA.* **Proc. AWRA Symposium on GIS and Water Resources, Sept. 22-26, Ft. Lauradale, Florida.**

Bohn, B.A. and J.L. Kershner 2002. *Establishing aquatic restoration priorities using a watershed approach.* **Journal of Environmental Management**. 64: 355–363

Breiman, L., J. H. Friedman, R.A. Olshen and C. J. Stone. 1984. **Classification and Regression Trees**. Wadsworth, Inc. Belmont, California. 358 pp.

Brylinsky, M and K.H. Mann. 1973. *Analysis of factors governing productivity in lakes and reservoirs.* **Limnology and Oceanography.** 18:1-14.

Canfield, D.E., J.R. Jones, S.O. Ryding and D. Ullham 1989. *Factors and processes affecting the degree of eutrophication* pp.65-84. In Ryding, S.O. and W. Rast (Ed). **The Control Of Eutrophication Of Lakes And Reservoirs. Man And Biosphere Series**, UNESCO. Vol. 1, 314 pp.

Carlson, R. E. 1977. *Trophic state index for lakes.* **Limnology and Oceanography**. 22:361-369.

Carpenter, S.R, N.F. Caraco, D.A. Correll, R.W. Howarth, A.N. Sharpley and V.H. Smith. 1998. *Nonpoint pollution of surface waters with phosphorus and nitrogen.* **Issues in Ecology.** 3:1-11

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

Cooke, G.D., E.B. Welch, S.A. Peterson, and P.R. Newroth. 1993. **Restoration and Management of Lakes and Reservoirs (2nd Ed.).** Lewis Press, Boca Raton, Florida. 584 p.

Davies, N.M., R.H. Norris and M.C. Thoms. 2000. *Prediction and Assessment of local* *stream habitat features using large scale catchment characteristics.* **Freshwater Biology**. 45:343-369.

De' ath, G and K.E. Fabricius. 2000. *Classification and regression trees: a simple yet powerful technique for ecological data analysis.* **Ecology.** 8(11):3178-3192.

Dekker, A.G and S.W.M. Peters. 1993. *The use of Thematic Mapper for the analysis of eutrophic lakes: a case study in the Netherlands.* **International Journal of Remote Sensing**. 14(5):799-821.

DeNicola, D. M. E. de Eyto, A Wemaere and K. Irvine. 2004. *Using epilithic algal communities to assess trophic status in Irish lakes.* **Journal of Phycology.** 40 (3): 481 – 495.

Detenbeck, N.E., C.M. Elonen, D.L. Taylor, L.E. Anderson, T.M. Jicha, and S.L. Batterman. 2004. *Region, landscape, and scale effects on Lake Superior tributary water quality.* **Journal of the American Water Resources Association.** 40 (3): 705 – 720.

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

Earth-Observing-1, 2002. Mission Overview. http://eo1.gsfc.nasa.gov/overview/eo1Overview.html. Accessed in May 8, 2002

Edmundson, J. A. and A. Mazumder (2002). *Regional and hierarchical perspectives of thermal regimes in sub arctic, Alaskan lakes.* **Freshwater Biology**. 47:1–17.

Emmons, E.E., M.J. Jennings and C. Edwards. 1999. *An alternative classification method for northern Wisconsin lakes.* **Canadian Journal of Fisheries and Aquatic Sciences.** 56 (4):661-669.

EPA (U.S. Environmental Protection Agency). 1996. **Environmental Indicators of Water Quality in the United States.** Report No. EPA-841/R-96-002. Washington, D.C.

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

EPA (U.S. Environmental Protection Agency). 2001. **Development and Adoption of Nutrient Criteria into Water Quality Standards.** Memo No. WQSP-01-01. Washington, D.C.

Frissell, C.A., W.J. Liss, C.E. Warren and M.D. Hurley. 1986. *A hierarchical framework for stream habitat classification: viewing streams in a watershed context.* Environmental Management. 10:199-214.

Garbrecht, J. and L. W. Martz. 2003. *Assessing the performance of automated watershed delineation process from digital elevation models*, pp 17-24 in Lyon, J.G. **GIS for Water Resource and Watershed Management**. 266 pp.

Gesch, D., M. Oimoen, S. Greenlee, C. Nelson, M. Steuck and D. Tyler. 2002. *The National Elevation Dataset.* **Photogrammetric Engineering And Remote Sensing**. 68(1): 5-11.

Gibbs, R.J. 1970. *Mechanisms for controlling world water chemistry.* **Science** 180:71-73.

Golterman, H.L. 1973. *Natural phosphate sources in relation to phosphate budget: a contribution to the understanding of eutrophication.* **Water Research**. 7:3-17.

Gorham, E. 1961. *Morphometric control of annual heat budget in temperate lakes.* **Limnology and Oceanography**. 9: 525 – 529.

Grigg, D. 1965. *The logic of regional systems***. Annals of the American Association of Geographers**. 55: 465-491.

Hargrove, W.W. and R.J. Luxmoore. 1998. *A New High-Resolution National Map of Vegetation Ecoregions Produced Empirically Using Multivariate Spatial Clustering.* **ESRI ARC/INFO User Conference**. http://gis.esri.com/library/userconf/proc98/PROCEED/TO350/PAP333/P333.HTM

Hatch, L.K., A. Mallawatantri, D. Wheeler, A. Gleason, D. Mulla, J. Perry, K.W. Easter, R. Smith, L. Gerlach, and P. Brezonik. 2001. Land management at the major watershed-agroecoregion intersection. Journal of Soil and Water Conservation. 56 (1):44-51.

Hawkins, C.P., R.H. Norris, J. Gerritsen, R.M. Hughes, S.K. Jackson, R.K. Johnson and R. J. Stevenson. 2000. *Evaluation of landscape classifications for the prediction of freshwater biota: synthesis and recommendations.* **Journal of the North American Benthological Society**. 19(3): 541-556.

Heiskary, S. A. and C. B. Wilson. 1989. *The regional nature of lake water quality across Minnesota: an analysis for improving resource management.* **Journal of Minnesota Academy of Science.** 55:71-77.

Heiskary, S. A. 2000. *Ecoregional Classification of Minnesota Lakes.* Pages B4-B5 *in* EPA (U.S. Environmental Protection Agency). 2000. **Nutrient Criteria Technical Guidance Manual for Lakes and Reservoirs**. Report No. EPA-822-

B00-001. Washington, D.C.

Hewlett, J.D. 1961. *Watershed management.* Pages 61-66 *in* **Annual Report for 1961**. Forest Service, U.S. Department of Agriculture. Southeastern Piedmont Experimental Station. Ashville, N.C. (Microfiche).

Heywood, R.B., H.J.G. Dartnall and J. Piddle. 1980. *Characteristics and classification of lakes of Signy Island, Outh Orkney Islands, Antarctica.* **Freshwater Biology.** 10:47–60.

Hughes, R.M. 1995. *Defining acceptable condition.* Pages 31 – 41 *in* W.S. Davis and T.P. Simon (eds.). **Biological and Nutrient Criteria. Tools for Water Resource Planning and Decision making**. Lewis Publishers, Boca Raton, Florida.

Hughes, R.M., S.A. Heiskary, W.J. Matthews, and C.Q. Yoder. 1994. *Use of ecoregions in biological monitoring.* Pages 125-151 *in* S.L. Loeb and A. Spacie (Eds.). **Biological Monitoring of Freshwater Ecosystems**. Lewis Press, Chelsea, Michigan.

Hutchison, G.E. 1957. **A Treatise On Limnology. I. Geography, Physics And Chemistry**. John Wiley and Sons. New York. 1015 pp.

Jenerette, G.D., J. Lee, D. Waller and R.E. Carlson. 2002. *Multivariate analysis of the Ecoregion delineation for aquatic ecosystems.* Environmental Management. 29 (1): 67- 75.

Johnson, G.D., W.L. Myers, and G.P. Patil. 2001. *Predictability of surface water pollution loading in Pennsylvania using watershed-based landscape measurements.* **Journal of the American Water Resources Association**. 37(4):821-835.

Jones, J. R. and R. W. Bachmann. 1978. *Trophic status of Iowa lakes in relation to origin and glacial geology.* **Hydrobiologia.** 56:267-273.

Kratz T.K., K.E. Webster, C.J. Bowser, J.J. Magnuson and B.J. Johnson. 1997. *The influence of landscape position on lakes in northern Wisconsin.* **Freshwater Biology.** 37: 207 – 217.

Koponen, S., J Pulliainen, K. Kallio and M. Hallikainen 2002. *Lake water quality classification with airborne hyperspectral spectrometer and simulated MERIS data.* **Remote Sensing of Environment**. 79 (1): 51 – 9

Kost, J., and Kelly, G. 2001. *Watershed delineation using the National Elevation*

Dataset and semiautomated techniques, in Proc., **Twenty-First Annual ESRI International User Conference, San Diego, California, July 9-13, 2001**. Redlands, California, Environmental Systems Research Institute, Inc. (CD-ROM).

Lampert, W. and U. Sommers. 1997. **Limnoecology: The Ecology of Lakes and Streams**. Translated by J.F. Haney. Oxford University Press, New York, NY. 382 p.

Larson, D.W. 1970. **On reconciling lake classification with the evolution of four oligotrophic lakes**. PhD. Dissertation. Oregon State University, Oregon.

Lathrop (Jr.), R.G. 1992. *Landsat Thematic Mapper monitoring of turbid Inland water quality*. **Photogrammetric Engineering and Remote Sensing.** 54:465-470.

Lomnicky, G.A. 1995. **Lake Classification in the Glacially Influenced Landscape of the North Cascade Mountains, Washington, USA**. PhD. Dissertation. Oregon State University, Oregon.

Magnuson J.J. and T.K. Kratz. 2000. *Lakes in the Landscape: approaches to regional limnology*. **Verhandlungen Internationale Verreingung Für Limnologie.** 27:1-14.

Maxwell, J.R., C.J. Edwards, M.E. Jensen, S.J. Pautian, H. Parrot and D.M. Hill. 1995. **A hierarchical framework of aquatic ecological units of North America (Nearctic Zone).** North Central Experiment Station, Forest Service, U.S. Department of Agriculture. General Technical Report NC-176. St. Paul, MN.

Momen, B., and J.P. Zehr. 1998. *Watershed classification by discriminant analysis of lake water chemistry and terrestrial characteristics.* **Ecological Applications.** 8:497–507.

Mortimer, C.H. 1942. *The exchange of dissolved solids between mud and water in lakes.* **Journal of Ecology**. 30:147 – 201.

Nelson, S.A.C, P. A. Sorano, K.S. Cheruvelil, S.A. Batzli and D.A. Skole. 2003. *Regional assessment of water clarity using satellite remote sensing.* **Journal of Limnology** 1(27):27-32

Niles, R.K., D.L. King and R. Ring. 1996. *Lake classification systems - part I.* **The Michigan Riparian**. http://www.mslwa.org/lkclassif1.html.

Olmanson, L.G., Kloiber, S.M., Bauer, M.E., Day, E.E., and Brezonik, P.L. 2001. **Upper Great Lakes RESAC lake water quality, Image processing protocol**. Upper Great Lakes Regional Earth Science Applications Center. University of Minnesota, St. Paul, Minnesota. http://resac.gis.umn.edu/lakeweb/index.htm

Omernik, J.M., 1987, *Ecoregions of the Conterminous United States.* **Annals of the Association of American Geographers.** 77:118-125.

Omernik, J.M. C.M. Rohm, R.A. Lillie, and N. Mesner. 1991. *Usefulness of natural regions in lake management: analysis of variation among lakes in Northwestern Wisconsin, USA.* **Environmental Management.** 15:281-293.

Omernik, J.M., 2003. *The misuse of hydrologic unit maps for extrapolation, reporting and ecosystem management.* **Journal of the American Water Rcsources Association**. 39(3):563–573.

Omernik, J.M. and R.G. Bailey. 1997. *Distinguishing between watersheds and ecoregions.* **Journal Of The American Water Resources Association.** 33(5):935–949.

Paulsen, S.G., R.M. Hughes, and D.P. Larsen. 1998. *Critical elements in describing and understanding our nation's aquatic resources.* **Journal of the American Water Resources Association.** 34 (5): 995 – 1005.

Ponce, V.M. 1989. Engineering **Hydrology: Principles And Practices**. Prentice-Hall, Inc. New Jersey. 627p.

Reira J.L., J.J. Magnuson, Kratz T.K. and K.E. Webster. 2000. *A geomorphic template for the analysis of lake districts applied to Northern Highland Lake District, Wisconsin, USA.* **Freshwater Biology.** 43:301-318.

Rohm, C.M., J.M. Omernik, A.J. Woods and J.L. Stoddard. 2002. Regional characteristics of nutrient concentrations in streams and their application to nutrient criteria development. **Journal of the American Water Resources Association.** 38 (1): 213 – 239.

Satterlund, D.R. and P.W. Adams. 1992. **Wildland Watershed Management.** 2nd Ed. J. Wiley and Sons, New York, N.Y. 436p.

Schindler, D.W. 1971. *Light, temperature and oxygen regimes of selected lakes in experimental lakes area, Northwestern Ontario.* **Journal of Fisheries Research Board of Canada**. 28:157- 169.

Thorton, K.W., B.L. Kimmel, and F.E. Payne (eds.). 1990. **Reservoir Limnology: Ecological Perspectives.** John Wiley, New York. 260 p.

Van Sickle, J. and R.M. Hughes. 2000. *Classification strengths of ecoregions, catchments and geographic clusters for aquatic vertebrates in Oregon.* **Journal of the North American Benthological Society**. 19:370-384.

Verbyla, D.L. 1987. *Classification trees: a new discrimination tool.* **Canadian Journal**

of Forestry Research. 17:1150–1152.

Verdin, K.L. and J.P. Verdin. 1999. *A topological system for delineation and codification of the Earth's river basins.* **Journal of Hydrology**. 218:1 – 12.

Verdin, K. 2000. *Development of the National Elevation Dataset-Hydrologic Derivatives (NED-H)*, in Proc., **Twentieth Annual ESRI International User Conference, San Diego, California, July 10-14, 2000**. Redlands, California, Environmental Systems Research Institute, Inc. (CD-ROM).

Warren, C. E. 1979. **Toward classification and rationale for watershed management and stream protection.** EPA - 600 / 3-79-059. United States Environmental Protection Agency. Corvallis, Oregon. 143p.

Welch, D.M. 1978. **Land/Water Classification. A Review of Water Classifications and Proposals for Water Integration into Ecological Land Classification.** Ecological Land Classification Series, No.5. Environment Canada. Ottawa. 54p.

Wetzel, R.G. 1983. **Limnology.** Saunders College Publishing, Philadelphia, PA. 767p.

Wetzel, R.G. 1965. *Nutritional aspects of algal productivity in marl lakes with particular reference to enrichment bioassays and their interpretations. p. 139 – 157* in **Primary Productivity in Aquatic Environments,** C.R. Goldman (ed.), University of California Press, Berkeley, California.

Wetzel, R.G. and G.E. Likens, 2000. **Limnological Analysis.** 3rd Ed., Springler Verlag, New York. 432p.

Whittier, T.R., D.P. Larson, S.A. Peterson and T.M. Kincaid. 2001. *A comparison of impoundments and natural drainage lakes in the Northeast USA.* **Hydrobiologia.**

Willen, E., S. Hajdu and Y. Pejler. 1990. *Summer phytoplankton in 73 nutrient poor Swedish lakes: classification, ordination and choice of long term monitoring projects.* **Limnologica.** 20:217–228.

Winter, T.C. 1999. *The relation of streams, lakes, and wetlands to groundwater flow systems.* **Hydrogeology Journal.** 7:28–45.

Winter, T.C. 2001. *The concept of hydrologic landscapes.* **Journal of the American Water Resources Association**. 37(2):335–349.

Wolock, D.M., G.M. Hornberger, K.J. Bevan and W.G. Campbell. 1989. *The relationship between topography and soil hydraulic characteristics to lake alkalinity in northeastern United States.* **Water Resources Research.** 25 (5):829-837.

Wolock, D. M., T. C. Winter, and G. McMahon. 2000. **Delineation of hydrologic**

setting regions in the United States using geographic information system tools and multivariate statistical analyses. Unpublished manuscript. Denver, Colorado. U.S. Geological Survey. 25pp

Yang, M., C. J. Merry, R.M. Sykes. 1999. *Integration of water quality modeling, remote sensing and GIS.* **Journal of the American Water Resources Association**. 35(2):253-264.

Zhou, Y., S. Narumalani, W.J. Waltman, S.W. Waltman and M.A. Palecki. 2003. *A GIS-based spatial pattern ecoregion mapping and characterization.* **International Journal of Geographical Information Science**. 17(5):445-462.

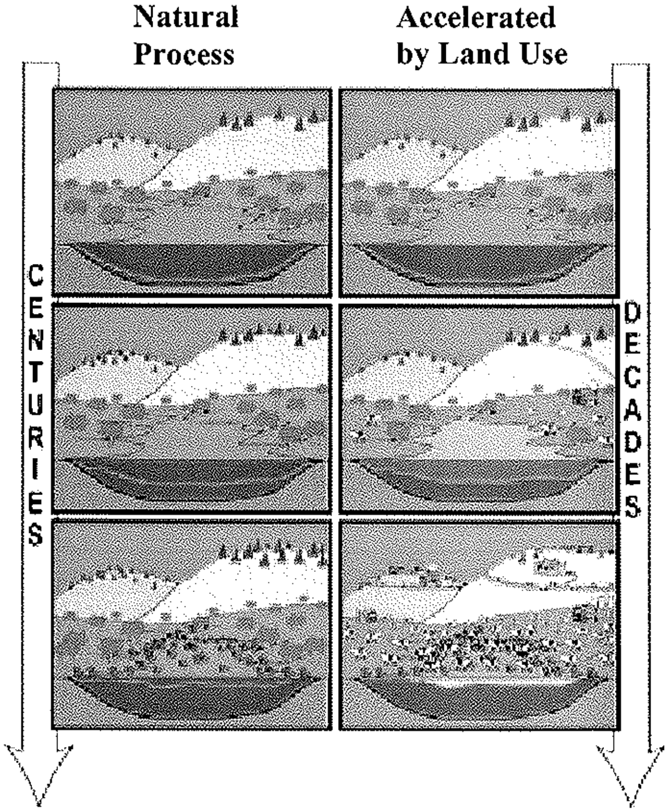


Figure 2.1. Lake eutrophication (or aging) process. The natural process takes place over centuries, but this process can be accelerated to occur in a few decades due to increased land use activities, e.g. agricultural land use. *Modified after Carpenter et al. (1998).*

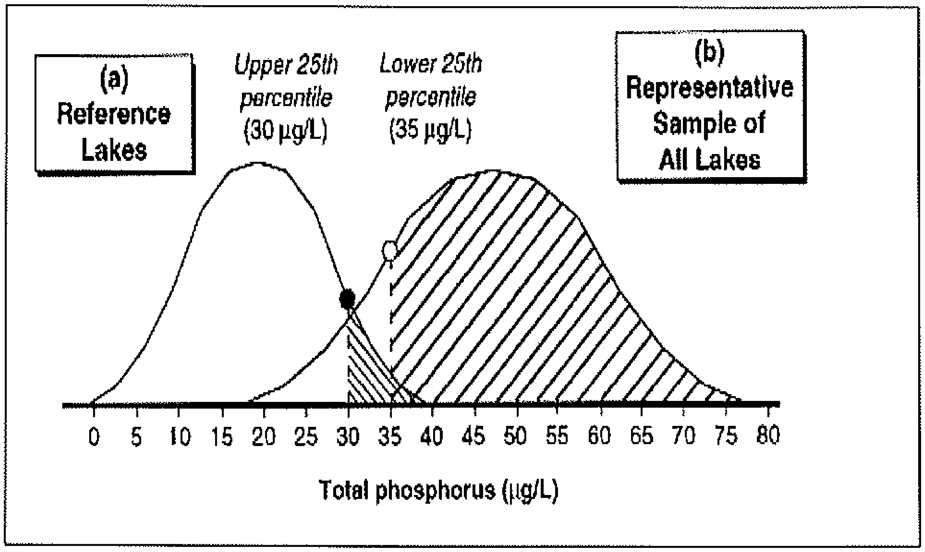


Figure 2.2. Sample distribution approach to establishing lake reference conditions; hypothetical example based on total phosphorus (EPA, 2001)

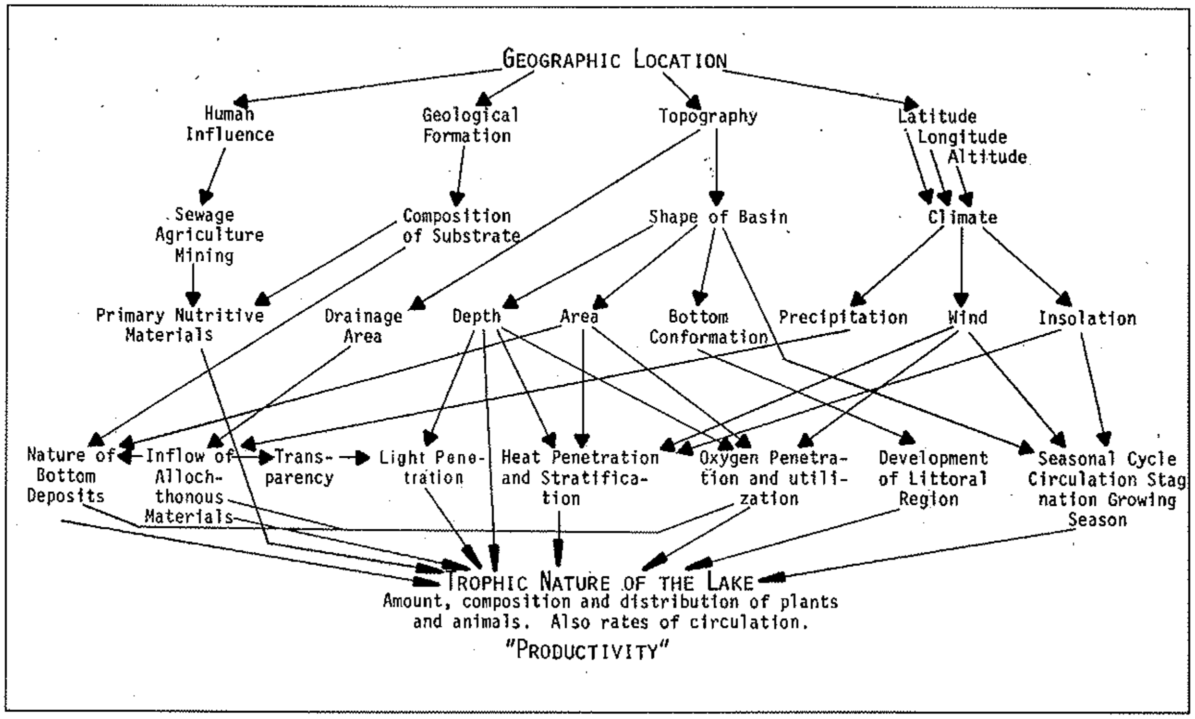

Geographic Location
Human Influence
Geological Formation
Topography
Latitude
Longitude
Altitude
Sewage
Agriculture
Mining
Composition of Substrate
Shape of Basin
Climate
Primary Nutritive Materials
Drainage Area
Depth
Area
Bottom Conformation
Precipitation
Wind
Insolation
Nature of Bottom Deposits
Inflow of Allochthonous Materials
Transparency
Light Penetration
Heat Penetration and Stratification
Oxygen Penetration and utilization
Development of Littoral Region
Seasonal Cycle Circulation Stagnation Growing Season
Trophic Nature of the Lake
Amount, composition and distribution of plants and animals. Also rates of circulation.
"Productivity"

Figure 2.3. Interrelationships of factors that affect lake productivity (Canfield *et al.*, 1989).

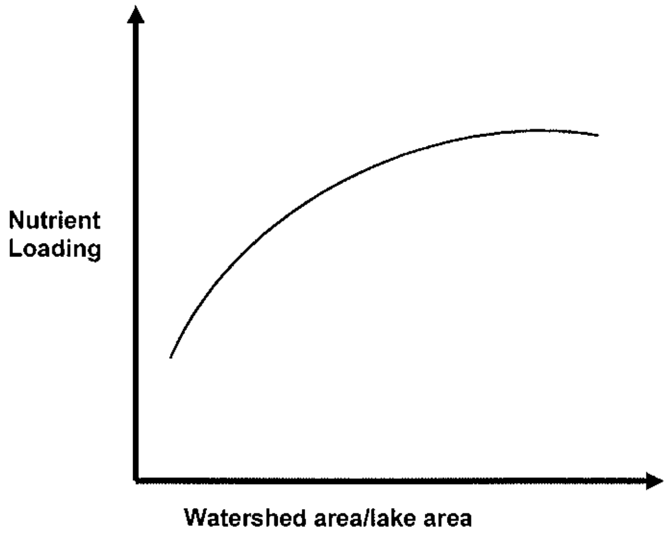


Figure 2.4. Typical relationship between nutrient loading and the ratio of watershed area to lake area.

CHAPTER 3. DATASET DEVELOPMENT AND PRELIMINARY ANALYSIS

3.1. Introduction

The validity and broad application of the results of any assessment depends on the quality of data used in that analysis. Hence, it is important to obtain or develop accurate datasets that are relevant to the watershed based lake classification and understand the nature of their variations. The work reported in this chapter represents the geospatial dataset development process and preliminary analyses of the patterns of variation and associations of the dataset. This chapter is divided in four sections that reflect: (a) development on an up-to-date and comprehensive digital map of Nebraska lakes in order identify reservoirs in the state; (b) delineation of watershed boundaries for selected Nebraska reservoirs; (c) assessment of whether the sampled reservoirs, used in delineating watershed boundaries are representative of all Nebraska reservoirs that are at least 4 hectares (or 10 acres) in surface area: and (d) preliminary analysis of spatial patterns of variations and correlation analysis of the geospatial dataset of watershed characteristics.

3.2. Mapping Nebraska reservoirs

Accurate identification of Nebraska reservoirs is necessary to delineate their watershed boundaries and develop management criteria for groups or classes of reservoirs based on their potential water quality. However, there has been no existing digital geospatial dataset that provides a complete depiction of the number and locations of all reservoirs in Nebraska. In order to develop a geospatial dataset of Nebraska reservoirs, it was important first to map all Nebraska lakes (Figure 3.1). The types and sources of data that were used in this work are summarized in Table (3.1).

Initially all water features from the latest version of the NRCS (Natural Resources Conservation Service) Soil Survey Geographic Database (SSURGO) were extracted and used as baseline dataset (or coverage) of water features in a Geographic Information Systems (GIS). The baseline dataset was edited to remove any stream-like features and artifacts of the SSURGO data capture process (e.g., small polygons associated with many large lakes). This GIS coverage was then updated using other data sources, including the U.S. Geological Survey (USGS) National Hydrography Dataset (NHD), USGS National Land Cover Data (NLCD) and U.S. Census Bureau TIGER (Topologically Integrated Geographic Encoding and Referencing) data, to fill gaps in counties where there were no available SSURGO data at the time of dataset development. All the datasets were projected into Albers Conformal coordinate system and same datum (NAD1983) to reduce distortions at the edges of the data and thus ensure that the data overlay properly (ESRI, 1997).

Polygons in lakes GIS coverage were filtered to remove polygons less than 0.8 hectares (2 acres) in surface area. This threshold was used to remove additional artifacts of all the input data sets (e.g., digitizing errors in TIGER data). The choice of 0.8 hectare threshold was based on the fact that it generally reflects the maximum size of polygons included in the data as a result of digitizing errors or slivers from data transformations during the lake mapping process. The filtered coverage was then edited to generate a draft map of Nebraska lakes (Figure 3.2).

3.2.1. Assigning attributes to lake features

With a draft digital map of Nebraska lakes assembled (hereafter referred to as Nebraska lakes-1), the next step was to characterize the lake polygon features by "lake type" in order to segregate the reservoirs from natural lakes and sand pits. The natural lakes are found mostly in the Nebraska Sand Hills region, a unique ecological area of grass covered sand dunes (Figure 3.3). The sand pits on the other hand were the results of land excavations to provide aggregate material for road constructions in the mid-twentieth century. The primary water source for both natural lakes and sand pits is ground water, so these lake types were excluded from any further consideration.

The Nebraska Dams inventory dataset (from U.S. Army Corps of Engineers) and a sampled Nebraska lakes water quality dataset (Holz, 2002) were used as initial sources of reservoir information. However, the projection parameters of the Dams inventory data were found to be inaccurate which led to incorrect alignment with Nebraska lakes-1, therefore the Dams inventory dataset was recreated. This was done using geographic coordinate information (latitude and longitude) of the original Dams inventory dataset, and the projection was reestablished to the Universal Transverse Mercator (UTM) coordinate system. This revised Dams inventory dataset was then reprojected into Albers Conformal coordinate system and it aligned better with Nebraska lakes-1. Once the two data layers were overlaid correctly, a *spatial join* function in ArcMap® GIS software was used to extract lake type attributes (i.e. reservoirs) from the Dams inventory dataset.

Next, a polygon coverage of the Nebraska Sand Hills region was overlaid on to the Nebraska lakes-1 coverage. All lakes within the Sand Hills region were identified as natural lakes, except those already designated as reservoirs or sand pits based on dams inventory and sampled lakes datasets. The remaining lakes (everything excluding the Sand Hills Region) were either reservoirs or sand pits and were exported into new lakes coverage (Nebraska lakes-2). A preliminary size restriction of 4 hectares was applied to separate this new coverage into two sub-layers, because the reservoir classification work was the part of an Environmental Protection Agency (EPA) effort to develop nutrient criteria for lakes larger than 4 hectares (EPA, 2001). Thus, the sub-layer with lakes larger than 4 hectares was processed first (Nebraska lakes-2a). Unidentified lakes in this layer that intersected with Nebraska streams data were identified as reservoirs while the rest were identified as sand pits. The same approach was applied to the sub-layer containing lakes smaller than 4 hectares (Nebraska lakes-2b). Again, lakes that did not intersect with streams were identified as sand pits while the rest were identified as reservoirs.

Both Nebraska lakes-2a and 2b were merged and the resulting coverage (Nebraska lakes-3) was panned through, on county by county basis, to verify the correct assignment of lake types. Other ancillary data, e.g. coordinate information in tabular lakes dataset from Nebraska Department of Environmental Quality, were used to aid the revision process. In all, 17 lakes that were initially identified as sand pits were reassigned as reservoirs, while 6 reservoirs were also reassigned as sand pits. Some of the lakes were identified as "oxbow lakes", and so a new category of lake type was created. In the final step, Nebraska lakes 1 and Nebraska lakes 3 were combined into an up-to-date digital map of Nebraska lakes.

3.2.2. Final Map of Nebraska lakes

The final and updated digital map of all Nebraska lakes that are at least 0.8 hectares (2 acres) is shown in Figure 3.4. This Nebraska lakes map comprises of 13,520 lakes (i.e. 6796 reservoirs, 3644 natural lakes, 3068 sand pits and 12 oxbow lakes). This map is believed to be the most comprehensive and accurate representation of Nebraska lakes in a GIS coverage, compared to other Nebraska lake datasets that were available at the time of the dataset development (Figures 3.5). All reservoirs larger than 4 hectares were extracted from the updated map of Nebraska lakes (Figure 3.6). The 4 hectares size restriction reflects the minimum threshold required for EPA lake nutrient criteria development (EPA, 2001). The extracted reservoir coverage (Figure 3.6) was then used in the watershed boundary delineation process described below.

3.3. Delineating reservoir watershed boundaries

A simple and effective means to delineate watershed boundaries is required for the watershed-based reservoir classification. Previous efforts to delineate reservoir watershed boundaries for water resource management were limited by artifacts of county-based digital elevation models (DEM) such as seams. An existing database that is commonly used as framework for characterizing lake watersheds is the Hydrologic Unit Coverage (HUC). The most comprehensive and nationally available HUC's are based on a 8-digit standardized coding system that divides the United States into four hierarchical levels, i.e. regions, sub-regions, accounting units (or basins) and cataloging units (or sub-basins) (U.S. Geological Survey, 1982; Seaber *et al.,* 1987). However, the current 8-digit HUCs do not provide sufficient detail in order to easily extract or delineate watershed boundary of reservoirs for water quality assessment (Omernik and Bailey, 1997; Verdin and Verdin, 1999; Verdin, 2000; Kost and Kelly, 2001; Omernik, 2003).

Seamless digital elevation model (DEM) derivatives are available from parallel United States Geological Survey (USGS) projects; namely, the National Elevation Dataset (NED) and Elevation Derivatives for National Applications (EDNA). According to Gesch *et al.* (2002), the NED was the result of efforts by the USGS to provide 1:24,000-scale DEM data for the conterminous United States. The NED was developed by merging the highest resolution and best quality elevation data available across the United States into a seamless raster format. The USGS Elevation Derivatives for National Applications (EDNA) project, previously known as the National Elevation Dataset-Hydrologic derivatives (NED-H), was aimed at a systematic derivation of standard hydrologic derivatives (http://edna.usgs.gov/; Kost and Kelly, 2001).

EDNA datasets are available for all of the conterminous United States of America and they include synthetic streams, sub-catchments (i.e. contributing drainage area for each stream reach), and a revised hydrologic unit coding system, all generated from 30-meter DEM (http://edna.usgs.gov/). The sub-catchments and hydrologic coding system were based on a system that uses the Pfafstetter stream numbering scheme for codifying river basins (Pfafstetter1989; Verdin and Verdin, 1999; Verdin, 2000). The Pfafstetter stream numbering scheme is a self-replicating numbering system based on the topology of the drainage network and the size of the surface area drained by that network. This allows for identification numbers of the smallest sub-basins and inter-basins extractable from a DEM (Pfafstetter1989; Verdin and Verdin, 1999). According to Verdin and Verdin (1999), "the appeal of the Pfafstetter's scheme is due to its economy of digits, the topological information that the digits carry, and the global applicability of this approach".

3.3.1. Automated delineation of watershed boundary

EDNA datasets, obtained from EROS Data Center (EDC), included Pfafstetter sub-catchments, modified hydrologic unit boundaries, synthetic (i.e. DEM generated) streamlines, flow direction and shaded relief data for all areas that drain into water bodies of Nebraska (Pfafstetter, 1989; Verdin and Verdin, 1999; Gesch *et al.*, 2002; http://edna.usgs.gov/). The DEM-based EDNA datasets were used in ArcView® GIS to delineate the watersheds of 88 Nebraska reservoirs (Figure 3.7). These reservoirs were selected because their location and type have been verified and they form part of an existing lake water quality database obtained from the School of Natural Resources, University of Nebraska – Lincoln (Holz, 2002).

Watershed boundaries of these reservoirs were delineated using EDNA stage-2 ArcView® GIS extension together with the ArcView "Hydro" extension (Olivera *et al.*, 2000; USGS, 2001). This process identifies a reservoir's watershed based on stream network, stream flow direction and sub-catchments information available in the EDNA dataset (Verdin and Verdin, 1999). The flow direction data is the primary DEM derivative that is used in delineating sub-catchment and watershed boundaries.

After the DEM data was processed to remove spurious sinks, i.e. areas or depressions where water enters but cannot exit, a flow direction grid was generated. The flow direction grid was comprised of cell values (integers) that indicated the direction of water flow from each cell, based on DEM elevation values (Figure 3.8). The direction of water flow for each cell in the elevation grid was determined and a value is assigned to the flow direction grid. There were eight valid output water flow directions, with respect to the 8-cell neighborhood surrounding each cell (ESRI, 1992; ESRI, 1997). The flow direction grid was then used to generate a flow accumulation grid, which was also used to derive the synthetic stream network. The synthetic stream network is created by identifying flow accumulation grid cells that had high cell or flow accumulation values (ESRI, 1992 ESRI, 1997).

The EDNA stage-2 ArcView tool was used to aggregate the sub-catchments based on the Pfafstetter coding systems (Verdin and Verdin, 1999). Where there were difficulties in aggregating the sub-catchments (as was the case with some small reservoirs in low relief areas), ArcView® GIS "Hydro" extension was used, together with the flow accumulation data, to delineate the reservoir watershed boundary (Olivera *et al.,* 2000). The EDNA-derived stream network was useful for locating outputs (or pour-points) from which the sub-catchments were delineated. The watershed boundaries of sampled Nebraska reservoirs that were derived from EDNA DEM are shown in Figure 3.9.

3.3.2. Assessing the accuracy of automated watershed boundary delineation

The watershed boundaries of selected reservoirs were overlaid on digital raster graphics (DRG) and manually digitized watershed boundaries, obtained from the Nebraska Department of Natural Resources (DNR), to compare the effectiveness of the watershed delineation process. The DRG data consisted of scanned images of 1:24000 scale topographic maps. Therefore, the DRG were used as background data for the comparisons. For example, an overlay of the DEM derived and DNR digitized watershed boundaries of Harry Strunk reservoir on DRG showed relatively little disagreements in boundary outline (Figure 3.10). However, it is worth noting that the DEM-derived watershed boundary was closer to the dammed portion of the reservoir than the DNR boundary. This is particularly important because the USGS guideline for watershed boundary delineation emphasizes the need to take into account the dam structure of reservoirs (NRCS, 2002).

Also, the percentage deviation of the DEM-generated watershed boundaries from the digitized watershed boundaries was determined based on watershed topologic, geometric and hydrologic parameters such as total drainage area, catchment slope, mean drainage density, and total and mean drainage length (Garbrecht and Martz, 2003). This was done to ascertain the effectiveness of the automated watershed boundary delineation process. This was important to ensure the validity of any subsequent analyses based on the watershed boundaries.

Watershed area was computed from the DEM-derived and DNR-digitized watershed boundaries datasets for 18 randomly selected reservoirs (representing about a quarter of the all the DEM-derived watershed boundaries). Other watershed parameters were obtained by overlaying the watershed boundaries of the 18 selected reservoirs on raster (or grids) datasets of slope, drainage network, and drainage density. Summary or "zonal" statistics for each watershed parameter (e.g., maximum, minimum, and mean catchment slope) were generated for reservoir watershed boundaries using ArcMap GIS software. This was done for both DEM-derived and DNR-digitized watershed boundaries datasets. The DNR-digitized watershed boundaries were used as validation datasets. The percent deviation of DEM-derived watershed boundaries from the validation datasets, based on total drainage area for example, was computed as follows:

$$\text{Percent Deviation (\%)} = \frac{\text{ABS}(\text{Area}_{\text{DNR}} - \text{Area}_{\text{DEM}})}{\text{Area}_{\text{DNR}}} \text{ x } 100 \qquad (3.1)$$

where ABS is a function used to transform the difference into absolute values.

Results of the comparison of topographic, topologic and hydrologic parameters for the 18 selected watersheds showed less than 10 percent deviation of DEM derived watershed boundaries from DNR-digitized watershed boundaries (Table 3.2). The watershed parameter values in Table 3.2 represent average values from 18 selected watersheds of small, medium and large reservoirs. Each watershed was considered as a lumped unit; so the values do not reflect spatial variations within individual watersheds. For example, the percent deviations based on total drainage area, drainage density, and mean watershed slope were 1.79, 4.12, and 1.84, respectively. It is important to note that the deviations of DEM-derived watersheds from DNR validation watersheds were less than 5 percent. This is because total drainage area, drainage density, and mean watershed slope are critical to the transport of sediment and agricultural pollutants via streams to reservoirs (Satterlund and Adams, 1992).

The aforementioned comparisons indicate that the automated watershed boundary delineation process, combined with the EDNA datasets, was effective in delineating watershed boundaries for Nebraska reservoirs. The seamless nature of the EDNA datasets for the entire area that drains into water bodies in Nebraska also ensured that the watershed boundaries conformed to the topography of the state. Despite the fact that the watershed boundaries have not been field-checked and standardized, they still provide sufficient conformity with local terrain.

3.4. Assessing representativeness of sampled Nebraska reservoirs

The watershed boundary delineation process, discussed in section 3.3, was based on reservoirs that were sampled to assess selected water quality parameters including secchi depth, chlorophyll, and total phosphorus (Holz, 2002). These reservoirs were sampled without any particular statistical sampling design. Also, watershed boundaries of some of the sampled reservoirs extend beyond the Nebraska state line and were excluded from any analysis. This raises two key questions; namely (i) were sampled reservoirs whose watersheds fall within Nebraska (adjusted sampled data) different from boundary reservoirs whose watersheds fall outside Nebraska? and (ii) were the adjusted sampled reservoirs data (mentioned above) different from all reservoirs (larger than 4 hectares) whose watersheds fall within Nebraska?

Nebraska reservoirs that were categorized as follows; Groups 1and 2 consist of sampled reservoirs whose watershed boundaries fall outside Nebraska (8), and within (80) Nebraska, respectively; and, Group 3 consists of all reservoirs whose watershed boundaries fall within Nebraska and were at least 4 hectares in surface area (954). The sampled reservoirs (Groups 1 and 2) make up 9.22 percent reservoirs in Group 3. When the reservoir dataset was adjusted to exclude boundary reservoirs, the proportion of sampled reservoirs (Group 2) to reservoirs Group 3 declined to 8.39 percent. Knowledge of the proportions of the sampled reservoirs to all Nebraska reservoirs that were at least 4 hectares, provide a context for developing water quality standards. This information on the proportion of sampled reservoirs to all Nebraska reservoirs addresses a key requirement for the development of lake nutrient criteria guidelines (EPA, 2001).

Initially the distributions of all three datasets (Group 1, 2, and 3) were compared using box-whisker plots. Box-whisker plots can provide a concise picture of the distribution of the datasets (Tukey, 1977). The central line in each box represents the median value (50th percentile) while the edges of the box represent the first quartile (25th percentile) and third quartile (75th percentile). The mean area of reservoirs in Groups 1, 2, and 3 were 2457, 249.35, and 10.21 hectares, respectively (Table 3.2). The average area of reservoirs in Group 3 was relatively small compared to both Groups 1, and 2 reservoirs. This was due to the large number of small reservoirs in the Group 3 as reflected in the median, upper and lower quartile values of the box-whisker plots (Figure 3.11).

The next step in assessing the representation of sampled reservoirs was done to determine the significance of the abovementioned differences between the datasets; specifically Group 1 vs. Group 2, and, Group 2 vs Group 3. The "Npair1way" non-parametric procedure in SAS® was used to test the significance of the aforementioned differences based on Wilcoxon Signed-Rank and Kruskal-Wallis test statistics (SAS Institute, 2000). For example, the normal approximation of Wilcoxon Signed-Rank paired test (z) for a variable (T) is given as:

$$z = \frac{\mathrm{T} - \mu_T}{\sigma_T} \qquad (3.2)$$

where, $\mu_T = \dfrac{n(n+1)}{4}$ and $\sigma_T = \sqrt{\dfrac{n(n+1)(2n+1)}{24}}$

The null hypothesis (H_o) was that there were no differences between the datasets (e.g., H_o: Group 1µ = Group 2µ, where μ is the mean reservoir area). There were three options for stating the alternative hypotheses (H_a). One option was a non-directional H_a, also known as two-tailed test (e.g., H_a: Group $1\mu \neq$ Group 2μ) (Sheskin, 2000). There were two possible directional or one-tailed alternative hypotheses (i.e. H_a: Group $1\mu>$ Group 2μ or H_a: Group $1\mu <$ Group 2μ). According to Sheskin (2000), the directional H_a does not require as large a difference in order to reject the H_o, as compared to the non-directional H_a. So, the directional H_a was used in this assessment.

Results of the assessment of differences between the datasets are shown in Table 3.3. Since the aim of the analysis was to determine whether the sampled reservoirs could be used to approximate the distribution of Nebraska reservoirs, it was important that the risk of making a Type-I error in wrongly rejecting the null hypothesis (Ho) was maintained at the barest minimum. Of the possible confidence levels (95.0%, and 99.0%) the 99.0% confidence level offers the least opportunity for rejecting the H_o when there was no significant difference between the two datasets. Table 3.3a shows results of comparison between Groups 1 and 2, where a directional H_a ($p = 0.01$) was used to test the significance of the differences. Based on the Wilcoxon (two-sided Pr > |z|) and Kruskal-Wallis (Pr > Chi-Square) test there was not be enough difference to reject the Ho. Hence the observed differences between Groups 1 and 2 may be due to chance. Also, results of the comparison between Groups 2 and 3 are shown Table 3.3b. The difference between the two samples was significant enough, based on Wilcoxon (two-sided Pr > |z|) and Kruskal-Wallis (Pr > Chi-Square) test, to reject the H_o.

For this reason, it was anticipated that surface area distribution of the adjusted sampled reservoir dataset was different than the surface area distribution of all Nebraska reservoirs that were at least 4 hectares in size. When other factors such as density of reservoir distribution, climate divisions, and ecological regions were considered, it was shown that the adjusted sampled reservoir datasets were well distributed across Nebraska and hence could be used to in this study (Figures 3.12a and 3.12b).

3.5. Derivation of watershed characteristics

Factors that affect reservoir water quality are usually interrelated and complex. A key premise of the watershed-based approach to developing a classification system for Nebraska reservoirs is that the system must be designed to assess the potential reservoir conditions. According to Warren (1979), such a system should be based on environmental variables that underlie, determine, and explain the patterns of change in physical, chemical or biological water quality performances over seasonal or annual cycles.

Available geospatial datasets for these environmental characteristics were extracted for each watershed boundary (see Figure 3.9 and Table 3.1). The datasets included 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 (e.g., precipitation, temperature and humidity).

Watershed area was computed from the watershed boundary data while slope and relief data were derived from 30-meter digital elevation models (DEM) obtained from the USGS EROS Data Center (EDC) in Sioux Falls, South Dakota. Soil erodibility, soil infiltration rate, soil organic matter, soil reaction (pH), soil cation exchange capacity and soil carbonate data were derived from the USDA/NRCS State Soil Geographic Database (STATSGO) (Soil Survey Staff, 1993; Bliss, 1995).

Climate data (e.g., precipitation, temperature and humidity) were obtained from climatological summaries for the conterminous United States web site (www.daymet.org). Daymet is a model designed to interpolate and extrapolate from ground-based meteorological stations, an 18-year daily dataset (1980 – 1997) of temperature, precipitation, humidity and radiation, over large regions at 1 km resolution (Thornton *et al.,* 1997). The climate data for Nebraska were extracted from a much larger database of daily weather parameters based on the 1-kilometer grids for the entire conterminous United States (Thornton *et al.,* 1997). This was necessary to ensure that results of the watershed based reservoir classification could be applicable to other parts of the United States. The climate data were then clipped to the Nebraska state boundary.

A subset of 80 reservoir watershed boundaries, i.e. reservoirs in the GIS database whose watersheds fall within Nebraska, was used to extract the watershed characteristics from the STATSGO, DEM and climate datasets. This was necessary because the STATSGO data are tiled by states. Attempting to extract these data for the states that surround Nebraska (Colorado, Kansas, South Dakota and Wyoming) was beyond the time-frame available for this study.

All data were rasterized and, when required, resampled to the 30m resolution of the DEM data. The watershed boundary coverage was then used to clip the raster layers for each watershed characteristic (e.g., soil erodibility). Next, summary or "zonal" statistics for each watershed characteristic (e.g., maximum, minimum, and mean erodibility values) were generated for the reservoir watershed boundaries using ArcMap® GIS. The above process was repeated to derive summary statistics for climate variables such as total precipitation, precipitation intensity and maximum temperature. All the summary statistics were then appended to the watershed boundary dataset and the resultant information was converted into spreadsheets for further statistical analyses. An examination of the spreadsheet data indicated that two reservoirs (Skyview and Box Butte) had no summary data and were excluded from any further consideration; so only 78 reservoirs were used in subsequent analyses.

3.6. Preliminary analyses of reservoir watershed characteristics datasets

It is important to understand the patterns of variation in the geospatial dataset that were employed in the watershed based reservoir classification process. Analyses of the spatial patterns of variation, as well as, sample distribution of each watershed characteristic were done to provide insights into appropriate statistical approach and a perspective for interpreting the reservoir classification results. Watershed characteristics that were examined included watershed size, mean 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 variables, such as precipitation, temperature and humidity (Table 3.1). For example, a GIS map of nine categories of soil infiltration rate shows that the highest soil infiltration rates occur in the Sand Hills area of Nebraska while most of the state has moderate to low infiltration rates (Figure 3.13).

Histograms were also used to explore patterns of distributions in the datasets. The summary statistics of watershed characteristics data was used (in SAS® software) to generate histograms or bar charts for each dataset. Most of the dataset distributions were skewed or multimodal, as evident in the sample distribution of soil infiltration rate and watershed relief (Figure 3.14).

Subsequently, a Spearman's ranked correlation was also performed for 78 sampled reservoirs to identify any associations and possible redundancies in the watershed data. Results of the correlations analysis showed that soil permeability, for example was highly correlated to soil infiltration rate, while watershed relief (difference between maximum and minimum elevation) was correlated with mean watershed slope (Table 3.5a). The correlation analyses of climate data showed that the growing degree days (base of 10 °C) was highly correlated to all other climate data (precipitation intensity, total precipitation, maximum and minimum temperature, and humidity) ((Table 3.5b). Humidity was also highly correlated to minimum temperature ($r = 0.99$, $p < 0.001$). These patterns show that the climate variables are highly interdependent.

Since we don't know which of these variables has the most significant impact on lake water quality, all the variables that were used in the preliminary analysis could be retained in any further analyses in order to explore their relative impacts on the reservoir classification process, as described in Chapters 4 and 5. This is due to the complexity of possible interactions among the variables that could affect lake water quality.

3.7. Summary

Results summarized in this chapter include the development of an up-to-date and comprehensive GIS map of Nebraska lakes, a vital step in identifying Nebraska reservoirs, delineating watershed boundaries, and extracting watershed characteristics data. The watershed boundaries were delineated from EDNA datasets, which ensures that the results of the watershed-based reservoir classification could be integrated with other geospatial datasets in the state and across the conterminous United States. Comparisons of DEM-derived watershed boundaries with manually digitized DNR watershed boundaries showed less than 10 percent deviation, based on watershed parameters such as drainage area and drainage density. Despite the fact that the watershed boundaries have not been field-checked and standardized, the extent of percentage deviation reflected sufficient conformity with the terrain.

Another important feature of the geospatial database development process was determination of the ratio of sampled reservoirs to Nebraska reservoirs, as well as a test of statistical significance of any difference between the two datasets. This was identified as key information needed to provide a context for the development lake nutrient criteria guidelines (EPA, 2001). The comparisons of the surface area of sampled reservoirs with all Nebraska indicated that there was a difference between the sampled reservoirs and Nebraska reservoirs larger than 4 hectares. When other factors such as density of reservoir distribution, climate divisions and ecological regions were considered, it was shown that adjusted sampled reservoir datasets were spatially well distributed with respect to Nebraska reservoirs that were at least 4 hectares. Therefore, use of the adjusted sampled reservoir data to characterize the Nebraska reservoirs should be viewed in the context of the data employed.

Zonal or area summary statistics for each watershed characteristics were derived for the watersheds characteristics dataset and the resultant information were converted into spreadsheets. Preliminary analyses on these data showed that the watershed characteristics were not normally distributed. Consequently, non-parametric statistical approaches become essential for ensuing reservoir classification analyses described in Chapters 4 and 5.

References cited

Bliss, N.B. 1995. **Processing STATSGO in ARC/INFO**. Hughes STX Corporation, EROS Data Center, Sioux Falls, South Dakota.

EPA (U.S. Environmental Protection Agency). 2001. **Nutrient Criteria Technical Guidance Manual for Lakes and Reservoirs.** Report No. EPA-822-B00-001. Washington, D.C.

ESRI 1992. **Cell-based modeling with grid 6.1.** Envrionmental Systems Research Institute, Inc. Redlands, CA. p 309-327.

ESRI 1997. **Using ARCGRID with ARC/INFO.** Envrionmental Systems Research Institute, Inc. Redlands, CA. p 7.3-7.33.

Garbrecht, J. and L. W. Martz. 2003. *Assessing the performance of automated watershed delineation process from digital elevation models*, pp 17-24 in Lyon, J.G. **GIS for Water Resource and Watershed Management**. 266 pp.

Gesch, D., M. Oimoen, S. Greenlee, C. Nelson, M. Steuck and D. Tyler. 2002. *The National Elevation Dataset.* **Photogrammetric Engineering And Remote Sensing**. 68(1): 5-11.

Holz, J.C. 2002. **Lake And Reservoir Classification In Agriculturally Dominated Ecosystems.** EPA 2002 Aquatic Ecosystem Classification Workshop, Denver, CO, September, 2002, oral presentation, invited.

Kost, J., and Kelly, G. 2001. *Watershed delineation using the National Elevation Dataset and semiautomated techniques*, in Proc., **Twenty-First Annual ESRI International User Conference, San Diego, California, July 9-13, 2001**. Redlands, California, Environmental Systems Research Institute, Inc. (CD-ROM).

NRCS, 2002. **Federal standards for delineation of Hydrologic unit boundaries.** http://www.ftw.nrcs.usda.gov/HUC/HU_standards_v1_030102.doc. Accessed on December 22, 2003

Olivera, F., S. Reed and D. Maidment. 2000. **HEC-PrePro version. 2.0: An ArcView Pre-Processor for HEC's Hydrologic Modeling System.** University of Texas at Austin - Center for Research in Water Resources. Austin, Texas. http://www.ce.utexas.edu/prof/olivera/esri98/p400.htm. Accessed on April 4, 2002.

Omernik, J.M., 2003. *The misuse of hydrologic unit maps for extrapolation, reporting and ecosystem management.* **Journal of the American Water Resources Association**. 39(3):563–573.

Omernik, J.M. and R.G. Bailey. 1997. *Distinguishing between watersheds and ecoregions*. **Journal Of The American Water Resources Association.** 33(5):935–949.

Pfafstetter, O. 1989. *Classification of hydrographic basins: coding methodology.* **Unpublished Manuscript. DNOS**, August 18, 1989, Rio de Janeiro. Translated by J.P. Verdin, U.S. Bureau of Reclamation, Brasilia, Brazil, September 5, 1991.

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

Satterlund, D.R. and P.W. Adams. 1992. **Wildland Watershed Management.** 2nd Ed. J. Wiley and Sons, New York, N.Y. 436p.

Seaber, P.R., F. P Kapinos, and G. L. Knapp. 1987. **Hydrologic Unit Maps**. Water Supply Paper 2294. United States Geological Survey, Denver, Colorado.

Sheskin, D.J. 2000. **Handbook of Parametric and Non-Parametric Statistical Procedures.** 2nd Ed. Chapman and Hall, New York, N.Y. 982 p.

Soil Survey Division Staff. 1993. **Soil Survey Manual**. Soil Conservation Service, U.S. Department of Agriculture. Handbook 18. Washington, DC.

Thornton, P.E, Running, S.W. and White, M.A., 1997. *Generating surfaces of daily meteorological variables over large regions of complex terrain.* **Journal of Hydrology.** 190: 214-251.

Tukey, J.W. 1977. **Exploratory data analysis**. Addison-Wesley, Reading, MA. 688p.

U.S. Geological Survey. 1982. *A U.S. Geological Survey data standard - codes for the identification of hydrologic units in the United States and the Caribbean outlying areas.* **U.S. Geological Survey**, Circular 878–A. 115 p.

U.S. Geological Survey. 2001. **EDNA Stage 2 Tool Overview.** http://edcnts12.cr.usgs.gov/ned-h/stage2/stage2.htm. Accessed on December 22, 2003.

Verdin, K.L. and J.P. Verdin. 1999. *A topological system for delineation and codification of the Earth's river basins.* **Journal of Hydrology**. 218:1 – 12.

Verdin, K. 2000. *Development of the National Elevation Dataset-Hydrologic Derivatives (NED-H)*, in Proc., **Twentieth Annual ESRI International User Conference, San Diego, California, July 10-14, 2000**. Redlands, California, Environmental Systems Research Institute, Inc. (CD-ROM).

Warren, C. E. 1979. **Toward classification and rationale for watershed management and stream protection.** EPA - 600 / 3-79-059. United States Environmental Protection Agency. Corvallis, Oregon. 143p.

	DEM derived watersheds *	**DNR digitized watersheds ***	**Percent deviation (%)**
Total drainage area (ha)	52895	53629	1.79
Mean drainage length (m)	2565	2597	3.55
Total drainage length (m)	405595	404085	6.99
Mean drainage density (m^{-1})	1.5926×10^{-4}	1.5921×10^{-4}	0.996
Drainage density (m^{-1})	9.79×10^{-2}	9.84×10^{-2}	4.123
Maximum catchment slope (%)	21.39	22.37	5.59
Mean catchment slope (%)	3.41	3.42	1.84

Table 3.1. Comparison of DEM derived watersheds to DNR digitized watershed boundaries for selected Nebraska reservoirs.

** Data represents average values from 18 selected watersheds of small, medium and large reservoirs.*

	Group 1	Group 2	Group 3
Number of observations	8	80	954
Mean Area (ha)	2457	249.35	10.21
1st quartile (ha)	92.14	14.88	4.44
2nd quartile (ha)	347.66	36.68	5.91
3rd quartile (ha)	3470	112.26	9.47
Standard deviation	4162	721.07	20.81
Variance	17320147	519947	433
Range	11814	5812	389
Coefficient of % variation	169.37	289.18	204
Skewness	2.06	6.26	12.82
Kurtosis	4.12	46.07	210

Table.3.2. Descriptive statistics of Nebraska reservoir datasets. Groups 1 and 2 consist of sampled reservoirs whose watershed boundaries fall outside and within Nebraska, respectively; and, Group 3 consists of all reservoirs whose watershed boundaries fall within Nebraska.

a. Group 1 vs. 2

Group	**N**	**Sum of Scores**	**Expected Under H_0**	**Std. Dev. Under H_0**	**Mean Score**
1	8	532	356	68.9	66.50
2	80	3384	3560	68.9	42.30

Wilcoxon Two-Sample Test			**Kruskal-Wallis Test**	
S**	z	Pr > |Z|	Chi-Square	Pr > Chi-Square
532	2.55	0.0109	6.5258	0.0106

b. Group 2 vs. 3

Group	**N**	**Sum of Scores**	**Expected Under H_0**	**Std. Dev. Under H_0**	**Mean Score**
2	80	71883.5	41400	2565.66	898.54
3	954	463211.5	493695	2565.66	485.55

Wilcoxon Two-Sample Test			**Kruskal-Wallis Test**	
S**	z	Pr > |Z|	Chi-Square	Pr > Chi-Square
71883.5	11.88	<.0001	141.17	<.0001

Table 3.3. Comparison of difference between sampled reservoirs and Nebraska reservoirs: (a) tests if the distribution of surface area for sampled reservoirs whose watersheds within Nebraska, is different from that of boundary reservoirs watersheds which fall outside Nebraska; and (b) tests if the distribution of surface area for sampled reservoirs whose watersheds within Nebraska is different from that of all reservoirs whose watersheds fall within Nebraska and are at least 4 ha in size.

*S** = Wilcoxon Signed Rank statistic.*

Dataset	**Data Source***	**Scale**
Lake boundaries	NRCS - (*SSURGO*) USGS - (*NHD& NLCD*), USCB - (*TIGER*)	1:24,000 (baseline)
Lake surface area	Calculated	1:24,000
Lake elevation	USGS/EDC - (*EDNA*)	1:100,000
Stream network	USGS/EDC - (*EDNA*)	1:100,000
Lake landscape position (lake order)	USGS/EDC - (*EDNA*)	1:100,000
Hydrologic unit coverage (HUC)	USGS/EDC - (*EDNA*)	1:100,000
Pfafstetter sub-catchments	USGS/EDC - (*EDNA*)	1:100,000
Flow direction data	USGS/EDC - (*EDNA*)	1:100,000
Shaded relief data	USGS/EDC - (*EDNA*)	1:100,000
Digital elevation model (DEM)	USGS/EDC - (*EDNA*)	1:100,000
Watershed area	Calculated	1:100,000
Watershed slope	USGS/EDC - (*EDNA*)	1:100,000
Watershed relief	USGS/EDC - (*EDNA*)	1:100,000
Soil erodibility	USDA /NRCS - (*STATSGO*)	1:250,000
Soil permeability	USDA /NRCS - (*STATSGO*)	1:250,000
Soil infiltration rate	USDA /NRCS - (*STATSGO*)	1:250,000
Soil organic matter	USDA /NRCS - (*STATSGO*)	1:250,000
Soil carbonates	USDA /NRCS - (*STATSGO*)	1:250,000
Soil salinity	USDA /NRCS - (*STATSGO*)	1:250,000
Soil reaction (pH)	USDA /NRCS - (*STATSGO*)	1:250,000
Soil cation exchange capacity	USDA /NRCS - (*STATSGO*)	1:250,000
Soil available water holding capacity	USDA /NRCS - (*STATSGO*)	1:250,000
Ground water regions	CSD	1:24,000
Ecoregions (Omernik's Levels 3 and 4)	USEPA	1:250,000
Land use and land cover	USGS/EDC - (*NLCD*)	1:24,000
Natural vegetation	CSD	1:250,000
Potential natural vegetation (Kuchler)	USEPA	1:250,000
Nebraska county boundaries	USGS	1:24,000
Geology	CSD	1:24,000
Total precipitation	NTSG - (*DAYMET*)	1 km resolution
Precipitation frequency	NTSG - (*DAYMET*)	1 km resolution
Precipitation intensity	NTSG - (*DAYMET*)	1 km resolution
Air temperature	NTSG - (*DAYMET*)	1 km resolution
Solar radiation	NTSG - (*DAYMET*)	1 km resolution
Humidity	NTSG - (*DAYMET*)	1 km resolution
Growing degree days	NTSG - (*DAYMET*)	1 km resolution
Frost free days	NTSG - (*DAYMET*)	1 km resolution

Table 3.4. Geospatial datasets available in Nebraska lake classification database.

*Data Sources:
USGS – United States Geological Survey
USDA – United States Department of Agriculture
NRCS – Natural Resource Conservation Service
USCB – United States Census Bureau
USEPA – United States Environmental Protection Agency
NTSG – Numerical Terradynamic Simulation Group, University of Montana
DAYMET – Daily Surface and Climatological Summaries (www.daymet.org)
NHD – National Hydrography Dataset
STATSGO – State Soils Geographic Database
SSURGO– Soil Survey Geographic Database
EDNA – Elevation Derivatives for National Application
TIGER – Topologically Integrated Geographic Encoding and Referencing Database
CSD – Conservation and Survey Division, School of Natural Resources, University Of Nebraska – Lincoln

	Watershed Area (WA)	Lake Area (LA)	RATIO (LA:WA)	CaC03	CEC	Clay	Erodibility	OM	Permeability	pH	Infiltration	Salinity	Slope	Relief	Elevation	Drainage total	Drainage density
Watershed Area (WA)	1.000																
Lake Area (LA)	0.623	1.000															
	<.0001																
RATIO (LA:WA)	-0.343	0.383	1.000														
	0.002	*0.001*															
CaC03	0.068	-0.014	-0.042	1.000													
	0.554	*0.904*	*0.713*														
CEC	-0.358	-0.243	0.083	-0.048	1.000												
	0.001	*0.032*	*0.468*	*0.679*													
Clay	-0.560	-0.292	0.290	-0.149	0.470	1.000											
	<.0001	*0.010*	*0.010*	*0.194*	*<.0001*												
Erodibility	-0.469	-0.441	0.054	0.168	0.186	0.362	1.000										
	<.0001	*<.0001*	*0.637*	*0.142*	*0.104*	*0.001*											
OM	-0.200	-0.098	0.066	-0.540	0.513	0.483	0.093	1.000									
	0.080	*0.392*	*0.567*	*<.0001*	*<.0001*	*<.0001*	*0.420*										
Permeability	0.527	0.362	-0.172	0.277	-0.415	-0.879	-0.306	-0.365	1.000								
	<.0001	*0.001*	*0.133*	*0.014*	*0.000*	*<.0001*	*0.006*	*0.001*									
pH	0.079	0.157	0.231	0.690	-0.289	-0.105	0.224	-0.486	0.242	1.000							
	0.490	*0.170*	*0.042*	*<.0001*	*0.010*	*0.361*	*0.049*	*<.0001*	*0.033*								
Infiltration	0.444	0.316	-0.170	0.071	-0.376	-0.638	-0.002	-0.004	0.791	0.217	1.000						
	<.0001	*0.005*	*0.136*	*0.537*	*0.001*	*<.0001*	*0.983*	*0.973*	*<.0001*	*0.057*							
Salinity	-0.061	-0.097	0.046	0.713	0.103	0.093	0.113	-0.236	0.023	0.626	-0.099	1.000					
	0.599	*0.398*	*0.687*	*<.0001*	*0.371*	*0.418*	*0.326*	*0.038*	*0.844*	*<.0001*	*0.386*						
Slope	0.162	0.213	0.073	0.055	-0.152	0.100	0.111	0.165	0.058	0.415	0.346	0.03064	1.000				
	0.157	*0.061*	*0.524*	*0.632*	*0.183*	*0.385*	*0.333*	*0.150*	*0.616*	*0.000*	*0.002*	*0.79*					
Relief	0.787	0.506	-0.269	0.291	-0.378	-0.378	-0.323	-0.239	0.481	0.414	0.456	0.21292	0.503	1.000			
	<.0001	*<.0001*	*0.017*	*0.010*	*0.001*	*0.001*	*0.004*	*0.035*	*<.0001*	*0.000*	*<.0001*	*0.0613*	*<.0001*				
Elevation	0.135	0.034	-0.059	0.569	-0.329	-0.429	0.012	-0.785	0.341	0.603	0.078	0.39663	-0.088	0.224	1.000		
	0.237	*0.766*	*0.610*	*<.0001*	*0.003*	*<.0001*	*0.918*	*<.0001*	*0.002*	*<.0001*	*0.500*	*0.0003*	*0.443*	*0.049*			
Drainage total	0.433	0.217	-0.197	-0.144	-0.326	-0.379	-0.305	-0.207	0.320	0.009	0.248	-0.24558	0.175	0.349	0.071	1.000	
	<.0001	*0.057*	*0.083*	*0.209*	*0.004*	*0.001*	*0.007*	*0.069*	*0.004*	*0.940*	*0.029*	*0.0302*	*0.125*	*0.002*	*0.534*		
Drainage density	-0.281	-0.137	0.220	-0.339	0.082	-0.072	0.265	0.274	0.026	-0.061	0.176	-0.22581	-0.036	-0.361	-0.085	0.007	1.000
	0.013	*0.233*	*0.053*	*0.002*	*0.478*	*0.532*	*0.019*	*0.015*	*0.824*	*0.593*	*0.124*	*0.0468*	*0.754*	*0.001*	*0.460*	*0.949*	

Table 3.5a. Spearman ranked correlations of watershed and reservoir data (p-values are in italics).

	Maximum temperature	Minimum temperature	Average temperature	Precipitation intensity	Total precipitation	Humidity	Growing Degree Days
Maximum temperature	1						
Minimum temperature	**0.53742** *<.0001*	1					
Average temperature	**0.46942** *<.0001*	**0.68577** *<.0001*	1				
Precipitation intensity	**0.06711** *0.5594*	**0.71704** *<.0001*	**0.33531** *0.003*	1			
Total Precipitation	**0.29076** *0.0098*	**0.89345** *<.0001*	**0.5194** *<.0001*	**0.76343** *<.0001*	1		
Humidity	**0.46565** *<.0001*	**0.99115** *<.0001*	**0.65419** *<.0001*	**0.73839** *<.0001*	**0.91932** *<.0001*	1	
Growing Degree Days	**0.65346** *<.0001*	**0.97893** *<.0001*	**0.71035** *<.0001*	**0.6484** *<.0001*	**0.8392** *<.0001*	**0.95688** *<.0001*	1

Table 3.5b. Spearman ranked correlations of climate data (p-values are in italics)

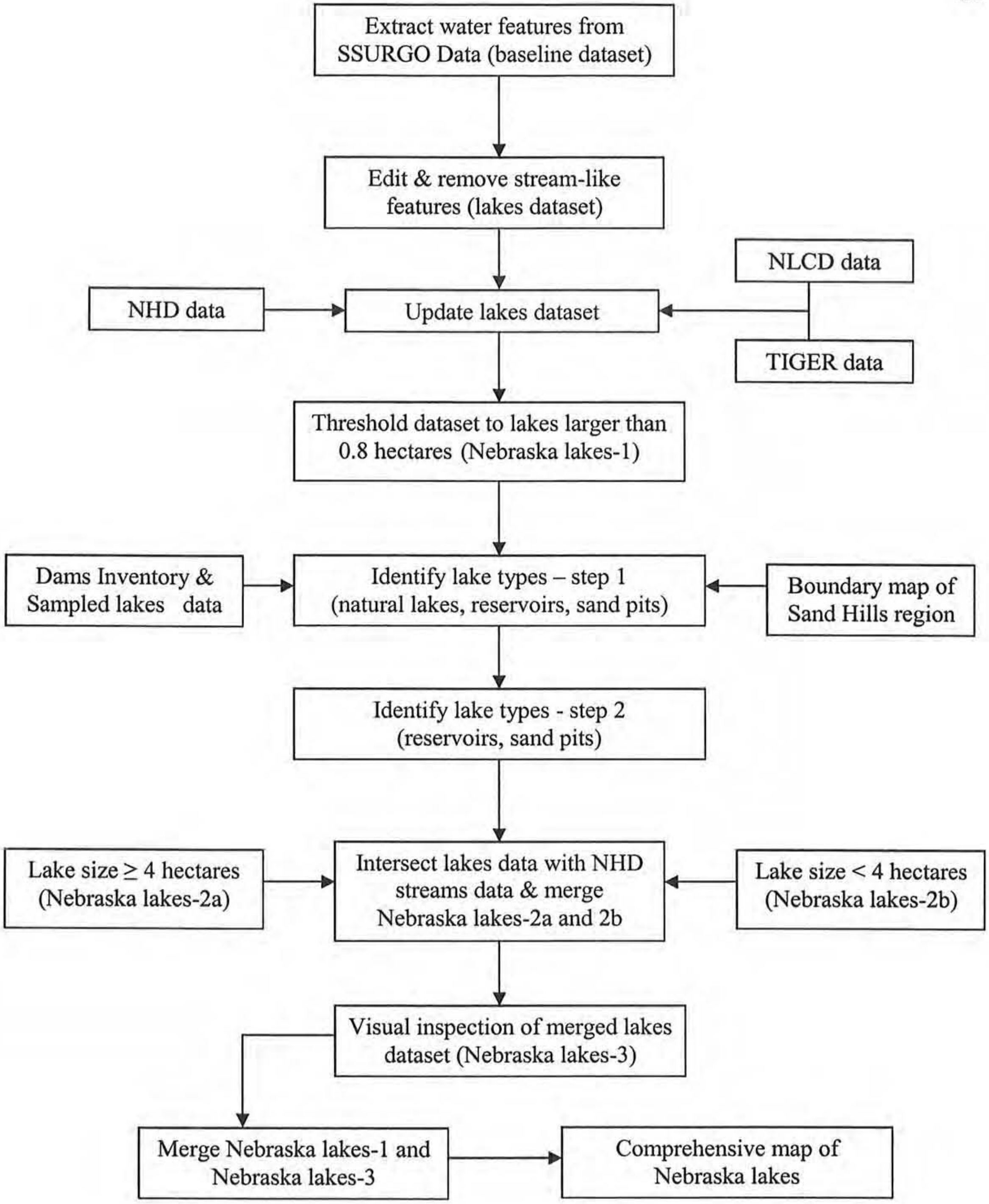

Extract water features from SSURGO Data (baseline dataset)
Edit & remove stream-like features (lakes dataset)
NLCD data
NHD data
Update lakes dataset
TIGER data
Threshold dataset to lakes larger than 0.8 hectares (Nebraska lakes-1)
Dams Inventory & Sampled lakes data
Identify lake types – step 1 (natural lakes, reservoirs, sand pits)
Boundary map of Sand Hills region
Identify lake types - step 2 (reservoirs, sand pits)
Lake size ≥ 4 hectares (Nebraska lakes-2a)
Intersect lakes data with NHD streams data & merge Nebraska lakes-2a and 2b
Lake size < 4 hectares (Nebraska lakes-2b)
Visual inspection of merged lakes dataset (Nebraska lakes-3)
Merge Nebraska lakes-1 and Nebraska lakes-3
Comprehensive map of Nebraska lakes

Figure 3.1. Process for developing a comprehensive map of Nebraska lakes

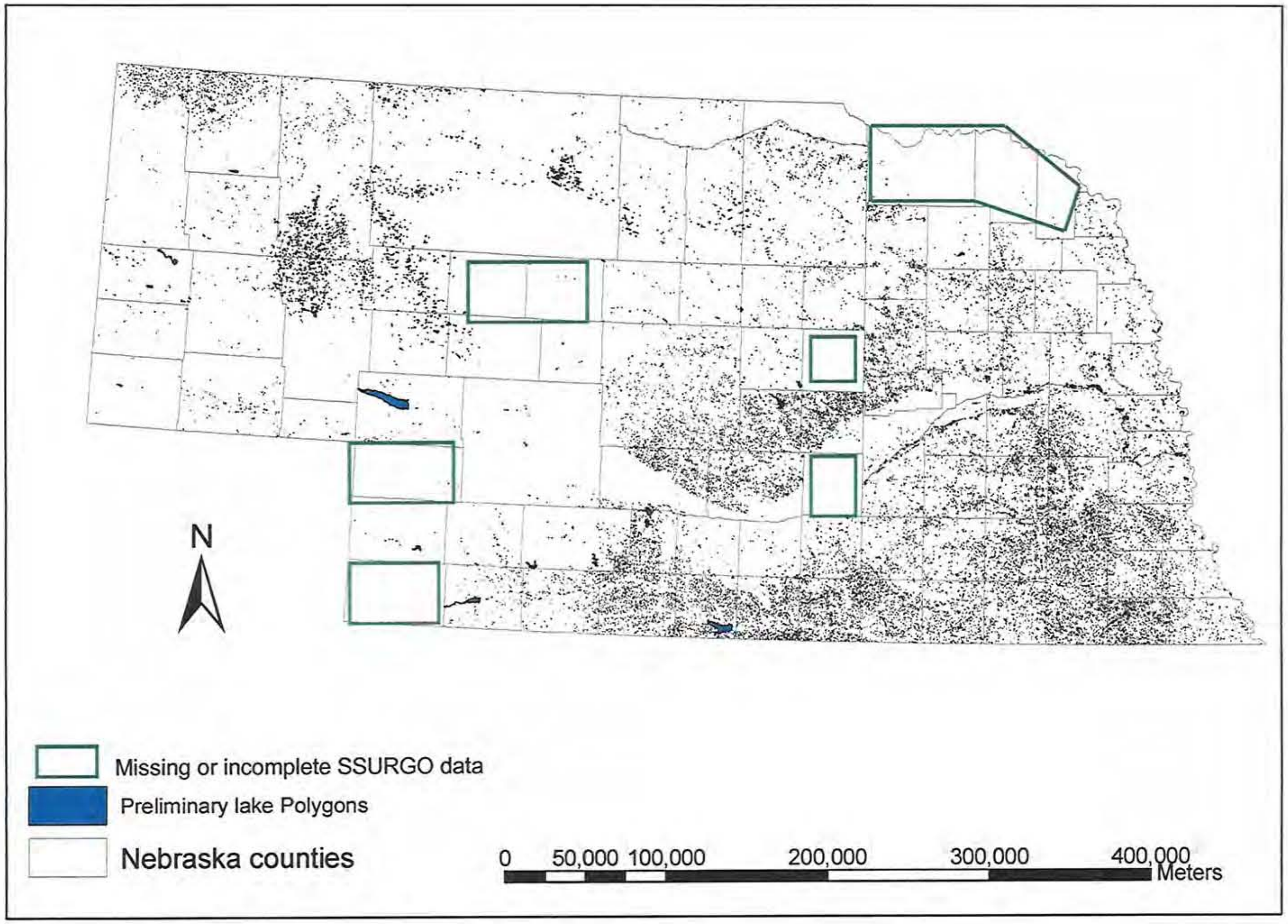


Figure 3.2. Lake features after editing SSSURGO data to remove stream-like features.

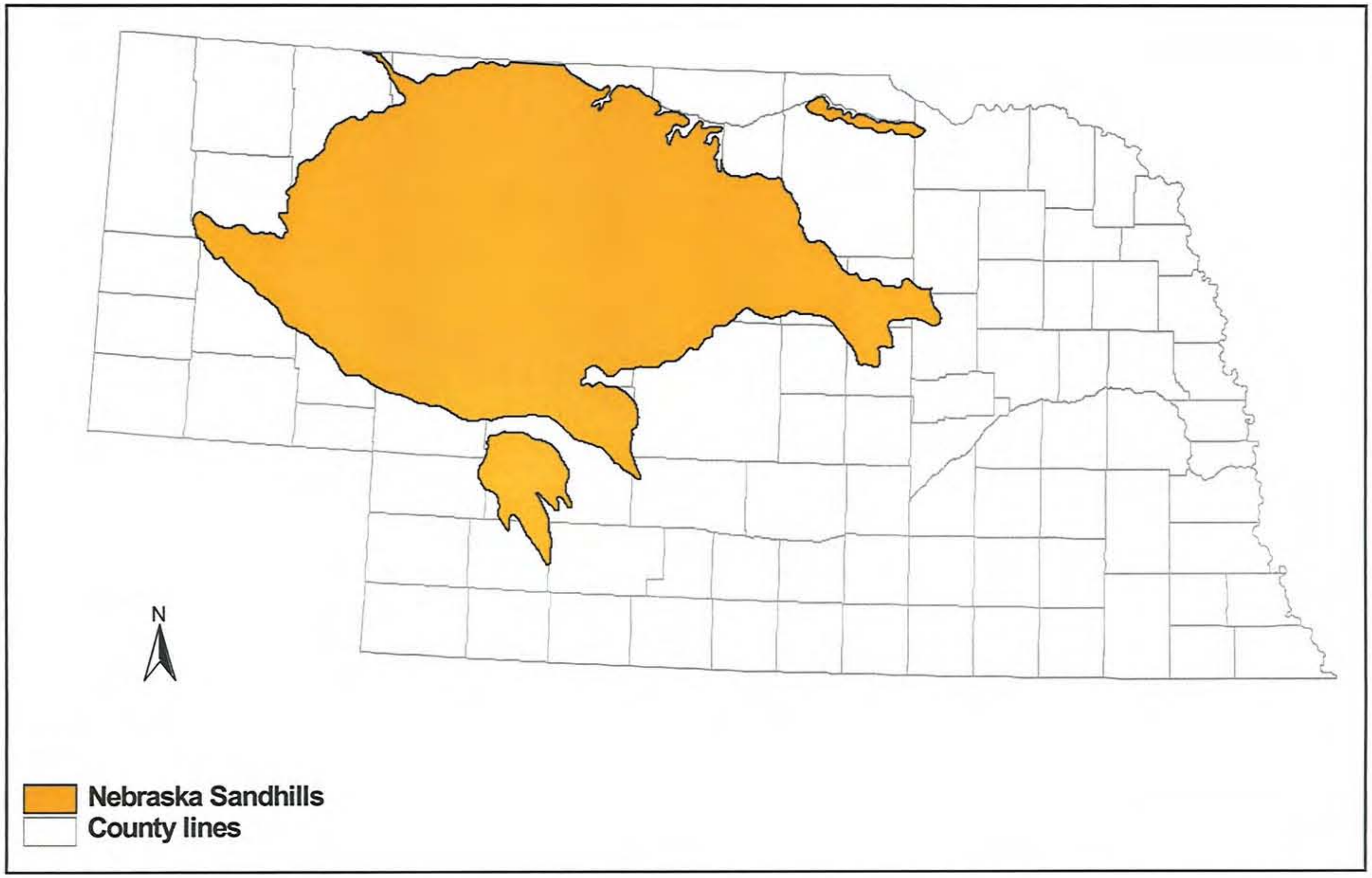


Figure 3.3. Map of the Sand Hills region of Nebraska

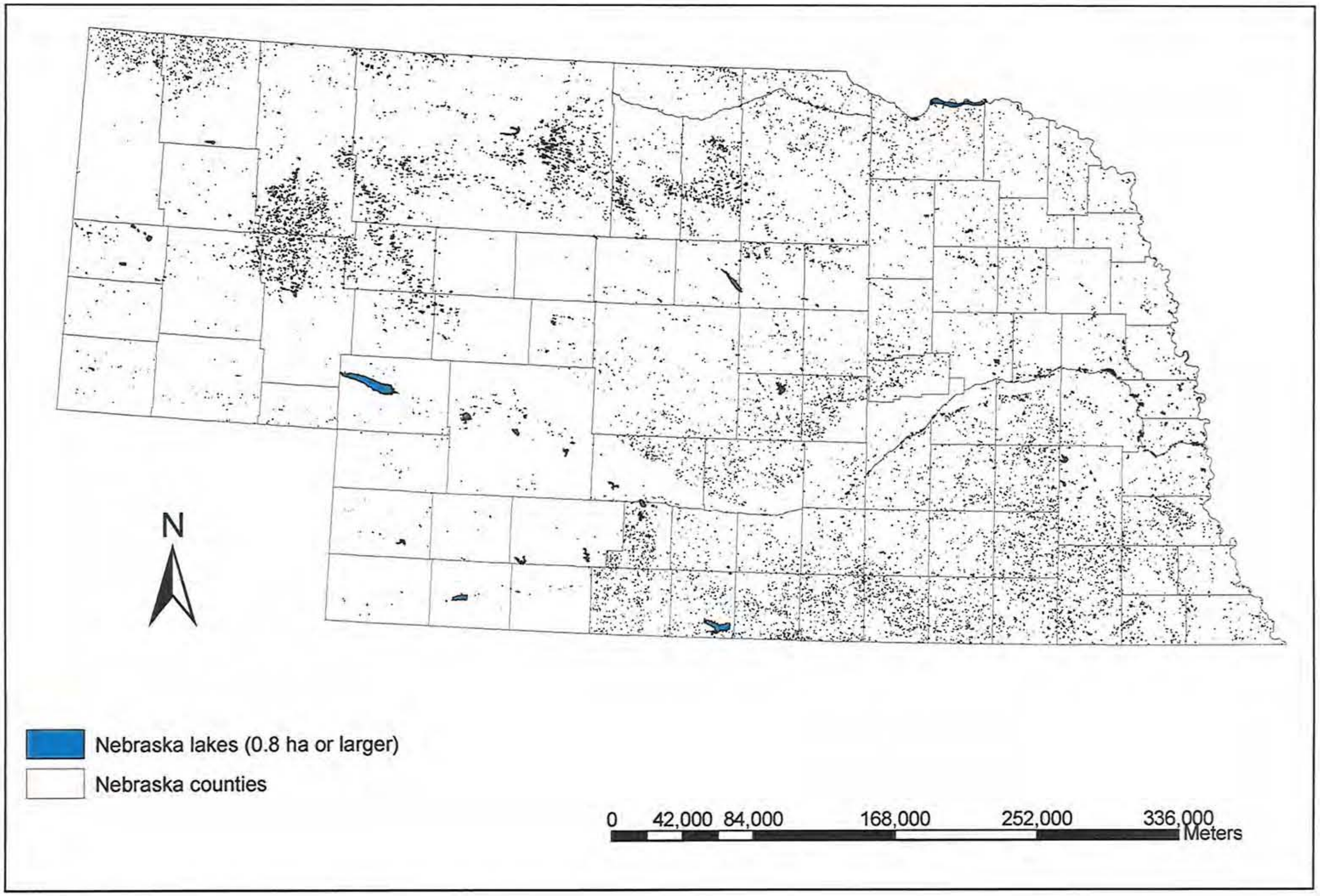


Figure 3.4. Final updated map of Nebraska lakes

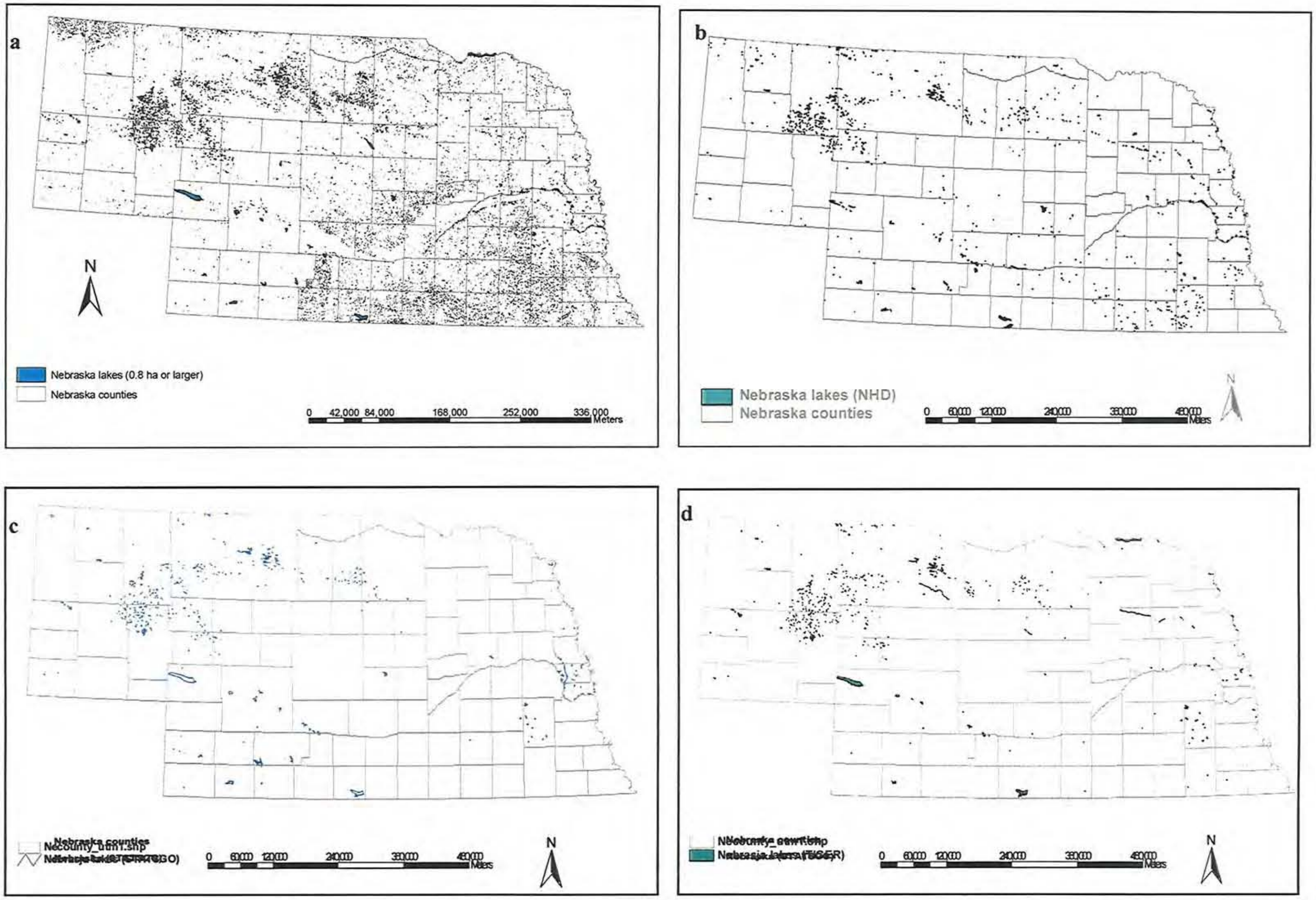


Figure3.5. Comparison of Nebraska lakes dataset: (a). Updated map map of Nebraska lakes; (b). NHD derived lakes; (c) STATSGO derived lakes; and (d) and TIGER derived lakes

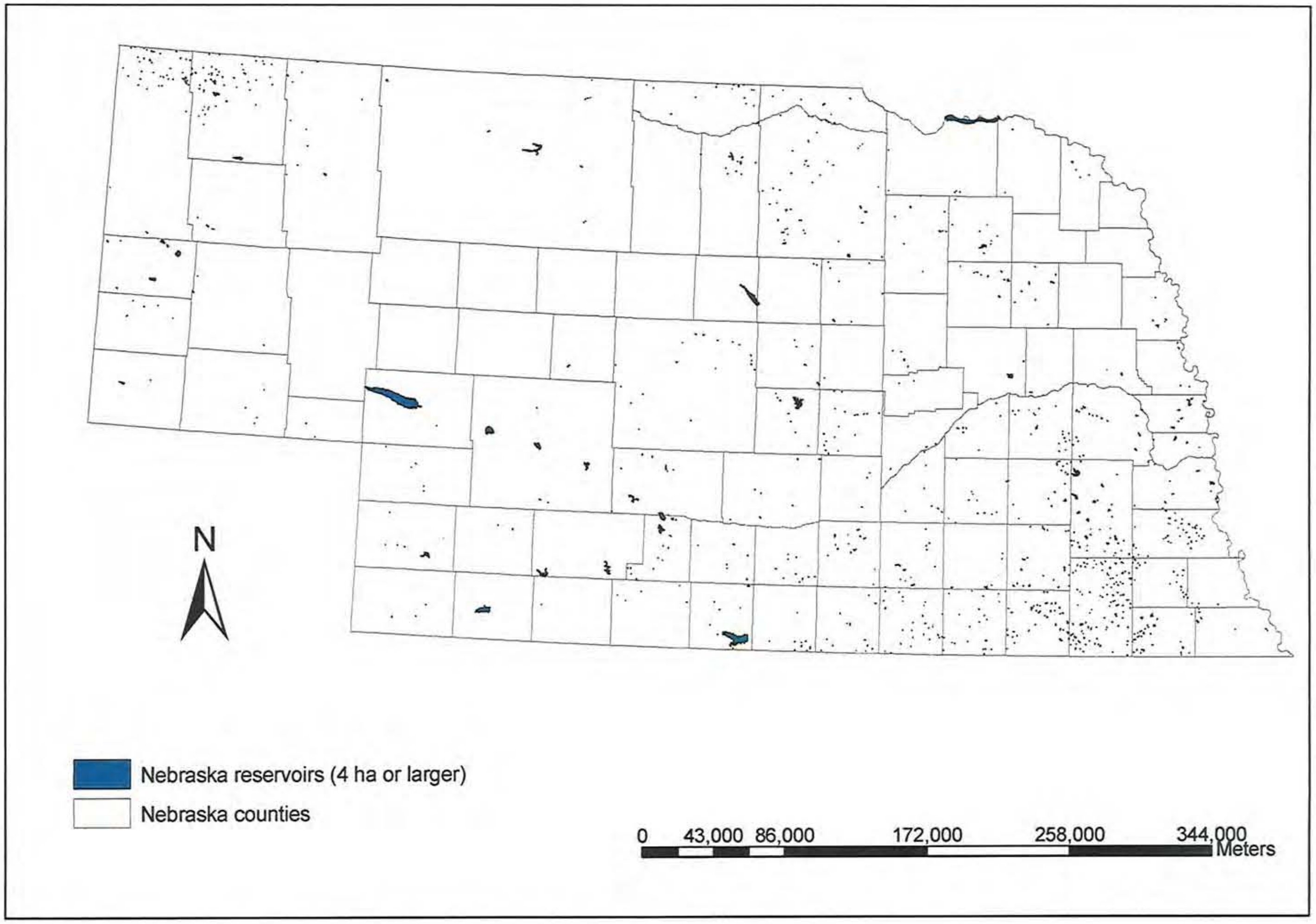


Figure 3.6. Nebraska reservoirs that are 4 hectares or larger

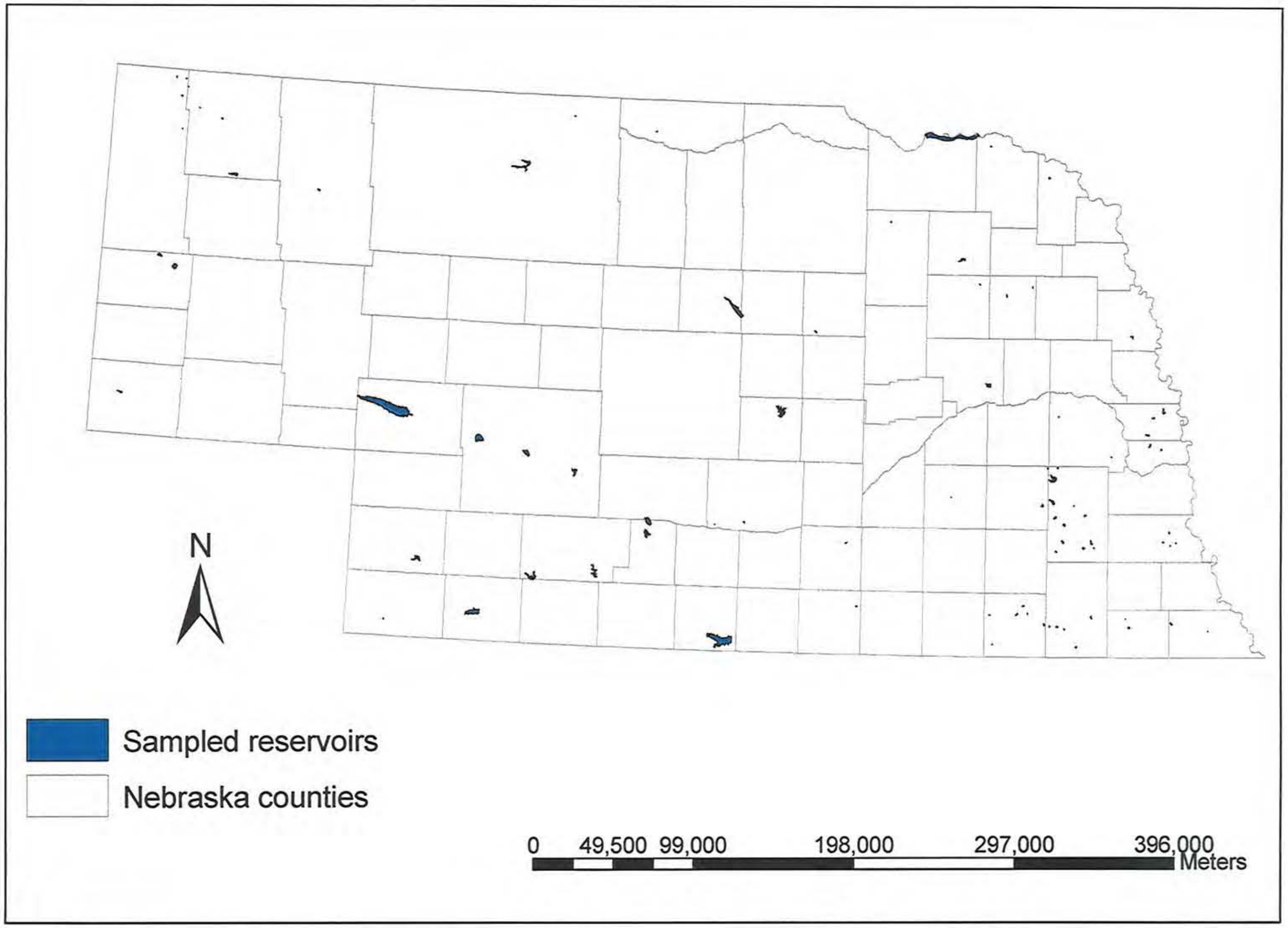
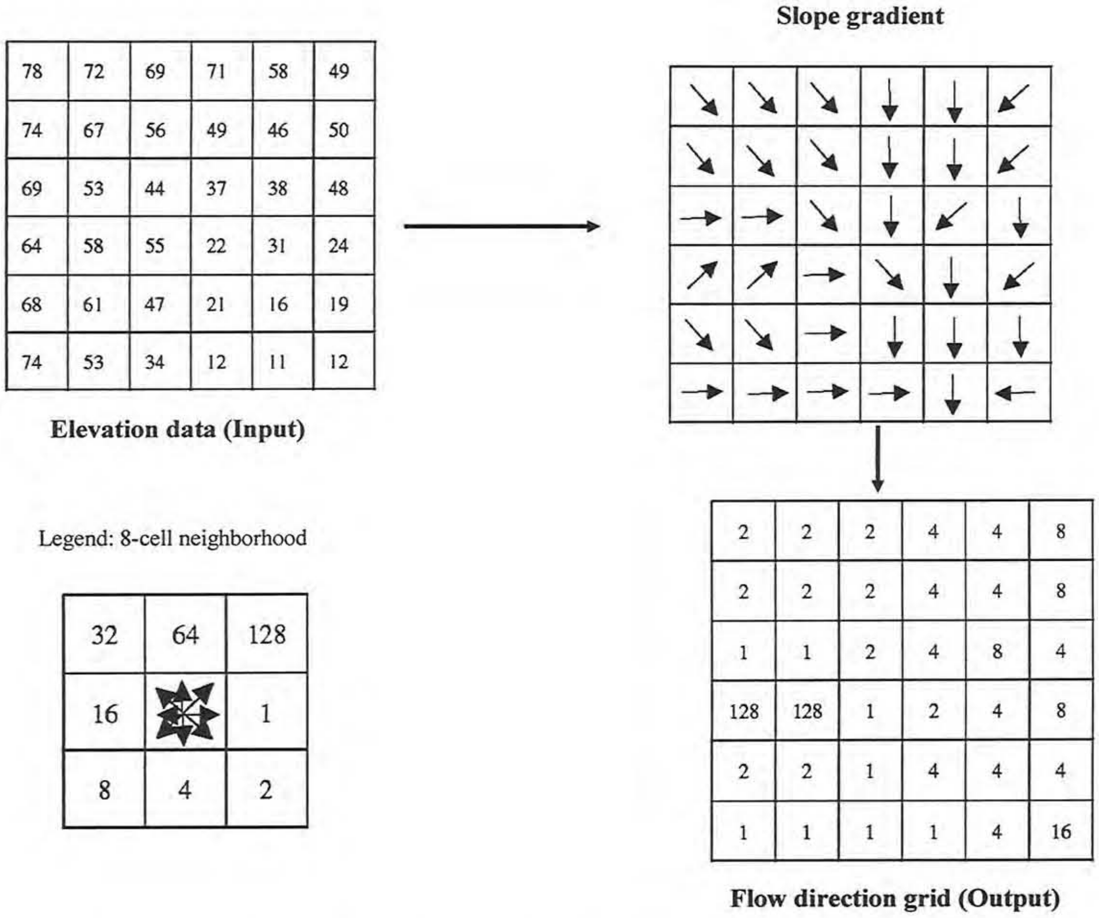


Figure 3.7. Locations of Nebraska reservoirs that were sampled between 1989 and 2001 for quality assessment



78	72	69	71	58	49
74	67	56	49	46	50
69	53	44	37	38	48
64	58	55	22	31	24
68	61	47	21	16	19
74	53	34	12	11	12

Elevation data (Input)

Slope gradient

↘	↘	↘	↓	↓	↙
↘	↘	↘	↓	↓	↙
→	→	↘	↓	↙	↓
↗	↗	→	↘	↓	↙
↘	↘	→	↓	↓	↓
→	→	→	→	↓	←

Legend: 8-cell neighborhood

32	64	128
16		1
8	4	2

2	2	2	4	4	8
2	2	2	4	4	8
1	1	2	4	8	4
128	128	1	2	4	8
2	2	1	4	4	4
1	1	1	1	4	16

Flow direction grid (Output)

Figure 3.8. Generating a flow direction grid based on 8-cell neighborhood (modified from ESRI, 1997)

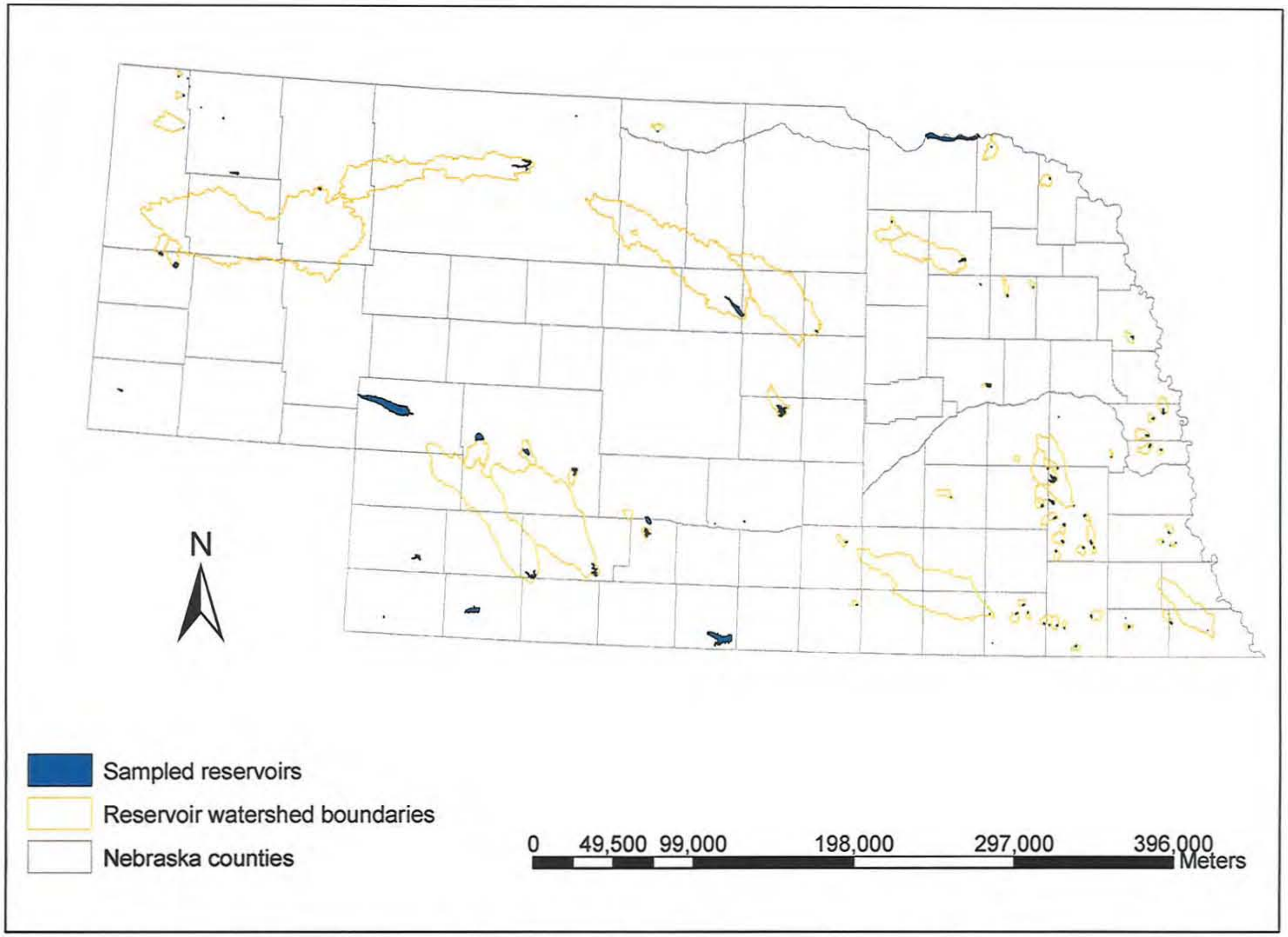


Figure 3.9. Watershed boundaries for 80 Nebraska reservoirs sampled between 1989 and 2003. Watershed boundaries that extend beyond the Nebraska state are not shown.

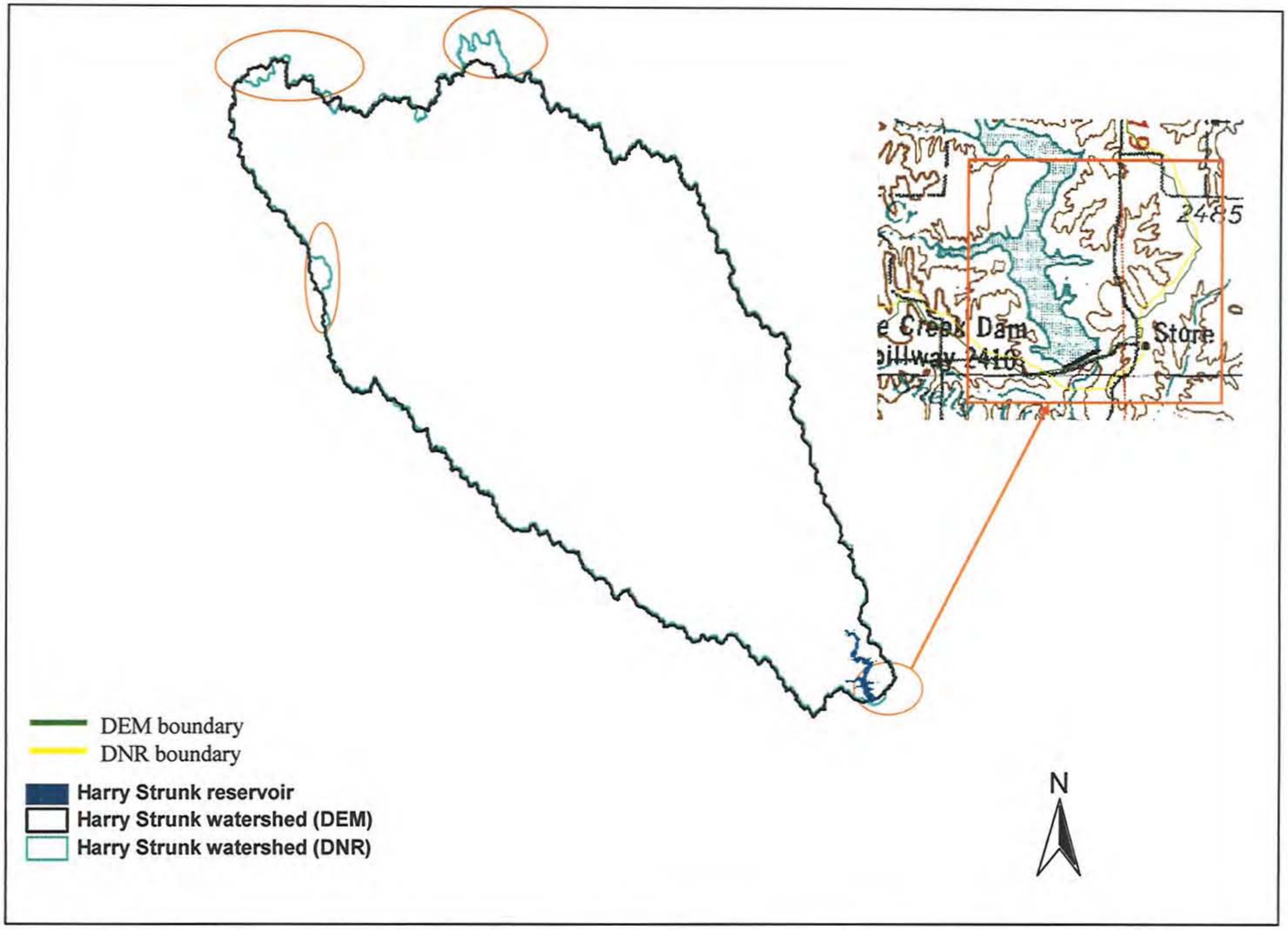


Figure 3.10. Comparison of DEM derived watershed boundary to digitized watershed boundary of Harry Strunk reservoir

Figure 3.11. Box plots: comparison of Nebraska reservoir datasets based on

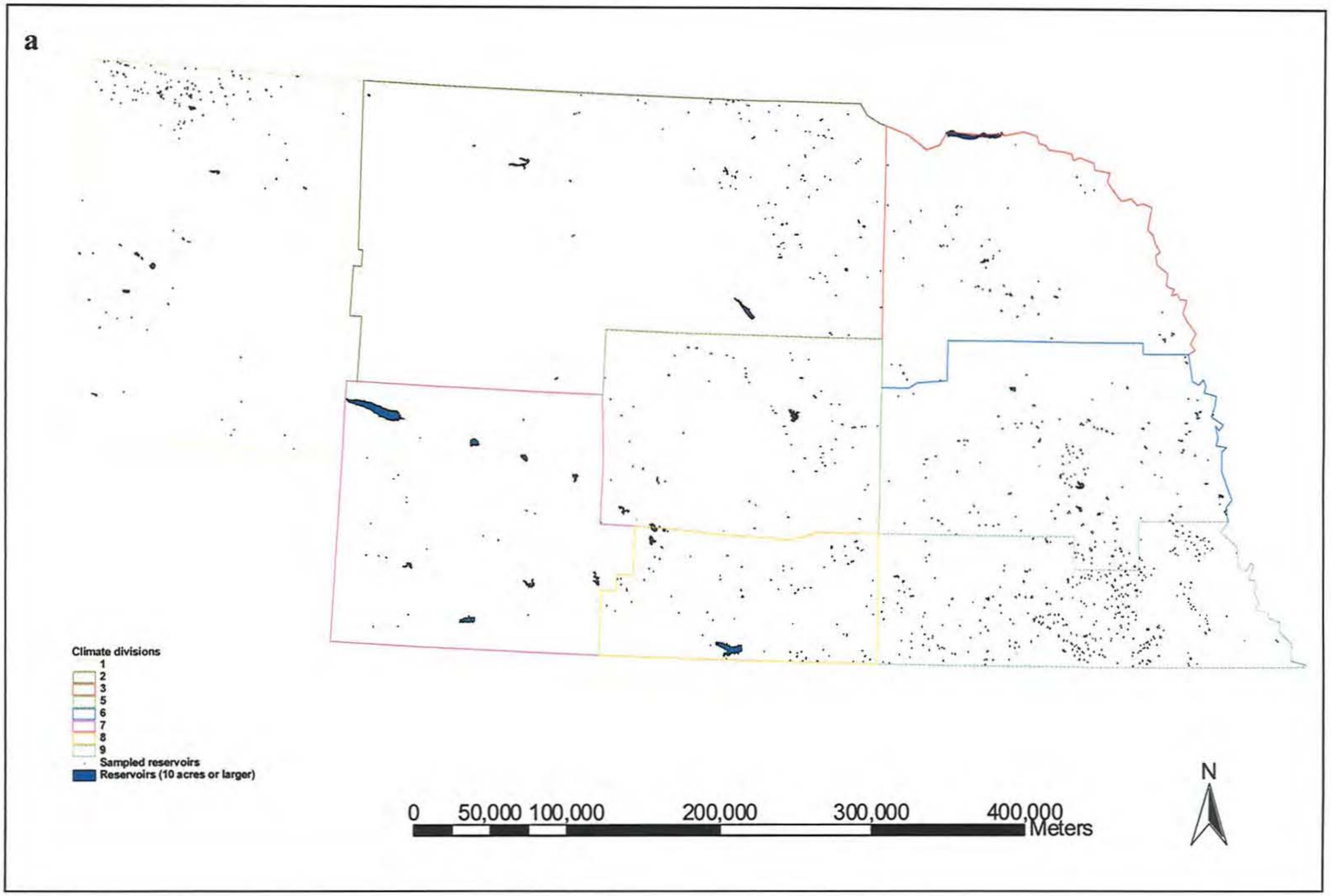
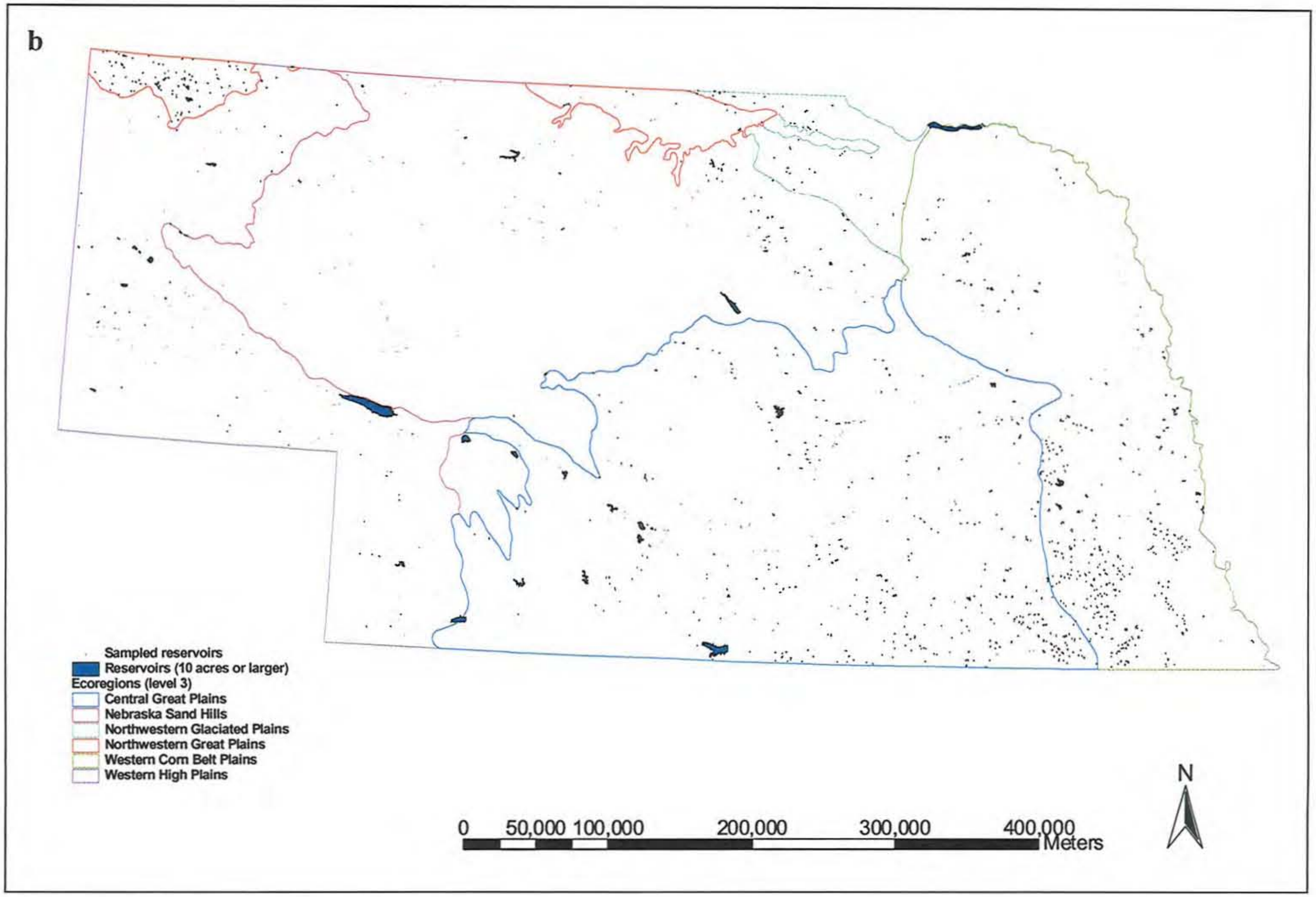


Figure 3.12. (a) Comparison of sampled Nebraska reservoirs with climate regions



3.12. (b) Comparison of sampled Nebraska reservoirs with Omernik Level II Ecoregions of Nebraska.

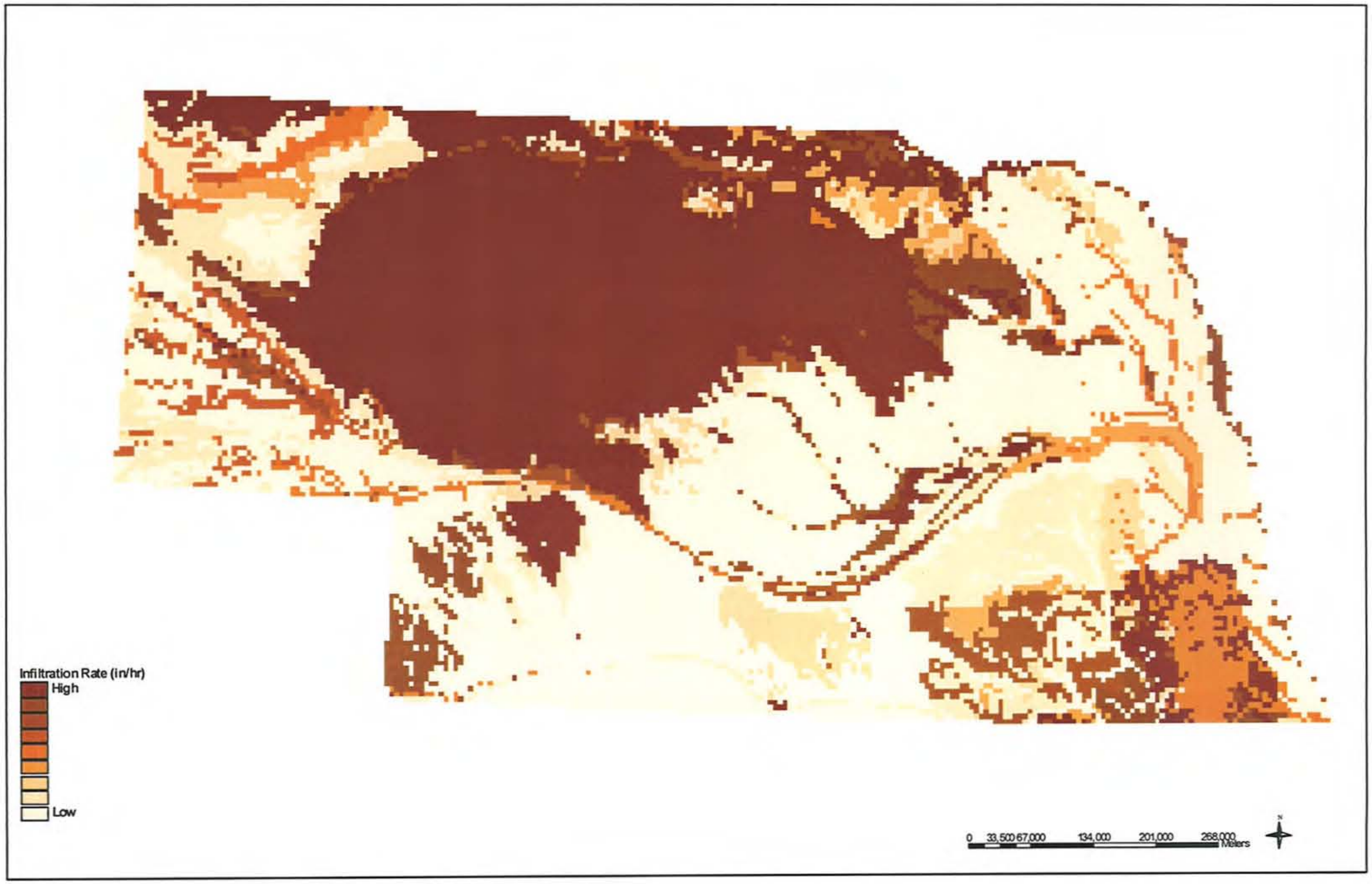


Figure 3.13. Map of spatial variations in soil infiltration rate in Nebraska

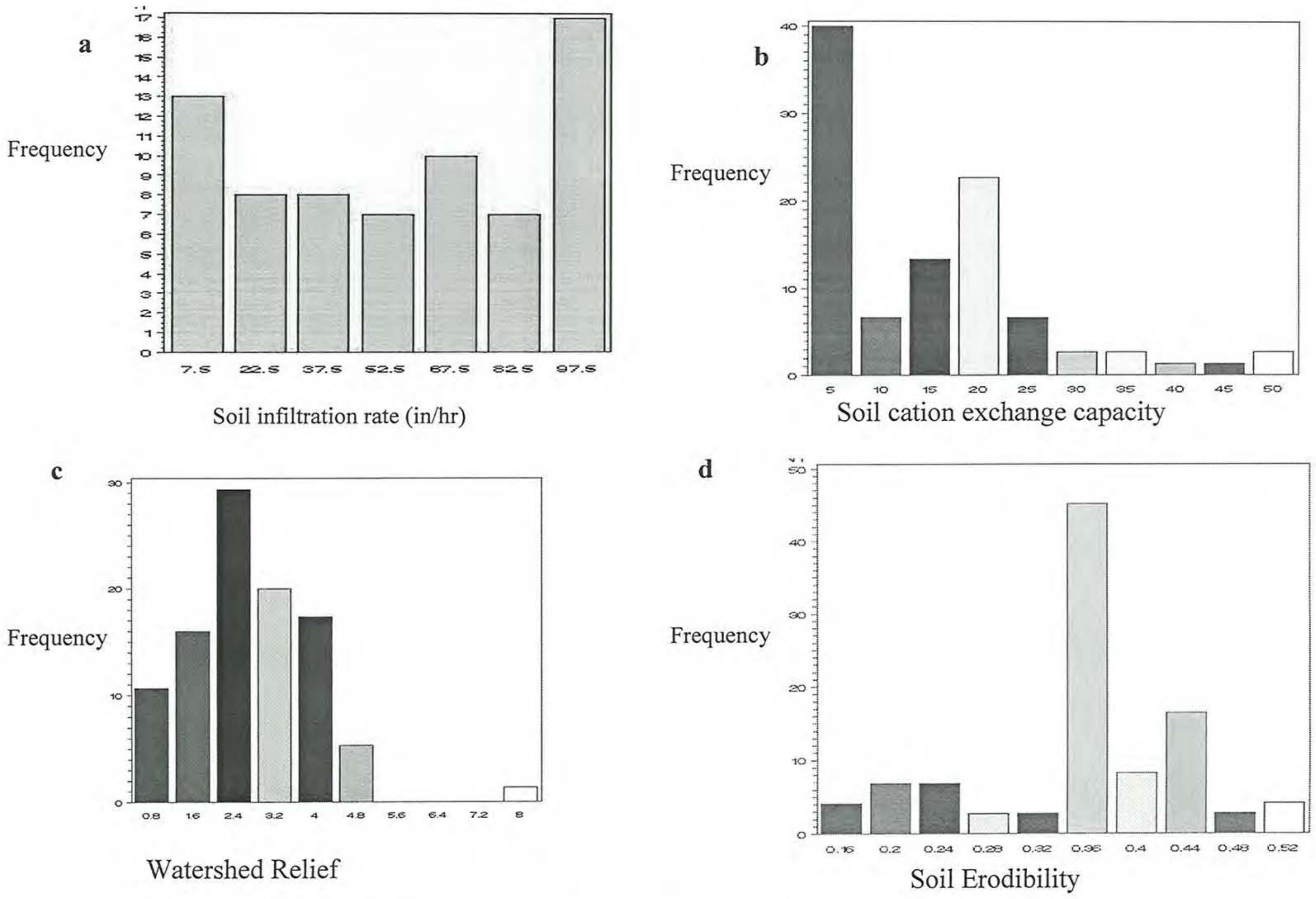


Figure 3.14. Histograms showing the distribution of selected watershed characteristics of Nebraska reservoirs: (a). Soil Infiltration Rate; (b). Soil cation exchange capacity; (c). Watershed Relief; and (d). Soil Erodibility

CHAPTER 4. IMPLEMENTATION OF A WATERSHED-BASED CLASSIFICATION SYSTEM FOR NEBRASKA RESERVOIRS

4.0. Introduction

An important component of managing reservoir water quality effectively is to segregate the reservoirs into similar "groups" or "classes", in terms of their potential to achieve certain water quality standards. However, information on the number of classes of Nebraska reservoirs is not available. This lack of knowledge limits our understanding of the biophysical characteristics of Nebraska reservoir classes and prevents accurate estimation of potential reservoir water quality. Such information is useful for many applications including predictive modeling of potential water quality impairment of reservoirs based on their class membership.

A vital step in developing a classification is to determine the optimal number of classes to be used. This requires partitioning a dataset such that the entities in one group are more similar to each other than to those in other groups. Similarity refers to the distance between two data entities, where the distance decreases for entities that are most alike (Gordon, 1999). Cluster analysis has been commonly employed to group data without prior knowledge of the class structure (Tou and Gonzalez, 1974; Hartigan, 1975; Hartigan and Wong, 1975; Jain and Dubes, 1988; Eldershaw and Hegland 1997; Legendre and Legendre, 1998; Gordon, 1999; Estivill-Castro and Houle, 2001). The most commonly used clustering techniques are the k-means and single linkage algorithms. The single linkage clustering algorithm is a non-iterative approach based on a local connectivity criterion (Jain and Dubes, 1988; Legendre and Legendre, 1998; Gordon, 1999). On the other hand, the k-means algorithm is an iterative and nonhierarchical clustering method that produces compact and non-overlapping clusters of a dataset (Tou and Gonzalez, 1974; Legendre and Legendre, 1998; Gordon, 1999). The k-means method aims to minimize the sum of squared distances between all points and the cluster centroid (Tou and Gonzalez, 1974; Legendre and Legendre, 1998; Gordon, 1999). The sum of squared distances (J) is given in as:

$$J = \sum_{j=1}^{n} \sum_{k=1}^{K} u \| x_j - z_k \|^2 \tag{4.1}$$

where $x = (x_1, x_2 ... x_n)$ is a set of data points; z = unknown cluster centers; and μ = crisp k x n partition matrix {1, 0}. Initially, the k cluster centers are assigned to k randomly chosen data points, which are then partitioned based on the minimum squared distance criterion. The cluster centers are subsequently updated to the mean of the points belonging to these clusters. The process of partitioning, followed by updating, is repeated until either the cluster centers do not change or there is no significant change in the J values of two consecutive iterations (Tou and Gonzalez, 1974; Legendre and Legendre, 1998; Gordon, 1999). The fundamental issue in any clustering approach is to determine which number of clusters best describes the class structure (or optimal number of classes) of the dataset (i.e. cluster validation).

Several approaches have been used to determine the optimal number of classes for a dataset (Milligan and Cooper, 1985; Xie and Beni, 1991; Gordon 1999; Theodoris and Koutroumbas, 1999; Halkidi *et al.*, 2002; Tibshirani *et al.*, 2001; Duda *et al.*, 2002; Ujjwal and Bandyopadhyay, 2002). These can be grouped into three main categories: use of internal criteria, external criteria, and relative criteria. The internal criteria approach to cluster validation involves analyzing the clustering results based on indices derived from the data, such as a proximity matrix, while the external criteria technique involves evaluating the clustering results based on a pre-defined structure that requires input from the analyst. These two approaches to cluster validation, however, are not appropriate for this research because they are based on statistical hypothesis testing which measures the degree to which a given dataset agrees with a specified scheme (Legendre and Legendre, 1998; Gordon, 1999; Theodoris and Koutroumbas, 1999; Halkidi *et al.*, 2002).

On the other hand, the relative criteria approach to cluster validation evaluates the clustering structure of a given clustering scheme by comparing it to other schemes that are based on the same algorithm, but with different parameter values (Gordon 1999; Theodoris and Koutroumbas, 1999; Halkidi *et al.*, 2002). For example, a comparison of different k-means cluster analysis based on different number of clusters fits the relative criteria scheme. Therefore, the relative criteria approach for determining the optimal number of clusters was used in this study.

Research summarized in this Chapter includes: (a) grouping of Nebraska reservoirs based on variables that underlie, determine, and explain the patterns of change in physical, chemical and biological water quality over seasonal or annual cycles rather than environmental stressors like land use; (b) statistical cluster validation techniques were employed to determine the optimal number of clusters that best describe the class structure of Nebraska reservoirs; (c) final determination of optimal number of clusters was based on both statistical inference and water resource management considerations, and (d) a watershed-based classification systems for Nebraska reservoirs was developed using decision tree inductive algorithms.

4.1. Methods

K-means cluster analysis was used to determine the optimal number of Nebraska reservoir watershed classes. For management purposes one would like to have the fewest number of classes that can be used to effectively distinguish lakes that have similar capacities to meet water quality standards. For a given a set of parameters (*p)* associated with a particular clustering algorithm the possible clustering scheme, C_i (i = 2, 3 … *p*), is defined by that clustering algorithm. The clustering algorithm (in this case, k-means clustering) is then run for all the clustering schemes, using the number of clusters between 2 and *n*. A plot of a clustering index (e.g., Calinski-Harabasz statistic, Dunn index, Cluster Distance, R-Squared, Hubert Γ statistic and Davies-Bouldin index) against the number of clusters usually highlights a point at which there is a significant local change in the clustering index (i.e. relative criteria approach to cluster validation). This change in value, which occurs as a “knee” in the plot, represents the "optimal" number of clusters (or classes) in the data set (Milligan and Cooper, 1985; Halkidi *et al.*, 2002). Milligan and Cooper (1985), Tibshirani *et al.* (2001) and Ujjwal and Bandyopadhyay (2002) examined different cluster validation indices and found that the Calinski-Harabasz statistic was one of the best performing indices. Thus, the Calinski-Harabasz statistic was used in this research.

A series of cluster analyses were performed for the 78 reservoir watersheds in Nebraska (Figure 4.1). Watershed characteristics that were used in the cluster analysis included watershed size, mean 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 variables including precipitation, temperature and humidity (Table 4.1).

The "FASTCLUS" procedure in SAS® was used to cluster the watershed data into numbers of classes ranging between 2 and 25. FASTCLUS finds disjoint and non-overlapping clusters of observations using *k*-means clustering method such that, observations that are very close to each other are usually assigned to the same cluster, while observations that are far apart are assigned to different clusters (SAS Institute, 2000). The maximum of 25 classes was chosen to reflect a reasonable uppermost limit of watershed management classes based on literature and several clustering attempts. The "FASTCLUS" procedure was used here because there is often no need to run the procedure to convergence.

The Calinski-Harabasz statistic (represented by "Pseudo F" in SAS® output) is defined as follows:

$$\text{Pseudo F} = \frac{[(R^2)/(c-1)]}{[(1-R^2)/(n-c)]} \qquad (4.2)$$

where R^2 = observed overall correlation; c = number of clusters; and, n = number of observations (Calinski and Harabasz, 1974; SAS Institute, 2000). The Pseudo F statistic was used to assess different clustering outputs based on the number of classes, by plotting Pseudo F values against the number of classes (hereafter referred to as NCL). The output of the cluster analysis showed that the potential NCL that were likely to reflect the class structure of Nebraska reservoirs were 3, 5, 13, 17, and 19. The class membership information from the SAS output for potential NCLs were exported into a spreadsheet and appended to the watershed characteristics dataset. The dataset was then used in

ArcMap® GIS software to generate maps showing reservoirs watershed classes for each of the potential NCLs.

Since there was more than one NCL that corresponded to the local changes in Pseudo F values, there was a need for further testing to identify a single NCL that best represented the optimal number of classes. Consequently, potential NCLs were evaluated using a predictive model, i.e. classification tree (See5® decision tree software), to refine the selection of the optimal NCLs based on their predictive effectiveness (Tibshirani *et al.* (2001). Although predictive accuracy can be based on a single training model, the accuracy is usually increased by using an averaged, weighted prediction error of several models as provided by cross-validation error (Breiman *et al.*, 1984; Ripley, 1996; Goute, 1997; De'ath and Fabricius, 2000).

Validation approaches for classification models in geosciences usually involve the use of contingency tables (confusion matrix) and Kappa statistic that is usually based on field samples (validation data) that are independent of the data used to develop the model or classification in question (Congalton, 1991; Congalton and Green, 1999). Fitzgerald and Lees, (1994) suggested that the Kappa statistic provides a better measure of the classifier model accuracy than the overall accuracy, since it considers inter-class agreement. The two approaches to accuracy assessment are employed to assess accuracy of a classifier model against independently collected and known validation datasets.

In many cases, there are limited sampled data for training and validation. Resubstitution estimates of prediction are commonly used to assess classifier or model accuracy based on the same data that are used to train the classifier. However, resubstitution estimates are usually optimistic and lead to generalization problem;

because the resubstitution approach gives little insight on how a classifier or model would perform on previously unseen data (Breiman *et al.*, 1984).

Stehman (2000), suggested a method to evaluate classification model accuracy using design-based sampling inferences, and it is also not susceptible to spatial autocorrelation. Henebry and Merchant (2001), noted however that despite it's usefulness in minimizing the confounding effects of spatial autocorrelation, design-based sampling inferences offer no means to predict the accuracy of unobserved data. According to Henebry and Merchant (2001), computer-intensive Monte Carlo error analyses can be used to compute the model reliability, which is estimated by the rate at which Monte Carlo predictions fall within a user-designated accuracy interval. Henebry and Merchant (2001), also highlighted the need to develop new approaches to validating models that are based on geospatial datasets.

Resampling techniques are other computer intensive approaches to evaluate model accuracy. Particularly, when dealing with non-linear models (such as classification trees) and geospatial datasets, it is important to obtain a good estimate of the generalization error, i.e. average error that a model will make on an infinite size and unknown test samples (ref). Resampling techniques provide a method for using all of the available data to train, yet still testing the classifier on unseen data (Stone, 1974; Breiman *et al.*, 1984; Efron and Tibshirani, 1993; Schaffer, 1993; Kohavi, 1995; Shao and Tu, 1995).

Resampling techniques that can compute the aforementioned generalization error include: hold-out which consist in removing data from the learning set and keeping them for validation; Monte-Carlo cross-validation (or simply cross-validation), where several hold-out validation sets are randomly and sequentially drawn from that dataset; k-fold cross-validation, where the initial set is randomly split into k roughly equal parts, each one being used successively as a validation set. A special case of k-fold cross-validation where the size of the validation set is 1 is called the "leave-one-out" method; and, the bootstrap which involves drawing validation sets with replacement from the original sample and using these sets to estimate the generalization errors. The recent bootstrap 632^+ is an improved version of the original bootstrap.

Although Stone (1977), suggested that the above-mentioned resampling methods of estimating generalization errors are asymptotically roughly equivalent, others have pointed out some exceptions and limitations as follows: leave-one-out is less biased but its variance is unacceptable; cross-validation is consistent (i.e. converges to the generalization error when the size of the sample increases) if the size of the validation set grows infinitely faster than the size of the learning set; cross-validation is almost unbiased; bootstrap is downward biased but has a very low variance most recent bootstrap method (632+) is almost unbiased and also has a low variance (Efron and Tibshirani, 1993; Kohavi, 1995; Shao and Tu, 1995).

According to Stone (1974), cross-validation simply consists of controlled or uncontrolled division of data sample into sub-sample. One sub-sample is used to compute a statistical predictor of the model, including any necessary estimation, and then the model performance is assessed by measuring its predictions against the other sub-sample. In this way, the accuracy of the classifier is tested on unseen data, and the estimates of classifier accuracy are more realistic than resubstitution estimates. Goute (1997) suggested that cross-validation error is a better indicator of model accuracy than that derived from the split-sample approach, especially when the sample size is relatively small (i.e. less than 100).

K-fold cross-validation is a more robust form of cross-validation. The data is divided into k equal subsets of independent training and test data, such that the first 1/kth subset of the data is assigned to the first test set, the second 1/kth subset is assigned to the second test set, and this is done for all k subsets. Thus, the test sets are completely independent of each other. For each test set, the remaining data are used to train the classifier or model, such that the test and training sets for each partition of the data are also independent. Estimates of prediction accuracy are computed for each of the k-fold partitions, and averaged to give overall prediction accuracy. A 10-fold cross-validation was used in this study because it is the typical number of subsets (or partitions) often used in k-fold cross-validations (Breiman, 1996; De'ath and Fabricius, 2000).

A k-fold cross-validated error, where *k* is 10 partitions, was employed to evaluate the predictive effectiveness of the classification tree model using the potential NCLs (3, 5, 13, 17, and 19) as the dependent variables. Watershed characteristics that were used in the cluster analyses were used here as independent explanatory variables in the classification tree model. For a given potential number of classes (NCLs), the See5® classification tree software was used to compute the error rates for each of 12 separate 10-fold cross-validation trials and the mean cross-validation error rates of these trials were calculated (RuleQuest, 2003).

Since 3 NCL was the minimum number of classes, the mean cross-validation error rates for the remaining NCLs (5, 13, 17, and 19) were normalized with respect to 3 NCL (i.e. the reference NCL). This was done to determine which increase in NCL resulted in least corresponding increase in mean cross-validation error. The normalized mean cross-validation error rates (NME) were computed as follows:

$$NME = \frac{\Delta ME}{\Delta NCL} = \frac{ME_{n=L} - ME_{n=3}}{L-3} \tag{4.3}$$

where *ME* is mean cross-validation error rate; NCL is number of classes, *n* is potential NCL; and *L* is the test NCL. Outputs of the above computation were plotted against the potential numbers of clusters and the optimal number of classes that exist among Nebraska reservoirs was identified.

4.2. Results and discussions

Pseudo F (Calinski-Harabasz statistic) values were obtained from the SAS cluster analyses outputs for each potential number of classes (NCL), *n* = 2, 3, 4 … 25 (Table 4.2). A plot of Pseudo F values against potential NCLs revealed that the NCLs that are likely to represent the structure of Nebraska reservoir classes are 3, 5, 13, 17, and 19 (Figure 4.2). As noted above, cross-validated mean error rates were derived from classification tree predictive models (Table 4.3). A plot of the normalized cross-validated mean error rates against the potential NCLs suggested that the optimal number of Nebraska reservoirs classes was 13 (Figure 4.3). In order to understand the relative importance of the number of classes, the map of 13 reservoirs classes was compared to maps that showed 3, 5, and 13 potential NCLs (Figure 4.4). The maps of 17 and 19 NCL were excluded because they did not result in any visible difference from the map of 13 NCL. The changes in spatial patterns of reservoir watershed classes in the maps appear to reflect major environmental conditions that affect lakes processes, as the number of clusters changed from 3 to 13 (Maxwell *et. al*, 1995).

The map of 3 NCL shows classes influenced mainly by climate (i.e. maximum temperature) and related vegetation patterns in Nebraska (Figure 4.4a). Reservoir watersheds in class 1 occupy the tall grass prairie in eastern Nebraska while class 2 and 3 reservoir watersheds are dominated by the Sand Hills prairie and the Niobrara shrub land, shortgrass sage-steppe prairie and Ponderosa pine, respectively. The spatial pattern of reservoir watershed classes in the map of 5 NCL reflects the influences of both climate and terrain characteristics (such as temperature and relief) on the watersheds (Figure 4.4b). The classes in 5 NCL map show additional segregation of classes in the map of 3 NCL.

The map of 13 NCL shows spatial patterns in the reservoir watershed classes that reflect the patterns of climate and terrain variability, as well as variations in soil characteristics across Nebraska (Figure 4.4c). Reservoir watersheds in the northeastern part of Nebraska belonged to class 2. The average size of reservoirs in this group was in the lower 25 percentile of the sampled reservoirs. The watersheds of these reservoirs are generally small and characterized by low relief, high soil erodibility, and high soil organic matter content (Table 4.4). Reservoir watersheds in classes 1 and 13 dominate southeastern Nebraska. Reservoirs in class 1 are, on average, smaller than those in class 13. Also, the average watershed size in class 13 appears to be larger than that of class 1. Both watershed classes have high soil organic matter content and relatively low soil erodibility; however, the watersheds in class 1 have steeper slope and higher relief compared to watersheds in class13.

Reservoir watersheds in northwestern Nebraska belong to classes 3, 9, 10, and 11. Classes 3 and 11 have only one watershed each, while classes 9 and 10 have seven and two watersheds respectively. Although classes 10 and 11 are adjacent, they are not similar. For example, class 11 watershed has larger area and traverse higher terrain relief than class 10 watersheds. In the north-central part of the state, there are two reservoirs in class 4. These reservoirs are characterized by large watersheds, with relatively low soil organic matter content and high relief. Reservoir watersheds in class 7 are aligned diagonally between the central and southwestern part of the state. These watersheds are similar to class 4 watersheds, except that they exhibit relatively lower relief. The central and southwestern portions of Nebraska are dominated by classes 12 and 8, respectively. Class 8 reservoirs are large and they have larger watershed size than reservoirs in class 12. Also, class 12 watersheds are found in low relief areas compared to those in class 8. The aforementioned descriptions of the spatial variability of watershed classes in the map of 13 NCL provide a synoptic overview of the general characteristics of these classes. Additional discussions with respect to how the watershed characteristics influenced the segregation of these classes are provided below.

Having identified an optimal number of Nebraska reservoir classes, a classification tree model was used to describe the structure of the different classes as well as the variables that contributed to the segregation of these classes (Figure 4.5). The rectangular boxes in figure 4.5 represent terminal nodes (i.e. there is no further division of the group) and are assigned a class number. The oval boxes represent non-terminal nodes and require further splitting. The cross-validation prediction error of the classification tree model for reservoir watersheds was 26.33 percent.

It can be seen that soil organic matter content was responsible for the initial split of watershed classes. Watersheds in classes 4, 7, 8, 9, 10, and 12 were relatively poor in organic matter, while watersheds in classes 1, 2, and 13 were rich in organic matter. The ability of soils to absorb agricultural effluents like pesticides decreases with a decrease in organic matter content (Kumada, 1987; Sparling *et al.,* 2003). Therefore, it is important to note that most of the reservoir classes (viz. classes 4, 7, 8, 9, 10, and 12) are inherently vulnerable to pollution from agricultural chemical effluents. Among these watersheds, soil cation exchange capacity (CEC) and drainage density were responsible for final splits into classes 9 and 12. Also, watershed relief, soil CEC and pH influenced the final splitting into classes 4, 7, and 8, 10. Classes 4 and 7 differed primarily in their respective watershed relief. Despite their low drainage density, both groups have relatively acidic soils with correspondingly low buffering capacity (CEC of less than 12.3). Specifically, the low relief reservoir watersheds in class 7 (i.e. relief less than 247 meters) are even more vulnerable to pesticides or herbicide effluents from agricultural activities in their watersheds.

The segregation of organic-rich reservoir watersheds into classes 1, 2, and 13 was influenced by soil erodibility, watershed slope and organic matter content respectively. The reservoirs in these watershed classes are relatively less susceptible to potential pollution from agricultural effluents like herbicides. The factors that influenced the final segregation of these classes emphasize the importance of land management practices that control soil erosion in these watersheds. This is particularly true for reservoirs in classes 1 and 2 that have relatively high mean watershed slope and soil erodibility.

A review of the terminal nodes in figure 4.5 revealed that only nine classes were represented by the classification tree instead of 13 classes. Reservoir classes 3, 5, 6, and 11 were missing from the classification tree. This is because the classification tree nodes (classes) that are not sufficiently compact are subsequently split or recombined into other nodes (Breiman *et al.*, 1984; De'ath and Fabricius, 2000). All four reservoir classes that were not represented in the classification tree had one watershed each and this is indicative of non-compact classes or classification tree nodes. Consequently, the class means for the respective watershed characteristics (Table 4.4) that were represented in the classification tree (Figure 4.5) were used in a principal component analysis (PCA). This was done to identify which classes in the classification tree were closest to the missing classes. The first and second principal components (PC) explained 65.1 percent of the variation in the data. A plot of PC 1 and PC2 showed that the missing classes (3, 5, 6, and 11) were closer to classes 4, 12, 13 and 8 respectively in the classification tree (Figure 4.6). This assertion is confirmed by the class distances obtained from clusters analysis based on 13 classes (Table 4. 5). Hence, the map of 13 NCL (Figure 4.4.c) was revised to reflect these similarities.

4.2.1. Nebraska reservoir watershed classes

ArcMap® GIS was used to update the attribute table of the map of 13 NCL by reassigning reservoir watersheds in class 3 to class 4; class 5 to 12; class 6 to 13; and class 11 to class 8. For example, watersheds classes 11 and 8 in western Nebraska were combined in the revised map (9 NCL) (Figure 4.7). Characteristics of the revised reservoir classes are described in summarized in Table 4.6. Additional information on class membership of each sampled reservoir used in this study is listed in Appendix 1. The revised map shows that the water quality of Nebraska reservoirs could be characterized based on nine classes.

Finally, inversion of the classification tree model (shown in Figure 4.5) was explored based on maps that could be generated by equations derived from the classification tree leaves or nodes (Appendix 2). ArcMap® GIS was used to generate maps for each node in the classification tree. Output maps of the model inversion showed that the classification tree model predictions were consistent with ArcMap generated reservoir classes based on equations derived from the classification tree leaves.

4.3. Summary

An approach to watershed based classification was developed and it was demonstrated to be effective in identifying the optimal class structure of Nebraska reservoirs as well as highlighting watershed characteristics that impact the segregation of the reservoir classes. Cluster analysis was performed on the watershed characteristics of 78 selected Nebraska reservoirs in order to determine the optimal number of Nebraska reservoir classes. A plot of the Pseudo-F statistic (obtained from the cluster analysis output) against the respective number of classes (NCL), suggested that the potential number of classes included 3, 5, 13, 17, and 19. Further analysis of the optimal number of classes (NCL) was based on the predictive strength of the potential NCL's using See5® classification tree software. The outcome of the classification tree modeling suggested that the optimal number of Nebraska watershed classes was 13 NCL. The classification tree was used to describe the structure of the Nebraska reservoir classes, and soil organic matter content was found to be the most important single variable for segregating the watersheds. The cross-validation prediction error of the classification tree model was 26.33 percent. Finally, the initial 13 NCL map was revised based on the classification tree and the revised map suggested that Nebraska reservoirs can be represented by nine optimal classes. The characteristics of the nine reservoir classes were subsequently described.

Although successful, this research clearly suggests the need for additional investigation. Additional work that needs to be done includes expanding the STATSGO datasets to incorporate watersheds that extend into neighboring states (Colorado, Kansas, South Dakota, and Wyoming). This will highlight the impact of large reservoirs on the classification results, since watersheds of most of the large reservoirs in the GIS database fall outside the Nebraska state boundary. It is also important to explore the potential advantages of higher resolution data (watershed characteristics derived from SSURGO database) on the lake classification process.

The use of k-means clustering has some limitations such as sensitivity to outliers or extreme values, susceptibility to the choice of starting points (cluster centroids), and tendency to produce classes with most data points concentrated in a few classes (Eldershaw and Hegland 1997; Legendre and Legendre, 1998; Gordon, 1999; Estivill-Castro and Houle, 2001). Further work to address these limitations and compare the performance of existing modifications or alternatives to k-means clustering is needed. Additional research in refining the classification tree splitting process could enhance the predictive effectiveness of the classification tree output models (Breiman *et al.*, 1984; De'ath and Fabricius, 2000). It is also important to test the reservoir watershed classification procedure by comparing the accuracy of the classification tree derived watershed classes to other classification approaches (e.g., discriminant analyses and Omernik's ecoregions).

Since the geospatial data employed in this study are available for the entire U.S. and the automated GIS-based procedures for watershed delineation are also nationally available, the watershed-based reservoir classification system described in this chapter has potential national application. Through model refinement, outputs of the classification tree procedure for watershed-based reservoir classification promises to provide water resources managers an effective decision-support tool in the management of reservoir water quality. For example the classification results could inform resource managers in the development of reservoir nutrient criteria.

References cited

Breiman, L., J. H. Friedman, R.A. Olshen and C. J. Stone. 1984. **Classification and Regression Trees**. Wadsworth, Inc. Belmont, California. 358p.

Breiman, L. 1998. **Arcing Classifiers.** Annals of Statistics. 26:801-824.

Calinski, R.B. and J. Harabasz. 1974. *A dendrite method for cluster analysis.* **Communications in Statistics**. 3:1-27.

Congalton, R. and K. Green. 1999. **Assessing the Accuracy of Remotely Sensed Data: Principles and Practices.** CRC/Lewis Press, Boca Raton, FL. 137p.

De' ath, G and K.E. Fabricius. 2000. *Classification and regression trees: a simple yet powerful technique for ecological data analysis.* **Ecology.** 8(11):3178-3192.

Duda, R.O., P.E. Hart and D.G. Stork. **Pattern Classification**, 2nd Edition. John Wiley and Sons. 680p.

Dunn, J.C. 1974. *Well separated clusters and optimal fuzzy partitions.* **Journal of Cybernetics.** 4: 95-104.

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

Eldershaw, C. and M. Hegland. 1997. Cluster Analysis Using Triangulation. *In*: B. J. Noye, M. D. Teubner, and A. W. Gill (Editors). Computational Techniques and Applications: CTAC97, World Scientific, Singapore. p 201-208.

Estivill-Castro, V. and M.E. Houle. 2001. *Robust Distance-Based Clustering with Applications to Spatial Data Mining.* **Algorithmica.** 30(2):216–242

Goute, C. 1997. *Note on free lunches and cross-validation.* **Neural Computation.** 9: 1211-1215.

Gordon, A. 1999. **Classification,** 2nd Edition. Chapman and Hall/CRC. London. 256p.

Halkidi, M, Y. Batistakis and M. Vazirgiannis. 2002. C*luster validity methods: part I.* **SIGMOD Record.** 31(2):40-45.

Hartigan, J.A. (1975). **Clustering Algorithms**. New York: John Wiley & Sons, Inc.

Hartigan, J.A. & Wong, M.A. (1979). *A K-means clustering algorithm: Algorithm AS 136.* **Applied Statistics.** 28, 126-130.

Henebry, G. M., and J. W. Merchant. 2001. *Geospatial data in time: limits and prospects for predicting species occurrences.* In **Predicting Species Occurrences: Issues of Scale and Accuracy**. Scott, J.M., P. J. Heglund, M.L. Morrison, J.B. Haufler, M.G. Raphael, W.A. Wall, and F.B. Samson (Eds). Island Press, Covello, CA. pp 291-302.

Jain, A. K. and R. C. Dubes. 1988. **Algorithms for Clustering Data.** Prentice Hall.

Kohavi, R. 1995. *A study of cross-validation and bootstrap for accuracy estimation and model selection.* Proceedings of 14th International Joint Conference on Artificial Intelligence. Vol. 2, Canada.

Kumada, K. 1987. **Chemistry Of Soil Organic Matter.** Elsevier, Amsterdam. 242p.

Legendre, P. and L. Legendre. 1998. **Numerical Ecology**.2nd *English edition.* Elsevier Science. BV, Amsterdam. 853p.

Maxwell, J.R., C.J. Edwards, M.E. Jensen, S.J. Pautian, H. Parrot and D.M. Hill. 1995. **A Hierarchical Framework of Aquatic Ecological Units of North America (Nearctic Zone).** North Central Experiment Station, Forest Service, U.S. Department of Agriculture. General Technical Report NC-176. St. Paul, MN.

Milligan, G.W. and M.C. Cooper. 1985. *An Examination of Procedures for Determining the Number of Clusters in a Data Set.* **Psychometrika.** 50:159-179.

RuleQuest Research. 2003. **See5: An Informal Tutorial.** http://rulequest.com/see5-win.html.

Ripley, B.D. 1996. **Pattern Recognition and Neural Networks.** Cambridge University Press. 403p

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

Schaffer, C. 1993. *Selecting a classification method by cross validation.* **Machine Learning.** 13:135-143.

Shao, J., and D. Tu. 1995. The Jackknife and Bootstrap. Springer Series in Statistics, Springer-Verlag, New York.

Sheskin, D.J. 2000. **Handbook Of Parametric And Non-Parametric Statistical Procedures.** 2nd Ed. Chapman and Hall, New York, N.Y. 982 p.

Sparling, G., R. L. Parfitt, A. E. Hewitt and L. A. Schipper. 2003. Three approaches to define desired soil organic matter contents. **Journal of Environmental Quality.** 32:760-766.

Stehman, S.V. 2000. *Practical implications of design-based sampling inferences for thematic map accuracy assessment*. **Remote Sensing of Environment**. 72:35-45.

Stone, M. 1977. *An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion.* Journal of the Royal Statistical Society. B39: 44–7.

Tibshirani, R, G. Walther and T. Hastie. 2001. *Estimating the number of clusters in a dataset using gap statistic.* **Journal of Royal Statistical Society Series B.** 63(2): 411-423.

Theodoris, S. and K. Koutroumbas. 1999. **Pattern Recognition**. Academic Press. 625p.

Tou, J.T. and R. C. Gonzalez. 1974. **Pattern Recognition Principles.** Addison-Wesley, Reading.

Ujjwal, M. and S. Bandyopadhyay. 2002. *Performance of some clustering algorithms and validity indices*. **IEEE Transactions on Pattern Analysis And Machine Intelligence**. 24 (12): 1650 – 1654.

Xie, X. L. and G. Beni. 1991. A validity measure of fuzzy clustering. **IEEE Transactions on Pattern Analysis and Machine Intelligence**. 13 (8):840-847.

Dataset	Abbreviation	Units	Source
Climate data (annual means)			
Maximum temperature	Temp_max	°C	DAYMET (www.daymet.org)
Minimum temperature	Temp_min	°C	DAYMET
Total precipitation	Ppt_tot	mm	DAYMET
Precipitation intensity	Ppt_intns	mm	DAYMET
Humidity	Humidity	mmHg	DAYMET
Growing degree days	GDD$_{\text{(base 10°C)}}$	degrees	DAYMET
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
Watershed relief	Elevation	m	EDNA DEM
Total drainage length	Drn_Tot	m	EDNA streams (edna.usgs.gov/)
Drainage density	Drn_Dnst	mm^{-2}	EDNA streams
Soils biophysical data			
Erodibility	Kfact	unitless	STATSGO (NRCS/USDA)
Clay content	Clay	% weight	STATSGO
Permeability	Perm	inhr^{-1}	STATSGO
Infiltration rate	Infilt	inhr^{-1}	STATSGO
Organic matter content	OM	% weight	STATSGO
Soils chemistry data			
Salinity	Sal	Mmhoss^{-1}	STATSGO
Soil reaction	pH	unitless	STATSGO
Cation exchange capacity	CEC	unitless	STATSGO
Soil carbonate	$CaCO_3$	% $CaCO_3$	STATSGO

Table 4.1. Some environmental characteristics that affect reservoir water quality. The variables listed above include only those that were used in the reservoir classification.

Number of clusters (NCL)	Pseudo F
2	21.49
3	31.33
4	27.62
5	28.06
6	23.93
7	22.97
8	22.86
9	22.34
10	22.03
11	21.69
12	21.57
13	23.16
14	21.04
15	20.5
16	20.41
17	24.63
18	24.00
19	25.51
20	24.82
21	26.86
22	24.09
23	25.47
24	23.37
25	25.02

Table 4.2. Output of cluster analysis using SAS "FASTCLUS" procedure. Pseudo F values were identified for each clustering output based on different number of clusters (NCL)

NCL	Cross-validation Mean Error (ME)	Change in ME	Normalized ME
3	4.77		
5	16.03	11.26	5.63
13	26.33	6.56	2.16
17	27.95	1.62	1.66
19	41.77	13.82	2.31

Table 4.3. Cross-validation errors derived from classification tree (See5®) predictive models.

Class	1	2	3	4	5	6	7	8	9	10	11	12	13
Watershed area (WA)	6639	2723	13547	202804	77	725	38478	204244	445	8743	462651	4164	13738
Lake area (LA)	473	160	56	9841	58	2430	2577	4108	106	2540	566	2797	136
RATIO (LA:WA)	0.071	0.059	0.004	0.049	0.751	3.351	0.067	0.020	0.239	0.291	0.001	0.672	0.010
CaC0$_3$	0.01	1.09	0.69	0.14	1.71	0.00	0.27	0.07	2.36	1.42	1.37	0.00	0.00
CEC	16.64	11.51	7.87	3.54	5.96	25.53	6.05	3.74	24.72	7.10	6.26	1.48	36.48
Clay	31.76	24.52	10.77	3.90	11.30	29.79	8.70	15.18	46.58	10.90	8.91	21.16	25.58
Erodibility	0.37	0.42	0.30	0.17	0.28	0.28	0.30	0.36	0.50	0.33	0.28	0.41	0.37
Organic matter	3.16	2.70	1.67	1.24	1.27	3.08	1.43	1.63	1.86	1.59	1.78	1.89	3.01
Permeability	0.65	1.55	3.71	12.71	16.66	0.91	9.74	4.83	0.27	4.17	8.17	1.26	0.78
pH	6.42	6.75	7.25	6.55	7.42	6.40	6.57	7.03	7.46	7.33	7.09	7.28	6.05
Infiltration	27.86	48.19	34.45	45.55	15.10	33.56	46.14	49.26	2.96	38.88	44.45	47.94	24.36
Salinity	0.01	0.31	0.18	0.00	0.19	0.90	0.04	0.00	0.87	0.05	0.46	0.00	0.00
Slope	3.20	3.22	7.45	3.44	1.16	0.43	1.69	3.58	2.29	3.51	2.25	5.02	1.54
Drainage total	2302	2365	3562	2729	2375	1229	2468	2633	1315	2902	2648	2878	2429
Drainage density	1.79	2.27	1.41	0.68	3.96	3.33	0.97	1.55	1.57	1.48	1.07	3.21	2.32
Relief	76	78	298	534	8	17	118	269	57	185	392	78	44
Elevation	401	531	1386	926	683	468	741	898	1143	1347	1257	763	441
Temp. (max.)	16.98	16.38	13.94	16.06	17.38	16.34	16.54	17.69	14.95	17.77	16.80	17.10	17.33
Temp. (min.)	4.17	2.68	-1.22	0.90	2.37	2.80	1.87	2.17	-0.65	0.75	0.45	2.14	4.03
Precipitation intensity	1.11	0.97	0.71	0.85	0.97	1.11	0.94	0.85	0.68	0.63	0.68	0.93	1.06
Precipitation total	78.73	67.14	41.69	52.35	54.81	68.49	58.26	50.49	40.87	35.65	38.13	57.19	75.11
Humidity	989	874	553	719	813	886	794	774	581	586	604	804	973
GDD (base 10°C)	4309	4021	2992	3660	4054	4015	3890	4039	3236	3769	3622	3979	4324

Table 4.4. Descriptive characteristics of different Nebraska reservoir watershed classes

Class	Number of reservoirs	RMS	Within-class distance (radius)	Nearest Class	Inter-class Distance
1	30	0.382	2.740	13	2.392
2	8	0.610	3.348	1	3.309
3	1	–	0	4	6.249
4	2	0.698	2.367	8	6.027
5	1	–	0	12	6.041
6	1	–	0	13	9.152
7	6	0.599	3.357	8	4.257
8	2	0.273	0.927	11	4.257
9	7	0.531	3.262	2	7.347
10	2	0.183	0.620	8	4.628
11	1	–	0	8	5.885
12	4	0.621	3.540	2	3.927
13	13	0.459	3.505	1	2.392

Table 4.5. Cluster analysis output (based on 13 classes) showing nearest classes and interclass distances.

Class	Number of reservoirs	RMS	Within-class distance (radius)	Nearest Class	Inter-class Distance
1	30	0.382	2.740	13	2.392
2	8	0.610	3.348	1	3.309
3	1	–	0	4	6.249
4	2	0.698	2.367	8	6.027
5	1	–	0	12	6.041
6	1	–	0	13	9.152
7	6	0.599	3.357	8	4.257
8	2	0.273	0.927	11	4.257
9	7	0.531	3.262	2	7.347
10	2	0.183	0.620	8	4.628
11	1	–	0	8	5.885
12	4	0.621	3.540	2	3.927
13	13	0.459	3.505	1	2.392

Table 4.5. Cluster analysis output (based on 13 classes) showing nearest classes and interclass distances.

NCL 9 classes	NCL 13 classes	No. of Reservoirs	Description
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 the 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.

Table 4.6. Nebraska reservoir classes derived from watershed-based classification

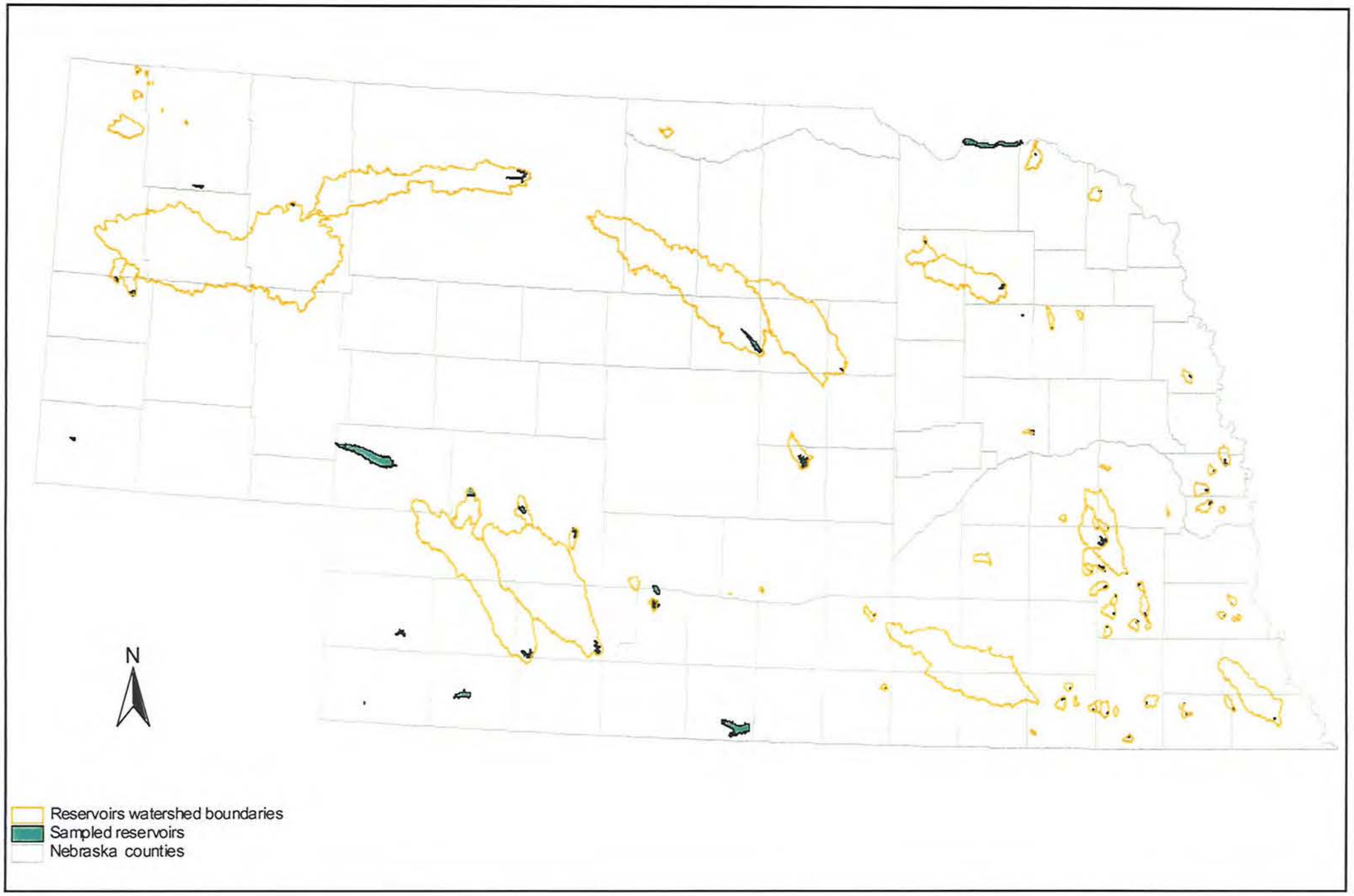


Figure 4.1. Watershed boundaries for 80 Nebraska reservoirs sampled between 1989 and 2003. Watershed boundaries that extend beyond the Nebraska state are not shown.

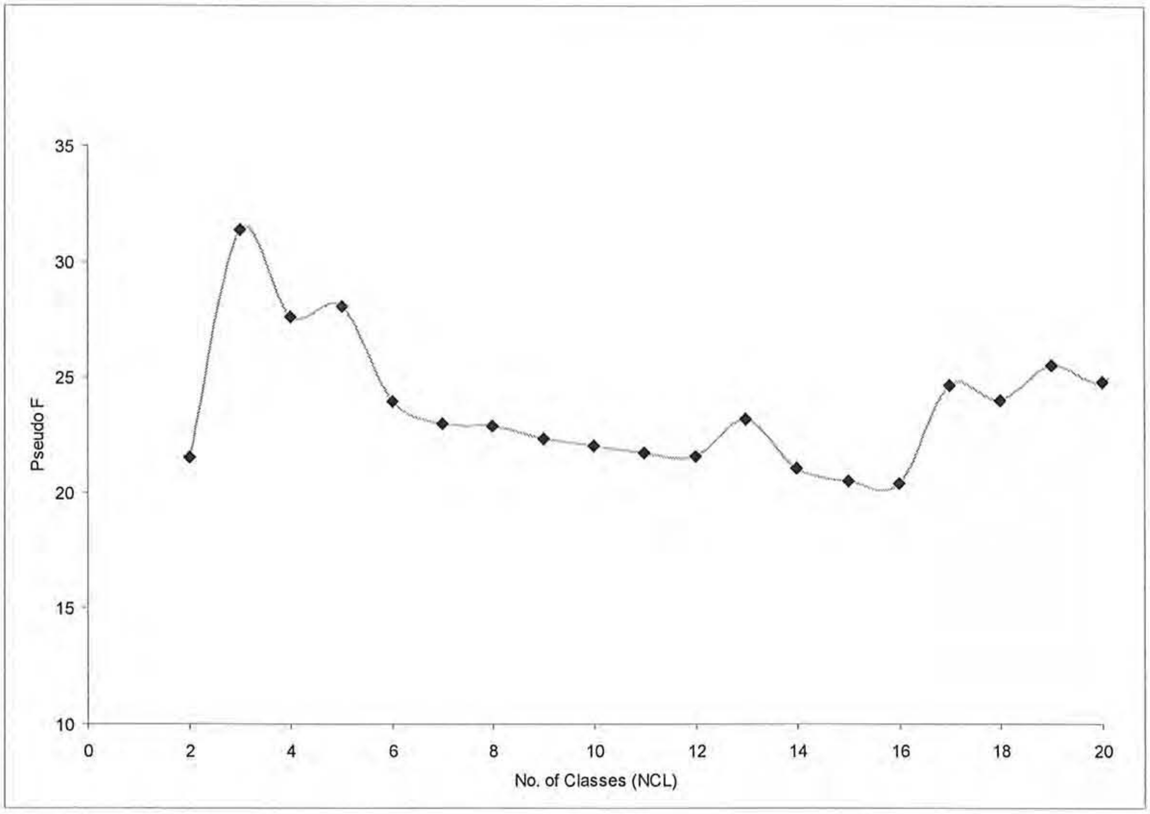


Figure 4.2. Plot of Pseudo-F and number of clusters (NCL) of Nebraska reservoir watersheds

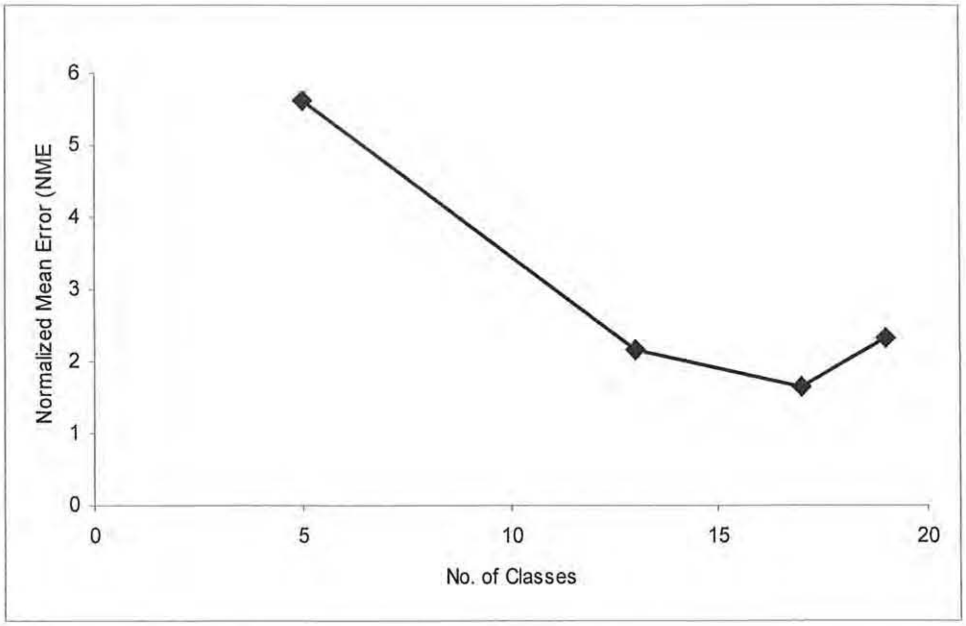


Figure 4.3. Plot of normalized mean cross-validation error (NME) and number of clusters

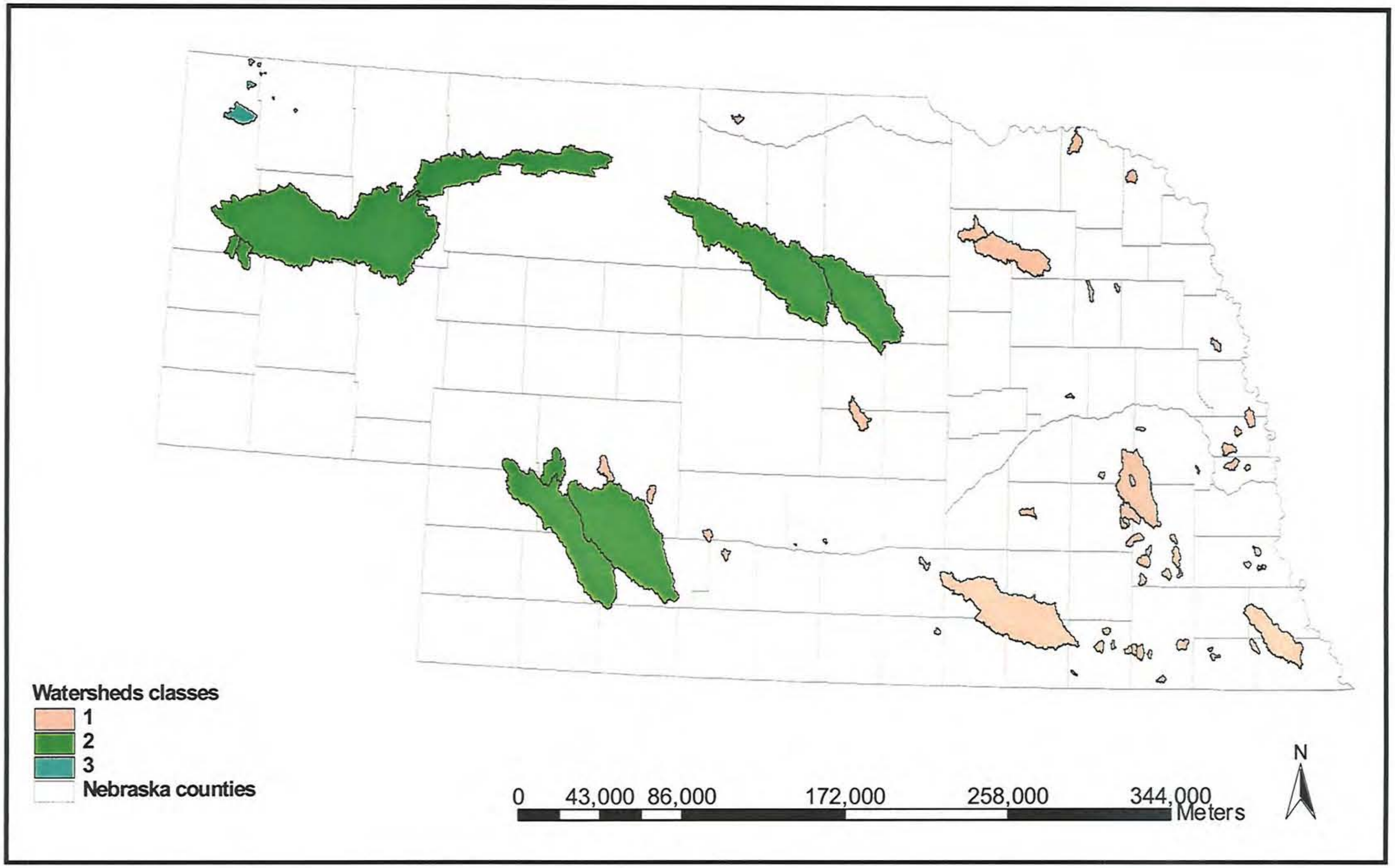


Figure 4.4a. Map of Nebraska reservoir watershed classes (3 classes).

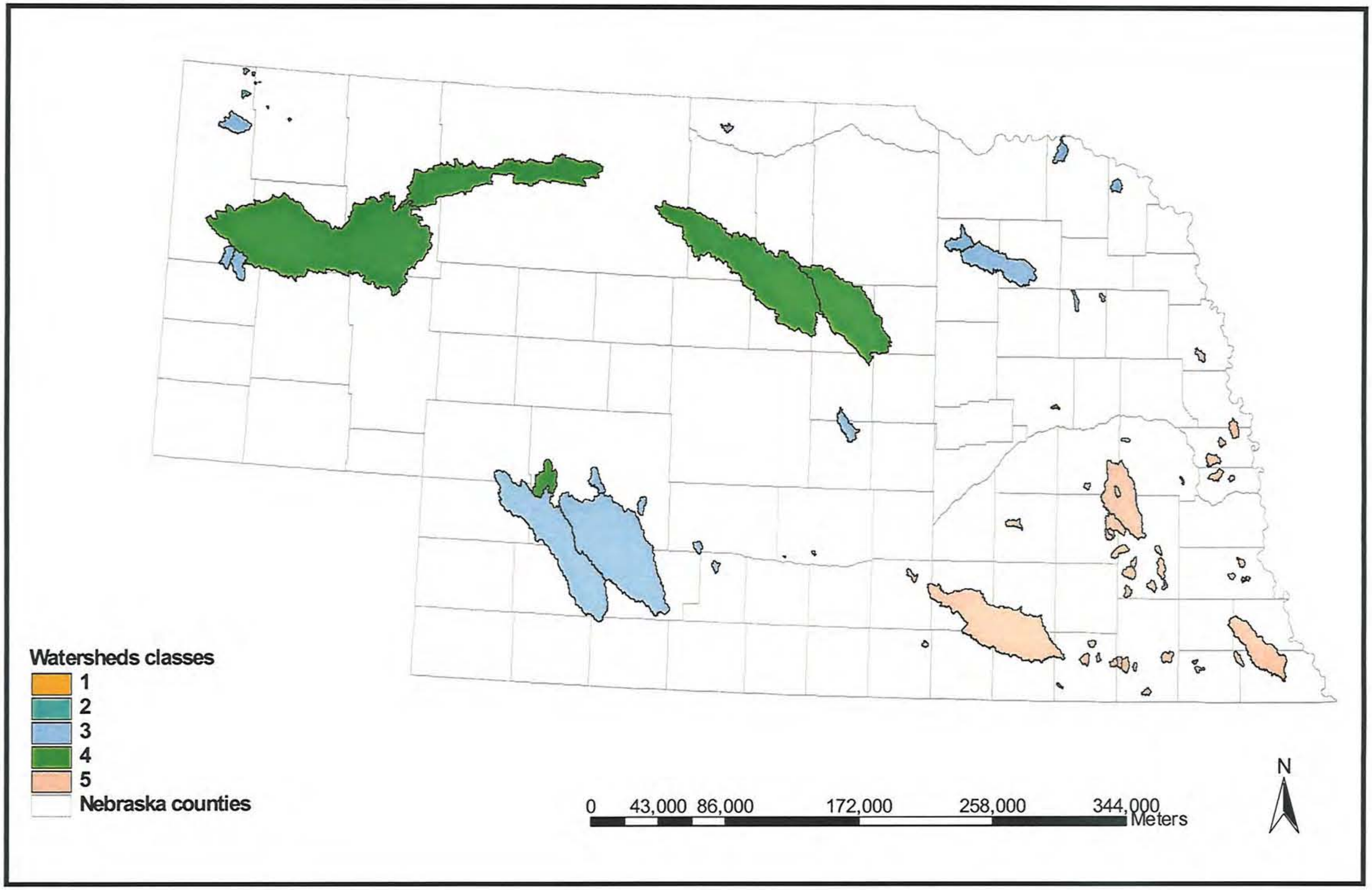


Figure 4.4b. Map of Nebraska reservoir watershed classes (5 classes).

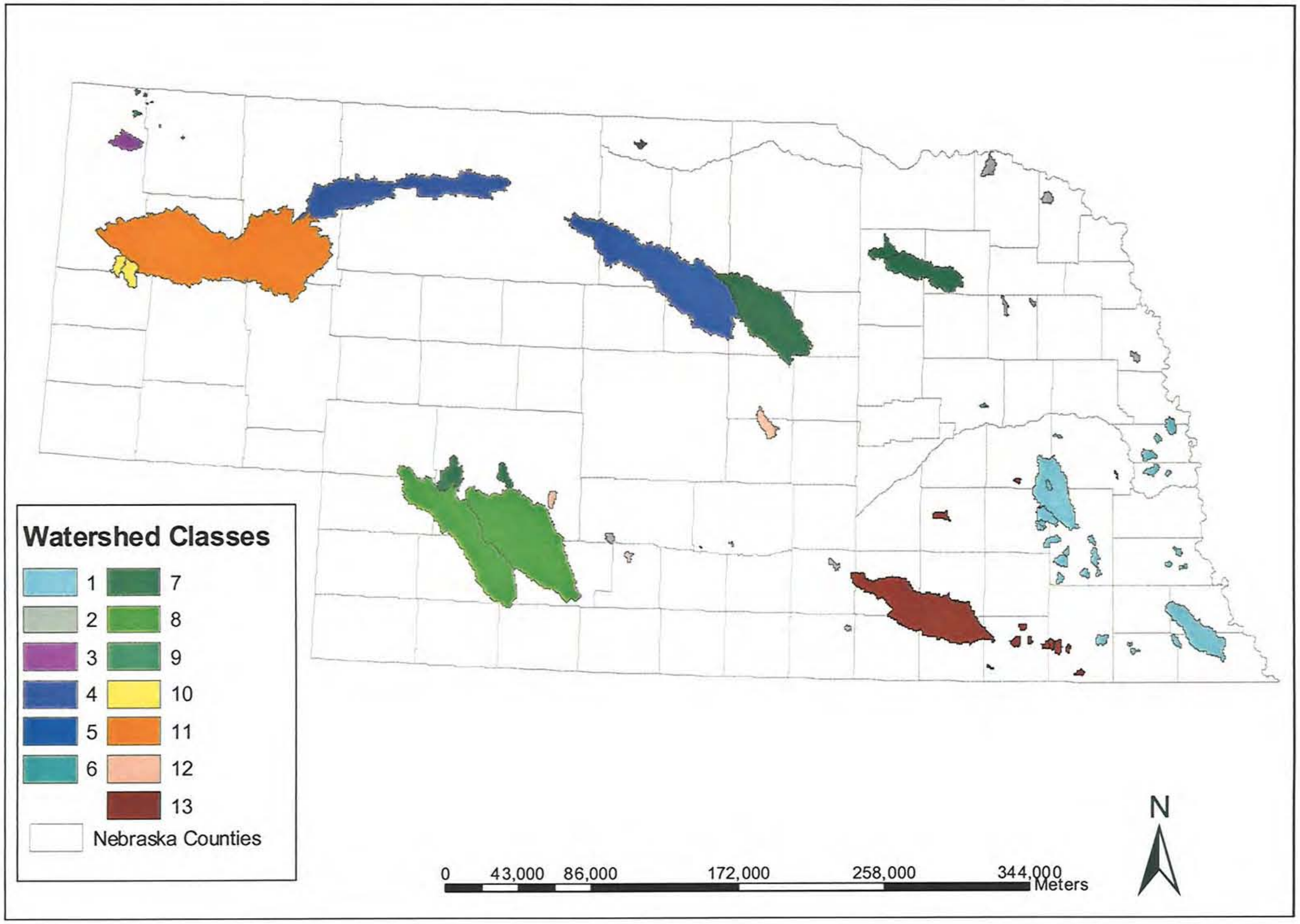


Figure 4.4c. Map of Nebraska reservoir watershed classes (13 classes)

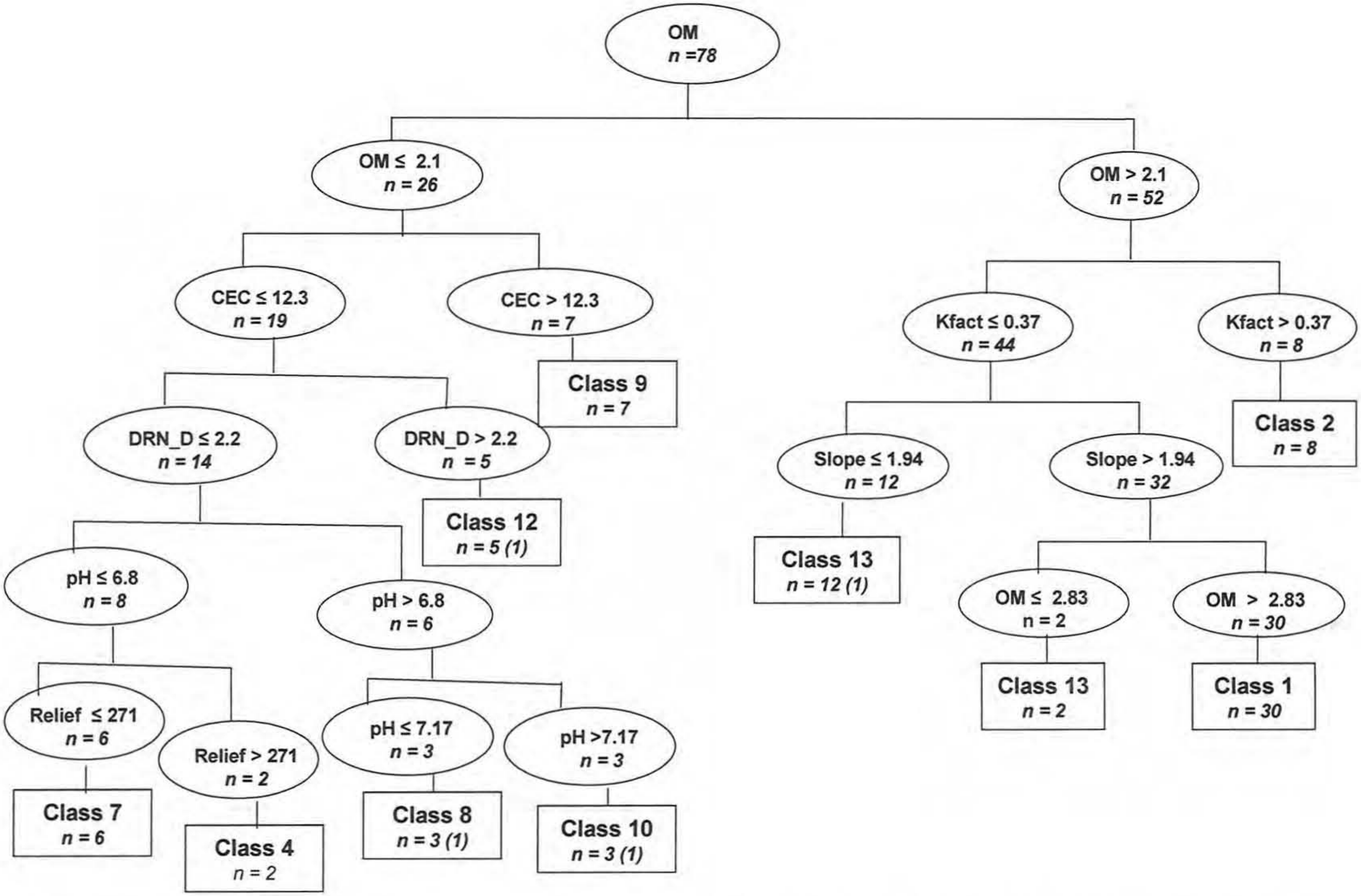


Figure 4.5. Classification tree for Nebraska reservoir classes. Rectangular boxes represent terminal nodes (classes); oval boxes represent non-terminal nodes that required further splitting.

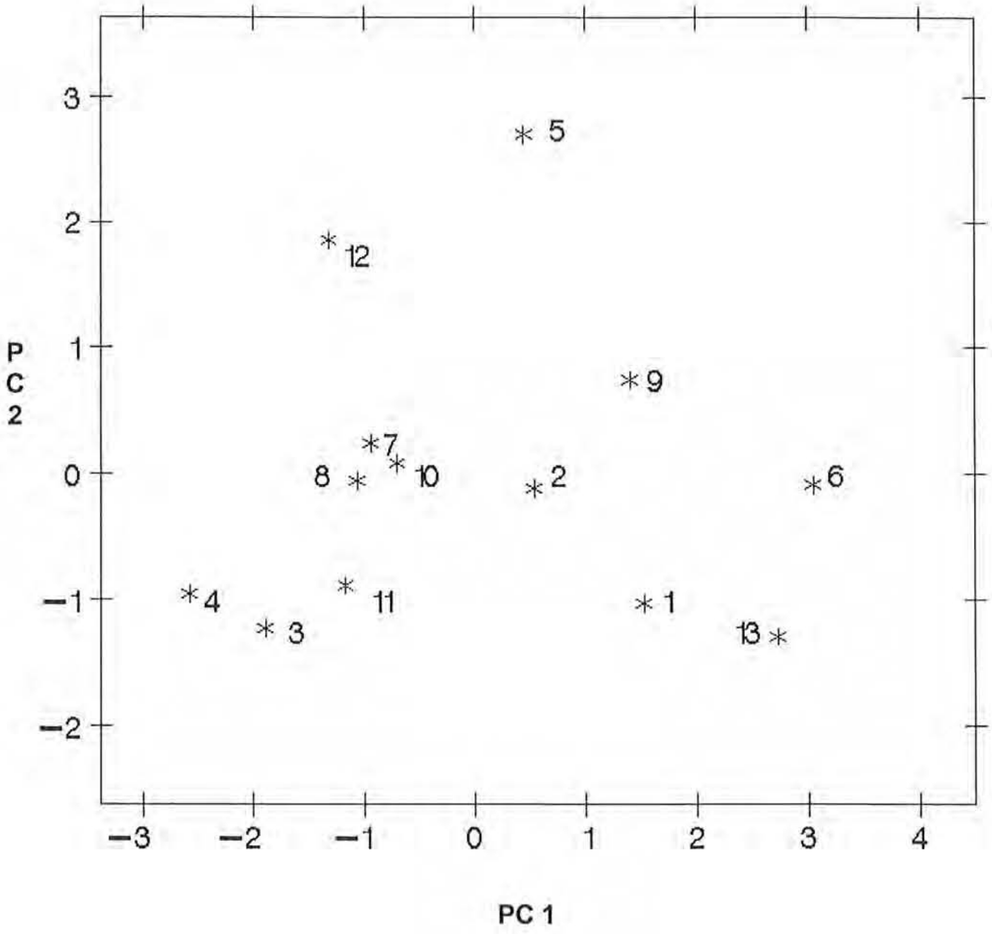


Figure 4.6. Plots of principal components (PC) of watershed classes showing classes 4, 12, 13 and 8 were closest classes 3, 5, 6 and 13 (missing in the classification tree) respectively. PC1 and PC2 represent principal components 1 and 2 respectively.

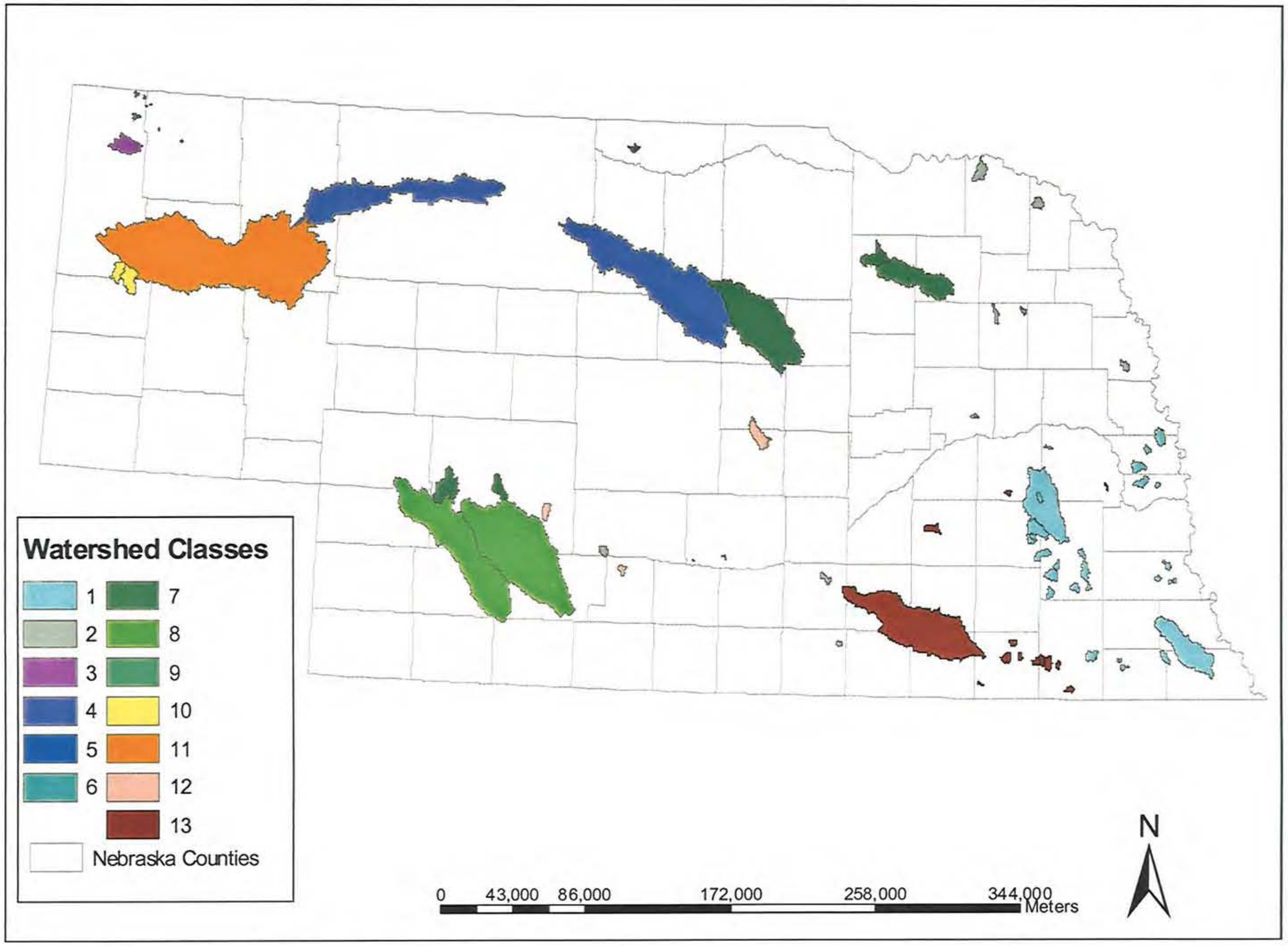


Figure 4.7a. Original map of Nebraska reservoir watershed classes (13 classes)

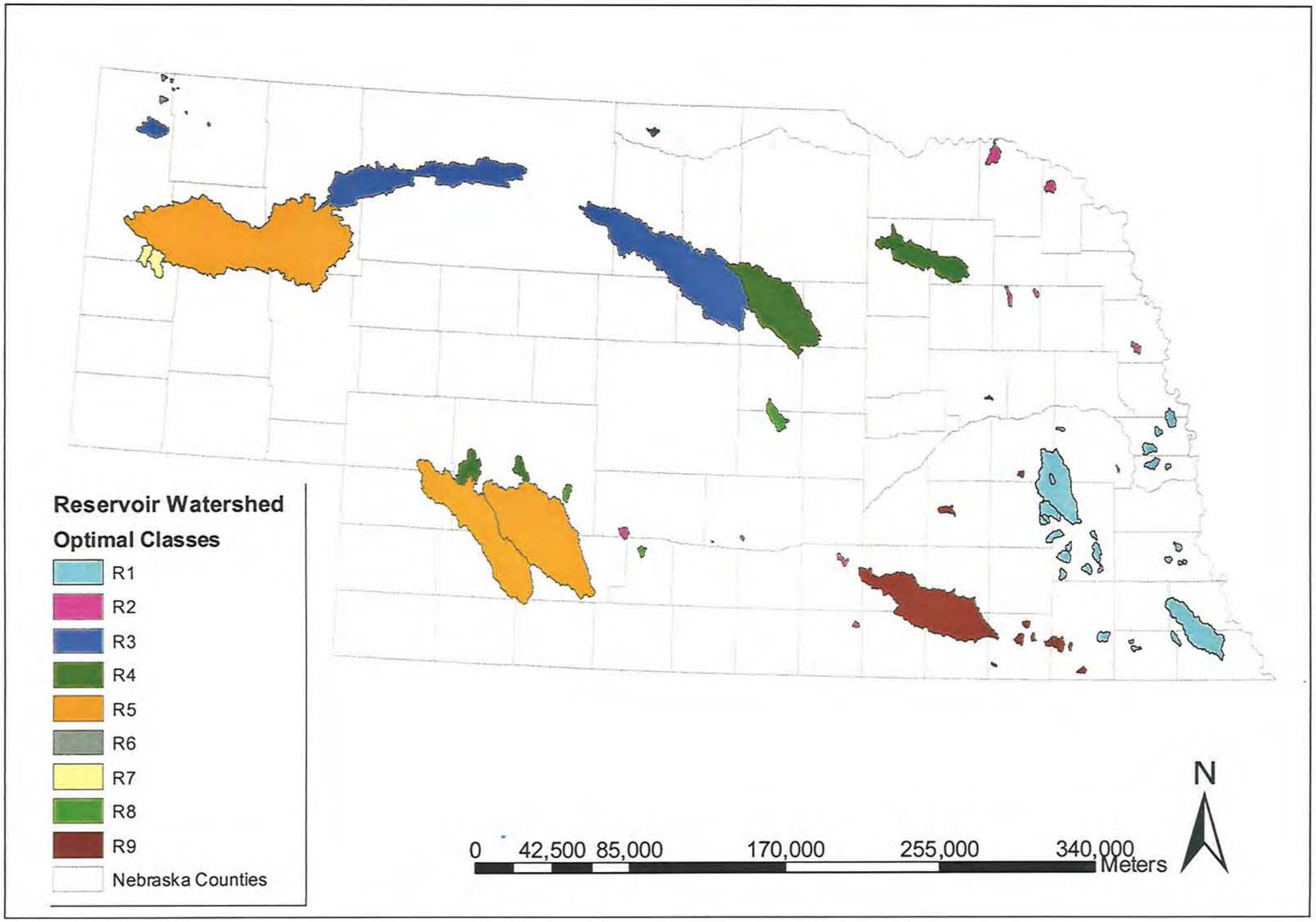


Figure 4.7b. Revised map Nebraska reservoir watershed classes (9 optimal classes)

CHAPTER 5. COMPARISON OF NEBRASKA RESERVOIR CLASSES ESTIMATED FROM WATERSHED-BASED CLASSIFICATION MODELS AND ECOREGIONS

5.0. Introduction

A lake classification can be used to group lakes into ecologically similar classes, enhance our understanding of complex systems, and improve management and decision-making processes (Conquest *et al.,* 1994; Hawkins *et al.,* 2000). Traditional statistical classification approaches, e.g. maximum likelihood classification and discriminant function analysis (DFA), have been used commonly in geosciences and ecological resource monitoring. However, ecological data that are employed in resource classifications are usually complex (with unequal variances) and often contain missing information for certain variables. Ecological data are also characterized by multimodal distributions, and the relationships among variables are non-linear and involve high-order interactions that render traditional statistical techniques ineffective for data exploration, pattern recognition and modeling (De'ath and Fabricius, 2000).

Concerns over the ability of traditional statistical classifiers to effectively classify complex ecological data have led to increasing interest in machine learning classification tools such as neural networks, genetic algorithms, and decision tree classifiers (German *et al.*, 1999). Machine learning involves the application of inductive algorithms to resolve classification problems. Decision tree algorithms (e.g., recursive partitioning) are more easily understood and less complicated than neural networks (e.g., Park *et al.,* 2003) and genetic algorithms (e.g., Chen, 2004), so the focus of this Chapter will be on evaluating decision trees as a potential modeling tool in watershed-based reservoir classification for water quality management. Research has shown that decision tree algorithms outperform traditional statistical approaches (including DFA classification) in accounting for variations in complex datasets for classification tasks (e.g. Breiman *et al.*1984; Quinlan 1986; Ripley 1996; Verbyla, 1987; Emmons *et al.*, 1999; German *et al.*, 1999; De' ath and Fabricius, 2000; Rogan *et al.,* 2003; Yang *et al.,* 2003; Lamon and Stow, 2004).

Ecoregions have frequently been used as natural geographic units for aquatic ecosystem management and assessment, e.g. to define *apriori* water resource classes with respect to potential lake water quality (Omernik 1987; Omernik and Bailey, 1997; EPA, 2002; Rohm *et al.,* 2002; Omernik, 2003). Ecoregions represent similar ecosystems and are based on land forms, land use, climate, potential vegetation and soils (Omernik, 1987; Omernik and Bailey, 1997; EPA, 2002). Previous research has shown that although Omernik's ecoregions are useful for general ecosystem management and analysis, they do not adequately account for the inherent variations among lake water quality data (e.g. Van Sickle and Hughes, 2000; Severn *et al.,* 2001; Winter, 2001; Jenerette *et al.,* 2002; Detenbeck *et al.,* 2003 and 2004). The objective of this Chapter is to compare the performance of the decision tree-based reservoir watershed classification model of Nebraska reservoirs developed in Chapter 4, to a discriminant function analysis (DFA)-based watershed classification system (Momen and Zehr, 1998) and Omernik's ecoregions derived reservoir classes (Omernik, 1987; EPA, 2002). The watershed-based reservoir classification is hypothesized to perform better than ecoregions in defining *apriori* classes of Nebraska reservoirs.

5.1. Background

Supervised, i.e. *apriori*, classification is useful once we have some knowledge of the class labels and the number of classes to be employed. Statistical classifiers attempt to identify an output class from a classification scheme (Π) to that of the input attributes (Ψ) for each explanatory variable and input vector (x) by defining the classification problem as:

$$\Psi^{m} \xrightarrow{\Gamma^{(n,m)}} \Pi^{k} \tag{5.1}$$

where *m* is the number of attributes, *k* is the number of classes, *n* is the number of samples and *Γ* is a transformation function. Dunteman (1984) modified equation 5.1 with respect to supervised classification following Bayes theorem as:

$$\rho(C_k \mid x) = \frac{\rho(x \mid C_k)\rho(C_k)}{\rho(x)} \tag{5.2}$$

where *k* is the number of classes; ρ ($C_k \mid x$) is the posterior probability of class *k* given the input vector x; $\rho(x \mid C_k)$ is the conditional probability of an input vector *x* given class k; ρ (C_k) is the probability that class *k* is present in the data; and, ρ(x) is the probability of an input vector *x* given any class (C_k).

The conditional probability function $\rho(x \mid C_k)$ is therefore required to compute ρ ($C_k \mid x$), however this probability is usually not available for most datasets including ecological data. Hence, $\rho(x \mid C_k)$ is usually computed from a training set as a probability density function *(pdf)* which is often used as a discriminant rule (or function) to identify the class membership of a given input vector *x* (reservoir sample). The type of *pdf* used in estimating $\rho(x \mid C_k)$ determines the type of approximation model. The maximum likelihood classification model is a variation of the equation 5.2; it generally uses the

Gaussian distribution to calculate the posterior probability of each of k classes, and then assigns a new input vector to the class with highest posterior probability (Dunteman, 1984; Huberty, 1994; Legendre and Legendre, 1998).

5.1.1. Discriminant function analysis

The maximum likelihood classifier uses the Bayesian approximation to model the volume of a particular class distribution. On the other hand, discriminant function analysis (DFA) uses empirical hypothesis testing approaches to determine which linear combination of input variables discriminate between two or more naturally occurring groups, i.e. models the surface of a class distribution (Dunteman, 1984; Kachigan, 1986; Ripley, 1996; Legendre and Legendre, 1998; Tabachnick and Fidell, 2001). The linear modeling used in DFA is similar to analysis of variance (ANOVA), multiple linear regression, and canonical analyses. The discriminant functions (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:

$$y = \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + ...\delta_p x_p = \delta' x \qquad (5.3)$$

The classification problem then reduces to identifying the appropriate function (δ) in equation 5.3.

Fisher's pairwise linear discriminant rule is among the commonly used and simplest modifications of the discriminant functions (δ) in equation 5.3. Fisher (1936) used the morphological characteristics of 150 iris specimens to translate multivariate inter-group distances into linear combinations of variables to assist in the segregation of three groups of irises. Since this study, others have developed variations of the concept to address classification problems involving multiple groups (Rao, 1952; Knapp, 1978;

Klecka, 1980; Huberty, 1994; Ripley, 1996). Where there are more than two independent variables, discriminant function analysis, like multiple regression, estimates the coefficients (discriminant functions) of a linear model of the classification matrix of explanatory variables that can best predict the response variable or classification criterion (Legendre and Legendre, 1998). This is followed by computing the discriminant score (or structure coefficient) for each observation based on the estimated coefficients. A classification rule is then developed by applying the Bayes Theorem to the discriminant scores. Further details of the discriminant analysis for ecological data are provided by Ripley (1996) as well as Legendre and Legendre (1998).

Breiman *et al.* (1984) and Quinlan (1993) discussed the limitations of DFA and these are summarized briefly. Effective use of DFA must meet the distributional assumption that: all the explanatory variables follow a multivariate normal distribution for each class of response variable; and, variance-covariance matrices for each class are equal. Although the assumption of normality is critical to DFA, the method is usually applied irrespective of whether the assumption is true for every explanatory variable employed in the analysis. Since the DFA classification method is 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 DFA method is not effective in using cases of missing explanatory variables and hence observations with missing variables are dropped from the analyses, leading to unintended bias due to elimination of variables that might otherwise be critical to developing an appropriate classification rule.

Common alternatives to the use of DFA in resolving classification problems include the logistic regression and probit models. These alternatives however have limitations that are similar to DFA in that they are also dependent on the assumption of normal distribution of explanatory variables, are only suitable for categorical data and may produce biased results when the data set contains missing variables. As such, all the preceding statistical classification methods are parametric and not well suited for ecological analyses. Studies by Breiman *et al.* (1984) and Quinlan (1986) provided impetus to interest in decision trees as suitable alternatives to discriminant analysis. Since then, decision tree approaches to ecological analysis and resource classification are becoming widespread (e.g., Michaelson *et al.*, 1987; Hansen *et al.*, 1996; Friedl and Brodley, 1997; Emmons *et al.*, 1999; German *et al.*, 1999; DeFries and Chan, 2000; De'ath and Fabricius, 2000; Friedl *et al.*, 2000; Witten and Frank, 2000; Rogan *et al.*, 2003; Yang *et al.*, 2003; Lamon and Stow, 2004). The work by De'ath and Fabricius (2000) on habitat types of coral taxa from Australian central Great Barrier Reef, in particular, focused attention on the potential of decision trees as *"powerful and yet simple technique for ecological applications"*.

5.1.2. Decision tree classifiers

Decision tree classifiers are usually implemented as rule-based classifiers (Hunt *et al.*, 1966; Breiman *et al.*, 1984; Quinlan 1986; Verbyla, 1987; Ripley 1996; Mitchell, 1997; De'ath and Fabricius, 2000; Witten and Frank, 2000). A simple form of rule-based classifier is a hierarchical construction (tree) with various levels (leaves) (Figure 5.1). At each level a test is applied which is comprised of simple questions, the answer to each of which traces a path down the tree. The classification or prediction made by the model is determined when a final point is reached. The prediction may be qualitative (e.g., least vulnerable lakes) or quantitative (e.g., temperature greater than 20 C°).

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). 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 class (Breiman *et al.,* 1 984; Quinlan 1986; Ripley, 1996; De'ath and Fabricius, 2000).

The advantages of the decision tree approach (e.g., classification tree) over discriminant function analysis (DFA) are summarized in Table 5.1. At first glance, the DFA and classification tree decision processes may seem alike due to the use of coefficients and splitting equations by both methods. However, they differ significantly based on the simultaneous decision-making process of DFA as opposed to the hierarchical decision-making process of classification tree. In general, classification tree approaches offer several advantages over DFA in dealing with complex ecological datasets. The classification tree methods are not limited by prior knowledge of dataset distributions, since modeling of these distributions is not required. Thus, classification tree algorithms can easily handle multimodal distributions and they have no restrictions on sample size, in contrast to Bayesian approximators such as maximum likelihood and DFA classifiers.

5.1.2.1. Classification tree building process

Previous authors (Breiman *et al.,* 1984; Quinlan, 1986 and 1993; Ripley, 1996; Mitchell, 1997; De'ath and Fabricius, 2000; Witten and Frank, 2000) have provided detailed descriptions of decision tree procedures. Here, only the classification tree method is reviewed because it was used to implement the watershed-based classification in this study. 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 (Figure 5.2). 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 or reservoir classes) (Breiman *et al.*, 1984; Quinlan, 1993). Attributes that do not seem to contribute in defining ultimate terminal nodes are usually excluded in the final tree structure, leaving only those attributes that influence the overall classification process (Quinlan, 1993).

There are three basic components of classification tree building process: a set of questions; splitting criteria; rules for assigning a class at a terminal node. The set of questions could be in the form of a continuous explanatory variable (is pH ≤ 6.48?) or a categorical explanatory variable (is $z = b$?). The splitting criteria generally involve impurity function or information gain (or entropy) approach. The impurity function approach was developed by Breiman *et al.* (1984) and seeks to increase within group homogeneity by minimizing their impurity. The commonly used impurity measures are Gini diversity index, twoing rule or linear combination splits (Breiman *et al.,* 1984). For example, the Gini index (i) is defined as:

$$i(t) = 1 - S \tag{5.4}$$

where t = tree node, and the impurity function (S) is $S = \Sigma \rho^2(j/t)$,

$j = 1, 2, 3 \ldots\ldots k$ classes, such that;

i(t) is maximum, if ρ (1/t) = ρ (2/t)= ….. = ρ (j/t)

i(t) is minimum, i(t) =0, if all cases at a node belong to one class

Given *S* (splitting function) at tree node *t*, then a goodness-of-fit criterion (decrease in impurity) is applied as:

$$\Delta i(s,t) = i(t) - \{\rho L[i(tL)] + \rho R[i(tR)]\} \tag{5.5}$$

where s = a particular split;

ρL = the proportion of the cases at node t that go into the left child node (t L);

ρR = the proportion of cases at node t that go into the right child node (t R)

i (t L) = impurity of the left child node;

i (t R) = impurity of the right child node.

The impurity function approach is used as the primary rule in Breiman's CART software (Breiman, 1984; Salford Systems, 1998). Class assignment involves the use of either the plurality rule (assign terminal node to a class for which ρ (j/t) is maximum) or assigning terminal node to a class for which the expected misclassification cost is minimum. Since Breiman's classification tree process depends on probabilities of classification, it sometimes tends to mimic parametric statistical approaches (Quinlan, 1993).

Information gain (or entropy) criterion, on the other hand, involves the use of least amount of information (in bits) to describe each splitting decision at a node in the classification tree, based on the frequency of each class at that node (Shannon, 1948; Hunt *et al.*, 1966; Quinlan, 1993; Shannon and Weaver, 1999). According to Quinlan (1993), for any subset (S) of a population the number of observations in *S* that belong to class (C_j) can be described as *freq*(C_j, S). A "communication", indicating that a randomly selected observation belongs to some class C_j, has the probability {*freq* (C_j, S) / |S|}, where |S| is the absolute number of observations in the subset *S*. The information transmitted by the communication is defined as:

$$-\log 2\{\frac{freq(C_j,S)}{|S|}(bits) \tag{5.6}$$

A summation over the classes with respect to their frequencies in *S*, gives the expected information (in bits) on class membership from such a message as:

$$\text{info}\,(S) = -\sum(\frac{freq(C_j,S)}{|S|})*\log_2(\frac{freq(C_j,S)}{|S|}) \tag{5.7}$$

When equation 5.7 is applied to a training set of observations, *info*(T) provides a measure of the average amount of information required to identify the class of an object in *T*. This amount is also referred to as the entropy of the set *T*. Again, taking into account a similar measurement after *T* has been partitioned in accordance with *n* outcomes of a test *X*; then the expected information is computed as a weighted sum over the *T* subsets:

$$\text{info}_x(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|}.\text{info}(\text{T}_i) \tag{5.8}$$

Based on equation (5.8) above, the information that is gained by partitioning *T* in accordance with the test *X* is measured as:

$$gain\ (\text{X}) = info(\text{T}) - info_x(\text{T}) \tag{5.9}$$

According to Quinlan (1993), the *gain criterion* aims at selecting a test to maximize the information gain. However, the *gain criterion* has significant limitation of bias since it favors tests with many outcomes (Quinlan, 1993). This anomaly is resolved by a *gain ratio criterion* in which the potential information is generated by normalizing *T* into *n* subsets (Quinlan, 1993). The splitting information in equation (5.7) is then modified as:

$$\text{split info}(X) = \pm\sum_{i=1}^{n} \frac{|T_i|}{|T|} * \log_2(\frac{|T_i|}{|T|}) \tag{5.10}$$

Then the proportion of information generated by the split that aids the classification process is given by:

$$gain\ ratio\ (X) = gain\ (X)\ /\ split\ info_x(X) \tag{5.11}$$

If the split is relatively insignificant, the split information will be small and the gain ratio will become unstable. Consequently, the gain ratio criterion selects a test to maximize the gain ratio (5.11), subject to the constraint that the information gain should be large (Quinlan, 1993).

5.1.2.2. Classification tree pruning

The use of splitting criteria intuitively suggests that splitting is only stopped when there is no further improvement in the gain ratio or impurity function. However, a stopping rule based on splitting criteria could result in an overlarge tree that "over fits" the data. Large trees are complicated to interpret and have poor generalizing ability. Secondly, too large a criterion could blur splits based on attribute interactions unless one of the associated main effects is large enough to generate a split (Breiman *et al.*, 1984; Quinlan, 1993; Esposito *et al.*, 1999). The tree pruning process involves removing branches and subtrees that are generated due to noise; and when done properly, can improve classification accuracy as well as produce more interpretable and simplified trees (Figure 5.3). Typical approaches to classification tree pruning are "cost-complexity" pruning (Breiman *et al.*, 1984) and "reduced error" pruning (Quinlan, 1993).

The effectiveness of pruning methods is constrained when the dataset set is small (e.g. less than 100 samples) in which case the original tree is constructed on a smaller training set. This problem is resolved by obtaining estimates of prediction error; the accuracy of these estimates are usually increased by using an averaged weighted prediction error of several models as provided by k-fold cross-validation error (Stone 1974; Breiman *et al.*, 1984; Ripley, 1996; Ronchetti *et al.*, 1997; Esposito *et al.*, 1999; De'ath and Fabricius, 2000; Bloch *et al.*, 2002).

5.1.2.3. Decision tree software

The suite of recursive partitioning decision tree algorithms and software that have been developed over the last two decades include CHAID (Chi-squared Automatic Interaction Detection) (Kass, 1980), FACT (Loh and Vanichsetakul, 1988), Breiman's

CART® (Breiman ,1984; Salford Systems, 1998), C4.5 and C5/See5 (Quinlan, 1994; RuleQuest 2003) and OC1 (Murphy *et al.*, 1994). According to Lim *et al.* (2000), C4.5 is one of the best performing classification tree algorithms, based on comparisons of classification accuracy, training time and number of leave nodes for 32 different decision trees algorithms. Hence, the Microsoft Windows version of C4.5 (i.e. See5®) was employed in this study.

5.2. Methods

Previous research work (see Chapter 4) showed that 9 classes may be optimal in describing the inherent structure of Nebraska reservoir classes, and that soil organic matter was the key watershed characteristic that contributed to the segregation of these classes. Once the numbers of underlying reservoir groups were identified, a classification tree predictive model was used to describe the reservoir class structure and also to develop the rule-based classification for Nebraska reservoirs as a model for agriculturally dominated ecosystems. In this chapter, the classification tree-based watershed classification developed in Chapter 4 was compared to Omernick's Level IV ecoregions (Omernik, 1987; EPA, 2002) and discriminant function analysis (DFA)-based watershed classification methods (Momen and Zehr, 1998). The comparison was a two step process: first, the watershed-based classifications were compared to ecoregions to determine their abilities to account for variations in water quality parameters of Nebraska reservoirs; second, the classification tree-based reservoir classification was compared to DFA-based classification with respect to classification accuracy. Comparing different 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 DFA (implemented in SAS software) were considered without any accuracy enhancements (e.g. prior probabilities for DFA and boosting for classification tree respectively).

5.2.1. Ecoregions and water quality datasets

Ecoregions of Nebraska were extracted from a dataset of Omernik Level IV ecoregions of the Conterminous United States (Omernik, 1987; EPA, 2002). The United States ecoregions dataset was clipped to GIS polygon coverage of Nebraska using ArcMap GIS software. The water quality data for 78 sampled reservoirs were derived from existing sampled Nebraska lakes water quality dataset that was obtained from the School of Natural Resources, University of Nebraska – Lincoln (Holz, 2002). A GIS "point" coverage of the sampled reservoirs was overlaid on Omernik's ecoregions of Nebraska in order to identify those ecoregions that corresponded to the 78 sampled reservoirs (Figure 5.4).

The water quality data (collected between 1988 and 2003) were summarized into annual means, corresponding to sampling data obtained between May and August of each year. For each of the ecoregions (identified in figure 5.4.) and corresponding reservoirs, the mean value was determined for the candidate reference water quality parameters that have been proposed by the U.S. Environmental Protection Agency (EPA) for use in developing lake nutrient criteria. The candidate reference water quality parameters are chlorophyll-a, Secchi depth, total phosphorus, total nitrogen and alkalinity of lake waters (EPA, 2001; Severn *et al.,* 2001). In addition to the preceding water quality parameters, two potential agrochemical herbicide pollutants (Atrazine and Alachlor) were included in the analysis because the outcome of this study also has 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.

5.2.2. DFA-based reservoir classification

Discriminant analysis (DFA) was performed on watershed characteristics of 78 sampled reservoirs that were used in the classification tree-based watershed classification in Chapter 4. The distributional assumptions of DFA (that all the explanatory variables must follow a multivariate normal distribution for each class of response variable; and, that of equal variance-covariance matrices for each class) limit the validity of prediction error in assessing the accuracy of DFA classification. Besides, substantive interpretation of statistically significant discriminant functions requires structure coefficients, i.e. correlation of each explanatory variable with the discriminant functions (similar to factor loadings) (Bray and Maxwell, 1982; Legendre and Legendre, 1998).

Currently, rules of thumb such as structure coefficients greater than 0.3 or 0.4, are used to determine which variables load on a discriminant function (Legendre and Legendre, 1998; Gordon, 1999). However, the condition for including a variable in the interpretation of a discriminant function is that its structure coefficients must be significantly different from zero. For sampled ecological datasets with multimodal distributions and unequal variances, the structure coefficients may have a large and apparently important value but this may not be significantly different from zero (Legendre and Legendre, 1998; Johnson, 1999). It is, therefore, critical to employ other means for statistical tests of significance in order to ensure that the accuracy of DFA for sampled ecological datasets is valid for generalization.

Resampling approaches (jackknifing, bootstrapping and cross-validation) offer non-parametric means to perform statistical significance test of structure coefficients of DFA (Stone, 1974; Efron, 1979; Efron and Gong, 1982; Breiman *et al.* 1984; Wu, 1986; Efron and Tibshirani 1993; Shao, 1993; Ronchetti *et al.,* 1997; Legendre and Legendre, 1998; Good, 1999; Johnson, 1999; Efron, 2003). The jackknife and bootstrapping methods are used to compute Spearman's ranked correlations, bias-corrected estimates of standard error and confidence intervals, irrespective of the sampling distribution of the dataset. The jackknife approach involves resampling without replacement, while bootstrapping involves resampling with replacement (Efron, 1979; Efron and Gong, 1982; Efron and Tibshirani, 1993; Good, 1999; Davison *et al.,* 2003; Efron, 2003). 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 (Stone, 1974; Shao, 1993 Ronchetti *et al.,* 1997; Efron, 2003; Wehberg and Schumacher, 2004).

For this reason the cross-validation resampling technique was employed in DFA method that was used in this study. The DFA was implemented in SAS® software using "Discrim" procedure with cross-validation option (SAS Institute, 2000). The DFA was performed using output of the cluster analysis of watershed characteristics datasets based on 13 and 9 classes respectively (see Chapter 4). This was done to explore the effectiveness with which the DFA could handle the more complicated 13-class data (involving 13 classes and single object classes) as compared to the less complicated 9-class dataset. Accordingly, the cross-validation prediction errors were determined for both 13 and 9 class datasets respectively.

The predicted reservoir classes (13 and 9 classes respectively), derived from the DFA-based watershed classification, were then extracted and ArcMap GIS was used to append this information to a watershed characteristics dataset that included predicted reservoir classes with respect to classification tree based watershed classification (13 and 9 classes) and ecoregions. The dataset also included annual mean summaries, of growing season index period, for water quality parameters (chlorophyll-a, Secchi depth, alkalinity, total phosphorus, total nitrogen, Atrazine and Alachlor).

5.2.3. Comparison of classification methods

The watershed-based classifications methods, DFA and classification tree (See5®), were compared to ecoregions regarding their abilities to account for variations in water quality parameters of Nebraska reservoirs. This was done using the concept of classification strength, which measures of how strongly different landscape classification approaches separate reference 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 as:

$$CS = \frac{\varpi}{\beta} \tag{5.12}$$

Where:

ß is variability in reference water quality conditions between classes

ϖ = variability in reference conditions within classes; ϖ is the overall weighted mean of within class variances (ϖ_i) (modified from Van Sickle and Hughes, 2000). The variance in mean annual water quality is given as:

$$\sigma^2 = \sum \frac{(x_i - X)^2}{n-1} \qquad \text{i} = 1, 2 \ldots n \text{ reservoirs} \qquad (5.13)$$

where

x = the annual mean value of water quality (e.g. chlorophyll-a) for each reservoir

X = the sample mean

n = 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:

i. Biophysical water quality (chlorophyll-a and Secchi depth)

ii. Chemical nutrient water quality (total phosphorus, total nitrogen and alkalinity)

iii. Agrochemical herbicide effluents (Atrazine and Alachlor)

Since the aim of the dissertation research 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 represents an increase in interclass heterogeneity or increase in within-class homogeneity. Consequently, the classification approach with the lowest CS value for the respective water quality categories was considered to be most optimal.

5.3. Results and discussions

A map of the sampled reservoirs overlaid on Omernik's Level IV ecoregions of Nebraska is shown in figure 5.4. There were 20 out of the 27 Nebraska ecoregions that corresponded to the sampled reservoirs locations. However, only 9 of these ecoregions had sufficient water quality data or more than one reservoir per ecoregion class. As such, 9 ecoregion classes were used in the comparisons of classification methods.

5.3.1. Comparison of classification methods

The classification strength (CS) of ecoregions, DFA and classification tree (See5®) based classifications are shown in table 5.2.a. These results were summarized for comparisons with respect to biophysical, chemical nutrient and agrochemical herbicide effluents water quality categories (Table 5.2.b). 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 DFA) were more effective than ecoregions in accounting for the variations in water quality characteristics of Nebraska reservoirs. The DFA method was most effective in segregating 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 5.2), the relatively high CS value for alkalinity lessens the effectiveness of ecoregions. 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, 1983; Wetzel and Likens, 2000). The above results were 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 discrimination between lakes of different water quality and suggested that the spatial distribution of lake ecosystems is more complicated than that presented by ecoregion boundaries.

Geospatial data employed in this study are available for the entire United States, e.g., US Geological Survey's Elevation Derivatives for National Applications (EDNA) datasets which are based on a seamless 30-meter resolution DEM available for the conterminous United States (Verdin and Verdin, 1999; Gesch *et al.*, 2002; http://edna.usgs.gov/). Thus, the comparison between watershed-based classifications and ecoregions derived reservoir classes has potential national applications that can address the concerns of Omernik and Bailey (1997) regarding incompatible comparisons between ecoregions and other ecological boundaries.

Is it important to note however that, the use of classification strength is assessing the effectiveness of classification methods is dependent sampled water quality data. Box-whisker plots were generated for each water quality parameter that was used in the classification strength comparisons ecoregions, DFA, and classification tree methods respectively (Appendix III). Log transformations of the water quality parameters helped alleviate the asymmetric distributions of the water quality data. The box-plots highlight the extent of variation in the water quality data. In general, the plots in Appendix III could provide a useful context for any interpretation of the classification strength comparisons between watershed-based reservoir classification methods and ecoregions-derived reservoir classes. Hence classification strength is to some extent affected by limitations of sampling in-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 See5® classification tree and DFA classifications was based on their cross-validation prediction errors (Table 5.3). The results showed that the classification tree method was more effective in handling the 13-class dataset than the DFA classification method. Also, the differences in prediction error rate between the 13-class and 9-class datasets are 9.49 and 30.30 for classification tree and DFA methods respectively. Despite the smaller prediction error for DFA with regards to the 9-class dataset, the significant jump in prediction error from the 13-class dataset shows how perturbations or complexities in a dataset can reduce the predictive effectiveness of the DFA method.

Thus the above results confirm the assertion that classification trees are most useful in dealing with complex datasets, such as ecological data (Breiman *et al.*, 1984; Quinlan 1993; German *et al.*, 1999; De'ath and Fabricius, 2000). Also, the results of comparisons between DFA and classification tree methods were in agreement with previous analyses on water quality, geospatial datasets (images and maps), and soft coral datasets (Emmons *et al.,* 1999; German *et al.*, 1999; De' ath and Fabricius, 2000). 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 DFA-derived classes.

5.3.2. Interpretive classification interface

Based on the classification tree models (Figures 5.5), an interpretive classification interface was developed to predict the classes to which different reservoir samples belong (Figures 5.5). This interface is particularly useful to water resource managers interested in identifying the class of a particular lake. A "classifier" button in See5® classification tree software invokes the interpreter interface; using the most recent and relevant classification tree, interpreter interface prompts for information about the new case to be classified (RuleQuest Research, 2003). For example, the classification interface was used to predict the class membership of Yankee Hill reservoir and it showed that Yankee Hill reservoir will belong to class 1 with 72 percent probability based on soil organic matter, erodibility and mean watershed slope (Figure 5.6).

5.4. Summary

A theoretical basis for comparing classification methods was described in this chapter. A classification tree-based reservoir watershed classification, developed and described in Chapter 4, was compared to Omernik's Level IV ecoregions and discriminant function analysis (DFA)-based watershed classification methods. The comparison was done to first evaluate the abilities of watershed-based classifications and ecoregions to account for variations in water quality parameters of Nebraska reservoirs; and second, to determine the predictive effectiveness of classification tree and DFA based reservoir watershed classification methods.

Sampled Nebraska reservoirs (78) were grouped into various classes using the above-mentioned classification approaches; namely, classification tree and DFA based reservoir watershed classifications and ecoregions derived reservoir classes. Also, annual mean summaries for water quality parameters (chlorophyll-a, Secchi depth, alkalinity, total phosphorus, total nitrogen, Atrazine and Alachlor) were generated and appended to classification tree, DFA, and ecoregions derived reservoir classes respectively. A classification strength metric (measures of how strongly different landscape classification approaches separates reference water quality water conditions) was used to evaluate the effectiveness of watershed-based reservoir classifications and ecoregions derived reservoir classes. The results suggested that both watershed-based classification approaches (classification tree and DFA) were more effective than 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 DFA-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 reservoir classes. The above results confirm previous conclusions that decision trees are more suited for the ecologically complex datasets than traditional statistical approaches (e.g., DFA) to resource classification.

However, 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). This limitation can be overcome by exploring expert systems (e.g., conditional probability networks) to incorporate prior knowledge of watershed characteristics and water quality parameters in a post-classification process to refine the results of the decision tree classification (Lauritzen and Spiegelhalter, 1988; Neapolitan, 1990; Heckerman, 1997).

Much of the known relationships between watershed characteristics and water quality parameters have been derived using parametric statistical methods such as correlation and linear regression analysis. Regression trees non-parametric approach (e.g., Cubist® by RuleQuest Research) 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 (watershed characteristics) in assessing relationships or associations among the variables of interest.

Results of such associations can be used in either pre-processing the input variables of classification tree modeling to enhance the splitting process or incorporated into post-classification expert systems to refine the classification tree modeling results. The regression tree derived associations between watershed characteristics and water quality parameters can also be used to rank reservoir watersheds using ArcMap GIS weighted combination method (ESRI, 2001). The ranking may be from most vulnerable to least impacted watersheds for determining reference conditions in each predetermined reservoir class. Water quality standards or "targets" can then be developed based on reference water quality conditions, for reservoirs in each class. The lake reference conditions are quantitative descriptions of "ideal" lake conditions used as standard of comparison. Although reference conditions are intended to portray pristine environmental conditions, it is generally recognized that they realistically portray least impacted or most sustainable conditions (Hughes, 1995; EPA, 2000; EPA, 2001). The least impacted watersheds are indicative of candidate lakes sites for the development of reference water quality conditions, e.g. via paleolimnological coring.

An important feature of the See5® classification tree software is the interpretive classification interface that was developed to predict the classes to which new cases belong. A "classifier" button in See5® classification tree software invokes the interpreter interface; using the most recent and relevant classification tree, interpreter interface prompts for information about the new case to be classified (RuleQuest Research, 2003). 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 in question.

References

Bloch, D.A., R.A. Olshen and M.G. Walker. 2002. *Risk estimation for classification trees*. **Journal of Computational and Graphical Statistics**. 11(2):263-288.

Bray, J.H. and S.E. Maxwell. 1982. *Analyzing and interpreting significant MANOVAs*. **Review of Educational Research**. 52:340-367.

Breiman, L., J. H. Friedman, R.A. Olshen and C. J. Stone. 1984. **Classification and Regression Trees**. Wadsworth, Inc. Belmont, California. 358p.

Chen, L. 2003. *A study of applying genetic programming to reservoir trophic state evaluation using remote sensor data*. **International Journal of Remote Sensing.** 24(11): 2265-2275.

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

Davison, A.C., D.V. Hinkley and G.A. Young. 2003. *Recent developments in bootstrap methodology*. **Statistical Science**. 18(2):141-157.

De' ath, G and K.E. Fabricius. 2000. *Classification and regression trees: a simple yet powerful technique for ecological data analysis.* **Ecology.** 8(11):3178-3192.

DeFries, R.S., and J. Chan. 2000. *Multiple criteria for evaluating machine learning algorithms for land cover classification from satellite data*. **Remote Sensing of Environment**. 74: 503-515.

Detenbeck, N.E., C.M. Elonen, D.L. Taylor, L.E. Anderson, T.M. Jicha, and S.L. Batterman. 2004. *Region, landscape, and scale effects on Lake Superior tributary water quality*. **Journal of the American Water Resources Association.** 40 (3): 705 – 720.

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

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

Efron, B. 2003. *Second thoughts on the bootstrap*. **Statistical Science**. 18(2):135-140.

Efron, B. 1979. *Bootstrap methods: another look at the jackknife*. **Annals of Statistics**. 7: 1-26

Efron, B. and G. Gong, 1982. *A leisurely look at the bootstrap, jackknife and cross-validation.* **The American Statistician**. 82:171-185.

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

Emmons, E.E., M.J. Jennings and C. Edwards. 1999. *An alternative classification method for northern Wisconsin lakes*. **Canadian Journal of Fisheries and Aquatic Sciences.** 56 (4):661-669.

EPA (U.S. Environmental Protection Agency). 2001**. Nutrient Criteria Technical Guidance Manual for Lakes and Reservoirs.** Report No. EPA-822-B00-001. Washington, D.C.

EPA (U.S. 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.

Fisher, R.A. 1936. *The use of multiple measurements in taxanomic problems*. **Annals of Eugenics**. 7:179-188.

Friedl, M.A., C. Woodcock, S. Gopal, D. Muchoney, A.H. Strahler and C. Barker-Schaaf. *2000. A note on procedures used for accuracy assessment in land cover maps derived from AVHRR data.* **International Journal of Remote Sensing**. 21 (5): 1073 – 1077.

Friedl, M.A. and C.E. Brodley. (1997). *Decision tree classification of land cover from remotely sensed data*. **Remote Sensing of Environment**. 61(3): 399-409.

German, G. W. H., G. A.W. West and M. G. Gahegan. 1999. **Statistical and AI Techniques in GIS Classification: A comparison.** Proc. of the 11th Annual Colloquium of the Spatial Information Research Centre, University of Otago, Dunedin, New Zealand, December 1999. CD-ROM.

Gesch, D., M. Oimoen, S. Greenlee, C. Nelson, M. Steuck and D. Tyler. 2002. *The National Elevation Dataset*. **Photogrammetric Engineering and Remote Sensing**. 68(1): 5-11.

Good, P. 1999. **Resampling Methods. A Practical Guide to Data Analysis**. Birkhauser, Boston, MA.

Gordon, A. 1999. **Classification,** 2nd Edition. Chapman and Hall. London. 256p.

Hansen, M., R. Dubayah, and R.S. DeFries. 1996. *Classification trees: An alternative to traditional land cover classifiers*. **International Journal of Remote Sensing.** 17(5): 1075-1081

Hawkins, C.P., R.H. Norris, J. Gerritsen, R.M. Hughes, S.K. Jackson, R.K. Johnson and R. J. Stevenson. 2000. *Evaluation of landscape classifications for the prediction of freshwater biota: synthesis and recommendations*. **Journal of the North American Benthological Society.** 19(3): 541-556.

Heckerman, D. E. 1997. *Bayesian Networks for Data Mining.* **Data Mining and Knowledge Discovery**. 1:79-119.

Holz, J.C. 2002. **Lake And Reservoir Classification in Agriculturally Dominated Ecosystems.** EPA 2002 Aquatic Ecosystem Classification Workshop, Denver, CO, September, 2002, oral presentation, invited.

Huberty, C.J. 1994. **Applied Discriminant Analysis.** John Wiley and Sons, New York.

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

Jenerette, G.D., J. Lee, D. Waller and R.E. Carlson. 2002. *Multivariate analysis of the Ecoregion delineation for aquatic ecosystems*. **Environmental Management.** 29 (1): 67- 75.

Johnson, D.H. 1999. The insignificance of statistical significance testing. **Journal of Wildlife Management**. 63 (3):763-772.

Kachigan, S.K. 1986. **Statistical Analysis**. Radius Press. New York.

Kass, G.V. 1980. *An exploratory technique for investigating large quantities of categorical data*. **Applied Statistics**. 29: 119-127.

Klecka, W.R. 1980. **Discriminant Analysis**. Sage Publications, Beverly Hills, California.

Knapp, T.R. 1978. *Canonical correlation analysis: a parametric significance testing systems*. **Psychological Bulletin**. 85:410-416.

Lamon, E.C. and C.A. Stow. 2004. *Bayesian methods for regional-scale eutrophication models*. **Water Research.** 38(11): 2764-2774.

Lauritzen S. L. and D. J. Spiegelhalter. 1988. *Local computations with probabilities on graphical structures and their application to expert systems.* **Journal of the Royal Statistical Society**. Vol. 50 (2): 157-224.

Legendre, P. and L. Legendre. 1998. **Numerical Ecology.** 2^{nd} English edition. Elsevier Science. BV, Amsterdam. 853 pp.

Lim, T.S., W.Y. Loh and Y.S. Shih. 2000. *A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms.* **Machine Learning**. 40 (3): 203-228

Loh, W.Y and N. Vanichsetakul. 1988. *Tree structured classification via generalized discriminate analysis.* **Journal of American Statistical Association.** 83:715-728.

Michaelson, J., F. Davis, and M. Borchert. 1987. *Non-parametric methods for analyzing hierarchical relationships in ecological data.* **Coenoses.** 1: 97-106.

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

Mitchell, Tom. M. 1997. **Machine Learning.** New York: McGraw-Hill. 414p.

Momen, B. and J.P. Zehr. 1998. *Watershed classification using discriminant analyses of lake water-chemistry and terrestrial characteristics.* **Ecological Applications.** 8 (2):497-507.

Murphy, S.K., S. Kasif and S. Salzberg. 1994. *A system for induction of oblique decision trees.* **Journal of Artificial Intelligence Research**. 2:1-32.

Neapolitan, R. E. 1990. **Probabilistic Reasoning in Expert systems: Theory and Algorithms.** John Wiley and Sons, New York. 448p.

Omernik, J.M., 2003. *The misuse of hydrologic unit maps for extrapolation, reporting and ecosystem management.* **Journal of the American Water Resources Association**. 39(3):563–573.

Omernik, J.M., 1987. *Ecoregions of the Conterminous United States.* **Annals of the Association of American Geographers.** 77:118-125.

Omernik, J.M. and R.G. Bailey. 1997. *Distinguishing between watersheds and ecoregions.* **Journal of the American Water Resources Association.** 33(5):935–949.

Park, Y.S., P.F.M. Verdonschot, T.S. Chon and S. Lek. 2003. *Patterning and* *predicting aquatic macroinvertebrate diversities using artificial neural network.* **Water Research**. 37(8): 17749-1758.

Quinlan, J. R. 1986. *Induction of decision trees.* **Machine Learning.** 1(1): 81–106.

Quinlan, J. R. 1993. **C4.5: Programs for Machine Learning.** Morgan Kaufmann Publishers Inc., CA.302p.

Rao, 1952. **Advanced Statistical Methods in Biometric Research**. John Wiley and Sons, New York. 390p

Ripley, B.D. 1996. **Pattern Recognition and Neural Networks.** Cambridge University Press. 403p

Rogan, J., J. Miller, D. Stow, J. Franklin, L. Levien and C. Fischer. 2003. *Land-cover change monitoring with classification trees using Landsat TM and ancillary data.* **Photogrammetric Engineering and Remote Sensing**. 69(7): 793-804

Rohm, C.M., J.M. Omernik, A.J. Woods and J.L. Stoddard. 2002. Regional characteristics of nutrient concentrations in streams and their application to nutrient criteria development. **Journal of the American Water Resources Association.** 38 (1): 213-239.

Ronchetti, E., C. Field and W. Blanchard. 1997. *Robust linear model selection by cross-validation.* **Journal of the American Statistical Association.** 92(439):1017-1023.

RuleQuest Research. 2003. **See5: An Informal Tutorial**. http://rulequest.com/see5-win.html.

Salford Systems 1998. **CART® Software**. Salford Systems, San Diego, CA.

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

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

Shannon, C. E. 1948. *A mathematical theory of communication.* **Bell System Technical Journal.** 27: 379-423 and 623-656.

Shannon, C. and W. Weaver. 1999. **The Mathematical Theory of Communication**. University of Illinois Press. 5th Ed. Urbana, IL, 144p.

Shao, J. 1993. *Linear model selection by cross-validation.* **Journal of the American Statistical Association.** 88(422):486-494.

Stone, M. 1974. *Cross-validatory choice and assessment of statistical predictions.* **Journal of the Royal Statistical Society**. Series B Vol. 36: 111-147.

Tabachnick, B.G. and L.S. Fidell. 2001. **Using Multivariate Statistics.** 2nd Edition. Harper and Row. New York.

Van Sickle, J. and R.M. Hughes. 2000. *Classification strengths of ecoregions, catchments and geographic clusters for aquatic vertebrates in Oregon.* **Journal of the North American Benthological Society**. 19:370-384.

Verbyla, D.L. 1987. *Classification trees: a new discrimination tool.* **Canadian Journal of Forestry Research.** 17:1150–1152.

Verdin, K.L. and J.P. Verdin. 1999. *A topological system for delineation and codification of the Earth's river basins.* **Journal of Hydrology**. 218:1 – 12.

Wehberg, S. and M. Schumacher. 2004. *A comparison of nonparametric error rate estimation methods in classification problems.* **Biometrical Journal**. 46(1): 35-47.

Wetzel, R.G. 1983. **Limnology.** Saunders College Publishing, Philadelphia, PA. 767p.

Wetzel, R.G. and G.E. Likens. 2000. **Limnological Analysis.** 3rd Ed., Springler Verlag, New York. 432p.

Winter, T.C. 1999. *The relation of streams, lakes, and wetlands to groundwater flow systems*. **Hydrogeology Journal.** 7:28–45.

Witten, I. H., and E. Frank. 2000. **Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.** Morgan Kaufmann. San Diego, CA. 371p.

Wu, C.F.J. 1986. *Jackknife, bootstrap and other resampling methods.* **Annals of Statistics**. 14: 1261-1350.

Yang, L.M., C.Q. Huang, C.G. Homer, B.K. Wylie and M.J. Coan, 2003. *An approach for mapping large-area impervious surfaces: synergistic use of Landsat-7 ETM+ and high spatial resolution imagery.* **Canadian Journal of Remote Sensing.** 29 (2): 230-240.

Decision tree classification	Discriminant function Analysis (DFA)
Inherently nonparametric: makes no assumptions of the distribution of the values of predictor (explanatory) variables	Inherently parametric: assumes normally distributed data; variance and covariance matrices for each class must be equal
Can handle numerical data of explanatory variables that are highly skewed or multimodal	Explanatory variables must follow a multivariate normal distribution for each class of response variable
Can handle categorical data with either ordinal or non-ordinal structure	Only suitable for continuous (numerical) predictor variables
Not influenced by outliers, collinearities, and heteroskedaticity in datasets	Sensitive to data anomalies, e.g. outliers
Deals effectively with cases of missing values of explanatory or predictor variables by making use of collinear variables in "surrogate" splits	Not effective in accounting for cases of missing explanatory variables, hence variables with missing values are usually dropped from analyses
Identifies splitting variables based on an exhaustive search of all possible alternatives	Segregation of variables is based only on linear combinations of explanatory variables
The inverted tree structure makes output classes simple to understand and interpret	Interpretation of statistically significant discriminant functions requires structure coefficients (or discriminant scores)
Handles hierarchical and non-linear relationships among predictor variables very well	Can only handle linear relationships among predictor variables
Has no restriction on sample size	Prediction accuracy usually decreases after a minimum threshold of sample size is reached
Can detect and reveal salient variable interactions	Variable interactions must be explored using other analyses prior to discriminant analyses

Table 5.1. Differences between classification tree and discriminant function analysis (DFA) classification algorithms

	Classification Strength (CS = W/B*)		
	DFA Watershed Classes	Ecoregions	See5 Watershed Classes
Water Biophysical Parameters			
Secchi Depth	1.520	1.053	1.538
Chlorophy-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**

Table 5.2.a. Comparison of classification strength of reservoir classification methods
* - W is within class variation
B is between class variations
** - Adjusted mean value of total nitrogen was used in this analysis

	Mean Classification Strength		
	DFA* Watershed Classes	Ecoregions	See5® ** Watershed Classes
Number of classes	8	9	8
Water Biophysical Parameters	2.305	3.523	3.215
Water Chemistry Parameters	1.961	2.649	1.897
Agrochemical herbicide effluents	1.649	1.472	1.310

Table 5.2.b. Summary of mean classification strength values for reservoir classification methods
* DFA was implemented using SAS® "Discrim" procedure (SAS Inc., 2000)
** Classification tree was implemented using See5® software (RuleQuest, 2003)

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

Table 5.3. Comparison of prediction strength for watershed-based reservoir classification methods

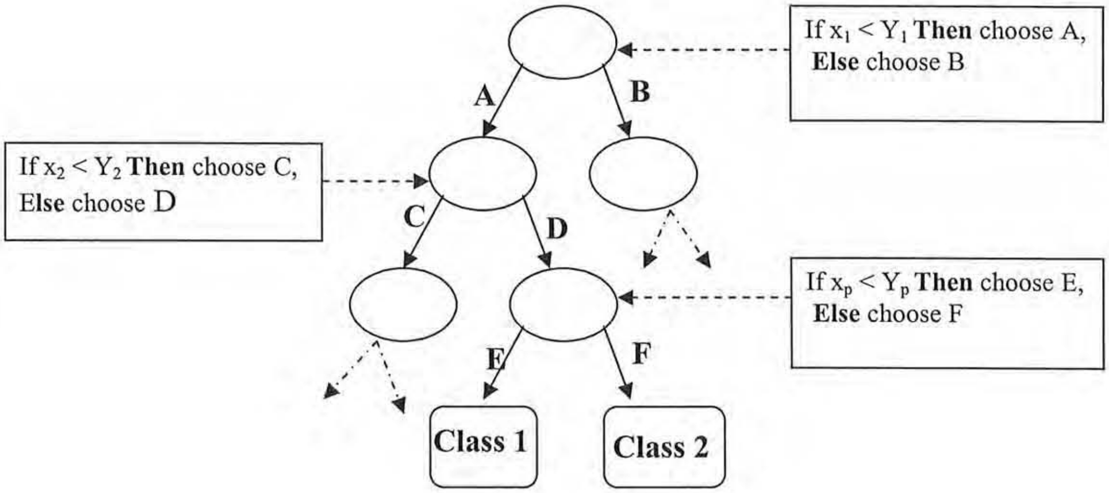

If $x_1 < Y_1$ **Then** choose A,
Else choose B
A
B
If $x_2 < Y_2$ **Then** choose C,
Else choose D
C
D
If $x_p < Y_p$ **Then** choose E,
Else choose F
E
F
Class 1
Class 2

Figure 5.1. Schematic representation of a simple decision tree process

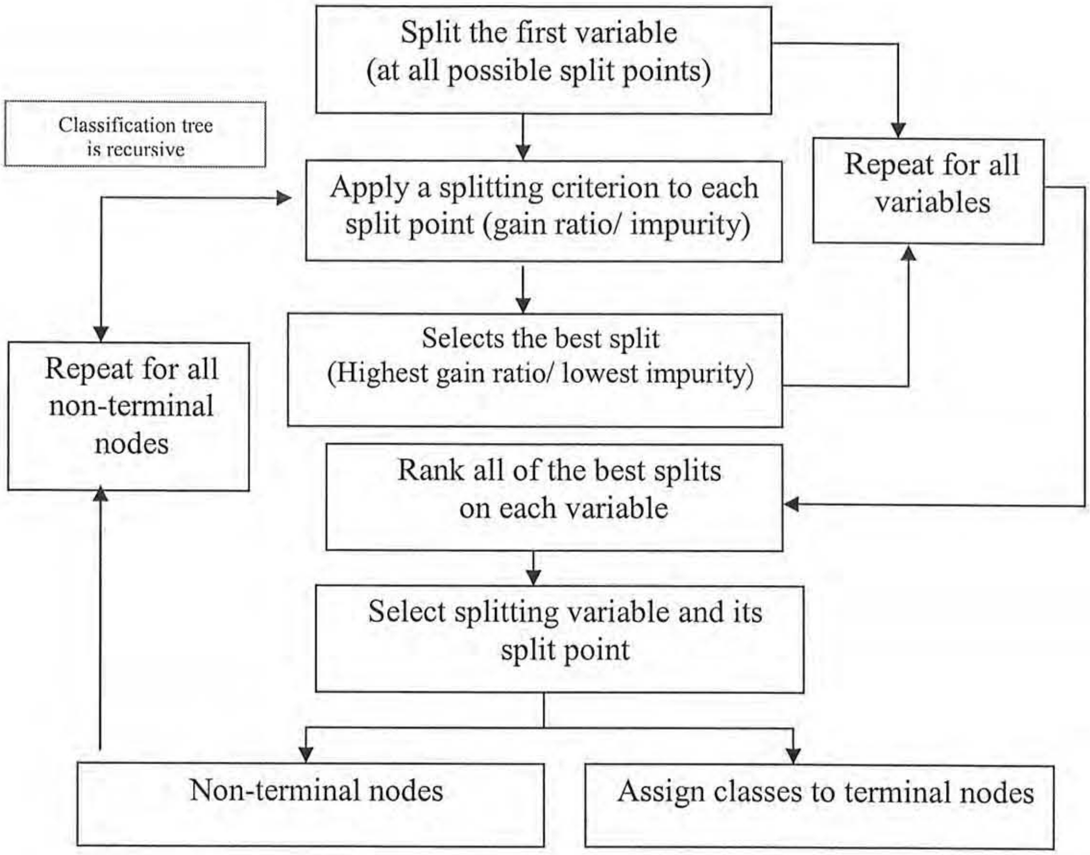
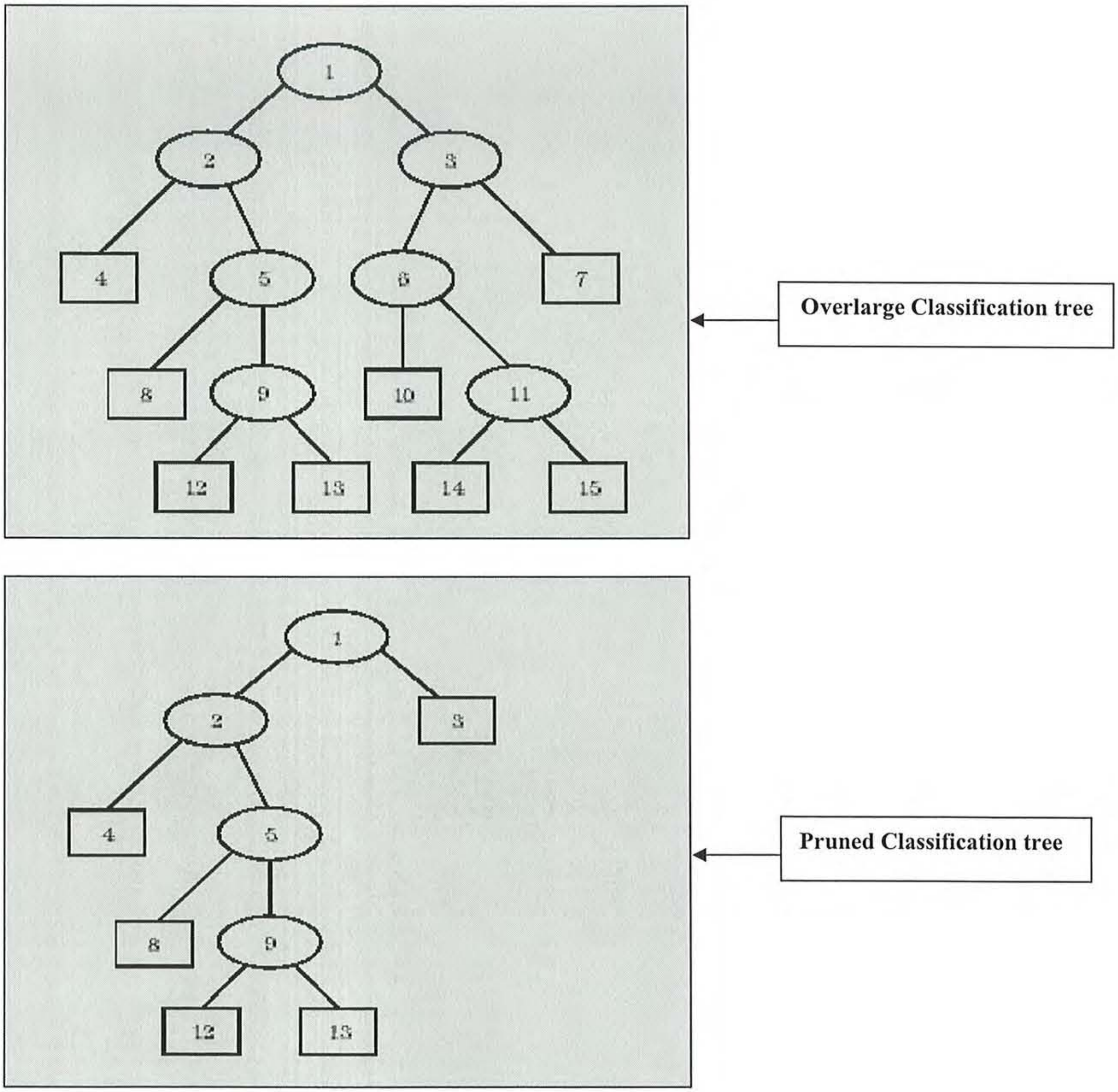

Split the first variable
(at all possible split points)
Classification tree
is recursive
Apply a splitting criterion to each
split point (gain ratio/ impurity)
Repeat for all
variables
Repeat for all
non-terminal
nodes
Selects the best split
(Highest gain ratio/ lowest impurity)
Rank all of the best splits
on each variable
Select splitting variable and its
split point
Non-terminal nodes
Assign classes to terminal nodes

Figure 5.2. Schematic representation of the recursive partitioning procedure of classification tree algorithms.



Overlarge Classification tree

Pruned Classification tree

Figure 5.3. Schematic diagrams showing an example of classification tree pruning process

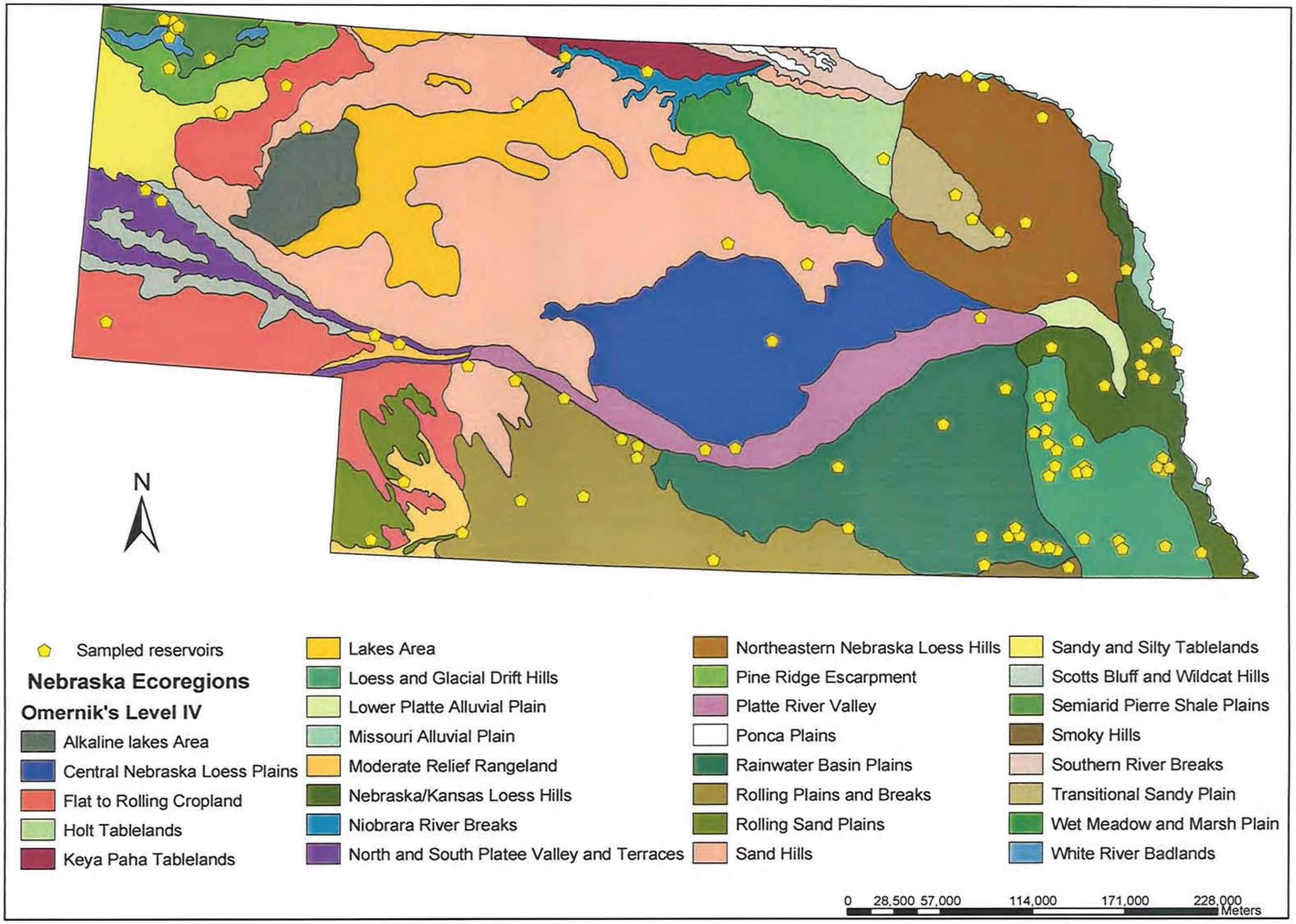


Figure 5.4. Sampled reservoirs sites overlaid on Omernik's Level IV Ecoregions of Nebraska

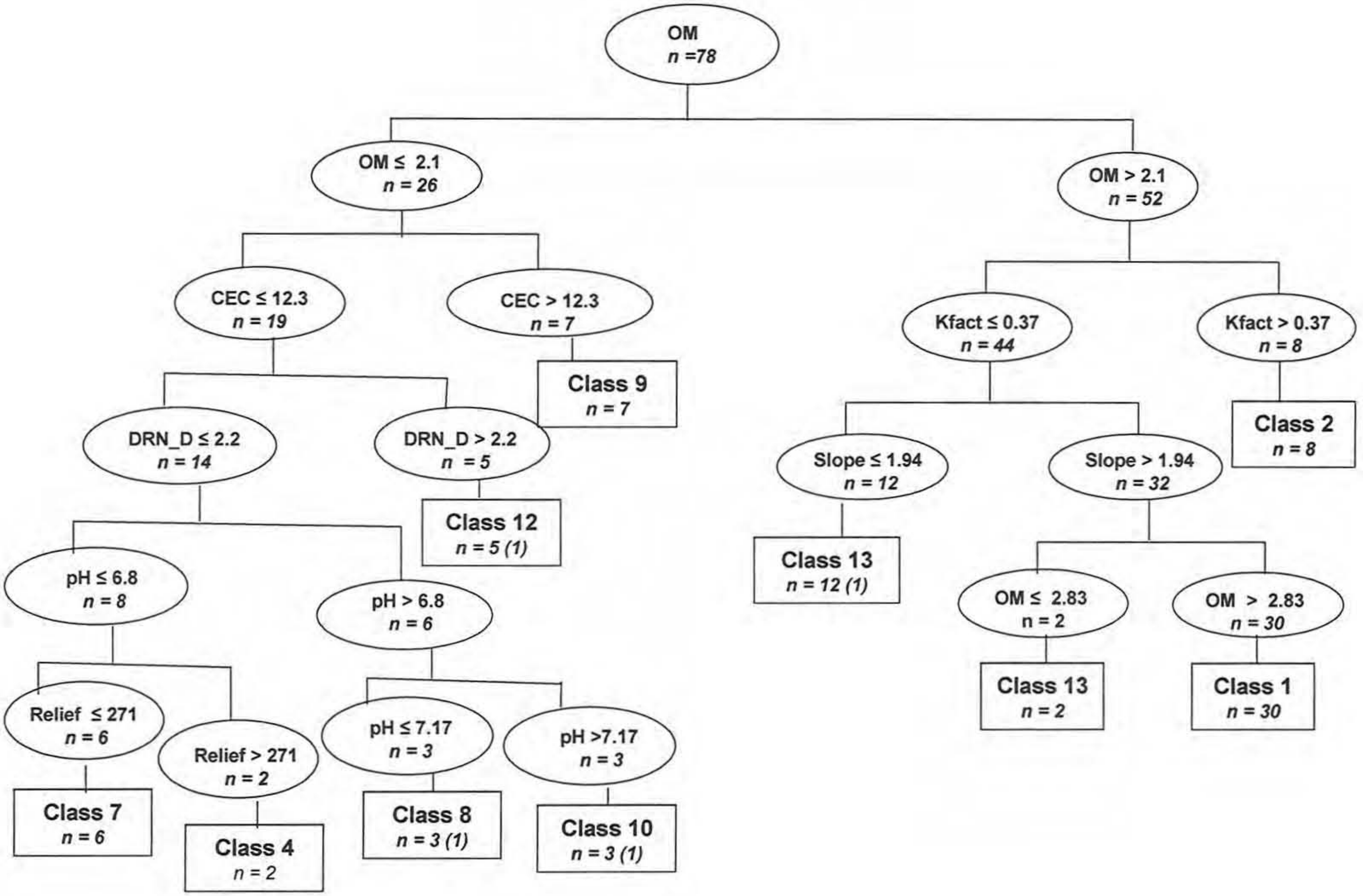
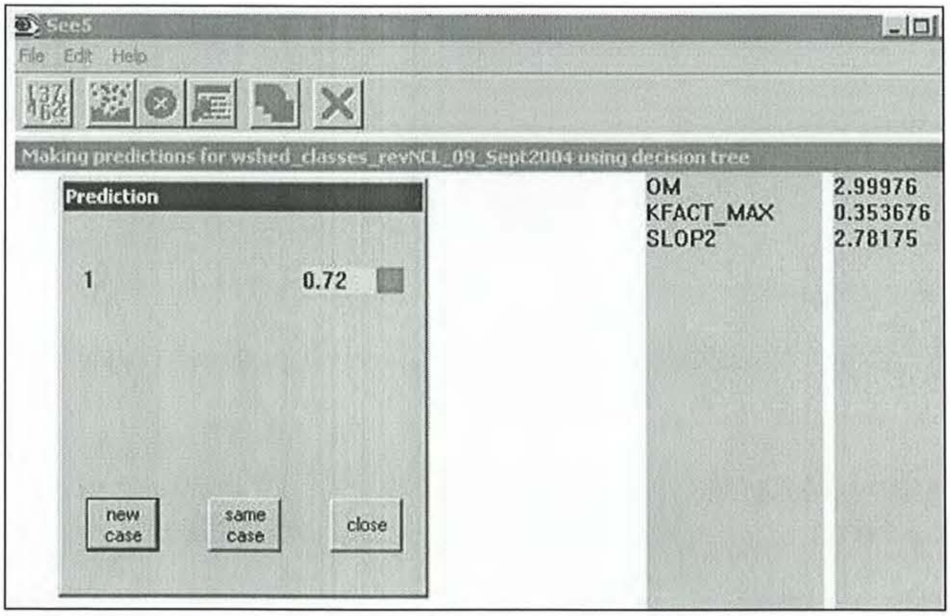


Figure 5.5. Classification tree for Nebraska reservoir classes. Rectangular boxes represent terminal nodes (classes); oval boxes represent non-terminal nodes that required further splitting.



Making predictions for wshed_classes_revNCL_09_Sept2004 using decision tree

Prediction

1 0.72

OM	2.99976
KFACT_MAX	0.353676
SLOP2	2.78175

Figure 5.6. Example of interpretative classification interface used to predict class membership of Yankee Hill Reservoir in southern part of Lincoln, Nebraska.

CHAPTER 6. CONCLUSIONS AND RECOMMENDATIONS

"The watershed approach is one of the most important environmental guiding principles...; failure to fully incorporate the watershed approach into program implementation will result in failure to achieve our environmental objectives in many of our nation's waters".

Assistant EPA Administrator G. Tracy Mehan, III (Mehan, 2002)

6.1. Summary

Public agencies such as the U.S. Environmental Protection Agency (EPA) are charged with establishing reasonable attainable water quality standards. The principal objective of this research was to define and test a watershed-based classification procedure for identifying groups of lakes that have similar potential capacity to meet proposed water quality standards. This dissertation research focused on reservoirs in Nebraska, an agriculturally-dominated area of the United States.

In this dissertation research, I proposed an approach describing the class structure of Nebraska reservoirs based classification of the reservoir watershed conditions. This approach was based on the premise that, in the absence of human interference, lake ecosystems evolve in response to physical, chemical and biological processes in their watersheds. Since my interest was in modeling reasonable attainable water quality standards for groups of lakes that are considered to share similar potential capacity to meet these standards, human factors such as land use were excluded from the analysis.

A watershed-based, decision tree classification procedure was developed. Results suggest that Nebraska reservoirs can be represented by 9 classes, and that soil organic matter content in the watershed was the most important single variable for classifying the watersheds. Comparison of the watershed-based decision tree classification approach that was employed in this study with other methods showed that: overall the watershed-based classification approach performed better than Omernik's Level IV ecoregions in accounting for variations in water quality characteristics of Nebraska reservoirs; and that the decision tree classification method was more effective in handling complex reservoir data than a discriminant analysis-based watershed classification method.

Information on the number and structure of lake classes is useful to water quality managers for many applications, including predictive modeling of potential impairment of reservoirs water quality based on their class membership. In addition, the findings of this research are important to the EPA nutrient criteria (water quality standards) development process as they demonstrate an intuitive method to identify lake classes and the environmental conditions that are pertinent to these classes. The interpretive classification interface provides a simple graphic user interface that eliminates the need for in-depth background in statistical analysis in order to use the decision tree-classification method.

The classification procedure is also useful for other applications, such as determining the categories of other resource and mapping problems. By using the cluster validation approach to determine optimal number of classes, as described in this dissertation, researchers and GIS analysts can reduce the extent of arbitrary selection of the number of resource classes.

6.1.1. Geospatial dataset development and preliminary analysis

This study focused on classification of reservoir watersheds and hence accurate identification of the locations of Nebraska lakes, thus it was necessary to delineate their watershed boundaries. Because there was no existing map that provided a complete and accurate depiction of the number and locations of lakes in Nebraska, an updated database of lake locations was developed using several data sources and ArcMap GIS software. I also employed a simple automated means to delineate reservoir watershed boundaries that has potential national applications. Although the dataset development process was not a primary objective of the dissertation research, it was critical to the study.

Outputs of the dataset development process include an up-to-date and comprehensive Geographic Information System (GIS) coverage of Nebraska lakes, a vital product needed to correctly identify Nebraska reservoirs, delineate watershed boundaries using digital elevation models (DEMS) and extract data on watershed characteristics. Comparisons of DEM-derived watershed boundaries with manually digitized watershed boundaries, obtained from the Nebraska Department of Natural Resources (DNR), showed less than 10 percent deviation based on such watershed parameters as drainage area and drainage density. This implies that the automated watershed delineation method produced watershed boundaries that were as good as manually-derived boundaries. Geospatial data employed delineating watershed boundaries are available for the entire U.S. and the automated GIS-based procedures for watershed delineation are also nationally available. Thus, the watershed-based decision tree reservoir classification described in the preceding chapters has potential national applications.

The sampled reservoirs made up 8.39 percent of all Nebraska reservoirs that are at least 4 hectares in size. Comparisons of sampled reservoirs with all Nebraska (based on lake area, climate divisions and ecological regions) indicated that there was no significant difference between the sampled reservoirs and Nebraska reservoirs larger than 4 hectares. Therefore, conclusions of the study could be assumed to be applicable to most Nebraska reservoirs.

6.1.2. Implementation of a classification system for Nebraska reservoirs

Determining the optimal number of classes is a vital step in developing an effective classification strategy. This is because most often, water resource classifications are based on arbitrary choices of the number of classes to be employed. Such a practice limits our understanding of the inherent structure and hence the biophysical characteristics of the lake classes. The use of cluster validation techniques coupled with predictive strength evaluation of the number of lake classes provide a quantitative basis for identifying the optimal number of classes. Decision trees were very useful supervised classification tool and in describing the structure of the reservoir classes, as well as identifying key watershed characteristics that contributed to the segregation of the lake classes.

A cluster analysis was performed on the watershed characteristics of 78 sampled Nebraska reservoirs in order to determine the optimal number of Nebraska reservoir classes. A plot of the Pseudo-F statistic (obtained from the cluster analysis output) against the respective number of classes (NCL), suggested that the potential number of classes included 3, 5, 13, 17, and 19. Further analysis of the optimal NCL was done based on the predictive strength of the potential NCL's using See5® classification tree software. The outcome of the classification tree modeling suggested that the optimal number of Nebraska reservoir watershed classes was 13 NCL. The cross-validation prediction error of the classification tree model for reservoir watersheds was 26.33

percent. The classification tree was later used to describe the structure of the Nebraska reservoir classes, and soil organic matter content was found to be the most important single variable for segregating the watersheds. Finally, the initial 13 NCL was revised based on the number of nodes in the classification tree, indicating that Nebraska reservoirs can be represented by nine optimal classes. The spatial distribution of these reservoir watershed classes was described, reflecting the hydrogeological and biogeographical pattern of terrain, soil and climate conditions of Nebraska.

6.1.3. Comparison of reservoir classification methods

Classification tree-based reservoir watershed classification was compared to Omernick's Level IV ecoregions and discriminant function analysis (DFA)-based watershed classification methods; first, the watershed-based classifications were compared to ecoregions to determine their abilities to account for variations in water quality parameters of Nebraska reservoirs; second, the classification tree-based reservoir classification was compared to DFA-based classification with respect to classification accuracy.

A classification strength metric was used to evaluate the effectiveness of watershed-based reservoir classifications and ecoregions derived reservoir classes. The results suggested that both watershed-based classification approaches (classification tree and DFA) were more effective than Omernik's Level IV ecoregions in accounting for the variations in water quality characteristics of Nebraska reservoirs. This result was in agreement with previous findings that, despite their usefulness in structuring environmental and natural resource research and management, Omernik's Level IV ecoregions may not adequately account for variations in lake water quality parameters.

Also, the classification tree and DFA-based watershed classification methods were compared with respect to their cross-validation prediction errors. This comparison showed that the classification tree method was more effective than DFA-method in handling complex watershed characteristics dataset and reservoir classes. These results confirm previous observations that decision trees are more suited for ecologically complex datasets than traditional statistical approaches (e.g., DFA) to resource classification.

Even though comparing classification methods can be problematic due to the different ways each classifier can be set up, the two-step comparison process that was employed in this study provided a substantive means to determine how classification methods can account for variations in water quality parameters, as well as a measure of classification accuracy. It was apparent that the classification tree method is a promising new tool for classifying lakes in order to set water quality standards and explore management implications these lakes.

However, 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). Therefore, it is important to explore means to incorporate meaningful associations between watershed characteristics and water quality parameters in order to improve the results of a classification tree analysis. Also, the classification tree (decision trees or "inductive" machine learning in general) concept is relatively new and users are subject to some limitations including skepticism of decision tree methodologies based on unrealistic claims and poor performance of earlier models.

6.2. Conclusions

The first objective of this dissertation was met through the use of a novel cluster validation procedure, coupled with predictive strength analysis of the potential number of classes (NCL), to determine the inherent groups of Nebraska reservoirs. A classification tree provided further insights into the structure of the watershed-based reservoir classes and the environmental conditions that contributed significantly to the segregation of these classes. The classification tree was also informative in highlighting the characteristics of the watershed-based reservoir classes, as well as the need to revise the number of classes from 13 to 9.

The second research objective was achieved through the classification tree-based watershed classification algorithm to predict the class membership of new reservoir cases and comparisons with ecoregions and traditional statistical approaches to reservoir classification. The results of these comparisons substantiated the premise of the watershed-based reservoir classification and also provided further proof that classification trees are more suitable than discriminant analysis in handling ecologically complex datasets. It is also important to note that there are options (e.g., boosting) available to the analyst to improve the prediction accuracy of the See5® classification tree. However, only the default options of See5® were used in this study in order to ensure a pragmatic comparison with other classification methods.

Although successful, there are some factors that could limit broad applications of the results of this study. These limitations include:

i. The use of small scale STATGO dataset to extract watershed characteristics information, because the more detailed SSURGO datasets were not completed for Nebraska at the time of this research.

ii. K-means clustering algorithms have inherent tendencies to aggregate most of the observations into a few classes, are sensitive to outliers and susceptible to the choice of cluster centroids (starting points)

iii. No options in classification tree algorithms to incorporate prior knowledge of known relationships between watershed characteristics and lake water quality, in order to improve the classification results

iv. Although the watershed boundary delineation method employed in this study proved to be effective, it is only applicable to stream fed lakes (mostly reservoirs). As such, groundwater-fed lake types were excluded from this study because the hydraulic divide of groundwater table does not coincide with the topographic divide; some natural lakes and sand pits are therefore likely to have relatively small or negligible surface watersheds. However, the delineation of hydraulic-divide of groundwater is limited by the lack of a detailed map of the water table. The process of converting current bore-hole water levels to groundwater hydraulic-divide would require a major project to complete (Gosselin and Chen, *pers. comm.*).

v. Despite the graphic user interface provided by the See5® classification tree software, some resource managers and research analysts do not have in-depth background in machine learning algorithms. This may limit the incorporation of classification trees as part of the suite of decision support systems.

6.3. Recommendations for future research

The aforementioned limitations clearly suggest the need for additional investigations. Additional work will be required to:

i. Compare the advantages of using higher resolution watershed characteristics datasets, i.e. 1:24,000 scale Soil Survey Geographic (SSURGO), to the 1:250,000 scale STATSGO derived watershed characteristics dataset that was employed in this study. The SSURGO dataset is currently being completed nationwide, so such a comparison may highlight the potential improvements in lake classification results and merits of the new SSURGO datasets in establishing lake nutrient water quality standards across the United States.

ii. Address limitations of k-means clustering and compare the performance of existing modifications or alternatives to k-means clustering is needed. This is because k-means clustering has limitations such as sensitivity to outliers or extreme values, susceptibility to the choice of starting points (cluster centroids), and tendency to produce classes with most data points concentrated in a few classes.

iii. Determine quantitative relationships between watershed characteristics (both categorical and numeric explanatory variables) and water quality (numeric dependent variables) using regression trees, e.g. Cubist® regression tree software (RuleQuest, 2003). Such information from regression trees analysis can be used in either pre-processing the input variables of classification tree modeling to enhance the splitting process or incorporate into post-classification expert systems to refine the classification tree modeling results.

iv. Explore the use of expert systems (e.g., conditional probability networks) to incorporate prior knowledge of explanatory variables (watershed characteristics) and water quality (dependent variables) in a post-classification process to refine results of classification tree analysis.

v. Explore options to estimate the contributing catchment areas for ground water fed lakes (natural lakes and sand pits) through the integration of existing groundwater well data, GIS models and remotely sensed ground water level datasets. Once this is done, the classification procedure described in this study can be applied for natural lakes and sand pits.

vi. The See5® classification tree software already provides a user-friendly graphic user interface for the prediction of the class membership of new reservoirs. There is a need to integrate See5® classification tree interface and ArcMap® GIS to develop a user-friendly suite of "one-shop" decision support tools for water resource managers and GIS analysts. This can be done using the open source codes for incorporating classification tree procedure into other applications.

References

Mehan, G.T. 2002. **Committing EPA's Water Program to Advancing the Watershed Approach.** EPA memo to Regional Water Division Directors. December 3, 2002. http://www.epa.gov/owow/watershed/memo.html. Accessed February 27, 2004.

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

RuleQuest Research. 2003. **See5: An Informal Tutorial.** http://rulequest.com/see5-win.html.