A Learning Vector Quantization Based Geospatial Modeling Approach for Inland WQ Remote Prediction

Sudhanshu S Panda¹, Indrajeet Chaubey ², Debu Mishra³

Abstract:
Quick and accurate quantification of lake water quality (WQ) is essential for its management and improvement. Use of geotechnology (remote sensing, GIS, and GPS) applications is a step forward in improving our ability to effectively quantify and manage the WQ of ungauged lakes. Beaver Reservoir, a drinking water source for over 280,000 people in northwest Arkansas, is facing increased chlorophyll-a (Chl-a) and suspended matter (SM) content in the lake. This study is designed to qualitatively predict the Chl-a and SM content in the lake on a spatial basis from Landsat-TM image digital information. A Learning Vector Quantization (LVQ) classification neural model was used to predict the qualitative (oligotrophic, moderately oligotrophic, and mildly trophic) classes for several spatial positions in the lake. The geostatistical tool in ArcGIS was used to spatially map the Chl-a and SM extent around the lake. The LVQ classification model predicted the Chl-a extent with more than 90% accuracy having only one point misclassified out of 14 testing points. The LVQ model prediction for SM resulted in four misclassified points out of 14 testing points with prediction accuracy of 72%. The final spatial zoning maps for Chl-a and SM extent in the entire lake could be used to help water use managers and end users design management strategies, and also demonstrates a relatively low cost WQ prediction mechanism that can be applied in developing countries and elsewhere when detailed in-situ monitoring is not feasible.

Key words: LVQ; Universal Kriging; geotechnology (remote sensing, GIS, GPS); chlorophyll; suspended matter

¹ Professor; Institute for Environmental Spatial Analysis, University of North Georgia, 3820 Mundy Mill Road, Oakwood, GA 30566, USA; Tel.: 1-678-507-3594; e-mail: Sudhanshu.Panda@ung.edu
² Dean; College of Agriculture, Health And Natural Resources, University of Connecticut, 1376 Storrs Road, Storrs, CT 06269 USA
³ Professor; College of Engineering and Mines, University of Alaska Fairbanks, PO Box 755800, Fairbanks, AK 99775-5800 USA
Quick and accurate quantification of lake water quality (WQ) is essential for its management and improvement. Reservoir WQ assessment usually involves expensive, cumbersome and time consuming in-situ water sample collection measurement, and subsequent laboratory analyses (e.g. see Gubala et al, 1993, Panda et al., 2005). Due to these limitations, sampling often cannot cover the entire water body. Time and cost constraints associated with in-situ measurements of lake WQ often limit assessment of spatial and temporal trends of WQ. Optical and radar remote sensing has potential to monitor WQ over a greater range of temporal and spatial scales as a quick and inexpensive approach (Dekker et al., 1996; Zhang et al., 2003; Panda et al., 2005). The use of geospatial tools including geographic information systems (GIS) and global positioning system (GPS), along with digital databases of archived remote sensing data, is a step forward in improving our ability to effectively quantify and manage the WQ of ungauged lakes.

Chlorophyll-a (Chl-a), suspended matter (SM), and dissolved organic matter are optically active parameters of lake WQ. CChl-a, the primary photosynthetic pigment of all oxygen evolving photosynthetic organisms and present in all algae, is a biomass indicator of productivity in both terrestrial and aquatic ecosystems (Dall’Olmo et al., 2003). It has two in vitro absorption bands at 660-665 nm and 430 nm of visible spectrum and in the near infrared (NIR) region (Wetzel, 1983). The highest absorption happens in 675 nm and 770 nm band for CChl-a presence in water (Dall’Olmo and Gitelson, 2005). Optical and radar remotely sensed data have been used widely to estimate major WQ variables such as Chl-a, turbidity, SM concentration, surface water temperature, wave height, and sea surface roughness. Several investigators have successfully used Landsat-MSS/TM and AVHRR imagery in inland and estuarine WQ monitoring (Lathrop and Lillesand, 1986, 1991; Ritchie et al., 1990; Lathrop, 1992; Braga et al., 1993; Dekker and Peters, 1993; Choubey, 1994; Keiner and Yan, 1998; Woodruff et al., 1999; Brivio et al., 2001; Baruah et al., 2001; Wang and Ma, 2001). Gholizadeh et al. (2016) in their review paper, detailed that various properties (spectral, spatial and temporal, etc.) of the more commonly employed spaceborne and airborne sensors are being used to estimate chlorophyll-a(Chl-a), colored dissolved organic matters (CDOM), Secchi disk depth (SDD), turbidity, total suspended sediments (TSS), water temperature (WT), total phosphorus (TP), sea surface salinity (SSS), dissolved oxygen (DO), biochemical oxygen demand (BOD) and chemical oxygen demand (COD) with efficiency.

In the case of inland water, the water dynamics are too complex to have a linear relationship between the satellite spectral signatures and the WQ parameters (Panda et al., 2005). There is considerable scattering (in all visible and even in near-IR bands) from lake waters with high sediment and chlorophyll content (Baruah et al., 2001; Zhang, 2005). For nonlinear environmental processes, artificial neural networks (ANN) are useful because of their ability to model nonlinear geophysical transfer functions. Other reasons to use ANN to estimate Chl-a and SM in lakes are: 1) there is no single spectral channel onboard remote sensors in which the effect of a single absorption (and backscattering) component can be independently dominating on the effect of other water components; and 2) ANN with a hidden layer can simulate various complex functions. Indeed, ANNs are able to model non-linear radiative transfer functions with higher accuracy than those algorithms of
traditional regression analyses (Ranaweera et al., 1995; Keiner, 1999; Krasnopolsky et al., 2000, Panda and Panigrahi, 2000; Panda et al., 2005). According to Lacroix et al. (1997) and Haykin (1999), ANNs have the power to approximate any non-linear relationship that exists between a set of inputs and their corresponding set of outputs through learning procedures. Dalmini et al. (2016) assessed the feasibility of integrating remote sensing and in-situ measurements in monitoring water quality status of Lake Chivero, Zimbabwe. They used MODIS images to quantify the lake chlorophyll_a concentrations and obtained high R² value of 0.89 and a root mean square error (RMSE) value of 0.003 μg/L with prediction results based on the relationship between observed and predicted chlorophyll_a (Dalmini et al., 2016). Hansen et al. (2017) confirmed the ability of remote sensing in estimating inland lake water quality through a process of selecting appropriate sensors that are suitable for the spatial and temporal variability in the water body. They also determined appropriate uses of near-coincident data in empirical model calibration along with recognizing potential limitations of remote sensing measurements which are biased toward surface and near-surface conditions (Hansen et al., 2017).

Back-propagation neural networks (BPNN) and Radial basis function networks (RBFN) have found increased application in pattern recognition, signal processing, load forecasting, environmental modeling, and WQ estimation, due to their structural simplicity and training efficiency (Goodman, 1993; Ranaweera et al., 1995; Walczak and Massart, 1996; Lee et al., 1996; Keiner and Yan, 1998; Wan and Harrington, 1999; Baruah et al., 2001; Panda et al., 2005). Anthwal and Pandey (2016) has used ANN to effectively mapping the eutrophication level inland water bodies. These ANNs are mostly used for quantitative or discrete modeling. However, in addition to quantitative information, qualitative spatial modeling can be an essential requirement for water managers. Learning Vector Quantization (LVQ) ANN models can be used for qualitative WQ modeling of lake or other terrestrial water bodies on a non-discrete scale. The LVQ neural technique is a classification network which assigns vectors to one of several classes (Kohonen, 1986 and 1988) and hence can be correlated to various WQ scenarios. Thus, a spatial Chl-a and SM estimation mapping would be possible within the lake using the LVQ neural technique, helping water managers in WQ management on specific portions of the lake.

Application of LVQ modeling in quantifying Lake WQ has not been done. However, LVQ was used by Bellantone et al. (2002) to classify the daily air quality of Rome, Italy. It is hypothesized that the LVQ architecture will provide satisfactory results to predict Lake WQ using satellite imagery information. The objective of this study was to develop an LVQ based spatial modeling technique to predict the spatial distribution of Chl-a and SM concentrations using the Landsat-TM satellite imagery, validated through simultaneously collected spatially distributed in-situ WQ samples from Beaver lake.

**LVQ algorithm application process in the study**

An LVQ network is an ANN that contains a Kohonen layer that learns and performs classification by assigning input vectors to one of several classes (Neural Ware, 2005). LVQ provides equal numbers of processing elements (PEs) for each class in the Kohonen layer. Neural Works (Neural Ware, Carnegie, PA) includes an implementation of LVQ networks that contains an input
layer, a Kohonen layer, and an output layer. A typical LVQ architecture is shown in Figure 1. For further explanations and descriptions of the LVQ model algorithms refer to DeSieno (1988) and Neural Ware (2005). One application of LVQ neural modeling is to use LVQ2 (a secondary aspect of LVQ model after LVQ1) along with LVQ1 or LVQ1 with conscience, since LVQ2 improves the boundary between regions where misclassifications occur. LVQ2 is applied only in situations where the closest prototype vector is in the wrong class and the second closest prototype vector is in the correct class, with respect to the input vector. In this study, both LVQ1 and LVQ2 are concurrently used for the Chl-a and SM prediction/classification.

**Fig. 1** Architecture of the LVQ network used in the study.

**METHODOLOGY**

The study was conducted in the Beaver Reservoir located in NAD 83 UTM coordinates of 389237 to 449127 m N and 3952319 to 4038348 m E (Figure 2). Beaver Reservoir is the primary drinking water source for more than 280,000 people in northwest Arkansas, one of the most rapidly growing metropolitan areas in the U.S. Eutrophication of the reservoir is a great concern for long-term Beaver Reservoir WQ management.
Fig. 2 Location of the study area, watershed stream network, USGS gauging stations, and other water sample collection stations in Beaver Lake in northwest Arkansas.

Image and WQ data acquisition

Ten Thematic Mapper (TM) cloud free images from two path/row combinations (25/35 and 26/35) covering Beaver Reservoir were used in this study (Table 1).

Table 1 Landsat TM-5 image acquisition and the corresponding water sampling dates.

<table>
<thead>
<tr>
<th>Year</th>
<th>Dates Water sampling</th>
<th>Dates Image acquisition</th>
<th>Path/Row</th>
<th># of sampling points</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>April 18</td>
<td>April 17</td>
<td>25/35</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>June 13</td>
<td>June 13</td>
<td>26/35</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>October 16</td>
<td>October 19</td>
<td>25/35</td>
<td>2</td>
</tr>
<tr>
<td>2002</td>
<td>July 10</td>
<td>July 9</td>
<td>26/35</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>July 24</td>
<td>July 21</td>
<td>26/35</td>
<td>2</td>
</tr>
<tr>
<td>2003</td>
<td>July 21</td>
<td>July 21</td>
<td>25/35</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>August 6</td>
<td>August 7</td>
<td>25/35</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>December 19</td>
<td>December 19</td>
<td>26/35</td>
<td>12</td>
</tr>
<tr>
<td>2004</td>
<td>February 21</td>
<td>February 21</td>
<td>26/35</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>April 4</td>
<td>April 2</td>
<td>25/35</td>
<td>10</td>
</tr>
</tbody>
</table>
The TM images were precision corrected including radiometric and geometric corrections. Six spectral bands: band 1 (TM1), 0.45 – 0.52 \( \mu m \); band 2 (TM2), 0.52 – 0.60 \( \mu m \); band 3 (TM3), 0.63 – 0.69 \( \mu m \); band 4 (TM4), 0.76 – 0.90 \( \mu m \); band 5 (TM5), 1.55 – 1.75 \( \mu m \); and band 7 (TM7), 2.08 – 2.35 \( \mu m \) of TM were used to determine the Chl-a and SM concentrations in water. Band 6 (57 m resolution) was not used because of the image resolution dissimilarity with other bands that were 28.5 m in resolution.

Since 2003 water samples were collected on cloud free dates from different spatial positions of Beaver Reservoir coinciding with TM acquisition dates of the Beaver Reservoir scenes. In addition, Chl-a measured data of two United States Geological Service (USGS) gauging stations (07049500 and 07049691) located in the Beaver Reservoir were also used in this study. The TM scenes were acquired coinciding with cloud free sampling dates of USGS gauging station in 2001 and 2002. The spatial positions of each water sample collection point including the USGS gaging stations are shown in Figure 2.

Water samples were collected from three depths (surface, 1m below surface, and 2 m below surface) for each location. For each water sample, 4 L of water was collected and stored in the dark on ice until returned to the laboratory for analyses. Upon returning to the laboratory 1000 to 1500 ml of each sample was filtered using Gelman GF/C 47-mm filters for Chl-a and SM determination. Filters for Chl-a extraction were macerated in 5 ml 90% acetone. Extracts were then cleared by centrifugation and analyzed spectrometrically (APHA, 1998). Filters for determination of SM were preweighed. After sample filtration, filters were oven dried at 105° C for 24 h and reweighed to determine SM. Average concentrations of Chl-a and SM from three depth samples were used for modeling.

The spatial positions of water sample points were recorded with a Trimble ProXRS GPS with real-time and differential correction after the data were collected. One of the major sources of GPS error is loss of accuracy caused by satellite geometry, generally known as dilution of precision (DOP). For this study DOP in position (PDOP) and DOP in horizontal position (HDOP) were kept to average minimums of close to 2, ideal for water sample collection. Water sampling points were also selected at distances from the edge of the reservoir necessary to dispel the probability of the edge effects (mixing of different land uses) in the satellite images.

Compared to lakes, reservoirs typically have a dendrite shape and longer shorelines. Beaver Reservoir also has a dendrite shape (Figure 2) and its width is small at many of our sampling points. Therefore, we used single pixel gray values instead of 3x3 window values from each WQ sampling to avoid pixels falling outside the water body. Single pixel gray values (digital numbers, DN) of the collected water sample spatial position were extracted from the TM data and used in the modeling effort. A studentized t-test was applied for outlier detection, and one outlier sample data point was eliminated from the database. The gray value of the outlier data was exceptionally high for each band of the image as compared to the other digital numbers. Geomatica 9.1 (PCI Geomatics, Richmond Hill, Ontario, Canada) and IDRISI 32.2 (IDRISI Production, Worcester, MA) software were used for image processing. Figure 3 describes the image processing and the input dataset preparation procedure for LVQ modeling.
**Fig. 3** Image processing and data preparation for LVQ model development procedure.

**Input and output dataset preparation**

Input data sets were prepared using DN values of different bands of TM images for both the Chl-a and SM prediction modeling. The dataset included six DN values: TM1, TM2, TM3, TM4, TM5, and TM7. ANN models using the simple DN values from the TM bands proved better for Chl-a and SM discrete prediction than the models using band indices, or a combination of simple DN values and band indices together (Panda et al., 2005). Therefore, no other datasets were used for this LVQ WQ prediction spatial modeling purpose. For LVQ modeling the input dataset was scaled down to values between zero and one using min-max scaling option of the software. The equation used for the scaling technique was

\[
X = \frac{X_i - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}
\]  

where \(X_i\) is the value of the raw input variable, \(X\), for the \(i^{th}\) training case; \(X_{\text{max}}\) is the maximum value of a training case; and \(X_{\text{min}}\) is minimum value of a training case in the dataset.

For Chl-a modeling 64 observations were available for analysis: 50 were randomly selected as training and 14 were selected as testing observations. For SM modeling only 54 observations were available. Out of these, 40 observations were randomly selected for model training and the remaining 14 were selected for model testing.

The LVQ ANN technique has the ability to predict qualitative output classes (or groups) from input data. The model learns from user-defined input training datasets which relate to any one of the output classes. Therefore, it was essential to create Chl-a and SM classes based on an empirical technique. Carlson (1977) devised the trophic state index (TSI) for water bodies. Table 2 shows different trophic states based on TSI generated using the Chl-a concentration in the water body. Equations 2 and 3 were used to generate TSI for each sample point with Chl-a and secchi depth amounts, respectively.

\[
\text{TSI}_{\text{Chl-a}} = 30.6 + 9.81 \times \ln \text{[Chl-a]} \, (\mu \text{g/L})
\]  

\[
\text{TSI}_{\text{SM}} = 60 - 14.41 \times \ln \text{[Secchi depth]} \, \text{(in meters)}
\]
Based on the TSI for Chl-a laboratory testing data, only three classes were present in the dataset which include both training and testing data. The classes were class1 (oligotrophic) with TSI <40, class2 (moderate) with TSI 40-50, and class3 (mildly eutrophic) with TSI 50-60 (Table 2). We used laboratory generated SM data for our LVQ spatial modeling and divided the SM output into three classes based on our practical experience. The output classes were class1 (oligotrophic) with SM amount from <1.5 mg/l, class2 (moderate) with SM amount 1.5-3.0 and class3 (eutrophic) with SM amount >3.0. All these outputs (Chl-a and SM) were transformed into three different classes (CL1, CL2, and CL3) based on the discussed algorithm and procedure.

<table>
<thead>
<tr>
<th>TSI</th>
<th>Trophic Status Index &amp; WQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 30</td>
<td>Oligotrophic; clear water; high DO throughout the year in the entire hypolimnion</td>
</tr>
<tr>
<td>30-40</td>
<td>Oligotrophic; clear water; possible periods of limited hypolimnetic anoxia (DO =0)</td>
</tr>
<tr>
<td>40-50</td>
<td>Moderately clear water; increasing chance of hypolimnetic anoxia in summer; fully</td>
</tr>
<tr>
<td></td>
<td>supportive of all swimmable/aesthetic uses</td>
</tr>
<tr>
<td>50-60</td>
<td>Mildly eutrophic; decreased transparency; anoxic hypolimnion; macrophyte problems; warm-water fisheries only; supportive of all swimmable/aesthetic uses but &quot;threatened&quot;</td>
</tr>
<tr>
<td>60-70</td>
<td>Blue-green algae dominance; scums possible; extensive macrophyte problems</td>
</tr>
<tr>
<td>70-80</td>
<td>Heavy algal blooms possible throughout summer; dense macrophyte beds; hypereutrophic</td>
</tr>
<tr>
<td>&gt; 80</td>
<td>Algal scums; summer fish kills; few macrophytes due to algal shading; rough fish dominance</td>
</tr>
</tbody>
</table>

LVQ network architecture

Fifty training observations and 14 testing observations were prepared for the LVQ Chl-a prediction modeling in the Neural Ware Professional II software. The initial LVQ model architecture was 6-5-3, i.e., comprised of 6 input neurons (IPE), 5 Kohonen nodes, and 3 output neurons (OPE). Following a protocol of the software, the 5 Kohonen nodes represent 10% of the 50 training data observations (Neural Ware, 2005). The 3 output nodes represent the three classes of Chl-a concentration in the reservoir. As a Neural Ware software protocol, the number of initial iterations used for the learning process of the network was fixed at 10,000 with 1,000 iterations for LVQ2. The LVQ2 phase was also activated for the learning iterations. Again, the LVQ protocol suggested that the initial learning rate for the LVQ1 be less than 0.1. Therefore, an initial learning rate of 0.06 and 0.03 was used for LVQ1 and LVQ2, respectively. A global learning rate schedule (Table 3) was created to show the learning rates reduced over each learning phase. The conscience factor for the network was selected as one and the frequency estimation was used as 0.001. The min-max table was selected to reduce the data range between 0 and 1.

<table>
<thead>
<tr>
<th></th>
<th>For LVQ1</th>
<th>For LVQ2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning count</td>
<td>4000</td>
<td>7000</td>
</tr>
<tr>
<td>Attraction rate1 (Initial learning rate)</td>
<td>0.05</td>
<td>0.025</td>
</tr>
<tr>
<td>Attraction rate2</td>
<td>0.05</td>
<td>0.025</td>
</tr>
<tr>
<td>Repulsion rate</td>
<td>0.05</td>
<td>0.025</td>
</tr>
<tr>
<td>Conscience factor</td>
<td>1.0</td>
<td>0.05</td>
</tr>
<tr>
<td>Frequency estimation</td>
<td>0.001</td>
<td>0.0005</td>
</tr>
<tr>
<td>LVQ2 width</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
An LVQ model optimization technique was followed to obtain the optimum classification result using the dataset. Initially, the LVQ network was optimized by changing the number of nodes in the Kohonen layer while the other network parameters were kept constant. When the Kohonen node optimization was obtained, the initial learning rate was optimized. Based on the same procedure, the network iterations were also optimized. The network parameter optimization was observed based on the root mean square error (RMSE), desired and predicted class correlation coefficient, and network classification rate that is calculated based on the following equations:

\[
RMSE = \sqrt{\frac{\text{SSE}}{n}}
\]  

(4)

where \(n\) is the number of observations, and \(\text{SSE}\) is sum of squared error.

\[
\text{classification rate} = \frac{\text{Number of correctly classified observations}}{\text{Total number of observations in the class}}.
\]  

(5)

Information on correct classification rate and misclassification rate was generated by the software. The confusion matrix (desired-actual) information generated by the software was also used as another method for choosing the optimum LVQ network parameter for model building. The best model efficiency in Chl-a prediction was achieved when the correct classification rate was highest, RMSE was lowest, and correlation coefficient (desired-actual) was highest. The global learning rate schedule was modified each time for model parameter optimization. A similar procedure was followed for LVQ-SM prediction modeling. The only difference was in the training and testing dataset preparation. Forty training and 14 testing observations were used for the model building. Therefore, the initial network architecture was 6-4-3.

Spatial modeling

Spatial variability of Chl-a or SM can be described based on the following definitions in geostatistics (Goovaerts, 1997). For a given Chl-a or SM \(Z\), the experimental semivariogram \(\gamma(h)\) measures the dissimilarity between \(Z(u_i)\) and \(Z(u_i+h)\) separated by a vector \(h\), called lag distance (equation 6).

\[
\gamma(h) = \frac{1}{2N(h)} \sum_{\alpha=1}^{N(h)} [z(u_i) - z(u_i + h)]^2
\]  

(6)

where \(Z(u_i)\) and \(Z(u_i+h)\) are the data values of Chl-a or SM \(Z\) at spatial location \(u_i\) and \(u_i+h\), respectively. \(N(h)\) is the number of data pairs with lag \(h\) for omni-directional semivariogram. The autocorrelation can be examined and quantified while modeling the semivariogram. Once the semivariogram is obtained at different lag distances, it can be represented through different models such as nugget, spherical, exponential, and Gaussian models. The final model used in this study was a semivariogram model with different model structures at various ranges, such as Nugget = 0.6, Sill = 0.4 + 0.6, and a structure value of 0.6 that construct the spherical model. The description of the model
(equation) used in this study was \(0.21842 \times \text{Spherical}(0.21561) + 0.37741 \times \text{Nugget}\). This model equation was best suited for our semivariogram modeling.

Figure 4 shows the cartographic model used for the spatial qualitative mapping of Chl-a and SM amounts in Beaver Reservoir. All the randomly chosen testing data points from several water samples (Table 1) were appended together into a single point attribute layer. The Chl-a and SM predicted results (qualitative classes) were added to the attribute table of the point attribute layer in the geodatabase as separate attributes. Similarly, the actual Chl-a and SM class values were also added to the attribute table as separate attributes. The Geostatistical Wizard in ArcGIS 9.2 (ESRI, Redlands, CA) was used to generate individual Chl-a and SM prediction spatial zone maps. The Universal Kriging algorithm was used to create these maps using the semivariogram modeling equation stated above.

Fig. 4 Cartographic model of the spatial zone map creation for Chl-a and SM prediction in Beaver Reservoir.

**RESULTS AND DISCUSSION**

**LVQ model for Chlorophyll-a prediction**

The initial LVQ model architecture was formed with 6-5-3, i.e., six IPE, five Kohonen nodes, and three OPE. The testing classification accuracy was only 0.5 (50%) for each CL1, CL2, and CL3 of output. The training classification accuracies of 50% for CL1, 100% for CL2, and 100% for CL3 were obtained from the same model architecture. The testing RMSE obtained for that architecture was 0.58 with a testing output correlation coefficient of 0.22. The result of the LVQ prediction model with the initial architecture was not promising. After model optimization, the LVQ network architecture was obtained as 6-5-3 to provide highest classification rate (100% for CL1, 83% for CL2, and 100% for CL3 of testing output), lowest RMSE of 0.22, and the best corresponding desired versus predicted correlation coefficient of 0.90. The other optimal model parameters were, learning iterations of 21,000 and learning rate of 0.05. Other model parameters were the same as the initial model. The optimal LVQ model provided the RMSE for the learning (training) phase as zero with a desired versus predicted correlation coefficient of 0.97. The training classification accuracy was 100% for each CL1
and CL2, and 50% for CL3 of output. We were disappointed with this result because the training dataset was trained and tested with the same dataset to provide a good result. However, the result of the testing model was very good with only a single sample point (CL2) out of the 14 testing sample point predicted differently as CL1. The optimum model result is shown in Table 4.

Table 4 Optimum LVQ classification model result for chl-a prediction.

<table>
<thead>
<tr>
<th>Model efficiency test parameters</th>
<th>Training dataset</th>
<th>Testing dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMS error</td>
<td>1</td>
<td>0.22</td>
</tr>
<tr>
<td>Correlation coefficient</td>
<td>0.97</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.83&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>1&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.17</td>
</tr>
</tbody>
</table>

<sup>a</sup> CL3 classified correctly. <sup>b</sup> CL2 classified correctly. <sup>c</sup> CL1 classified correctly.

The additional row shows the misclassification into another respective class.

The LVQ Chl-a prediction model was trained with 21 data points from CL1, 24 data points from CL2 and only five data points from CL3 category. There were six data points from the CL1 and CL2 and only two data points in CL3 category for the testing dataset. Except for CL3, the model should learn better for other two classes. But the testing prediction for one point was misclassified. The global rate schedule could be changed to look for a better prediction in CL2. We have used the same global learning rate schedule (Table 3) which we used for the initial LVQ model architecture. However, the LVQ prediction model did provide very promising results.

LVQ model for SM prediction

For the LVQ modeling of SM prediction, the initial model architecture used was 6-4-3 with the learning rate of 0.06 for LVQ1 and 0.03 for LVQ2, with a conscience factor of 1.0 for LVQ1, 0.001 of frequency estimation value, learning iterations of 10,000 for LVQ1 and 1,000 for LVQ2, and LVQ2 width parameter of 0.02. However, the initial network could not provide good testing and training prediction results for each class. Later, network optimization was obtained with 6-9-3 model architecture, LVQ1 learning rate of 0.05, and other similar parameters as the initial network. The testing classification accuracy obtained with the optimized model was 0.88 (88%) for CL1, 50% for CL2, and 50% for CL3 of output. The training classification accuracies of 75% for CL1, 67% for CL2, and 100% for CL3 were obtained with the same model. The optimal testing RMSE obtained was 0.43 with a testing output correlation coefficient of 0.47. One sample point in CL1 was predicted as CL2, one sample point of CL2 was predicted as CL1, and two out of four sample points in CL3 were predicted as CL3. This mis-prediction could be because the training dataset of the SM LVQ model was very highly underrepresented for CL2 and CL3, with only eleven and seven sample points, respectively out of 40 total data points. The optimum model result is shown in Table 5.
Table 5 Optimum LVQ classification model result for SM prediction.

<table>
<thead>
<tr>
<th>Model efficiency test parameters</th>
<th>Training dataset</th>
<th>Testing dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMS error</td>
<td>0.41</td>
<td>0.22</td>
</tr>
<tr>
<td>Correlation coefficient</td>
<td>0.88</td>
<td>0.90</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>0.17</th>
<th>1&lt;sup&gt;a&lt;/sup&gt;</th>
<th>0</th>
<th>0</th>
<th>0.5&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation coefficient</td>
<td>0.25</td>
<td>0.67</td>
<td>0</td>
<td>0</td>
<td>0.5&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>0.16</td>
<td>0</td>
<td>1&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

<sup>a</sup> CL3 classified correctly. <sup>b</sup> CL2 classified correctly. <sup>c</sup> CL1 classified correctly.

The other row shows the misclassification into another respective class.

Spatial Chl-a and SM zoning

The universal kriging-prediction map technique in the geostatistical wizard in ArcGIS 9.2 was followed to create the spatial Chl-a prediction zoning in Beaver Reservoir watershed. The LVQ model predicted class values of Chl-a attribute was used for the purpose. Therefore 14 data points were used with aconst (estimated by Global Polynomial Interpolation) order of trend to develop the prediction map. The final Chl-a predicted spatial zoning map is provided in Figure 5 showing graduated shading from oligotrophic to mild trophic condition. A few portions of the lake could not be brought under the prediction mapping category because we did not have water sample points from those regions. The kriging map was created as a rectangle using the top-left and bottom-right points as the extent. Finally, the rectangular Chl-a prediction spatial map was clipped to the Beaver Reservoir polygon spatial extent. Using the similar technique, the spatial zone map of SM extent in the lake was created. The map is shown in Figure 6.

![Fig. 5 Spatial zoning map of Beaver Reservoir](image1)

![Fig. 6 Spatial zoning map of Beaver Reservoir SM](image2)
SUMMARY AND CONCLUSIONS

Watershed and water use managers are concerned with the spatial representation of lake water quality. A fast and inexpensive method of lake water quality extent determination is essential for any water use managers or end users. This study used the Landsat-TM digital information in determining the Chl-a and SM content of the lake in a fast and economical manner. The LVQ classification/prediction technique was used to predict the Chl-a and SM contents at different spatial positions of the lake using the Landsat-TM DN values. Finally, universal kriging was used to map the Chl-a and SM extent for the entire lake in an effort to allow water managers or end users to draw inferences about the lake water quality and undertake remedial measures.

The LVQ prediction model provided more than 90% accuracy in the prediction of Chl-a class and 72% accuracy in the prediction of SM class in the lake. It was observed from the Chl-a spatial zonal map of Beaver Reservoir that the Chl-a content is higher near river confluence points. But the lake was totally oligotrophic towards the dam where the reservoir capacity is much higher. A similar trend was observed for SM in the lake with the only exception that in the lower middle portion, where the lake is very narrow, the SM content is very high. But in both Chl-a and SM cases, the lake was never worse than mild trophic.

The spatial Chl-a and SM zone map could not be generated in some portions of the lake because of the absence of sample data collection from these areas. It could have been done with more sample point collections from the bank area. Although the datasets used in this study are little older, the process developed with LVQ modeling approach to map the inland lake water quality is unique. It is expected that other researchers may be able to replicate this application of LVQ modeling process in efficiently estimating inland lake water quality with freely available Landsat 7 ETM+ or Landsat 8 imageries.

ACKNOWLEDGEMENT

Funding for this project was provided by the Environmental Protection Agency, Arkansas Soil and Water Conservation Commission, and the University of Arkansas - Division of Agriculture. We acknowledge the help provided in field data collection by several undergraduate and graduate students of the Biological and Agricultural Engineering Department.

REFERENCES


Neural Ware. (2005) *Reference Guide of Neural Ware*. Neural Ware, Carnegie, Pennsylvania, USA.


