Journal of the Korean Society of Remote Sensing, Vol.15, No.2, 1999, pp.107~117

Comparison between Neural Network and Conventional Statistical Analysis Methods for Estimation of Water Quality Using Remote Sensing

Jung-Ho Im* and Jong-Chul Jeong**

Graduate School of Environmental Studies, Seoul National University*, Korea Ocean Research and Development Institute**

원격탐사를 이용한 수질평가시의 인공신경망에 의한 분석과 기존의 회귀분석과의 비교

임정호* · 정종철**

서울대학교 환경대학원*, 한국해양연구소**

Abstract : A comparison of a neural network approach with the conventional statistical methods, multiple regression and band ratio analyses, for the estimation of water quality parameters is presented in this paper. The Landsat TM image of Lake Daechung acquired on March 18, 1996 and the thirty in-situ sampling data sets measured during the satellite overpass were used for the comparison. We employed a three-layered and feedforward network trained by backpropagation algorithm. A cross validation was applied because of the small number of training pairs available for this study. The neural network showed much more successful performance than the conventional statistical analyses, although the results of the conventional statistical analyses were significant. The superiority of a neural network to statistical methods in estimating water quality parameters is strictly because the neural network modeled non-linear behaviors of data sets much better.

Key Words : Neural Network, Backpropagation, Remote Sensing, Water Quality, Chlorophyll-a, Suspended Sediment, Transparency

요 약 : 본 연구에서는 원격탐사를 이용하여 수질 파라미터들을 평가하는데 기존의 다중 회귀나 밴드비 회귀분석을 이용한 통계적인 방법과 신경망을 이용한 방법을 비교하였다. 사용된 영상은 1996년 3월 18일 대청호 유역의 Landsat TM 영상이며, 30개의 현장 실측치가 위성이 통과하는 시간대에 샘플링되었다. 적용된 신경망은 3개의 층으로 구성된 전향 신경망이며 훈련방법으로는 역전파를 사용하였다. 본 연구에서는 가용한 훈련 데이터 셀이 작으므로 cross-validation 방법이 적용되었다. 비록 기존의 회귀분석에 의한 결과도 어느 정도 유의하게 나왔지만, 신경망에 의한 결과가 훨씬 성공적인 수행을 보여주었다. 신경망을 이용한 수질평가는 신경망이 자료의 비선형적 속성을 잘 반영해주기 때문에 기존의 통계적 기법보다 훨씬 나은 결과를 제공한다고 판단된다.

Received 2 March 1999 ; Accepted 2 May 1999.

1. Introduction

Recently, water pollution has become a major global environmental problem. It is not just a simple water problem but a complex one that affects a wide range of our society from human health and to our ecosystem (Jeong, 1999). Among various water pollution problems, lakes are especially easily polluted because of their geographical feature of being a large closed inland body of the water and their stagnant nature. Once they are polluted, a considerable amount of time and money is required to clear the polluted water quality. In order to prevent worsening the water pollution, continual water quality monitoring is required. Major water quality parameters include chlorophyll-*a*, suspended sediments (SS), and transparency. The traditional method used to estimate these parameters has been in-situ sampling aboard ship followed by the laboratory measurements. However, large amounts of time and cost for the sampling cruises were limiting factors. Remote sensing method as the estimator of these water quality parameters, with vast spatial range and multi temporal range, has been a powerful tool as an alternative to traditional in-situ sampling method by ship (Khorram and Cheshire, 1983; Baban, 1993, 1997; Gitelson et al., 1996).

To estimate the levels of surface chlorophyll-*a*, SS and transparency by radiometric measurements, the transfer function which is the relationship between the optical properties of the parameters and the radiances received from satellite sensor must first be modeled. However, the transfer function is often hard to express theoretically because of its non-linear behaviors. In such cases, the transfer function must be modeled from the comparison between the measurements by in-situ sampling and the measurements by satellite sensor using regression analysis or regression-like techniques (Keiner, 1997).

To estimate water quality, regression analysis as the common empirical method of modeling the transfer function has been extensively studied since 1980s (Whitlock et al., 1982; Gordon et al., 1983; Lathrop and Lillesand, 1986; Tassan, 1993; Baban, 1993; and Pattiaratchi et al., 1994). Combinations of bands, or ratios of bands, are used to create empirical algorithms relating in-situ sampling data and radiances received from satellite sensor. However, regression analysis has limitations because of the non-linear property of these relationships (Gordon et al., 1983; Krasnopolsky et al., 1995). However, neural networks can flexibly model a variety of non-linear behavior and have been shown to be useful in modeling a large range of transfer functions (Thiria et al., 1993; Krasnopolsky et al., 1995).

In this study, we used a neural network to model the transfer function between the levels of chlorophyll-*a*, SS and transparency and the radiances received at the Landsat TM sensor. In addition, the comparison between the neural network and the conventional statistical methods including multiple regression and band ratio analyses has been performed.

2. Neural Networks

Artificial neural networks were originally developed to model the functioning of human brains. The principles found in the brain and used in neural networks are parallel and distributed processing which means that information is not processed serially and is not stored at one fixed location (Bischof et al., 1992). These days neural networks are used for many application fields

such as classification, pattern recognition, signal processing, etc., and even on remote sensing (Augusteijn and Warrender, 1998; Zhang and Scofield, 1994). In this study, we employ a three-layered and feedforward network by backpropagation training algorithm.

There are an input layer, one hidden layer, and an output layer, each containing at least one node, which is called neuron as it performs neuron-like functions. It has been proven that a neural network with one hidden layer, no matter how complex it is can represent any function. This is known as Kolmogorov theorem (Beale and Jackson, 1990). The input layer brings the information to be processed into the network. Neurons in the input layer are hypothetical in that they do not themselves have inputs, and they do no processing of any sort (Master, 1993). The term feedforward means that information flows in one direction only. Each neuron performs two functions: a summation function dealing with linear nature and an activation function handling non-linear behavior. A summation function can be stated as

$$Net_j = \sum_{i=1}^{n} W_{ij} X_i + B_j$$

where X_i are the inputs, W_{ij} are the weights associated with each neuron connection, and B_j is a bias associated with neuron j. These inputs to the neuron are multiplied by their associated weights, summed and added to the bias. This sum is used in an activation function as

$$Z_j = f(Net_j) = \frac{1}{1+e^{-Net_j}}$$

where f is the activation function, which is called squashing function. Although there is no theoretical limit to what the value of a neuron can be, the range of the activation function is usually limited (Krasnopolsky et al., 1995). The most common limits are (0, 1), while some range from (-1, 1).

Most current models use a sigmoid (S-shaped) activation function. A sigmoid function can be simply defined as a continuous, real valued function, whose derivative is always positive, and whose range is bounded. The most commonly employed sigmoid function is the logistic function. One advantage of this function is that its derivative is easily found (Master, 1993). Other sigmoid functions, such as the hyperbolic tangent, are sometimes used. In most cases, it has been found that the exact shape of the function has little effect on the ultimate power of the network, though it can have a significant impact on training speed (Gose et al., 1996). For this study, a logistic function was used for f.

3. Data Representation

To obtain the necessary data pairs used to train the neural network and to establish the regression algorithms, there must be in-situ data coincident

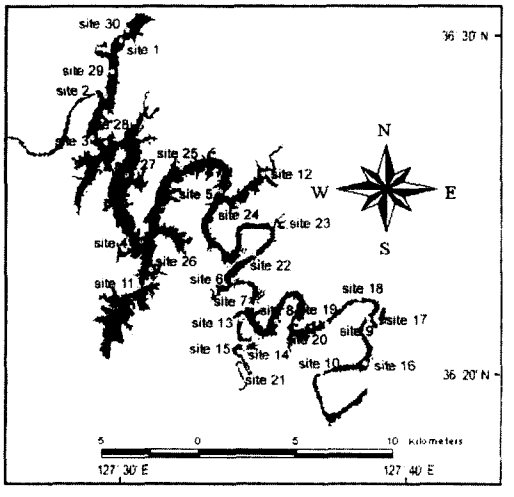

Fig. 1. Map of Sampling Stations

with Landsat TM overpass. In this study, there were thirty sets of training pairs (in-situ values and corresponding pixel values of each band) available for use to Lake Daechung (Kim, 1997). The sampling stations, which were determined by ship-equipped GPS, are illustrated in Figure 1.

The TM image used for this study was acquired on March 18, 1996. The first four bands(the visible and near IR range) were used in this study, since common algorithms estimating the concentrations of chlorophyll-*a* and SS, and transparency have been established with some of these four bands.

For the atmospheric correction, the normalization method using histogram adjustment known as common bulk correction was used (Ritchie et al., 1990; Pattriaratchi et al., 1994). The radiance received at the TM sensor consists of the water leaving radiance including the radiance due to atmospheric effects.

The TM image was geometrically corrected using ER Mapper software, referencing to a 1:25,000 standard map. The image was then resampled by nearest neighbor method, which was chosen because it does not change the original data values. The water area was extracted by the threshold of pixel values of the near IR band.

4. Network Architecture and Training

Figure 2 shows the network architecture used in this study. Each input neuron corresponds to three visible bands and a near IR band on the TM. The output values from the input layer are assigned to the three neurons in the hidden layer, where the summation and activation functions are performed. The output values of the hidden layer are then used as the input values of the output layer, which only performs the summation functions without activation functions. The output of this layer is the value of the parameters that we are looking for.

$$Output = \sum_{k=1}^{j} \omega_k Z_k + \beta$$

For the output layer, ω_k are the weights between the hidden layer and the output layer, β is the bias associated with the output layer. We did not use the activation function in the output layer, because the average error without activation function is lower than that with activation function.

The input values are fed into the network, and the network calculates the output. This output o is compared with known correct output t, and the difference between them is the network error. To modify the network output, the weights and

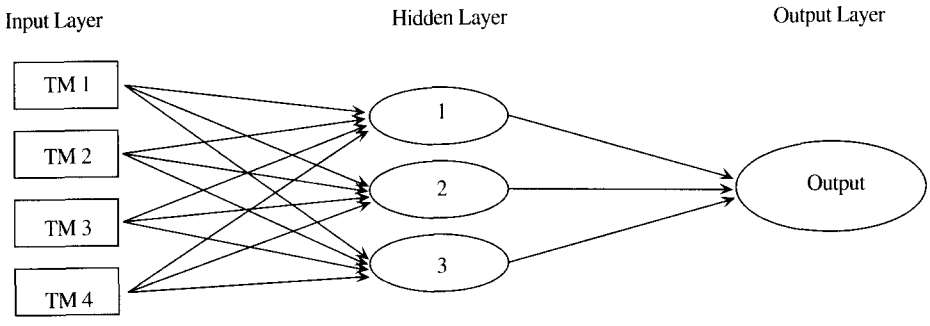

Fig. 2. Neural Network Architecture of this study

biases must be changed to decrease the network error. Finding the weights and biases that minimize this error is the goal of training. This is accomplished through backpropagation algorithm, which is that the network adjusts its weights and biases, working back through the network. We used the summed squared error (SSE) function, which is defined as

$$SSE = \frac{1}{2}\sum_{i=1}^{n}(t_i - o_i)^2$$

As the SSE is the function associated with the weights, the change of the weights is done by the following equation based on determining the direction of change that will decrease the SSE.

$$\omega_{ij}(t+1) = \omega_{ij}(t) - \eta\frac{\delta SSE}{\delta \omega_{ij}} + \alpha[\omega_{ij}(t) - \omega_{ij}(t-1)]$$

where η is a learning rate term and α is a momentum term. The learning rate η changes during training to accelerate convergence towards a minimum in the beginning; then it slows as the network gets close to a minimum, to prevent overshooting it. Using a momentum term in the backpropagation algorithm may increase stability when a large step size is used. The momentum term can decrease oscillations that may slow convergence. This training process operates by repeatedly using the data in the training set to change the network weights until an acceptable error level is reached (Beale and Jackson, 1990; Rao and Rao, 1995).

The neural network was trained several times with different training and validation sets of random initial weights. In some training sets the network became trapped in a local minimum and never converged under an acceptable error level. After several trials with different initial weights and biases, the combination of weights and biases with the most successful result was used with the entire TM image.

Before training, the input training sets were scaled between (0, 1) by min/max of water areas. This normalization helped the network converge faster. The same scaling factors were used when the actual TM image data was used with the network.

In this study, a simple form of cross-validation was used to validate the result of the training algorithm (Kwok and Yeung, 1995). Twenty-four pairs were used as the training set; the other six pairs were used as a test set for validation. After the network was trained, it was applied to the entire TM image.

5. Results and Discussion

1) Band ratio algorithms

Several band ratios were applied to find the best algorithm of estimating water quality. The most successful algorithm for chlorophyll-*a* was a third order function using the log (band 1/band 3) as a independent variable. In case of SS, it was a second order function using the same independent variable with chlorophyll-*a*. For the

Table 1. Band ratio algorithm of each case (x = log(Band1/Band3))

	Chlorophy II-α	SS	Transparency
Algorithm	log(C) = $20.418x^3 - 10.435x^2 + 0.0309x + 0.7472$	log(SS) = $9.1856x^2 - 6.0959x + 1.2501$	log(T) = $-5.7336x^2 + 4.2411x - 0.2857$
R^2	0.5912	0.6833	0.6812

transparency, it was also a second order function using the same independent variable with chlorophyll-*a*. The specific algorithms are shown in Table 1.

2) Multiple Regression Analysis

Both linear and log-linear regressions were performed for many different combinations of bands and band ratios, to find the best combination for estimating the parameters. The best combination for chlorophyll-*a* was the use of a linear equation including band 1, band 2, and the ratio of band 1 and 3. In case of SS, it was a linear equation consisting of band 1, band 2, and band 3. For the transparency, the most successful combination was a linear equation including band 1, band 2, and band 3. The algorithms are described in Table 2.

To indicate the significance of band ratio algorithms and multiple regression analysis several parameters such as the coefficient of determination (R^2), RMS (root mean square) error, and critical F-value at the 5% significance level were calculated. This is shown in Table 5.

3) Neural Network Results

The neural network was trained several times for the cases of chlorophyll-*a*, SS, and transparency. The weights and biases that were determined by the network to use for the entire image were shown in Table 3. Table 4 gives the training and test results of each case, which indicates that the network was well fitted, not overfitted. Figure 3 shows spatial distribution of the concentration of chlorophyll-*a* in Lake Daechung, which was determined by the neural network. Figure 4 and 5 show spatial maps of the SS concentration and transparency, respectively.

The range of chlorophyll-*a* in Lake Daechung using the results of the neural network was between 0.9 mg/m^3 and 7.9 mg/m^3, which was relatively low concentration except high concentration around the southeast in Lake Daechung. Studies of Kim (1997) showed that on the day of the TM overpass, Lake Dachung was in the early stage of a phytoplankton bloom dominated by the diatom *Fragilaria spp.* The spatial distribution of chlorophyll-*a* assumed a similar aspect of that of SS. However, in the upper stream of Lake Daechung, the concentration of chlorophyll-*a* appeared to be high at 8 mg/m^3, as expected, which resulted in the reflection and scattering by SS. The concentration of SS was lower than 8 mg/l in most areas of Lake Daechung, though it was very high about 40 mg/l around the southeast in Lake Daechung. Because in-situ sampling was two days after raining around upper stream area of Lake Daechung water system about 50 mm, the areas around entrances of main stream and branch streams had the inflows of water with soil. In case of transparency, secchi depth was between 3 and 4 m in most areas of Lake Daechung. However, transparency was lower than 1 m in the southeast areas of Lake

Table 2. Multiple regression algorithm of each case

	Chlorophyll-α	SS	Transparency
Algorithms	C = 5.51361+0.04949*band1 + 0.0421*band2 - 1.75787*(band1/3)	SS = -12.44642+0.26854*band1 - 0.24771*band2 + 2.50540*band3	T = 4.5768-0.12061*band1 - 0.0325*band2 - 0.10853*band3
R^2	0.5013	0.8636	0.5537

Table 3. Values of the weights and biases by the neural network

	Parameters	Chlorophyll-α	SS	Transparency
Weights	W11	21.5372	21.2385	-13.3527
	W12	-1.0333	9.4057	-0.0165
	W13	-3.8200	5.7937	-6.8620
	W21	21.2661	27.8670	-16.7912
	W22	-3.6101	5.3489	-13.2147
	W23	21.1133	-27.0098	-43.2107
	W31	6.2318	33.0859	-25.8765
	W32	24.1773	29.0915	-14.8333
	W33	51.4380	78.4254	-47.7606
	W41	-10.3755	-8.9348	-0.4941
	W42	15.0324	-6.0763	13.9421
	W43	12.3571	-6.3228	-11.4330
	ω1	-3.6598	27.7292	-9.2281
	ω2	3.2412	10.4117	3.3956
	ω3	2.3745	3.9296	8.8576
Biases	b1	-20.2249	-30.6077	15.2315
	b2	-22.1257	-19.8197	1.5426
	b3	-19.9884	-10.0535	28.1297
	β	2.2048	1.4700	0.3202

Table 4. Training and test errors per pattern in the neural network

per pattern / parameters	Training set error	Test set error	Total error
Chlorophyll-α	0.067	0.069	0.067
SS	0.218	0.233	0.221
Transparency	0.082	0.079	0.082

Daechung where the concentrations of chlorophyll-*a* and SS were high. In Figure 3, 4, and 5, the concentrations of chlorophyll-*a* and SS in the areas of low transparency were high, which clearly showed reverse-correlation of these water quality parameters, especially SS and transparency.

Table 5. Comparison among each performance

		Band Ratio Algorithm	Multiple Regression Analysis	Neural Network
Chlorophyll-α	R^2	0.6551	0.5013	0.9613
	RMS (mg/m^3)	0.747	0.885	0.252
	Critical F(F ratio)	$1.37*10^{-5}$ (27.65)	$2.82*10^{-6}$ (13.33)	
SS	R^2	0.7916	0.8636	0.9947
	RMS (mg/l)	5.125	4.086	0.897
	Critical F(F ratio)	$1.82*10^{-7}$ (29.12)	$2.2*10^{-11}$ (54.88)	
Transparency	R^2	0.4995	0.5537	0.9558
	RMS (m)	0.818	0.756	0.239
	Critical F(F ratio)	$1.99*10^{-7}$ (28.84)	$8.92*10^{-5}$ (10.75)	

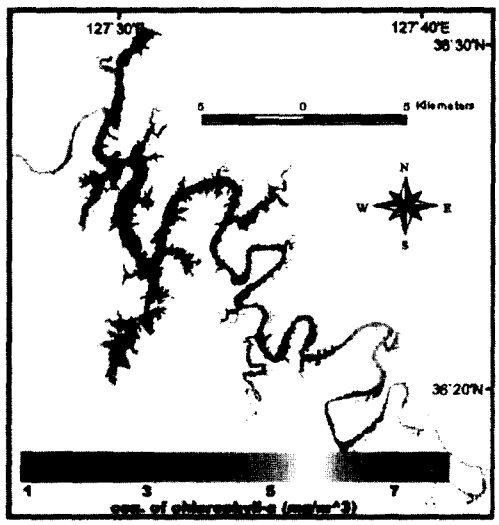

Fig. 3. Spatial distribution of the concentration of chlorophyll-*a* in Lake Daechung

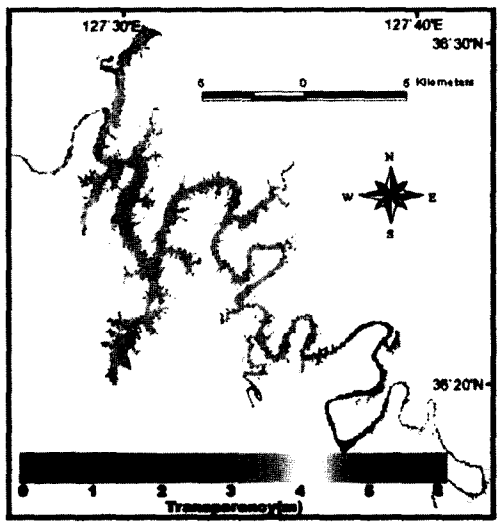

Fig. 5. Spatial distribution of the transparency in Lake Daechung

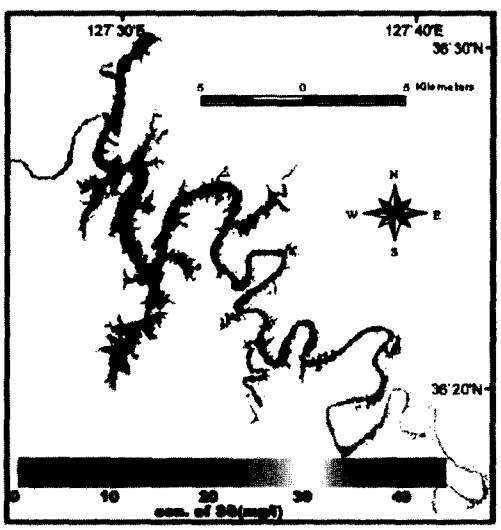

Fig. 4. Spatial distribution of the concentration of SS in Lake Daechung

4) Comparison

The statistics for the comparison among the results of the band ratio algorithm, the multiple regression analysis, and the neural network are shown in Table 5. The R^2 and RMS errors for the neural network were calculated in the same way as for the multiple regression analysis and the band ratio algorithm. The RMS errors from the neural network for chlorophyll-*a*, SS, and transparency were 7.8 %, 14.6 %, and 9.8 %, compared with 27.4 %, 66.7 %, and 31.1 % for the multiple regression analysis of chlorophyll-*a*, SS, and transparency, and with 23.1 %, 83.6 %, and 33.7 % for the band ratio analysis of chlorophyll-*a*, SS, and transparency, respectively. From critical F values at 5% significance level, the conventional regression analyses were significant, though they did poorer jobs than the neural network.

These statistics clearly show that conventional statistical analyses gave worse performance than the neural network in determining the relationship between the water quality parameters and the TM radiances for this study. Several reasons for this include errors in the in-situ sampling, errors in different time zone between in-situ sampling and TM image acquirement, and errors in the matchup data sets between the in-situ

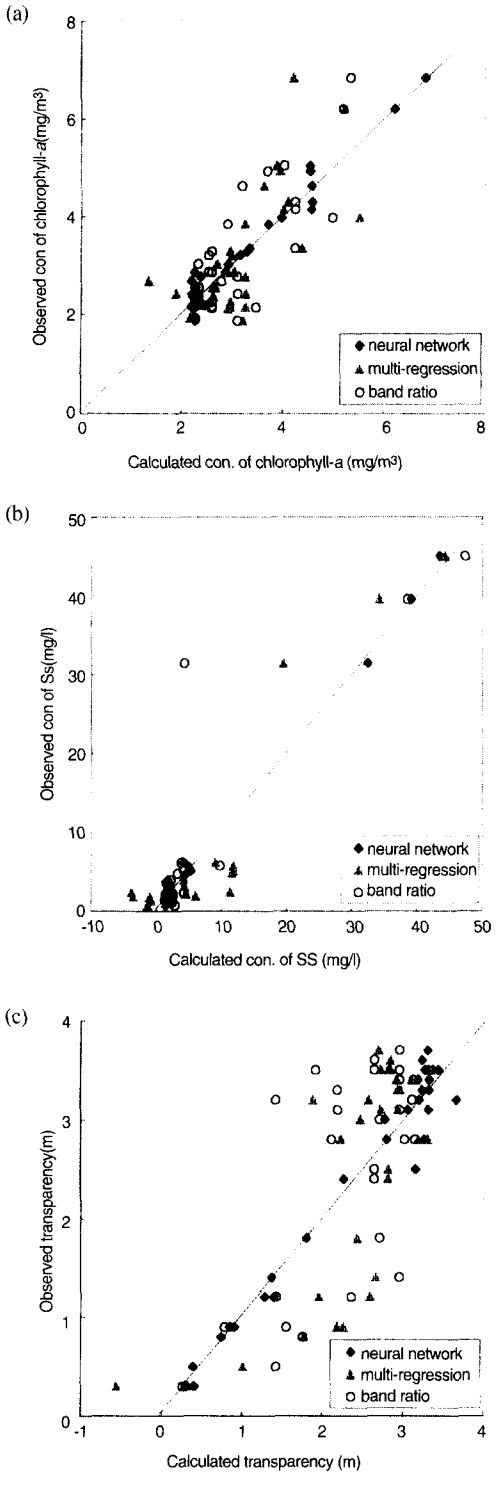

Fig. 6. Comparison of each result of (a) chlorophyll-*a*, (b) SS, and (c) transparency

sampling locations and the TM image locations. However, the main reason is the inability of the conventional statistical methods to model the non-linear behaviors of the transfer function.

A graphical comparison of these results is illustrated in Figure 6. It shows the comparisons between the in-situ measurements and calculated values : (a) chlorophyll-*a*, (b) SS, and (c) transparency. These results also clearly show that the neural network excelled multiple regression and band ratio for the analysis of these data. This is due to the non-linear property of the transfer function as above mentioned. The results of multiple regression analysis or band ratio analysis were significant for predicting water quality parameters, though they showed poorer performance than those of the neural network. This is because the data sets had small ranges, which made conventional statistical methods model non-linear behaviors partially.

6. Concluding Remarks

We set out in this paper to compare the neural network approach with the other methods, multiple regression and band ratio analyses, for the estimates of water quality parameters in Lake Daechung, one of major drinking water resources in Korea. From the results, the neural network has been proved to be more effective than the conventional statistical analysis to estimate water quality parameters. It mainly results from the non-linear properties of data set, which can be modeled well by neural network.

Our work shows that a neural network is a powerful tool for remotely sensed image analysis. However, it has a limitation which cannot apply to another TM image of same location or other

water area, because it was not fully trained over the whole range of usual in-situ sampling data. In other words, because a neural network is the greatest tool in estimating something within the range of training, it usually has weak generality. For the establishment of a more substantial neural network algorithm at a certain area, like Lake Daechung in this study, more matchup data sets of another imagery and in-situ sampling are needed.

Acknowledgement

The authors gratefully acknowledge the assistance of Dr. Kim who provided the data.

References

Augusteijn, M.F. and C.E. Warrender, 1998. Wetland classification using optical and radar data and neural network classification, *International Journal of Remote Sensing*, 19(8):1545-1560.

Baban, S.M.J., 1993. Detecting water quality parameters in Norfolk Broads, U.K., using Landsat imagery, *International Journal of Remote Sensing*, 14:1247-1267.

Baban, S.M.J., 1997. Environmental Monitoring of Estuaries; Estimating and Mapping Various Environmental Indicators in Breydon Water Estuary, U.K., Using Landsat TM Imagery, *Estuarine, Coastal and Shelf Science*, 44:589-598.

Beale, R. and T. Jackson, 1990. *Neural Computing: an Introduction*, Adam Hilger Co, Bristol, U.K.

Bischof, H., W. Schneider, and A.J. Pinz, 1992. Multispectral classification of Landsat-Images using neural networks, *IEEE Transactions on Geoscience and Remote Sensing*, 30(3):482-489.

Gitelson, A. et al., 1996. Chlorophyll estimation in the southeastern Mediterranean using CZCS images: Adaptation of an algorithm and its validation, *Journal of Marine System*, 9(4):283-290.

Gordon, H. et al., 1983. Phytoplankton pigment concentrations in the Middle Atlantic Bight: Comparison of ship determinations and CZCS estimates, *Applied Optics*, 22:20-36.

Gose, E., R. Johnsonbaugh, and S. Jost, 1996. *Pattern Recognition and Image Analysis*, Prentice Hall, NJ, USA

Jeong, J.C., 1999. *Water Quality Evaluation for Coastal Waters and Lake Sihwa Using Remote Sensing Techniques*, Seoul National University.

Keiner, L., 1997. *The Satellite Remote Sensing of Environmental Processes in Delaware Bay*, UMI, USA

Kim, T.K., 1997. *Assessment of Lake Water Quality Using LANDSAT TM Imagery Data*, Chonbuk National University.

Krasnopolsky, V., L. Breaker, and W. Gemmill, 1995. A neural network as a nonlinear transfer model for retrieving surface wind speeds from the special sensor microwave imager, *Journal of Geophysical Research*, 100:11033-11045.

Kwok, T.Y. and D.Y. Teung, 1995. Efficient cross-validation for feedforward neural networks, *IEEE Conference on Neural Networks Proceedings*, 5:2789-2794.

Lathrop, R. and T. Lillesand, 1986. Use of Thematic Mapper data to assess water

quality in Green Bay and central Lake Michiga, *Photogrammetric Engineering and Remote Sensing*, 52:671-680.

Masters, T., 1993. *Practical neural network recipes in C++*, Academic Press, CA, USA

Pattiaratchi, C. et al., 1994. Estimates of water quality in coastal waters using multi-data Landsat Thematic Mapper data, *International Journal of Remote Sensing*, 15:1571-1584.

Rao, V.B. and H.V. Rao, 1995. *C++ Neural Networks & Fuzzy Logic*, M & T, USA

Ritchie, J., C. Cooper, and F. Schiebe, 1990. The relationship of MSS and TM digital data with suspended sediments, shlorophyll and temperature in Moon Lake, Mississippi, *Remote Sensing of Environment*, 33:137-148.

Tassan, S., 1993. An Improved in-water algorithm for the determination of chlorophyll and suspended sediment concentration from Thematic Mapper data in coastal waters, *International Journal of Remote Sensing*, 14:1221-1229.

Thiria, S. et al., 1993. A neural network approach for modeling nonlinear transfer functions: Application for wind retrieval from spaceborne scatterometer data, *Journal of Geophysical Research*, 98:22827-22841.

Whitlock, C., C. Kuo, and S. LeCroy, 1982. Criteria for the use of regression analysis for remote sensing of sediment and pollutants, *Remote Sensing of Environment*, 12:151-168.

Zhang, M., and R.A. Scofield, 1994. Artificial neural network techniques for estimating heavy convective rainfall and recognizing cloud mergers from satellite data, *International Journal of Remote Sensing*, 15(16):3241-3261.