Polish J. of Environ. Stud. Vol. 18, No. 2 (2009), 151-160

Original Research

Comparative Prediction of Stream Water Total Nitrogen from Land Cover Using Artificial Neural Network and Multiple Linear Regression Approaches

B. J. Amiri*, K. Nakane

Division of Environmental Dynamics and Management, Graduate School of Biosphere Science, Hiroshima University, 1-7-1 Kagamiyama, Higashi-Hiroshima 739-8521 Japan

Received: 1 January 2008
Accepted: 4 August 2008

Abstract

Performance of two data-driven models that were developed using Artificial Neural Networks (ANNs) and Multiple Linear Regression (MLR) approaches were investigated in prediction of Total Nitrogen (TN) concentration in twenty-one river basins in Chugoku district of Japan. Comparison of TN concentration predictions, which were carried out using an ANN-based model and MLR-based model indicated that prediction of the former model (r^2=0.94, $p<0.01$) was more accurate than that of the latter model (r^2=0.85, $p<0.01$). Lack of a sufficient data set that might be considered an obstacle for cross-validating models that are developed was dealt with using a Monte Carlo-based sensitivity analysis of the developed models. This initiative could provide reliable information for judging predictive capacity of the developed models stochastically. Result of sensitivity analysis revealed that predictive capacity of the ANN-based model varied between 0-2 mg/L. Moreover, prediction of the negative outputs was not observed. using the ANN-based model for TN concentration in stream water.

Keywords: Artificial Neural Network, regression, water quality, modeling

Introduction

Scientists and environmental managers alike are concerned about broad-scale changes in land use and landscape patterns and their cumulative impacts on hydrological and ecological process that affect stream, wetland and estuary conditions [1]. Over the past few decades much attention has been paid to evaluation of the relative condition of water resources on regional and national scales by researchers [e.g. 2-4] through examining the possible relationship between land use types and different stream water quality variables.

Introducing analytical tools such as the geographical information system and multivariate statistics have enabled researchers to deal with spatial data and complex interactions in the environment [2]. Applying a multiple regression approach for specifying the relationship between land use and a given water quality variable provide not only information regarding the importance of spatial positioning of land cover, but also would be helpful in determining the relative importance of different land use types as nutrient contributors [5].

Statistical methods might be adequate to find an overall pattern of ecological systems but the non-linear behavior of ecosystems could not be efficiently reviewed by conventional linear methods. Most algorithms in machine learning,

*e-mail: j.amiri@yahoo.com

such as Artificial Neural Networks (ANNs), contain the models to deal with non-linear data based on adaptive or heuristic methods [6]. Artificial neural networks have become a popular and useful tool for modeling environmental systems [7] because of the ability of ANNs to find non-linear patterns in data [8].

Validating the models, which are developed based on either ANNs or MLR modeling approaches, might be considered a critical step in modeling. This task is conducted using two distinctive methods, namely residual-based validation and cross-validation. The first is the prevailing approach, which is applied for validating MLR-based models. The former has been widely considered by researchers involved in ANNs modeling. The modeling process might be halted for the application of cross-validation if the modeler has no access to sufficient data in number in order to allocate part of a data set for cross-validating the model. The objectives of this study were:

1) to predict streamwater total nitrogen as one of the important nutrients from land cover attributes (area % of different land covers such as urban, forest, agriculture, grassland and water body in the catchments) using ANNs and MLR approaches,
2) to compare performance of the developed ANNs-based model to that of the MLR modeling approach, and
3) to investigate the application of Monte Carlo Method for dealing with the scarcity of data set for cross validation of the developed models.

Materials and Methods

Study Site

The present study was carried out in the Chugoku district of Japan, on western Honshu island at (130º55'16" and 133º12'11") longitude and (33º57'40" and 35º23'34") latitude. It includes five prefectures (Hiroshima, Yamaguchi, Tottori, Shimane and Okayama), and covers 32,000 km^2 (Fig. 1). A spatial analysis of land cover map indicates that a high percentage (79.34%) of the study area is covered by forest. Other land cover classes, including urban, agriculture, grassland and water, make up 5.15, 8.33, 6.63 and 0.46% of the study area, respectively (Table 1). There are 7,732,499 inhabitants in the study area (Japanese Statistic Bureau of Ministry of Internal Affairs and Communication: http://www.stat.go.jp/data/kokusei/ 2000/final/zuhyou/008-01.xls) [9]. There are 54 major rivers in the Chugoku district, 21 of which were selected for this study based on the availability of satellite images and water quality data of the rivers (Fig. 1).

Materials

Data Sets

Annual mean of TN data in 2001 were used based on monthly values of TN concentration. Table 2 indicated median, maximum and minimum values of TN concentra-

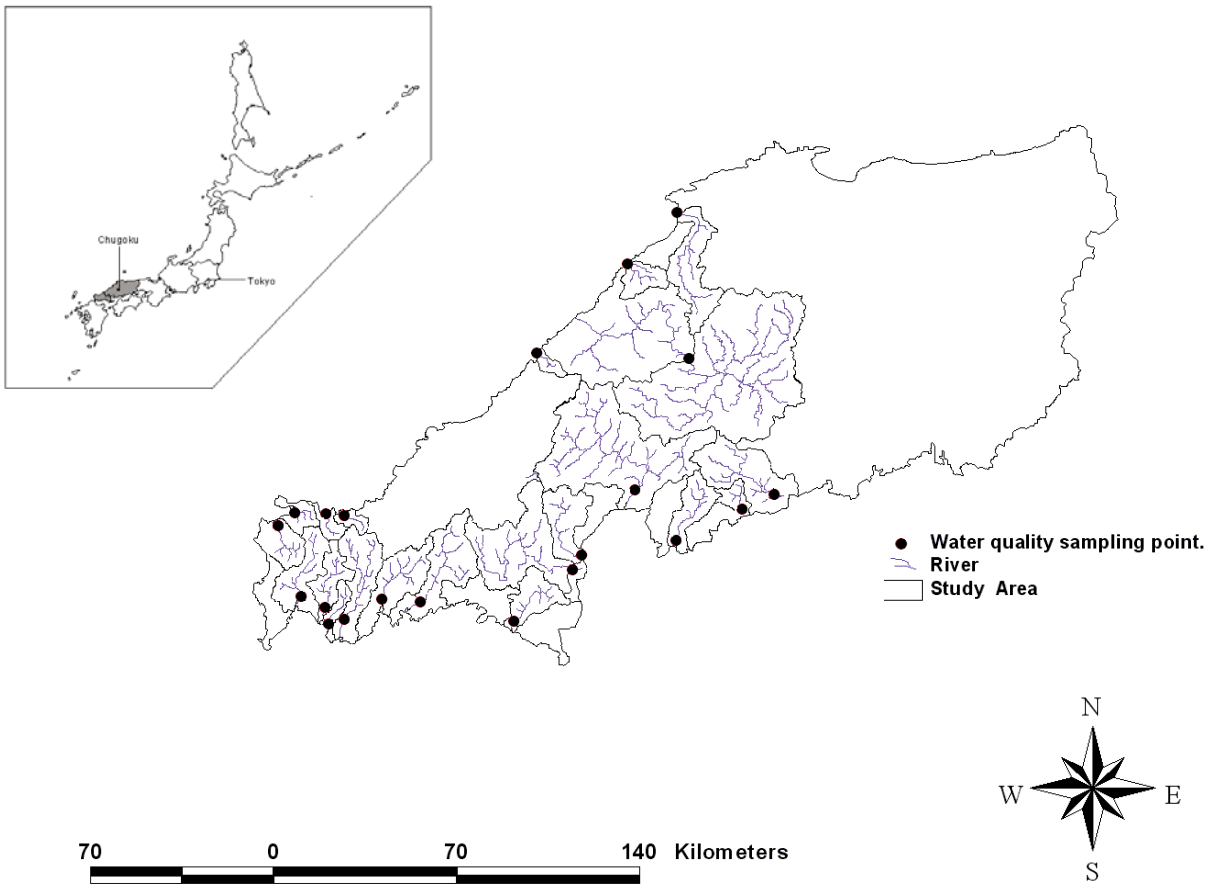

Fig. 1. Geographical situation of the study area.

Table 1. Physical features, land cover and population density in catchments of the Chugoku district.

Basin Number	River Name	Area (km^2)	Land uses (%)					Population density (Person/km^2)
			Urban	Forest	Agriculture	Grassland	Water body	
1	Awano	182	1.97	86.61	5.62	5.89	0.28	63
2	Kakefuchi	85	4.36	71.94	16.35	5.42	1.04	106
3	Fuka	72	6.39	84.17	5.01	4.27	0.10	150
4	Misumi	67	1.36	88.41	5.05	4.18	0.00	70
5	Hamada	253	8.48	78.59	3.72	6.87	1.01	176
6	Gonoo	2622	1.79	83.89	7.44	6.61	0.27	72
7	Shizuma	174	2.46	84.23	5.39	10.05	0.22	150
8	Kando	495	7.09	78.16	11.84	2.76	0.14	222
9	Numata	627	4.42	65.6	24.74	5.04	0.21	165
10	Kamo	98	3.79	78.92	7.62	9.63	0.00	211
11	Kurose	282	10.8	57.61	20.64	10.1	0.75	819
12	Ota	1700	4.74	85.22	4.29	4.71	0.29	386
13	Oze	354	3.01	86.91	3.31	6.15	0.56	183
14	Nishki	932	1.04	91.76	2.76	3.72	0.71	165
15	Shimada	284	5.91	77.75	8.14	7.72	0.43	267
16	Saba	572	2.62	88.35	2.89	5.61	0.53	225
17	Washino	300	13.99	72.84	6.96	5.89	0.30	434
18	Kotou	416	3.8	75.64	8.83	10.91	0.79	253
19	Ariho	98	9.23	72.76	9.24	8.34	0.44	477
20	Asa	226	6.24	78.09	7.14	7.94	0.52	109
21	Koya	299	4.67	78.67	7.97	7.46	1.07	362
Max.		2622	13.99	91.76	24.74	10.91	1.07	819.00
Min.		67	1.04	57.61	2.76	2.76	0.00	63.00
Mean		482.76	5.15	79.34	8.33	6.63	0.46	241.19

tion for 21 river basins in the study area. TN data were obtained from the five prefecture offices (Hiroshima, Yamaguchi, Shimane, Tottori and Okayama) in the study area that could be defined as a secondary database [2]. Water sampling of the stream and analysis are carried out monthly according to the Japanese Industrial Standard (TN: JIS K0102 45.2, 45.3 [10]. Details of sampling method and analysis procedures would be found in JSA JIS K 0102 [10]. Population data was obtained from the census in 2000 (http://www.stat.go.jp/data/kokusei/2000/final/zuhyou/008-02.xls) [9]. Topographical quadrangle maps (1:200,000) were obtained from the Japan Geographical Survey Institute (JGSI) and applied to delineate the catchments (Fig. 1). Satellite images (NASA Landsat-5 TM, 2000) were used to generate a land cover map.

Methods

GIS and Remote Sensing Analyses

Geographical Information System was established using ArcView 3.2 [11] for facilitating spatial analysis and determination of morphological attributes. Catchment boundaries were hand-digitized using JGSI topographic quadrangle maps (1:200,000) for all water-sampling stations illustrated in Fig. 1. A county-scale population database was linked with the digital map of counties for generating a human population density map. It was then overlaid by the catchments map and aggregated in order to find the catchment-scale population density map in the study area. All databases were transformed into a common

Table 2. Descriptive statistics of the annual median of TN in stream water of the study area.

Catchment No.	TN (mg/L)		
	Median	Min	Max
1	0.52	0.31	0.58
2	0.66	0.49	0.82
3	0.61	0.53	0.71
4	0.74	0.59	0.93
5	0.32	0.23	0.98
6	0.62	0.37	0.90
7	0.70	0.45	0.83
8	0.47	0.39	0.57
9	0.87	0.71	1.10
10	0.46	0.15	1.10
11	1.85	1.30	2.90
12	0.69	0.46	1.00
13	0.70	0.50	0.89
14	0.47	0.41	0.64
15	0.73	0.56	0.88
16	0.50	0.40	0.77
17	1.62	1.03	3.06
18	0.76	0.54	0.94
19	1.24	0.78	5.55
20	0.74	0.56	0.89
21	0.71	0.50	1.00

Table 3. Results of normality test for TN and compositional attribute of land covers.

Variable	Sharpio-Wilk statistics	
	Statistic	Sig.
TN	0.655	0.01
Population density	0.918	**0.336**
Urban	0.893	**0.167**
Forest	0.915	**0.313**
Agriculture	0.873	**0.079**
Grassland	0.939	**0.475**
Water body	0.961	**0.746**

*All bold values are significant at p <0.05.

digital format, projected onto a common coordinate system (UTM, zones 52 and 53).

Two scenes of NASA Landsat-5 TM (2000) were used for generating a land cover map in the study area. The satellite images were geo-referenced by the affine procedure. The supervised classification method was applied to classify land use, which included forest, agriculture, grassland, urban and water body (including natural wetlands and artificial lakes). The generated land cover map was verified using JGSI maps. The satellite data was prepared, interpreted and analyzed using the Integrated Land and Water Information System [12]. For calculating the real extent of each land cover for each catchment the generated land cover map was then superimposed with a catchments map, which was subsequently divided by the catchments area to drive the percentage of the catchments covered by each type.

MLR Modeling Approach

All TN and land cover data were tested for normality using the Sharpio-Wilk test with a p-value of less than 0.05 (Table 3). For determination of linkage between land cover-stream water TN concentrations, the MLR (linear, logarithmic, exponential and power) modeling approach was applied using a backward method in order to achieve the best fit model for TN variable. Inter-variable collinearity of the model was investigated referring to the Variance Inflation Factor (VIF). Normality of residuals of the models were then examined using Sharpio-Wilk test with a p-value<0.05 (Table 4). For TN variable, the appropriate model was selected based on regression statistics (r^2, p-value) and considering the significance of the coefficients of the model if the residual of the model was normally distributed. Finally, the goodness-of-fit of the statistically significant regression models was evaluated by scatter plot, and simple linear regression of observed versus equivalent model prediction [4]. Statistical analyses were performed using Excel add-ons (XLSTAT™ 2006), SPSS for Windows Release10 and EasyFit 3.2.

ANNs Modeling Approach

For determination of linkage between stream water TN concentration and land cover attributes, artificial neural network modeling was applied. Generally, a neural network consists of a number of elements, so called "nodes," and connection pathways linking them (Fig. 2) [13]. A network with back-propagation of error typically, at least, comprises three layers: input, hidden and output [14]. The input layer is first since the input data are applied there. In this study it includes six neurons relating to the five land cover variables and one neuron for human population density variable. The output layer is the last one since it is where the outputs are not only processed but also extracted there. In the case of this study, the output layer constitutes a single neuron relating to the value of dependent variable to be predicted (water quality variables). The hidden layers are those that are placed between the input and output. Based on Lek et al. [13] determination of a "state" or "activity level" for

each neuron is specified by the input received from precedent units in the network. Generally, in the hidden layer, the net input-to-unit *j* is of the form as follows:

$$H_j = \sum_{i=1}^{n} x_i v_{ji} \tag{1}$$

...where x_i is the output from unit *i*, (v_{j1}, v_{j2}, . . ., v_{jn}) is the weight vector of unit *j* and n is the number of neurons in the layer preceding the layer including unit *j*. For the output layer, the net input to unit *k* is of the form as follows:

$$I_k = \sum_{j=1}^{h} y_j w_{kj} \tag{2}$$

...where y_j is the output from unit *j*, (w_{k1}, w_{ka}, . . ., w_{kh}) is the weight vector of unit *k* and *h* is the number of neurons in the layer preceding the layer, including unit *k*. The most common transfer function is the sigmoidal as follows:

$$f(x) = \frac{1}{1 + \exp^{-\beta x}} \tag{3}$$

Before training, weights v_{ji} were initialized with random values in the range [0, 1]. Training the network to produce a desired output vector involves systematically changing the weights until the network produces the desired output. This is repeated over the entire training set. The learning process in the network would go on until the error at iteration t+1 becomes higher than the error at iteration t or the iteration target is achieved as given in Eqs 4 and 5 for output layer and hidden layer:

$$W_{kj}^{new} = W_{kj}^{old} + \Delta W_{kj}(t+1) \tag{4}$$

$$V_{kj}^{new} = V_{kj}^{old} + \Delta V_{kj}(t+1) \tag{5}$$

Moreover, change in error value is another criterion to stop the training of the network, which is applied by the application. If change in error value becomes less than 10^{-9}, the training process will be stopped. Computation of the weights and prediction of TN concentrations were performed by Backpropagation Neural Network 1.0 developed by Rudiyanto and Setiawan [15].

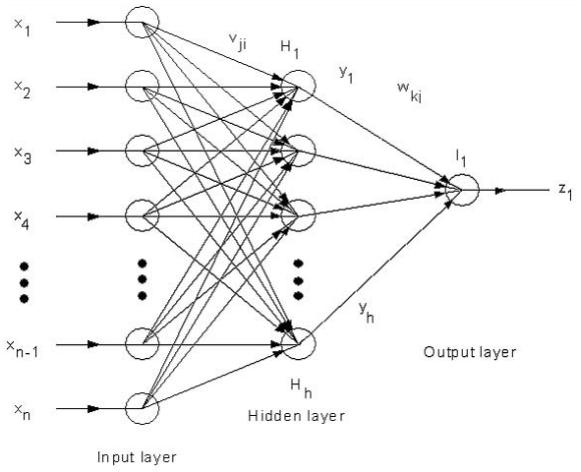

Fig. 2. General pattern of artificial neural network.

Based on the early stopping idea [16], a data set should be classified into three groups called training (60%), controlling (25%) and testing (15%) data sub-sets. The first subset is used to estimate the parameters. The second subset is called the validation set. The third subset, which acts as the error on the validation set, monitors the estimation process. When the network begins to overfit the data, the error on the validation set typically begins to rise. When the validation error increases for a specified number of iterations, the estimation process is discontinued, and the parameters estimated at the minimum of the validation error serve as final estimates [16].

Considering the number of rivers for which water quality data were available (21 basins), it seems that the number of rivers that should be kept for testing the trained network would be very few (three basins). Subsequently, reliable information could not be provided for judging on the trained network. Therefore, testing the trained networks gave up using testing data that was real but few in numbers. They were added to the controlling data set. Finally, twenty-one catchments were classified into two sets in proportion to 60% and 40% as training (12 river basins) and controlling (9 river basins) data, respectively. For addressing the shortage of the testing data, it was decided to use the Monte Carlo Method for sensitivity analysis of the trained network.

According to the Monte Carlo Method, 5000 random numbers were generated using the random number generator command in Microsoft Excel for six variables consisting of area (%) of urban, forest, agriculture, grassland and water body, and human population density. The trained network was tested using the generated random number data set based on the Monte Carlo procedure. In generating random numbers, three conditions were considered:

1) random numbers should vary between 0-100,
2) the summation of five variables for land use types to be equal to 100%, and
3) random numbers for each variable should follow a known distribution.

Initially, six variables were defined in Microsoft Excel. 5,000 random numbers were generated for each variable using the related command (randbetween (bottom) (top)) in it for meeting the first condition. Each row stands for a hypothetical river basin that consisted of five variables for land use types (urban, forest, agriculture, grassland and water body areas in percent) and the last variable for human population density. For meeting the second condition, the summation of five land use variables was calculated and each land use variable divided by the summation since in a real situation for a given river basin, land use types as land use variables cannot be independent and any change in one land use type should be at the expense of others.

For carrying out the modeling task of this study, Backpropagation Neural Network 1.0 [15] stopped the training process if the global optimum using controlling data set and measures for learning rate (0.1), moment (0.1) and gain (0.9) were met. This feature will avoid over-training phenomena while training the network, since the over-training is accounted as a potential limitation on the use of

Table 4. Results of MLR-based TN model.

Regression model	Variable		Standard Coefficient	Statistics				Sharpio-Wilk test	
	Dependent	Independent		S.E.	r^2	p	VIF*	Statistics	Sig.
MLR	TN	Constant		1.075	0.85	0.001		0.928	0.400
		Human population density	0.002	0.001			1.437		
		Forest	-0.029	0.012			1.591		
		Water Body	-0.637	0.210			1.143		

*VIF= Variance Inflation Factor

ANNs. It occurs when the capacity of the ANNs for training is too much because it is allowed too many training cycles. It degrades the prediction performance of ANNs significantly [17].

Although Maeir and Dandy [7] suggested a trial and error approach for determining the optimum number of nodes in the hidden layer, Hecht-Nielsen [18] proposed a general guideline for specifying upper limits for the number of nodes in the hidden layer as follows:

$$N^H \leq 2N^I + 1 \quad (6)$$

...where N^H is the number of nodes in the hidden layer and N^I is the number of input nodes.

For this study, a combination of trial and error approach and the general guideline which was proposed by Hecht-Nielsen [18] were used. Firstly, the upper limit of number of nodes in the hidden layer was calculated. In this study, it would be equal to 13 nodes. The optimum number of the nodes in the hidden layer was then determined using trial and error so that for a specific number of nodes training the network was repeated five times. Secondly, mean of correlation coefficient value between observed and predicted values for five trials were calculated. Finally, that number of nodes whose mean of the correlation coefficient was highest was selected as the optimum number of nodes in the hidden layer. Fig. 3 indicated that mean correlation coefficient between observed and predicted values reached to its maximum value for two nodes in the hidden layer and then it revealed a steady-state for other tials.

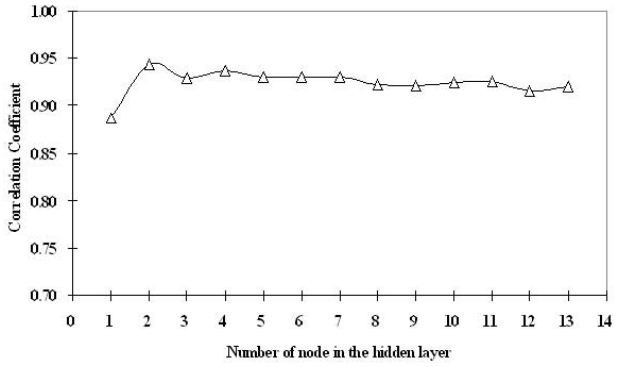

Fig. 3. Correlation coefficient value between the observed and predicted values for five time trials for TN with different node numbers in the hidden layer.

The training task of the network was then carried out using the optimum number of nodes in the hidden layer. The result of training was evaluated using a scatter plot, simple linear regression of observed versus equivalent model prediction and determination of the Pearson's coefficient of regression (r^2) [4].

Results and Discussion

Multiple Regression Models

The backward approach was applied to determine a final regression model representing the linkage between land cover and stream water TN concentration. For TN regression model, the initial fixed variables were area (%) of land cover variables (*urban, forest, agriculture, grassland and water body*), TN and *human population density*. The results of the MLR-based modeling are summarized in Table 4.

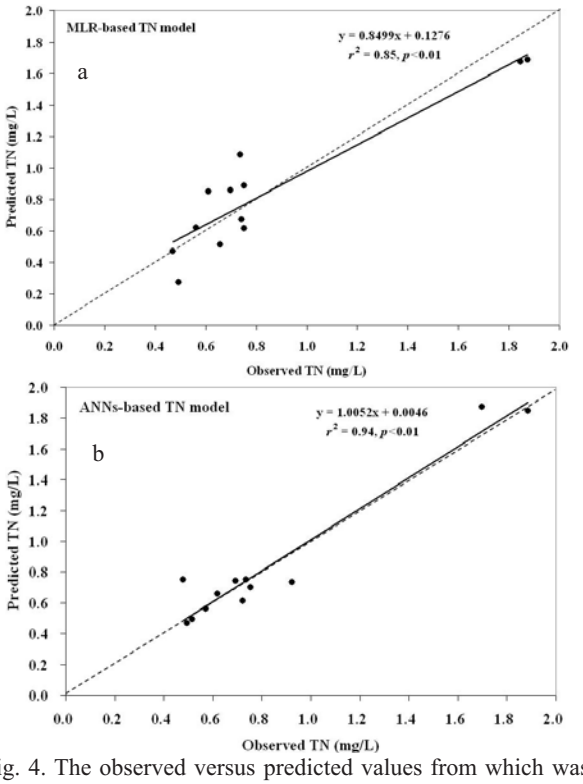

Fig. 4. The observed versus predicted values from which was generated by MLR-based (a) and ANNs-based (b) TN models.

In the TN regression model (r^2= 0.85, p<0.01), 85% of total variations in TN concentration would be explained by change in human population density (%), area (%) of forest and that of water body in the river basin. TN concentration would be decreased by an increase in area (%) of forest and water body. It would mean that these land use types could play as a sink role and trap different forms of nitrogen before entering into the river system. This is while a positive significant relationship was observed between TN concentration in the stream and human population density (%) in the river basin.

Inter-variable collinearity of the MLR-based model was investigated by referring to VIF (Table 4). While a VIF>10 could be considered as severe collinearity within variables in the model [19, 20], the model has revealed no collinearity (VIF<2). Normality of residuals of the model was tested using the Sharpio-Wilk test with a p<0.05 whether it follows a normal distribution (Table 4). The results of the test have suggested the residuals of all models were normally distributed at significance level of p<0.05. A one-by-one relationship between observed versus predicted values for the MLR-based model was illustrated in Fig. 4a.

Training the ANNs Models

The back-propagation algorithm [21] was applied for training the network. Input variables consisted of the area (%) of urban, forest, agricultural, grassland, water body and human population density, and the output of ANNs model was stream water TN concentration. Network architecture of the models includes six nodes (five nodes for land cover type and one for human population density), the optimum number of nodes in the hidden layer; and one output node for TN variable. The appropriate number of the hidden layer node was evaluated referring the correlation coefficient value between the observed and predicted values for five time trials for each node that varied between 0 to 13 (Fig. 3).

Table 5 indicated the results of network training for TN concentration. Formula for TN concentration would be generalized based on Fig. 2 as:

$$z_k = f(I_k) = f\left(\sum_j w_{kj} y_j\right) = f\left(\sum_j w_{kj} f(H_j)\right) = f\left(\sum_j w_{kj} f\left(\sum_i v_{ji} x_i\right)\right) \quad (7)$$

...where;

Z_k output value from output layer of node of zth,

I_k is the input value for activation function at output layer of node of kth,

W_{kj} is weight value from hidden layer of node jth to output layer of node kth,

Y_j is output value from hidden layer node of jth,

H_j is the input value for activation function at hidden layer of node of jth;

v_{ji} is weight from input layer of node ith to hidden layer of node jth, and

x_i is value of input x of node ith (input layer).

Table 5. Result of training the network for TN concentration variable.

Symbol	From		To		TN
	Layer	Node	Layer	Node	
Vji	1	1	2	1	-1.3970
Vji	1	2	2	1	3.0753
Vji	1	3	2	1	0.8101
Vji	1	4	2	1	0.4920
Vji	1	5	2	1	1.4827
Vji	1	6	2	1	-1.4872
Vji	1	1	2	2	0.9171
Vji	1	2	2	2	1.0774
Vji	1	3	2	2	0.3969
Vji	1	4	2	2	0.4689
Vji	1	5	2	2	0.6301
Vji	1	6	2	2	1.0117
Wkj	2	1	3	1	-3.1508
Wkj	2	2	3	1	1.9734
Iteration					11600
RMSE					0.0025

Residual-based validation of the ANNs model was evaluated by plotting the predicted versus observed values and calculating r^2 and the Root Mean Square Error (RMSE). Fig. 4b illustrates the one-by-one relationship between observed versus predicted values by the ANNs model. Table 5 and Fig. 4b indicate RMSE and r^2 for TN. The ANNs modeling approach could achieve a measure of 0.0025, 0.94 for RMSE and r^2 after 11,600 iterations using the 6-2-1 architectural pattern for it. Its architectural pattern stands for 6 input nodes, 2 hidden layer nodes and 1 output node.

Sensitivity Analysis of ANNs and MLR Models

Cross-validation [22] of the ANNs and MLR-based models were not carried out using real data because of limited access to TN concentration data in the study area. Instead, the Monte Carlo Method was considered for analyzing sensitivity of ANNs and MLR-based models that were developed in this study. Table 6 summarized statistics of the data set that was used for analyzing sensitivity of the ANNs and MLR-based models. Based on the Monte Carlo Method for sensitivity analysis of the ANNs and MLR based models, 5,000 random data sets were generated so that each set consisted of five variables representing area (%) of urban, forest, agricultural, grassland and water body in hypothetical river basins and human population density for each.

Table 6. Summary of the statistics of the data set that was used for sensitivity analysis of the ANN model.

Statistics	Urban	Forest	Agriculture	Grassland	Water body	Human Population Density
Mean	19.86	19.82	20.23	19.82	20.27	438.57
Standard Error	0.16	0.16	0.16	0.16	0.16	3.11
Median	19.80	19.82	20.18	19.83	20.18	441.00
Mode	0.00	0.00	0.00	0.00	0.00	404.00
Standard Deviation	11.32	11.52	11.61	11.53	11.63	219.79
Sample Variance	128.16	132.73	134.76	133.04	135.15	48305.84
Kurtosis	-0.30	0.07	-0.20	0.15	-0.09	-1.21
Skewness	0.29	0.39	0.32	0.42	0.36	-0.01
Range	61.29	72.26	70.00	71.43	72.07	754.00
Minimum	0.00	0.00	0.00	0.00	0.00	63.00
Maximum	61.29	72.26	70.00	71.43	72.07	817.00
Count	5000	5000	5000	5000	5000	5000
Confidence Level (95.0%)	0.31	0.32	0.32	0.32	0.32	6.09

Table 7. Results of the distribution fitting of the input variables that were generated for sensitivity analysis, and that of outputs of the developed models.

Variable		Kolmogorov-Smirnov Test	
		Statistics	Statistical Distribution
Input:	Urban	0.0395	Normal
	Forest	0.0425	Normal
	Agriculture	0.0405	Normal
	Grassland	0.0426	Normal
	Water body	0.0404	Normal
	Human Population Density	0.0510	General Extreme Value
ANNs Model Output:	TN	0.1051	Weibull (3P)
MLR Model Output:	TN	0.0254	General Extreme Value

*Critical value is 0.0192 for P<0.05.

In spite of the generation of random numbers that were varied between 0-100 (according to the first condition that was noted in 2.3.2), the approach used for meeting the second condition caused the range of input variables to be changed in 0-70%.

In order to specify what statistical distribution each of the randomly generated variable follows, distribution-fitting approach was applied. Because statistical distribution of variable(s) should be known, they are used for the Monte Carlo Method [23]. The distribution fitting of the variables was carried out by computing the Maximum Likelihood test with maximum 100 iterations and accuracy of 10^{-4}. For goodness of fit, the Kolmogrov-Smirnov test was carried out. The results of the distribution fitting of the input variables were summarized in Table 7. The results indicted that all the input variables except human population density were normally distributed at significance level of $p<0.05$. Following from that, these data sets were entered into the ANNs and MLR-based models that were developed for prediction of TN concentration separately. The output of each model was then computed. For the output of ANNs and MLR-based models, the distribution fitting was carried out one more time in order to determine what statistical distribution the output of the models follow. Moreover, another purpose for distribution fitting was to specify the probability of an event such as $Pr_{(output)}< 0$, since only positive values make sense in relation to the water quality variables in general and for TN in particular.

For determination of the probability of $Pr_{(output)} < 0$, the estimated cumulative distribution curve was plotted for the output of TN models that were developed. The results of the distribution fitting of the output of ANNs and MLR-based models were summarized in Table 7 and illustrated in Fig. 5a.

For the MLR-based TN model, results of the distribution fitting and Kolmogrov-Smirnov tests indicated that output of the model followed general extreme value distribution since the Kolmogrov-Smirnov measure (0.025) was significant ($p<0.05$) (Fig. 5a and Table 7). Although statistics of the model (Table 4) were high ($r^2=0.85$, $p<0.01$), the possibility of generation of negative outputs $Pr(TN<0)$ by the model is significantly high 90% (Fig. 5a) as well. This feature of the developed model has decreased reliability of this model for prediction of TN concentration in basins.

For the ANNs-based TN model, the results of distribution fitting revealed that output of the model has followed Weibull (3P) distribution (Table 7 and Fig. 5b). The event of Pr ($0 < TN > 2$) (mg/L) for the output of ANNs-based TN model was not observed using a plot of the cumulative distribution curve (Fig. 5b). It would be implied that ANNs-based TN model would not react properly to the TN concentration measures more than 2 mg/L in real situations. Therefore, it might be suggested that the capability of prediction of the developed ANNs model for TN would have a variation between 0-2 mg/L.

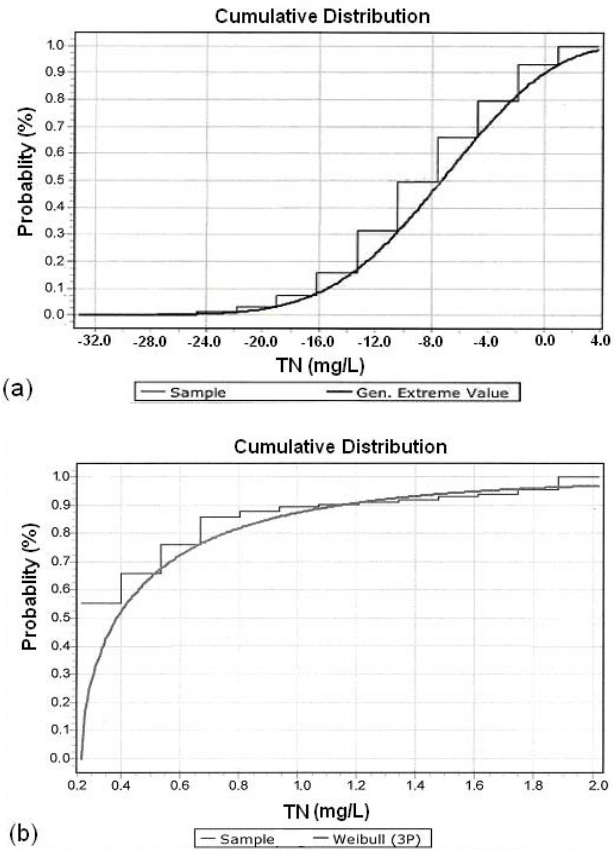

Fig. 5. The cumulative distribution of the output of MLR-based (a) and ANN-based (b) TN model.

Conclusions

Using the ANNs modeling approach could develop the model with a plausible measure of statistics ($r^2=0.94$, $p<0.01$) for TN comparing with that of MLR approach ($r^2=0.85$, $p<0.01$). Therefore, this model would be applied for predicting TN concentration in stream water in the study area.

This study has addressed the issue of the lack of access to a sufficient data set to test the trained network by generating 5,000 random number data sets as hypothetical river basins. Sensitivity of the output of the developed models was then successfully analyzed using those randomly generated data sets based on the Monte Carlo Method. The capability of the models that were developed in this study for prediction of the TN concentration was investigated using estimation of the cumulative distribution of the models' output.

The difference between the maximum value for each input variable (urban, agriculture, grassland, water body and human population density) of the generated data set with that of the real data set, which were used for training the networks, were deliberately set high in order to provide reliable information with evaluation of the behavior of the developed models in response to the conditions that are different from those on which they were based. This initiative provided us with some possibilities to determine upper and lower limits of the prediction capability of the ANNs and MLR-based models that were developed in this study.

Sensitivity analysis could specify that relying on statistics of a model could not provide sufficient information to judge whether it can be applied for prediction of environmental or ecological variables in other areas or river basins. For our case, although statistics of the MLR-based model were satisfactorily high, the result of sensitivity analysis indicated that the application of this model would result in implausible prediction of TN concentration in other river basins. Therefore, sensitivity analysis of any developed model should be carried out and considered as a final step in modeling environmental or ecological variables in further studies.

One of the most challenging tasks in applying data-intensive method (ANNs-based modeling) was tackled by using the Monte Carlo method-based sensitivity analysis. This approach could show how to model when there is serious lack of data. The procedure that was applied to tackle the lack of data would be useful when direct decisions are needed for a problem and we do not have the luxury to wait until data is gathered.

Acknowledgments

A postdoctoral fellowship from Hiroshima University for one of the authors (B.J.A.) is gratefully acknowledged. The authors would like to express their special thanks to Dr. John Elwyn from Cardiff University, England, and Mr. Rudiyanto from Mie University, Japan, for their help in carrying out our study.

References

1. JOHNES K. B., NEALE A. C., NASH M. S., VAN REMORTEL R. D., WICKHAM J. D., RIITTERS H., O'NEILL R. V. Predicting nutrient and sediment loadings to streams from landscape metrics: A multiple watershed study from the United States Mid-Atlantic Region. Landscape Ecology **16**, 301, **2001**.
2. SLIVA L., WILLIAMS D. D. Buffer zone versus whole catchment approaches to studying land use impact on river water quality. Water Research, **35** (14), 3462, **2001**.
3. JARVIE H. P., OGUCHI T., NEAL C. Exploring the linkage between river water chemistry and watershed characteristics using GIS-based catchment and locality analyses. Regional Environment Change, **3**, 36, **2002**.
4. AHEARN D.S., SHEIBLEY R.W, DAHLGREN R.A., ANDERSON M., JOHNSON J., TATE K.W., 2005.Land use and land cover influence on water quality in the last free-flowing river draining the western Sierra Nevada, California, Journal of Hydrology **313**, 234, **2005**.
5. BASNYAT P., TEETER L. D., FLYNN K. M., LOCKABY B. G. Relationships between landscape characteristics and nonpoint source pollution inputs to coastal estuaries. Environmental Management, **23** (4), 539, **1999**.
6. JEONG K. S., KIM D. K., CHON T. S., JOO G. J. Mashine learning application to the Korean freshwater ecosystems, Korean J. Ecol. **28** (6), 405, **2005**.
7. MAIER H. R., DANDY G. C. Neural network based modelling of environmental Variables: A systematic approach, Mathematical and Computer Modelling **23**, 669, **2001**.
8. HANSEN J. V., NELSON R. D. Neural networks and traditional time series methods: A synergistic combination in state economic forecasts. IEEE Transactions on Neural Networks **8**(4), 863, **1997**.
9. JAPANESE STATISTICS BUREAU of MINISTRY OF INTERNAL AFFAIRS AND COMMUNICATION: http://www.stat.go.jp/data/kokusei/2000/final/zuhyou/008-01.xls
10. JAPANESE STANDARD ASSOCIATION. Testing methods for industrial wastewater: K0102, Tokyo. **1998**.
11. ESRI (Environmental Systems Research Institute). ArcView GIS Software, Redlands, California, USA. **1999**.
12. INTEGRATED LAND AND WATER INFORMATION SYSTEM (ILWIS). The Remote sensing and GIS software. ITC, the Netherlands. **2004**.
13. LEK S, GUIRESSE, GIRAUDEL J. L. Predicting stream nitrogen concentration from watershed features using neural networks, Wat. Res. **33**(16), 3469, **1999**.
14. SPITZ F., LEK S. environmental impact prediction using neural network modeling: an example for wildlife damage, Journal of applied Ecology, **36**, 317, **1999**.
15. RUDI Y., SETAIWN I. Backpropagation Neural Network 1.0, Bogor Agricultural University. Indonesia. Personal Correspondence. **2003**.
16. RECH G. Forecasting with Artificial Neural Network Models, SSE/EFI Working Paper Series in Economics and Finance, No. **491**, 38 , **2002**.
17. ILIADIS L. S., MARIS F. An Artificial Neural Network model for mountainous water-resources management: The case of Cyprus mountainous watersheds, Environmental Modelling & Software **7** (22), 1066, **2007**.
18. HECHT-NIELSEN R. Kolmogorov's mapping neural network existence theorem. Proceedings of the First IEEE International Joint Conference on Neural Networks, San Diego, California, pp. 11-14, IEEE, New York. **1987**.
19. NETER J., KUNTER H.M., NACHTSHEIM C.J. WASSERMAN W. Applied Linear Statistical Models, Irwin, Chicago, Illinois, USA. **1996**.
20. CHATTERJEE S., HADI A.S., PRICE B., The Use of Regression Analysis by Example, John Wiley and Sons, New York, USA. **2000**.
21. RUMELHART D.E., HINTON G.E., WILLIAMS R.J. Learning internal representations by error propagation. In: Rumelhart, D.E., McClelland, J.L. (Eds.), Parallel Distributed Processing. MIT Press, Cambridge. **1986**.
22. KOZAK A., KOZAK R. Does cross validation provide additional information in the evaluation of regression models?. Can. J. For. Res. **33**, 976, **2003**.
23. WITWER J.W. **2004**. Monte Carlo Simulation Basics, http://vertex42.com/ExcelArticles/mc/MonteCarloSimulation.html, downloaded on 2007/11/15.