World Academy of Science, Engineering and Technology
International Journal of Bioengineering and Life Sciences
Vol:1, No:9, 2007

# Morphometric Analysis of *Tor tambroides* by Stepwise Discriminant and Neural Network Analysis

M. Pollar, M. Jaroensutasinee, and K. Jaroensutasinee

Open Science Index, Bioengineering and Life Sciences Vol:1, No:9, 2007 publications.waset.org/7530.pdf

***Abstract*—**The population structure of the *Tor tambroides* was investigated with morphometric data (i.e. morphormetric measurement and truss measurement). A morphometric analysis was conducted to compare specimens from three waterfalls: Sunanta, Nan Chong Fa and Wang Muang waterfalls at Khao Nan National Park, Nakhon Si Thammarat, Southern Thailand. The results of stepwise discriminant analysis on seven morphometric variables and 21 truss variables per individual were the same as from a neural network. Fish from three waterfalls were separated into three groups based on their morphometric measurements. The morphometric data shows that the nerual network model performed better than the stepwise discriminant analysis.

***Keywords*—**Morphometric, *Tor tambroides*, Stepwise Discriminant Analysis , Neural Network Analysis.

## I. Introduction

GEOGRAPHICAL isolation can result in the development of different morphological features between fish populations because the interactive effects of environment, selection, and genetics on individual ontogenies produce morphometric differences with in a species [1], [2]. The quantification of specific characteristics of an individual, or group of individual can demonstrate the degree of speciation induced by both biotic and abiotic conditions, and contribute to the definition of differrent stock of species [3]. The concept of geographical structure in fish population is fundamental for population dynamics and management of fisheries [1]-[4]. More recently, the image analysis systems play a major role in the development of morphometric techniques, boosting the utility of morphometric research in fish population identification [2]. Data on morphometric measurements are able to identify differences between fish populations [4]-[11], and used to describe the shape of each fish [8].

Multivariate techniques are widely used tools in ecological studies to assess the relationships between biological communities [12]. These techniques can yield information complementary to that derived from biochemical, physiological and life history studies [4]. A stepwise discriminant analysis of morphometric characters is a powerful technique to investigate the geographical variation of stocks [4]. This technique is a traditional multivariate of morphometric data [6].

Since a few years ago, the artifitial neural networks (ANNs) have become one of the most promising tools for predicting [13] and solving problems of differentiating between groups [14]. The ANNs are non-linear mathematical structures capable of representing the complex non-linear process that relates the inputs to the outputs of a system [13]. There is no need to specify a particular model. Rather, the model is an adaptive form based on the features present from the data. This data-driven approach is suitable for many empirical data sets where no theoretical guidance is available to suggest an appropriate data generating process [15]. ANNs models have been increasingly applied in many fields of science and usually providing highly satisfactory results [13].

This study aimed at examining the morphometric variability of *Tor tambroides* populations at Sunanta, Nan Chong Fa and Wang Muang waterfalls, Khao Nan National Park, Thailand using Stepwise Discrimant Analysis (SDA) and Neural Network Analysis (NNA).

## II. Materials and Methods

### A. Fish Biology

*Tor tambroides* is in the Cyprinidae family typically inhabited waterfalls and have a long large flat torso with a long mental lobe and small head, green brown colours, large scales and 15-20 cm in body. *Tor tambroides* are sexually mature when mall, silvery with yellow, orange, pink or pale red fins.

### B. Study Site

Nakhon Si Thammarat is a southern province bordering the Gulf of Thailand located at 8° 22′- 8° 45′ N 99° 37′- 99° 51′. Khao Nan National Park covers 436 km$^2$ encompassing a huge variety of wildlife, including mountains, forests, rivers and waterfalls (Fig. 1).

Manuscript received November 30, 2006. This work was supported in part by PPT Public Company Limited, the TRF/BIOTEC Special Program for Biodiversity Research and Training grant BRT T_148016, and CXKURUE, the Institute of Research and Development, Walailak University.

Pollar M. is with School of Science, Walailak University, 222 Thaiburi, Thasala District, Nakhonsithammarat 80161, Thailand and Ratjabhat Nakhonsithammarat University, Muang, Nakhon Si Thammarat 80280, Thailand (phone: +66 75 672 075; Fax: +66 75 672 038; e-mail: m_pollar @hotmail.com).

Jaroensutasinee M. is with School of Science, Walailak University, 222 Thaiburi, Thasala District, Nakhonsithammarat 80161, Thailand (phone: +66 75 672 005; fax: +66 75 672 004; e-mail: jmullica@wu.ac.th).

Jaroensutasinee K. is with School of Science, Walailak University, 222 Thaiburi, Thasala District, Nakhonsithammarat 80161, Thailand (phone: +66 75 672 005; fax: +66 75 672 004; e-mail: krisanadej@gmail.com).

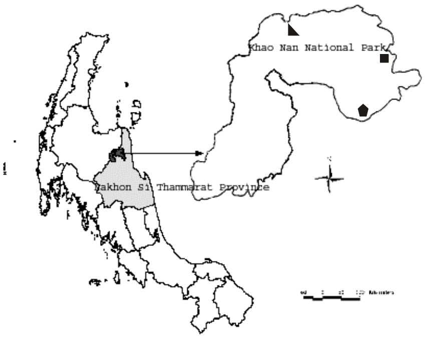


Fig. 1 Location of sample sites from (▲) Sunanta, (■) Nan Chong Fa and (♦) Wang Muang waterfalls

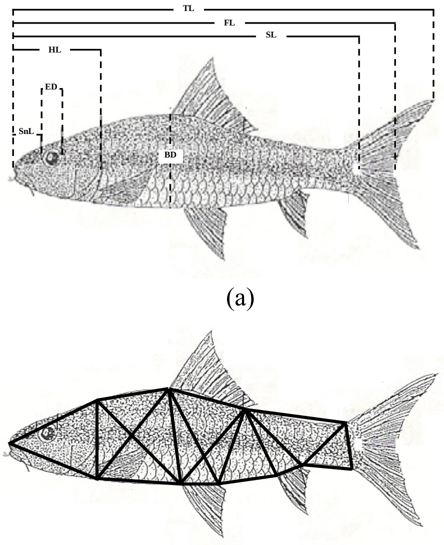
### C. Sampling Measurement

A total of 116 specimens of *Tor tambroides* were collected during July-August 2006 from three waterfalls: Sunanta, Nan Chong Fa and Wang Muang waterfalls. Digital photographs were taken with a Sony DSC-F717 on the left side of each fish. All measurements were taken with a digital caliper by *Mathematica* 5.2 at ± 0.01 mm. Each individual was measured using seven morphometric variables including Total length (TL), Fork length (FL), Standard length (SL), Head length (HL), Snout length (SnL), Eye diameter (ED), Body depth (BD), (Fig. 2a) and 21 truss variables (Fig. 2b).

Geographic variation of size was usually assessed by allometric analysis. This analysis provides a method to elucidate the relationship between process of growth and evolution. Most of the variability in a set of multivariate character is due to size. Thus, shape analysis should be free from the effect of size to avoid misinterpretation of the results. The morphometric data were transformed to common logarithms in order to increase linearity and multivariate normality [5]. Size-dependent variation was removed using an allometric approach [Res]. Data were transformed using (1)

$$M_{trans} = logM\text{-}b(logSL\text{-}logSL_{mean}) \qquad (1)$$

where $M_{trans}$ is the transformed measurement, $M$ is the original measurement, $b$ is the within-group slope regression of the *logM* versus *logSL*, *SL* is the standard length of the fish, and $SL_{mean}$ is the overall mean of the standard length.



(a)

(b)

Fig. 2 (a) Morphometric measurement per individual (b) Truss measurement per individual

### D. Stepwise Discriminant Analysis

Stepwise estimation is an alternative to the simultaneous approach. The stepwise approach begins by choosing the single best discriminating variables. The initial variable is then paired with each of the other independent variables one at a time. The variables are best able to improve the discriminating power of the function in combination with the first variable chosen. The third and any subsequent variables are selected in a similar manner. As additional variables are included in the model, some previously selected variables may be removed if the information they contained about group differences is available in some combination of the other variables included at later stages. If a stepwise method is used to estimate the discriminant function, the Mahalanobis $D^2$ and Rao's V measures is most appropriate. Both are measures of general distance. The Mahalanobis $D^2$ procedure is based on generalized squared Euclidean distance that adjusts for unequal variances [16].

A stepwise discriminant analysis was performed to investigate the integrity of the pre-defined groups. This statistical analysis builds a predictive model of group membership based on observed characteristics of each sample [1] (in each case the membership of each scale to sites Sunanta, Nan Chong Fa and Wang Muang waterfalls). Each individual was allocated to the group with nearest centroid, and the proportion of individuals allocated to each group was calculated.

A cross-validation testing procedure was performed to assess the ability of the selected variables to predict fish from the three sites [1]. In cross-validation, one individual is removed from the original matrix. The discriminant analysis is then performed from the remaining observations and used to classify the omitted individual. The proportion of individuals correctly re-allocated was taken as a measurement of the integrity of that group. Differences were tested with a stepwise discriminant analysis with variables entered in a forward manner using $F = 3.84$ for entering, and $F = 2.71$ for removal. Statistical differences were considered significant where $P$-value = 0.05 [1], [16]. The statistical treatment was carried out with SPSS 12.0 for Windows.

### E. Neural Network Analysis

The structure of a neural net consists of connected units referred to as "nodes" or "neurons". Each neuron performs a portion of the computations inside the net: a neuron takes some numbers as inputs, performs a relatively simple computation on these inputs, and returns an output. The output value of a neuron is passed on as one of the inputs for another neuron, except for neurons that generate the final output values of the entire system.

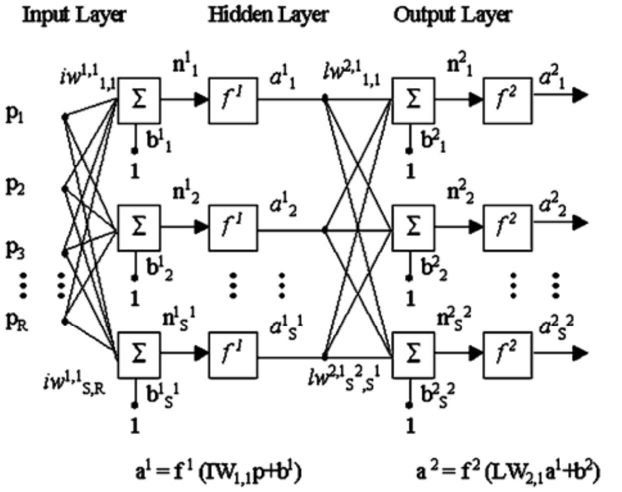


Fig. 3 A schematic diagram of an artificial neural network where *R* is the number of input variables and *S* is the number of neurons in the hidden layer [17]

A schematic diagram of an artificial network is shown in Fig. 3, where the small black circles are artificial neurons and the connections represent weights that describe the importance of the signal being transmitted along a given path [17]. Thus, the net input to a given neuron would be given by (2).

$$a = f\,(Wp + b) \tag{2}$$

where *p* is the total signal being transmitted from one neuron to the next, along a single axon *W* is the weight function for that connection, and *f* is a transfer function.

In ANN the scalar input *p* is transmitted through a connection that multiplies its strength by the scalar weight *W* to form the product *Wp* plus a bias *b*. This sum is the argument of the transfer function f that produces the scalar output *a*. The bias is much like a weight, except that it has a constant input of 1. The transfer function *f* is typically a step function or a sigmoid function, that takes the argument *Wp+b* and produces the output a. In most ANN applications, a network with one hidden layer is used in Fig. 3 as in (3).

$$a^2 = f^2(LW^{2,1}\, f^1(IW_{1,1}p + b^1) + b^2) \tag{3}$$

where $a^2$ is the total scalar output, $f^2$, $f^1$ the transfer functions, *LW* and *IW* the scalar weights, $b^1$, $b^2$ are the bias for the hidden and the input layer, respectively. The way in which each node transforms its input depends on the so-called "connection weights" (or "connection strength") and "bias" of the node that can be modified in the training process. The output of each node to another node or the external world then depends on both its weight strength and bias and on the weighted sum of all its inputs, which are then transformed by a weighting function (usually sigmoidal) referred to as its activation function. Specifically, NeuralTools 1.0 uses the hyperbolic tangent function [19].

Multi-Layer Feed forward Networks (MLFN) (also referred to as Multi-Layer Perceptron Networks") are systems capable of approximating complex functions, and thus capable of modeling complex relationships between independent variables and a dependent one. When MLFN nets are used for classification, they have multiple output neurons, one corresponding to each possible dependent category. A net classifies a case by computing its numeric outputs; the selected category is the one corresponding to the neuron that outputs the highest value.

When using NeuralTools 1.0 [19], neural networks are developed and used in four steps:

1) Data Preparation

A Data Set Manager is used to set up data sets so they can be used over and over again with your neural networks.

2) Training

A neural network is generated from a data set comprised of cases with known output values. These data often consist of historical cases for which you know the values of output/dependent variable.

3) Testing

A trained neural network is tested to see how well it does at predicting known output values. The data used for testing is usually a subset of your historical data. This subset was not used in training the network. After testing, the performance of the network is measured by statistics such as the % of the known answers it correctly predicted.

4) Prediction

A trained neural network is used to predict unknown output values. Once trained and tested, the network can be used as needed to predict outputs for new case data.

The variable impact analysis is used to measure the sensitivity of net predictions to changes in independent variables. This analysis is only done on training data. As a result of the analysis, every independent variable is assigned a "Relative Variable Impact" value; these are percent values and add to 100%. The lower the percent value for a given variable, the less that variable affects the predictions. The results of the analysis can help in the selection of a new set of independent variables, one that will allow more accurate predictions. For example, a variable with a low impact value can be eliminated in favor of some new variable. However, one needs to keep in mind that the results of the Impact Analysis are relative to a given net.

## III. Results and Discussion

### A. Stepwise Discriminant Analysis

When ten morphometric variables (i.e. FL, SL, ED, AB, CK, DE, DH, DI, DJ and HI) had been entered, Wilks'λ dropped to 0.064 with a significant difference among three waterfalls ($F$ = 30.616, $P$<0.001, Table I).

The first canonical discriminant function of the discriminant analysis explained 73.1% of the total variance while the second one accounted for 26.9% of the total variance. The plot of the two canonical variables shows a complete separation among three waterfalls (Fig. 4). The discriminant analysis correctly classified 111 of the 116 fishes (i.e. 95.7%), while the cross-validation testing procedure correctly classified 98 of the 116 fishes (i.e. 93.1%) of the fishes.

For both analyses, fish collected from Sunanta waterfall were the most correctly classified fish (Table II, and III), and followed by fish collected from Wang Muang waterfall (Table II, and III). The least corrected classification were fish collected from Nan Chong Fa waterfall (Table II, and III). The plot of the two canonical variables showed a complete separation between the three groups (Fig. 4).

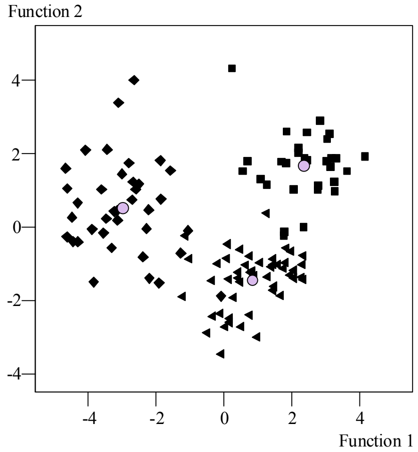


Fig. 4 Scatter plot of the canonical scores from the discriminant analysis of (▲) Sunanta, (■) Nan Chong Fa and (♦) Wang Muang waterfalls

### B. Neural Network Analysis

When morphometric variables were entered, the percent of relative variable impact reduced from 7.5% to 0.7% (Fig. 5).

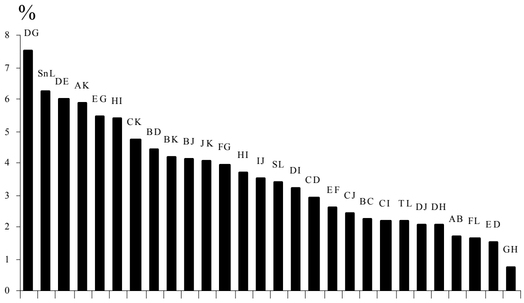


Fig. 5 Percent of Relative Variable Impact

The neural network analysis correctly classified 91 of the 93 fishes (i.e. 97.8%) from training procedure, 20 of the 23 fishes (i.e. 87.0%) from testing procedure and 111 of the 116 fishes (i.e. 95.7%) from predicting procedure (Table III-V). Fish collected from Wang Muang waterfall were gave the highest correct classification, followed by fish collected from Sunanta waterfall and Nan Chong Fa waterfalls (Table VI).

TABLE I
Summary of Stepwise Discriminant Analysis

| Step | Variable | Wilks'λ | $F$-statistics | Sig. |
|---|---|---|---|---|
| 1 | FL | 0.284 | 142.514 | 0.000 |
| 2 | SL | 0.175 | 78.055 | 0.000 |
| 3 | ED | 0.161 | 55.352 | 0.000 |
| 4 | AB | 0.122 | 51.279 | 0.000 |
| 5 | CK | 0.111 | 43.557 | 0.000 |
| 6 | DE | 0.097 | 39.779 | 0.000 |
| 7 | DH | 0.089 | 35.900 | 0.000 |
| 8 | DI | 0.079 | 33.913 | 0.000 |
| 9 | DJ | 0.070 | 32.505 | 0.000 |
| 10 | HI | 0.064 | 30.616 | 0.000 |

TABLE II
Classification Results for the Discriminant Analysis (Original)

| Site | Predicted Group Membership | | | Correct Total |
|---|---|---|---|---|
| | Sunanta | Nan Chong Fa | Wang Muang | |
| S | 36 | 0 | 1 | 37 |
| | (97.3) | (0.0) | (2.7) | (100.0) |
| N | 0 | 28 | 2 | 30 |
| | (0.0) | (93.3) | (6.7) | (100.0) |
| W | 1 | 1 | 47 | 49 |
| | (2.0) | (2.0) | (95.7) | (100.0) |

95.7% of selected original grouped cases correctly classified.

TABLE III
Classification Results for the Cross-Validate Testing Procedure

| Site | Predicted Group Membership | | | Correct Total |
|---|---|---|---|---|
| | Sunanta | Nan Chong Fa | Wang Muang | |
| S | 33 | 1 | 3 | 37 |
| | (89.2) | (2.7) | (8.1) | (100.0) |
| N | 0 | 28 | 2 | 30 |
| | (0.0) | (93.3) | (6.7) | (100.0) |
| W | 1 | 1 | 47 | 49 |
| | (2.0) | (2.0) | (95.7) | (100.0) |

93.1% of selected cross-validated grouped cases correctly classified.

TABLE IV
Classification Results for the Neural Network (Training Cases)

| Site | Predicted Group Membership | | | Correct Total |
|---|---|---|---|---|
| | Sunanta | Nan Chong Fa | Wang Muang | |
| S | 32 | 0 | 1 | 33 |
| | (97.0) | (0.0) | (3.0) | (100.0) |
| N | 1 | 24 | 0 | 25 |
| | (4.0) | (96.0) | (0.0) | (100.0) |
| W | 0 | 0 | 35 | 35 |
| | (0.0) | (0.0) | (100.0) | (100.0) |

TABLE V
CLASSIFICATION RESULTS FOR THE NEURAL NETWORK (TESTING CASE S)

| Site | Predicted Group Membership | | | Correct Total |
|---|---|---|---|---|
| | Sunanta | Nan Chong Fa | Wang Muang | |
| S | 3<br>(75.0) | 0<br>(0.0) | 1<br>(25.0) | 4<br>(100.0) |
| N | 0<br>(0.0) | 4<br>(80.0) | 1<br>(20.0) | 4<br>(100.0) |
| W | 0<br>(0.0) | 1<br>(7.14) | 13<br>(92.86) | 14<br>(100.0) |

TABLE VI
CLASSIFICATION RESULTS FOR THE NEURAL NETWORK (PREDICTING CASES)

| Site | Predicted Group Membership | | | Correct Total |
|---|---|---|---|---|
| | Sunanta | Nan Chong Fa | Wang Muang | |
| S | 35<br>(94.6) | 0<br>(0.0) | 2<br>(5.4) | 37<br>(100.0) |
| N | 1<br>(3.3) | 28<br>(93.3) | 1<br>(3.3) | 30<br>(100.0) |
| W | 0<br>(0.0) | 1<br>(2.0) | 48<br>(98.0) | 49<br>(100.0) |

The result of the multivariate analysis 7 morphometric variables and 21 truss variables showed differences between all three waterfalls. From the discriminant analysis and neural network analysis of three populations were distinguished, belonging to the three sampling sites. The first discriminant function opposed individuals from Sunanta waterfall and Nan Chong Fa waterfall, while the second discriminant function discriminates Nan Chong Fa waterfall from Wang Muang waterfall. A clear geographical gradient occurs among waterfalls suggesting that in some cases, the fish from these areas represent three separate groups.

Morphometric studies have been able to identify differences between fish populations. Therefore, morphometric measurement is a helpful tool for the discrimination of fish populations [3], [7], [8]. Morphometric measurements combining with image analysis are steps ahead to produce a better understanding of fish stock structures [3]. Within this context, our study highlights that morphology can be successful used to discriminate fish populations within waterfall as a fine spatial scale, i.e. a waterfall stretch [2]. The use of fish scale morphology is an easy-to-implement method, relatively rapid, inexpensive nor require fish sacrifice [2]. Since the identification of populations and their connectivity between each other is a major point for conservation and management of vulnerable species, the use of scale morphology to this purpose appears to be very promising [2].

## ACKNOWLEDGMENTS

We thank Mr.Michael Sullivan for comments on previous versions of this manuscript. This work was supported by PPT Public Company Limited, the TRF/BIOTEC Special Program for Biodiversity Research and Training grant BRT T_148016, and CX-KURUE, the Institute of Research and Development, Walailak University.

## REFERENCES

[1] N. Poulet, Y. Reyjol, H. Collier, and S. Lek, "Does fish scale morphology allow the identification of populatio *leuciscus burdigalensis* in river Viaur (SW France)," *Aquat. Sci*., vol. 67, pp. 122-127, 2005.

[2] S. H. Cardin and K. D. Friedland, "The utility of image processing techniques for morphometric analysis and stock identification," *Fisher. Research*, vol. 43, pp. 129-139, 1999.

[3] K. M. Bailey, "Structural dynamics and ecology of flatfish populations," *J. Sea Research*, vol. 37, pp. 269-280, 1997.

[4] A. G. Murta, "Morphological variation of horse mackerel (*Trachuvus trachurus*) in the Iberian and North African Atlantic: implications for stock identification," *J. Mar. Sci.*, vol. 57, 1240-1248, 2002.

[5] A. Pinheiro, C. M. Teixeira, A. L. Rego, J.F. Marques, H.N.Cabral, "Genetic and morphological variation of Solea lascaris (Risso, 1810) along the Portuguese coast," *Fisheries research*, vol. 73, pp. 67-78, 2005.

[6] A. Silva, "Morphometric variation among sardine (*Sardina pilchardus*) populations from the northestern Allantic and the Western Mediterranean," *J. Mar. Sci.*, vol. 60, pp. 1352-1360, 2003.

[7] F. Saborido-Rey and K. J. Nedreaas, "Geographic variation of *Sebastes mentella* in the Northeast Arctic derived from a morphological approach," *J. Mar. Sci.*, vol. 57, pp. 965-975, 2000.

[8] J. Palma and J. P. Andrade, "Morphological study of *Diplodus sargus*, *Diplodus puntazo*, and *Lithognathus mornurus* (Sparidae) in the Eastern Atlantic and Mediterranean Sea," *Fisher. Research*, vol. 57, pp.1-8, 2002.

[9] J. P. Salani, D. A. Milton, M. J. Rahman, and M. G. Hussian, "Allozyme and morphological variation throughout the geographic range of the tropical shad, hila Tenualosa ilisha," *Fisher. Research*, vol. 66, pp. 53-69, 2004.

[10] K. Vidalis, "Discrimination between population of picarel (*Spicara smaris* L., 1758) in the Aegean Sea, using multivariate analysis of phonetic characters," *Fisher. Research*, vol. 30, pp.191-197, 1997.

[11] W. R. Bowering, "An analysis of morphometric characters northwest Atlantic using a multivariate analysis of covariance," *Can. J. Fisher. Aquat. Sci.*, vol. 45, pp. 580-585, 1998.

[12] K. A. Smith, "A simple multivariate technique to improve the design of a sampling strategy for age-based fishery monitoring," *Fisher. Research*, vol. 64, pp. 79-85, 2003.

[13] I. Pulido-Calvo and M. M Portela, "Application of neural approaches to one-step daily flow forecasting in Portuguese watersheds," *J. Hydrol.*, to be published.

[14] G. Winterer, M. Ziller, B. Kloppel, A. Heinz, L. G. Schmidt, and W. M. Herrmann, "Analysis of quantitative EEG with Artificial neural networks and discriminant analysis – A methodological comparison," *Neuropsychobiol.*, vol. 37, pp. 41-48, 1998.

[15] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, pp. 159-175, 2003.

[16] J. F. Hair, E. A. Rolph, L. T. Roland, and C. B. William, "*Multivariate data analysis*," New Jersersy: Prentice Hall, 1995, ch. 5.

[17] Z. Ramadan, S. Xin-Hua, K. H. Philip, J. J. Mara, and M. S. Kate, "Variable selection in classification of environmental soil samples for partial least square and neural network models," *Anal. Chem. Acta*, vol. 446, pp. 233-244, 2001.

[18] D. P. Swain and C. J. Foote, "Stocks and chameleons: the use of phenotypic variation in stock identification," *Fisher. research*, vol. 47, pp. 113-128, 1999..

[19] *Guide to Using Neural Tools*, Palisade Corporation, New York, 2005.