Scientific Research and Essays Vol. 6(25), pp. 5279-5284, 30 October, 2011
Available online at http://www.academicjournals.org/SRE
DOI: 10.5897/SRE10.866
ISSN 1992-2248 ©2011 Academic Journals

Full Length Research Paper

A self organizing map (SOM) guided rule based system for freshwater tropical algal analysis and prediction

M. Sorayya[1]*, S. Aishah[1], B. Mohd. Sapiyan[2] and S. A. Sharifah Mumtazah[3]

[1]Institute of Biological Sciences (ISB), University Malaya, Kuala Lumpur, Malaysia.
[2]Faculty of Science Computer and Information Technology, University Malaya, Kuala Lumpur, Malaysia.
[3]Faculty of Engineering, Universiti Putra Malaysia (UPM), Serdang, Malaysia.

Accepted 23 May, 2011

This paper describes the feasibility study of applying a hybrid combination of Kohonen self organizing feature maps (SOM) and a rule based system in predicting the biomass of selected algae division (Chlorophyta) at tropical Putrajaya Lake (Malaysia). The system was trained and tested on an over five years of limnological time-series data sampled from Putrajaya Lake. Results from trained SOM were used to extract rules of relationships between input variables and the Chlorophyta biomass which was used to construct a rule based system. Selected input variables were water temperature, Secchi depth and nitrate nitrogen (NO_3-N). The rules extracted conformed to findings as postulated in literatures. The overall rule based system yielded an accuracy of 73%.

Key words: Kohonen self organizing feature maps, prediction, rule based system, chlorophyta.

INTRODUCTION

The study of algae biomass is crucial since it provides an indication to assess water quality in both moving and still water ecosystems. However, the study of algae biomass remains a difficult task since the temporal dynamics of algal communities are influenced by a complex array of biotic and abiotic factors operating through both direct and indirect pathways (Carrillo et al., 1995; Vanni and Temte, 1990; Sommer, 1989). Therefore, machine learning techniques such as artificial neural networks (ANNs) have help to unravel the complexity of algal population dynamics. ANNs has been proven to be more effective and more robust in modeling data with a high level of non-linearity as compared to the traditional linear regression approach (Lek et al., 1996). There are two types of ANNs training methods, namely supervised and unsupervised. Supervised type of ANNs models have been successfully implemented for eutrophication modeling and lake management in ecology (Melesse et al., 2008; Sorayya et al., 2009, 2010; Recknagel et al., 1997, 2006; Maier et al., 1998; Wilson and Recknagel, 2001). However, these models are mostly 'black box' in nature whereby the knowledge is hidden within the system parameters and little is made known in understanding the relationship of algae dynamics with regard to the environmental factors. Meanwhile, self organizing feature map (SOM) which is an unsupervised type of ANNs allows knowledge discovery. SOM reduces the dimensions of data of a high level of complexity and plots the data similarities through clustering technique (Kohonen, 2001). SOM has been used effectively in ecological modeling of temperate water bodies (Chon et al., 1996; Foody, 1999; Recknagel et al., 2005). But it is only limited to clustering and ordination of ecological data and not much coverage has been given in application of SOM for knowledge and pattern discovery (Ultsch and Korus, 1995). Thus, the aim of this study is to apply SOM for knowledge discovery of selected algae (Chlorophyta) pattern abundance at tropical Putrajaya Lake with regards to environmental parameters such as water temperature, nitrate nitrogen and Secchi depth. Although Chlorophyta are universally distributed, knowledge of the algae is mainly based on studies in temperate waters whereas studies on Chlorophyta in tropical waters have been limited to their taxonomy and distribution.

*Corresponding author. E-mail: sorayya@um.edu.my.

Table 1. Principal features of Putrajaya Lakes.

Catchment area	Water level	Surface area	Storage volume	Average depth	Average catchment inflow	Average retention time	Circulation type
50.9 km^2	RL 21 m	400 ha	26.5 million cubic meters	6.6 m	200 million liters per day	132 days	Warm polymictic (non stratified shallow lake)

Table 2. Summary statistics of limnological variables of Putrajaya Lake from 2001 to 2006.

Limnological variable	Mean	Minimum	Maximum
Water temperature, ℃	30.35	26.76	34.54
Nitrate nitrogen (NO_3-N) mg/L	1.14	0.00	4.88
Secchi depth (m)	1.15	0.01	2.40
Chlorophyta (cell/ml)	46	1	537

Chlorophyta comprises about 26% of algae population in Putrajaya Lake and amongst its most dominant genera include desmids group such as *Staurastrum*, *Cosmarium, Closterium* and *Pediastrum* and micro-green algae such as *Scenedesmus*, *Chlamydomonas* and *Chlorella.* Desmids are generally more common and diverse in oligotrophic lakes and ponds (Gerrath, 1993). They are modeled in this study as they are highly sensitive to changes in the environmental parameters that could be considered as bioindicators for monitoring water quality (Coesel, 1983, 2001).

MATERIALS AND METHODS

Experimental data preparation

Putrajaya Lake is located at the south of the densely inhabited Klang Valley, Malaysia. It is a non stratified shallow warm polymictic lake. Putrajaya Lake adopted the multi-cell and multi-stage design strategy, and was developed in 1997. The lake covers 400 ha area and was created by submerging the valley of Sungai Chua and Sungai Bisa. Studies carried out reported that major inflows from upstream sources and outside of Putrajaya development boundary contains high level of pollutants which is below the standard set by Perbadanan Putrajaya. Nutrient loading to Putrajaya Lake are predominantly caused by non-point sources from maintenance of plant and turf area by using agrochemicals and fertilizer and development of surrounding area which involves land clearing and soil disturbance. The removal of vegetation in surrounding area accelerates mineral lost through surface runoff water, which is released into the lake water bodies. The lakes were developed primarily for conservation, recreation and esthetical purposes (Putrajaya Corporation, 1998). Therefore it is necessary to maintain Putrajaya Lake water quality. Table 1 lists principal features of Putrajaya Lake (Putrajaya Corporation, 2000).

Limnological parameter data used in this study was compiled over a period of five years (August 2001 to May 2006). Data samples were collected monthly from 23 monitoring stations. The input variables used in this study are selected using correlation analysis. The selected variables are water temperature, Secchi depth and nitrate nitrogen (NO_3-N). The data was further categorized into two different sets which are used for training and testing purposes, to avoid biased results. The categorization is as follows:

Set A: Data for training were obtained from 15 sampling stations
Set B: Data used for testing are from data from 8 sampling stations

Sampling procedures including preservation for water quality parameters were carried out in accordance with WHO (1987) and APHA (1995). The analytical methods for the measured parameters were adopted from manual published by the American Public Health Association (APHA, 1995). Whereas, phytoplankton samples (cell ml^{-1}) were collected using 50 ml vials and preserved in four percent formalin. Net samples were obtained using plankton net with mesh size of about 30 µm. Species of algal were then identified by means of a Nikon light microscope (x1000) (Salleh, 1996) and phytoplankton counts were made using the sedimentation inverted microscope technique. Summary of all the variables used in this study is given in Table 2.

SOM for analysis

The SOM module was applied according to Kohonen (1995) and developed using the SOM toolbox in MATLAB (Matlab, 2006). As a result of the training of SOM by means of normalized data using logistic function to range of 0 to 1, the Euclidian distance between the inputs are calculated and visualized as distance matrix (U-matrix) and a partition map (K-means). The K-means algorithm is implemented in SOM to generate data clusters. Data from dataset A is being used for this purpose. The K-means algorithm divides the dataset into a number of clusters by minimizing energy function. K-means clustering in SOM in this study is carried out using Davies-Bouldin index (DB) (Davies and Bouldin, 1979).

SOM for propositional rule based system development

The data mining and knowledge discovery process starts with the

Table 3. Extracted rules for Chlorophyta rule based system.

Rule number	Cluster number	Extracted rule
1	1	If temperature > 30.4, secchi > 1.08 and nitrate ≤ 1.31, then Chlorophyta is medium
2	2	If temperature > 30.4, secchi ≤ 1.08 and nitrate > 1.31, then Chlorophyta is low
3	4	If temperature ≤ 30.4, secchi > 1.08 and nitrate ≤ 1.31, then Chlorophyta is low
4	3	If temperature > 30.4, secchi > 1.08 and nitrate > 1.31, then Chlorophyta is low
5	5	If temperature > 30.4, secchi ≤ 1.08 and nitrate ≤ 1.31, then chlorophyta is low
6	6	If temperature ≤ 30.4, secchi > 1.08 and nitrate > 1.31, then Chlorophyta is low
7	7, 9	If temperature ≤ 30.4, secchi ≤ 1.08 and nitrate ≤1.31, then Chlorophyta is low
8	8	If temperature ≤ 30.4, secchi ≤ 1.08, and nitrate > 1.31, then Chlorophyta is low

Table 4. Chlorophyta rule based system accuracy.

Variable	Training result	Testing result
Type of data set	Data set A	Data set B
Number of data set	536	230
Percentage of accuracy (%)	73	73

training of the SOM network. The process completes when the figurative rules are extracted. Cluster boundaries are interpreted in the form of rules which are explicit to the users. The rules to model and predict the algal biomass are extracted by mapping the clusters generated from the clustering map with the input variables component planes, respectively. The rule based system is developed based on the propositional IF…THEN … ELSE type of rules. These rules describe the characteristics of each cluster using Visual C++ programming. Data from dataset A is utilized to train the SOM that generates the component planes and cluster map. However, prior to the rules extraction, the input and output variables are labeled into two category. Putrajaya Lake trophic status is an oligotrophic lake, which is defined as having low productivity. The level of the water quality is well controlled as the diversity of the species is high with low number of individual phytoplankton. This limits the possible categorization of input and output variables to two ranges of categories only. The variables threshold range for each category is determined from the component planes of each variable are generated from SOM training. Threshold between less than 70 cell/ml and higher than 70 cell/ml were set for Chlorophyta biomass. True positive value of each extracted rule is calculated to determine the strength of the rule. Extracted rules are tested again with data Set A which is the training data. A different dataset which is not used for SOM training (namely dataset B) was used to test the effectiveness of the rule based system mainly to avoid producing biased testing results.

RESULTS

Figure 1 illustrates SOM cluster map and Figure 2 illustrates corresponding component planes generated from trained SOM using dataset A. Nine clusters are formed and total of eight rules are determined as cluster number 7 and 9 share similar properties. The extracted rules from mapping clusters formed in Figure 1 with input variables component planes in Figure 2 are depicted in Table 3. True positive value of each rule depicted in Table 3 is calculated and illustrated in Figure 3. True positive values calculated for each rule indicates that rule 8 is most prominent rule meanwhile rule 1 is the least prominent rule applicable to Putrajaya Lake. Rule based system performance for Chlorophyta biomass prediction for training and testing data set is depicted in Table 4.

DISCUSSION

The results demonstrated that SOM being an unsupervised artificial neural network has a potential as a tool for analyzing complex ecological relationships in tropical water bodies. The data used in this study illustrates non normality distribution pattern, which may not be effectively modeled through other conventional statistical techniques that are requires for normality. However, such a requirement is not necessary for artificial neural network, which makes ANN highly suitable in this scenario as it would be somewhat impractical to transform environmental data to suit normality that would results in altering the condition of the tropical lake. The study has also managed to illustrate the application of SOM in discovery of relationship between Chlorophyta biomass and selected water quality parameters (water temperature, Secchi depth and nitrate nitrogen). The Chlorophyta biomass was successfully predicted by yielding success rate of 73%. True positive value calculated for each identified rule reveals rule 8 as the most prominent rule and rule 1 as the least prominent

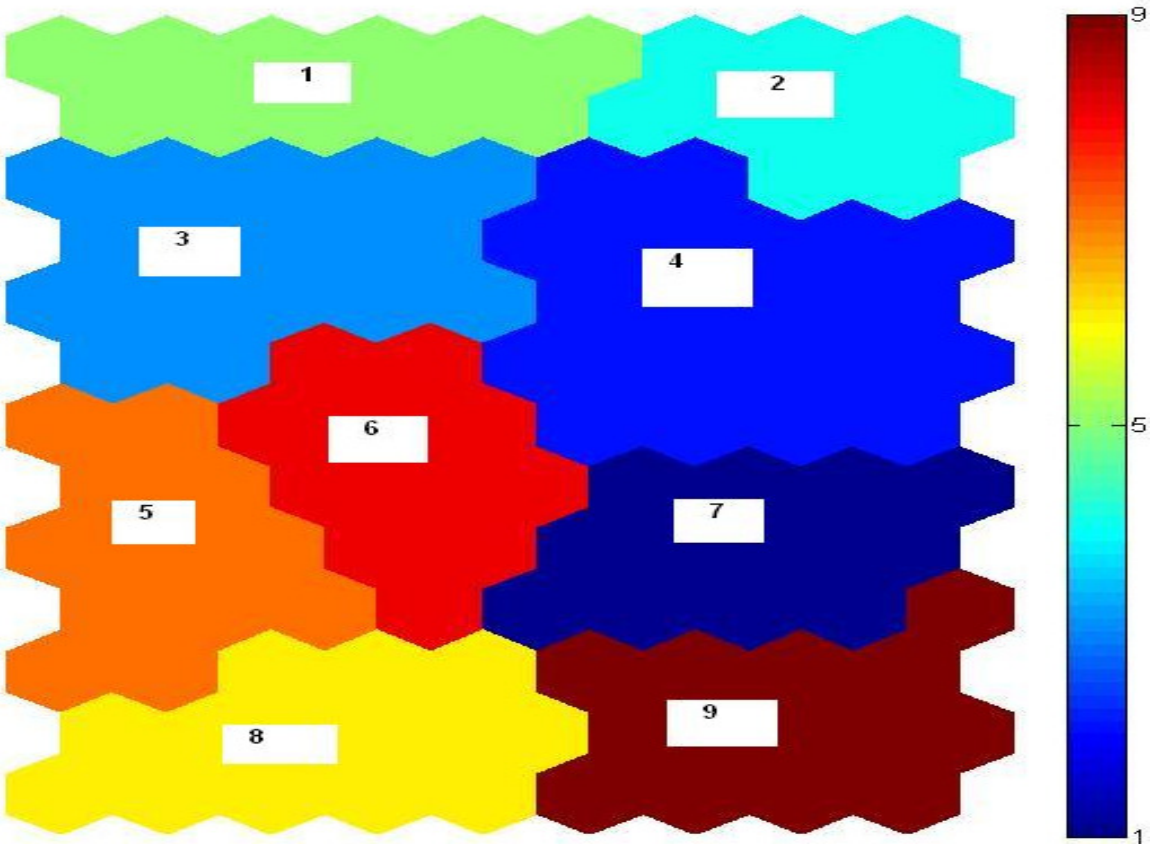

Figure 1. SOM K-means cluster map of Chlorophyta division.

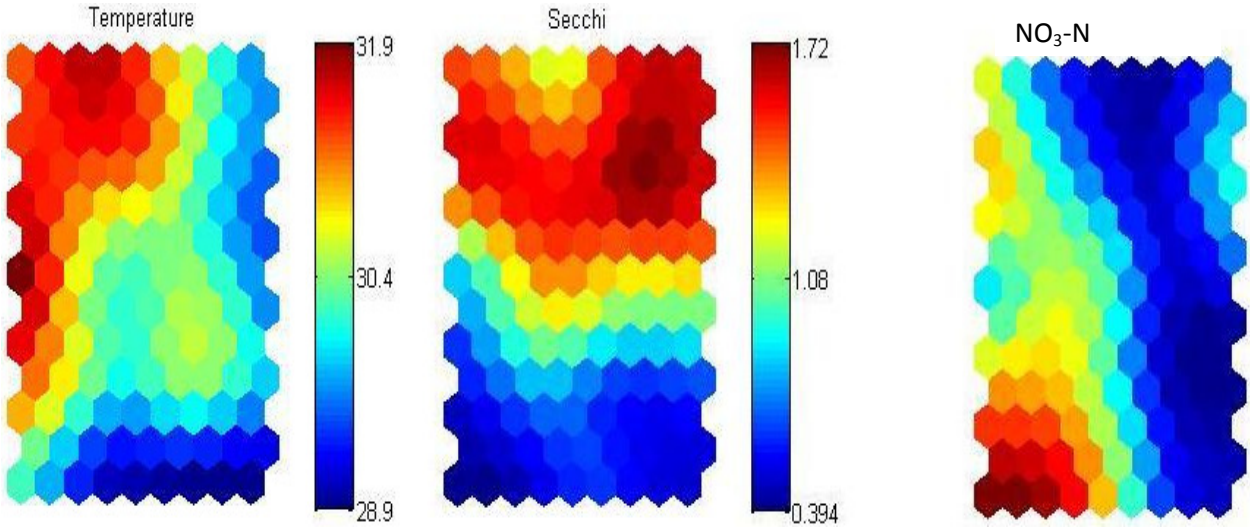

Figure 2. Component planes of input parameters.

rule. Rule 1 is least prominent as high abundance of Chlorophyta is not frequently reported in Putrajaya Lake as the lake is categorized as an oligotrophic lake having low productivity. Rules 2 to 8 describes low abundance of

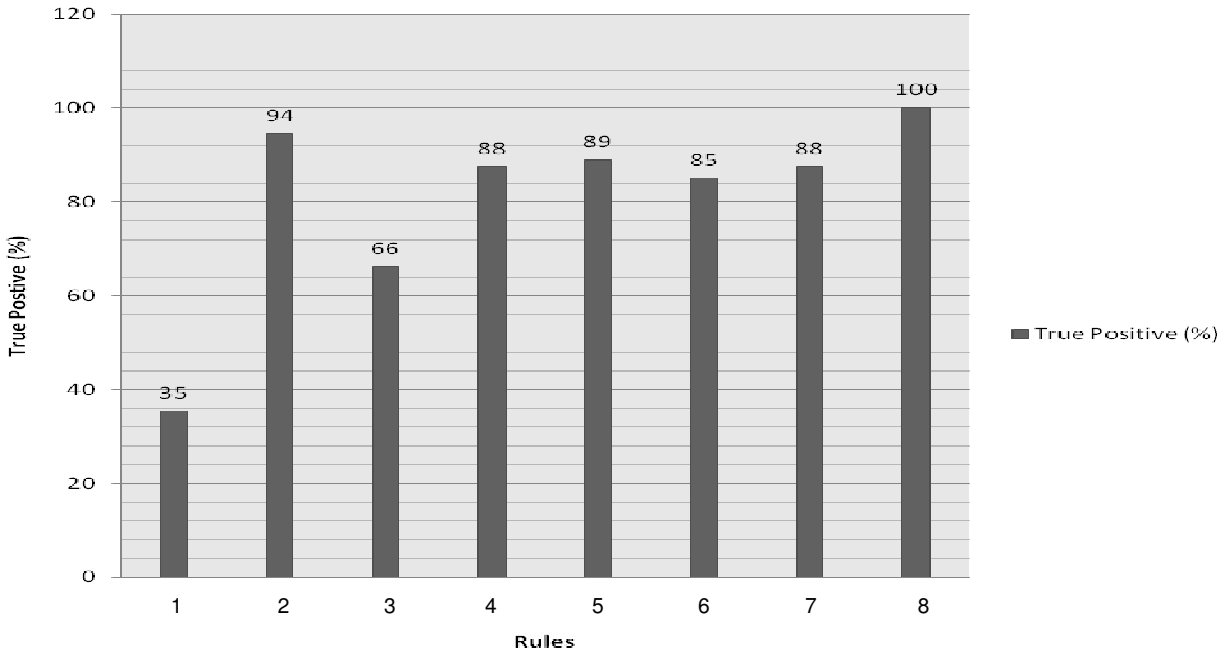

Figure 3. True positive value (%) of each rule.

Chlorophyta. Rules 3, 6, 7 and 8 describes condition of low water temperature below (30.3℃). These rules indicates low abundance of Chlorophyta when the water temperature is below (30.3℃), regardless of Secchi depth and nitrate nitrogen concentration. Priddle and Happy (1983) reported optimum growth condition of *Chlamydomonas, Chlorella* and *Scenedesmus* which belongs to category of micro-green algae that are also found in Putrajaya Lake are favored by warm water conditions and high light irradiance. This can be further supported by Craig (1992) where it is reported that maximum growth of larger green algae is most common in the natural environment during summer warm water conditions. Meanwhile rule 1, 2, 4 and 5 explains condition of high water temperature above 30.3℃; however, only rule 1 explains high abundance of Cholorophyta. Rule 1 explains optimum condition for Chlorophyta at tropical water bodies are high water temperature above 30.3℃, high Secchi depth above 1 m and low concentration of nitrate nitrogen below 1.31 mg/L. This is in conformity with the finding postulated by Blakar et al. (1990), where warm water condition and low concentration of nitrate nitrogen favors maximum growth of Chlorophyta. This is further supported by findings postulated by Reynolds (1984), where *Staurastrum* as one of the dominant species found in Putrajaya Lake is tolerant to enhanced optical depth and higher water temperature. To support that high abundance of Chlorophyta are related to condition of low nitrate nitrogen concentration, findings from this study is compared to Lake Chini Pahang, Malaysia's second largest inland lake. It is reported the most abundant division is at Lake Chini is Chlorophyta which comprises of 65% from species total species and concentration of nitrate nitrogen reported is around 1.1 mg/L (Kutty et al., 2001).

Conclusion

The current study has demonstrated that complex limnological data can be processed by SOM to unravel ecological relationships between Chlorophyta and water quality in tropical lakes. The extracted rules deployed by the rule based system to predict biomass of Chlorophyta species yielded an acceptably high accuracy rate and the rules extracted conforms to literature findings. These models can subsequently be used to monitor water quality for the tropical lakes. However, similar studies have to be carried out on more tropical lakes to obtain a more generic model that is applicable for all tropical lakes.

ACKNOWLEDGMENT

The authors wish to thank Mr. Akashah Hj. Majizat Environmental, Lake and Wetland Division, City Planning Department, Perbadanan Putrajaya, Putrajaya, Malaysia for the data used in this study. This work was supported by the University of Malaya research grant (PJP)/FS306/2008A.

REFERENCES

American Public Health Association (APHA) (1995). American Water Works Association (AWWA) and Water Environment Federation, Standard methods for the examination of water and waste water. 19th Edition. APHA, Washington, D.C.

Blakar I, Digernes I, Seip HM (1990). Precipitation and stream water chemistry at an alpine catchment in central Norway. In: The surface waters acidification programme (ed. Mason, B.J.) Cambridge University Press, pp. 69-73.

Carrillo P, Reche I, Sánchez-Castillo P, Cruz-Pizarro L (1995). Direct and indirect effects of grazing on the phytoplankton seasonal succession in an oligotrophic lake. J.Plankton Res. 17: 1363-1379.

Cherkassky V, Mulier F (1998). Learning from Data: Concepts, Theory and Methods. **Wiley**-**Interscience** New York**.**

Chon TS, Park YS, Moon KH, Cha EY (1996). Patternizing communities by using an artificial neural network. Ecol. Model. 90: 67-78.

Coesel PFM (1983) The significance of desmids as indicators of the trophic status of freshwaters. Schweiz. Z. Hydrol. 45:388–393.

Coesel PFM (2001). A method for quantifying conservation value in lentic freshwater habitats using desmids as indicator organisms. Biodivers. Conserv. 10:177–178.

Craig DS (1992).Growth and Reproductive Strategies of Freshwater Phytoplankton. Cambridge University Press, pp. 452.

Davies DL, Bouldin DW (1979). A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1(2) : 224-227.

Foody GM (1999). Application of the self-organizing feature map neural network in community data analysis. Ecol. Model. 120: 97-107.

Gerrath JF (1993). The biology of desmids: a decade of progress. In: Round FE, Chapman DJ (eds) Progress in phycological research, Biopress, Bristol. 9: 79–192.

Kitner M, Poulickova A (2003). Littoral diatoms as indicators for the eutrophication of shallow lakes. Hydrobiologia. 506/509: 519 – 524.

Kohonen T (1995). Self-Organizing Maps. Springer Series in Information Sciences (30). Springer, Berlin, Heidelberg,New York. 2nd edition..

Kohonen T (2001). Self-Organizing Maps. Springer Series in Information Sciences (30). Springer, Berlin, Heidelberg,New York.

Kutty AA, Ismail A, Fong CS (2001). A preliminary Study of Phytoplankton at Lake Chini Pahang. Pak. J. Biol. Sci. 4(3):309 – 313.

Lek S, Delacoste M, Baran P, Dimopoulos I, Lauga J, Aulagnier S (1996). Application of neural networks to modelling nonlinear relationships in ecology. Ecol. Model. 90: 39–52.

Maier HR, Dandy GC, Burch MD (1998). Use of artificial neural networks for modeling Cyanobacteria Anabaena spp. in the River Murray, South Australia., Ecol. Model. 105: 257 – 272.

Melesse AM, Krishnaswamy J, Zhang K (2008). Modeling Coastal Eutrophication at Florida Bay using Neural Networks. J. Coastal Res. 24 (2B): 190-196.

Priddle J, Happy-Wood CM (1983). Significance of small species of Chlorophyta in freshwater phytoplankton communities with special reference to five Welsh Lakes. J. Ecol. 71: 793 – 810.

Putrajaya Corporation (1998). Putrajaya Lake Management Guide.Putrajaya Corporation.

Recknagel F, French M, Harkonen P, Yabunaka KI (1997). Artificial Neural Network Approach for modellling and prediction of algal blooms. Ecol. Modelling. 96: 11-28.

Recknagel F, Kim B, Takamura , Welk A (2006). Artificial Neural Network approach to unravel and forecast algal population dynamics in two lakes different in morphometry and eutrophication. In: Recknagel, 2nd Edition, Ecological Informatics. Scope, Techniques and Applications, Springer-Verlag, Heidelberg, New York, pp. 325 – 345.

Recknagel F, Welk A, Kim B , Takamura N (2005). Artificial Neural Network Approach to Unravel and Forecast Algal Population Dynamics of Two Lakes Different in Morphometry and Eutrophication. In: Recknagel, F. Ecological Informatics. (2nd Ed.) Springer- Verlag, New York, pp. 325-345.

Reynolds CS (1984).The Ecology of Freshwater Phytoplankton. Cambridge University press. pp. 384.

Salleh A (1996). Panduan mengenali alga air tawar. Dewan Bahasa dan Pustaka, Kuala Lumpur.

Sommer U (1989). Nutrient status and nutrient competition of phytoplankton in a shallow,hypertrophic lake. Limnol. Oceanogr. 34:1162–1173

Sorayya M, Aishah S, Sapiyan B (2010). A Comparison between Neural Network Based and Fuzzy Logic Models for Chlorophlly-a Estimation. Second International Conference on Computer Engineering and Applications (ICCEA). 2 :340-343.

Sorayya M, Aishah S, Sapiyan B (2009). Prediction of Population Dynamics of Bacillariophyta in the Tropical Putrajaya Lake and Wetlands (Malaysia) by a Recurrent Artificial Neural Networks. International Conference on Environmental and Computer Science (ICECS). Dubai, IEEE computer society, pp.407-410

The Math Works Inc. Matlab (Version 6.5.1), 2006.

Ultsch A, Korus D (1995). Automatic acquisition of symbolic knowledge from subsymbolic neural nets, In: Proceedings of the 3rd European conference on intelligent techniques and soft computing, pp 326–331.

Vanni MJ, Temte J (1990). Seasonal patterns of grazing and nutrient limitation of phytoplankton in a eutrophic lake. Limnol. Oceanigr. 35: 697 – 709.

WHO (1987). UNEP/WHO/UNESCO/WMO Project on Global Environmental Monitoring. GEM Water Operational Guide.

Wilson H, Recknagel F (2001).Towards a generic artificial neural network model for dynamic predictions of algal abundance in freshwater lakes. Ecol. Modeling. 146: 69 – 84.