PREDICT OF RIVER WATER QUALITY MODELING USING ARTIFICIAL NEURAL NETWORK (ANN).

Norhafiza Mat Daud*, Latifah Abdul Manaf, Mohd Kamil Yusoff and Hafizan Juahir

M.Sc (GS19189)
5th Semester

Introduction:

Malaysia is a country which is endowed with an abundance of water resources such as from sea, river, lake, pond, dam and others. However, the rapid pace of socio-economic development in the country has begun to strain the available water resources. As a result a number of water related problems have emerged in recent years; these include source contamination, river pollution, water shortages and floods.

As water drains from the land surface, it carries the residues from the land. Surface runoff, especially under the first flush phenomena, is an important source of non-point source pollution. Runoff from different types of land use maybe enriched with different kinds of contaminants. Runoff from agricultural lands maybe enriched with nutrients and sediments. Likewise, runoff from highly developed urban areas maybe enriched with rubber fragments, heavy metal, as well as sodium and sulfate (Tong and Chen, 2002).

Diverse multivariate techniques have been used to investigate how environment variables are related to explain as the dependent variable, including several methods of ordination, canonical analysis and univariate or multivariate linear, curvilinear, or logistic regressions (Lek *et al.*, 1999).

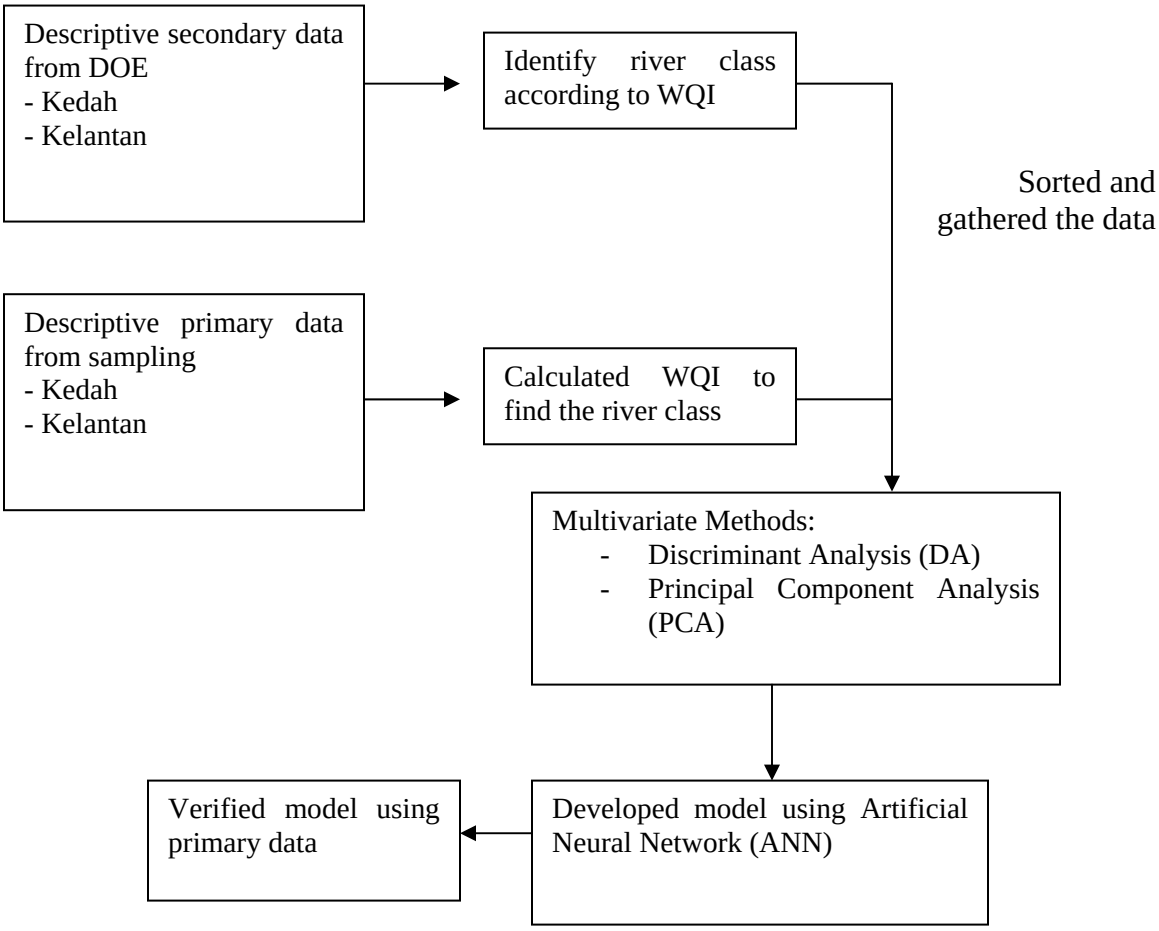
Research Objective:

This study on the water quality modeling was carried out in order to develop a water quality model for prediction of water quality parameter. Its main objectives are as follows:

1. To identify significant water quality parameters
2. To develop a water quality prediction model of water quality parameters
3. To verify water river classification using Artificial Neural Network (ANN).

Research Methodology:

Flow chart of Methods:



Results and Discussions:

The first step of the statistical work is analysis using Discriminant Analysis (DA). It was applied on raw data to predict a classification (X) variable based on known continuous responses (Y). In this study case, parameters as a classification (X) variable and river class a continuous response (Y). Discriminant Analysis is used to determine which parameters are the best predictors.

1- Discriminant Analysis for Secondary data of Kedah River:
There are 30 water quality parameters that have been analyzed at 9 stations with 56 total observations. 9 stations numbers were labeled by DOE with code 2KD01 -2KD09. After run Discriminant Analysis, only 17 parameters were chosen as best predictors in Kedah River. The parameters are shown in the table 1.

Table 1: Summary of Discriminant Analysis of Kedah

Stations	Selected Parameters
2KD01-2KD09	DO, COD, K, SAL, NH3-N, SS, E.COLI, BOD, Ph, NA, CA, COND, COLIFORM, FE, ZN, MG, DS

2- Discriminant Analysis for Secondary data of Kelantan River:

There are 30 water quality parameters that have been analyzed at 67 stations with 51 total observations data from Kelantan. The 67 stations numbers were labeled by DOE namely, 4KE01-4KE67. But, only 23 parameters were chosen as best predictors, and already shown in the table 2.

Table 2: Summary of Discriminant Analysis of Kelantan

Stations	Selected Parameters
4KE01-4KE67	BOD,COD,NH3-N, CR, FE, TEMP, SS, SAL, CL, NA, K, MG, NO3,Ph, AS, DS, COLIFORM, DO, COND, MBAS, E.COLI, CD, PB

The second step of statistical work is Principal Component are sorted by descending order of how much of the initial variability they represent (converted to %). To find the best result of the corresponding factor, eigenvalue should be is greater than 1.

1-Principal Component Analysis PCA is a very useful method to analyzed numerical data. To simplify matters, one usually starts with the eigenvalues. Each eigenvalue correspond to a factor, and each factor to a one dimension. Tables 3 below are shown the eigenvalues and the corresponding factor for secondary data of Kedah River.

Table 3: Principal Component Analysis for secondary data of Kedah River

Varimax Factor	Variability	Parameters	Loading Factor
VF1	27.926	DS	0.924
		TS	0.864
		CL	0.967
		CA	0.668
		MG	0.953
		NA	0.932
		MBAS	0.957
VF2	37.933	COND	0.930
		SAL	0.935
		NO3	0.867
		AS	0.846
VF3	46.843	SS	0.941
		TUR	0.926
VF4	54.865	BOD	0.710
VF5	60.160	PB	0.724
VF6	64.564	E.COLI	0.930
VF7	68.854	pH	0.802
VF8	72.685	CD	0.900
VF9	76.297	CR	0.845
		FE	0.787

Table 4: Principal Component Analysis for secondary data of Kelantan River

Varimax Factor	Variability	Parameters	Loading Factor
VF1	24.719	COND	0.882
		SAL	0.795
		DS	0.962
		TS	0.708
		CL	0.943
		CA	0.906
		K	0.883
		MG	0.925
		NA	0.942
VF2	41.858	pH	0.985
		NO3	0.956
		AS	0.992
		HG	0.899
		MBAS	0.984
VF3	49.987	SS	0.991
		TUR	0.988
VF4	56.207	CR	0.883
		PB	0.863

VF5	61.337	FE	0.731
VF6	66.228	PO4 E.COLI	0.678 0.801
VF7	70.893	BOD NH3-NL	0.735 0.842
VF8	74.318	COD	0.775

Artificial Neural Network

Based on the Discriminant Analysis, only 17 parameters meet the requirement as best predictors in Kedah. The hidden nodes were ranging from 1 to 10 to accurate data and optimal for generalizing the variable relations. Therefore, neural networks with 5 hidden nodes were used in this data. The coefficient of determination, R^2 was used to estimate model fit. R^2 for validation was 0.98. The result is shown in the table 5 below:

Table 5: R^2 values of ANN for Kedah

	Specify			
Hidden Nodes	5			
Overfit Penalty	0.001			
Number of Tours	20			
Max Iterations	50			
Converge Criterion	0.00001			
	Objective			
SSE	7.0593282628			
Penalty	4.3964155594			
Total	11.455743822			
N	365			
	96			
0	Converged At Best			
0	Converged Worse Than Best			
0	Stuck on Flat			
0	Failed to Improve			
20	Reached Max Iter			
Y	SSE	SSE Scaled	RMSE	RSquare
RIVER CLASS	2.5735783813	7.0593282628	0.09874438	0.9806

The 23 parameters meet requirement as good predictors in Kelantan. The number of hidden nodes taken 10 nodes gives the best results presented in table 6, the result according to R^2 values.

Table 6: R^2 values of ANN for Kelantan

	Specify			
Hidden Nodes	10			
Overfit Penalty	0.001			
Number of Tours	20			
Max Iterations	50			
Converge Criterion	0.00001			
	Objective			
SSE	473.36805968			
Penalty	31.66285597			
Total	505.03091565			
N	2023			
	241			
0	Converged At Best			
1	Converged Worse Than Best			
0	Stuck on Flat			
6	Failed to Improve			
13	Reached Max Iter			
Y	SSE	SSE Scaled	RMSE	RSquare
RIVER CLASS	134.84755787	473.36805968	0.34251351	0.7635

As the validation for the result above, comparison of ANN between raw data (using 30 parameters) and data after discriminate are shown in the table 7 below:

Table 7: Comparison of ANN between raw data and discriminate data.

River	Raw Data	Discriminate Data
Kedah	Hidden Nodes: 6 R^2: 0.94	Hidden Nodes: 5 R^2: 0.98
Kelantan	Hidden Nodes: 9 R^2: 0.84	Hidden Nodes: 10 R^2: 0.77

Significance of Finding:

The overall research study is to reduce the sampling cost and develop more intelligent computer aided tools in evaluating water quality. Besides, the use of ANN in this study is to reduce the number of parameters needed to carry out water quality prediction without much loss of information.