KINTA RIVER WATER QUALITY AND LAND USE PATTERNRECOGNITION AND MODELING USING ARTIFICIAL NEURAL NETWORK (ANN)

Nabeel M. Gazzaz*, Mohammad Kamil Yusoff, Ahmad Zaharin Aris, Hafizan Jauhir and Mohammad Firuz Ramli

Ph.D (GS 18306)
6th. Semester

---

(1) Introduction

The surface water quality (WQ) in a region largely depends on the nature and extent of land uses (LUs) and other anthropogenic activities in the catchments. In mixed use watersheds, the percentages of each type of land use (LU), i.e., mining, industrial, agricultural, and residential become important if they differ significantly in contribution to non-point source pollution.
Urban development in Kintan River basin has been rapid. It is impacting water resources through an increased demand on water and increased inputs of nutrients, organics, and microbes from human activities. But the relative impacts of different types of LUs on river WQ are yet to be ascertained and quantified. Hence, there is an increasing interest in establishing models which can predict the effects of LU and changes in it on river WQ.
This research uses a comprehensive, watershed-based approach to examine the effects of LU on river WQ at a local scale. Statistical analyses, tests, and modeling, as well as artificial neural network pattern recognition and modeling techniques will be employed to examine and model the relationships of LU with Kinta River WQ.
Environmental stewardship and restoration efforts in Kinta River basin necessitate prediction of WQ in response to LU patterns; current and projected. Study will develop a quantitative model and then use it to estimate impacts of future LU changes on the WQ of the river. Model predictions will allow preventive measures/ actions to be taken early during the development planning process to cope with any foreseen changes in river WQ in conjunction with planned development. In addition, the built model will allow reduction of the time, effort, and cost of gaining WQ data without compromising reliability and credibility of conclusions. Ultimately, it will contribute to the keen efforts of policy makers and decision takers in establishment of a balance between water resource use and extended, sustainable development.

(2) Study Goal and Objectives

The goal of this research is to recognize patterns in, and model relationships between, LU and WQ of Kinta River basin using chemometric and artificial neural network techniques.
Accordingly, the objectives of this study are to:
1- Recognize, model, and predict patterns in Kinta River WQ and LU data using chemometric and ANN pattern recognition and prediction techniques.

2- Explore and model the plausible relationships of land uses with Kinta River WQ using artificial neural network (ANN).
3- Predict Kinta River WQ in light of urban development plans in the watershed and projected and/ or planned changes in LU patterns.

(3) Methodology:

The study begins with examination of the rate of urbanization as well as the changing patterns of LUs within Kinta River basin between 1997 and 2006. It then provides analysis of the relationship between WQ and changes in the rate of urbanization and LUs. LU statistics will be extracted using suitable LU and digital maps. Digital maps of the L 8082 series will be used as reference maps and also for the purpose of delineation of the river basin.

3.1 Study Area

Kinta River is located in Perak, Ipoh (Malaysia). It flows from Gunung Korbu at Ulu Kinta, to Tanjung Rambutan. It is 100 km long and covers an area of 2540 km^2.
Kinta River is currently classified with an average Class III water quality and a water quality index (WQI) of 51.9 – 76.5. This means that the water is polluted and requires extensive treatment before it can be used for drinking purposes. Major causes of pollution are industrial and residential discharges, animal husbandry farms, sand-mining, and soil erosion.

3.2 The Data Set

The WQ dataset comprises 9180 entries derived from 36 measurements on 255 samples. The 36 monitored parameters encompass hydrologic, physico-chemical, chemical, and microbial variables. This dataset covers the period from March, 1997, to November, 2006, and documents the values of 32 pollution indicators for 8 monitoring locations along the river.
The LU dataset comprises LU statistics for Kinta River basin over 10 years covering the period 1997-2006. It encloses LUs in the year 1997 and changes in which in the years 2000, 2002, 2004, and 2006. LU maps will be employed for statistics extraction of priority LU; mining, industrial, animal husbandry, agricultural, logging, residential, and forest uses.

3.3 Statistical Analyses and Tests

The river WQ dataset was screened for anomalies; i.e., outliers, missing values, zeros, values below the detection limits, and inconsistent data. The data was also explored for normality (the Shapiro–Wilk's test) and for symmetry of variable distributions (Skewness test).
Spearman rank correlation test was used to test for associations between the studied variables. In addition, both this test and the Kruskal-Wallis test were used for spatial and temporal trend analysis. Linear regression was used to develop a model predictive of the WQI from the various WQVs. Thee chemometric techniques principal factor analysis,

hierarchical cluster analysis, and discriminant function analysis are adopted for pattern recognition and trend analysis and modeling purposes.

3.4 ANN Pattern Recognition, Trend Analysis, and Modeling

We propose to use multi-layer perceptron (MLP) neural networks to establish models that predict (i) the WQI from the WQVs, and (ii) the WQI from the LUs of interest in the river basin. Both for the river WQ and the LU data sets, we will use (i) self-organizing feature maps (SOFM) for pattern recognition and trend analysis, and (ii) radial basis function (RBF) neural networks for classification and class membership prediction.

4. Results and Discussion

4.1 Data Screening and Pretreatment

Main features of the WQ data archive are non-normal distribution and positive skewness of the WQVs, and presence of some missing values, entries below detection limits, structural zeros, and statistical outliers. Statistical outliers were used as is whereas other data anomalies were imputed using ordinary least square multiple linear regression (OLS MLR).

Shapiro–Wilk's test reveals that, except temperature and pH, the WQVs do not fit the normal distribution. This finding is normal since WQ data are typically non-normally distributed (Helsel, 1987; Tsirkunov *et al.*, 1992). Almost all studied WQVs show varying degrees of positive skewness. This is expected since most WQ and environmental data are characterized by positive skewness (Helsel and Hirsch, 2002). Only pH and Pb show slight negative skewness. The WQI has a slight negative skewness mostly arising from skewness in pH.

4.2 Monovariate Statistics

The various WQVs studied were explored for their ranges, minima, maxima, means, and standard deviations. Most of the variables have wide ranges thus indicating the wide heterogeneity of the river system and the multiplicity and variability of pollutant sources along its course. On the other hand, and with the exclusion of pH, the WQVs having the highest weight in DoE and the regression WQI formulae, that's, turbidity, SS, DO, BOD, and COD have maxima far exceeding the upper limit for unpolluted waters and are accordingly identified as the main culprits of deterioration of Kinta river WQ.

4.3 Bivariate Statistics:

The statistical analyses applied in this part aimed primarily at (i) exploring associations between the various WQVs and between each and the WQI, and (ii) exploring the spatial and temporal trends in the WQI.

We applied the Spearman rank correlation analysis to explore associations between the various WQVs, and between the WQVs each and the WQI, time and space. In addition, we applied the Kruskal-Wallis test to determine if the different months, years, and monitoring locations do, or do not, cluster into groups of equal medians. When the Kruskal-Wallis test revealed significant median differences, we used the significant

results on the Kruskal-Wallis test as justification to perform a set of Mann-Whitney U tests (also known as Wilcoxon rank-sum test) to determine the underlying regularity in the dependent variables.

4.3.1 Relationships between the Water Quality Variables and the WQI

Main features of relationships between the WQVs are: (i) lack of any very high associations between variables (highest $\rho_S < 0.90$), (ii) absence of any high negative associations, and (iii) month, PO_4^{3-}, Fe, and pH exhibited no significant correlation with any variable.
The DO concentration is the only WQV that has high, positive correlation with the WQI ($\rho_S = 0.73$). No WQ parameter has moderate positive correlation with the WQI while ten parameters show negative, moderate correlation with strengths in the order BOD> NH_3-N> COD> DS> K> Cl> TS> Ca> conductivity> Mg. These correlations point to the individual contribution of these variables to any model that may be designed for prediction of river WQ.

4.3.2 Trend Analysis

4.3.2.1 Vriable/ Parameter Trends. This class of trends is discussed under the chemometric pattern recognition/ principal factor analysis section.

4.3.2.2 Spatial Trends

The Correlation Approach

The WQI has a low, negative correlation with distance downstream the river ($\rho_S = -0.23$, $\alpha = 0.00018$). In other words, there is a trend of slight decline in the WQI with distance from headwaters. The DO concentration declines away from the headwaters at moderate pace. The major soluble ions (Na, K, Ca, Mg, Cl) increase with distance; Ca and Mg increase moderately while the rest major ions increase slightly. On the other hand, our data set reveals no significant spatial differences in NO_3^- and PO_4^{3-} concentrations.

The Group Median Comparisons Approach

Kruskal-Wallis test indicated lack of clear-cut, consistent spatial trends in river WQ. Nonetheless, the test allowed classification of the monitoring stations into four groups. Moreover, our results identify the surroundings of the monitoring stations 2PK25, 2PK34, and 2PK59 (the river stretch 33.52- 44.1 km) and 2PK60 as critical pollution areas that classify as priority zones for extensive research on the land uses and pollution sources in these areas, and for urgent remedial action to restore desired WQ.

4.3.2.3 Temporal Trends

The Correlation Approach

The WQI has a low, positive correlation with year ($\rho_S = 0.14$, $\alpha = 0.026$) meaning that it tends to increase with time, but at slow pace. On the other hand, the WQI has no significant relationship with sampling month ($\rho_S = 0.11$, $\alpha = 0.086$) thus indicating that the temporal trends in WQI can not be detected at the month level.

No WQV tends to decrease highly with time. Hg is the only variable that decreases moderately with time whereas a number of variables show a tendency of slight decrease. The concentrations of Cd and Pb are the only variables that show a high tendency to increase with time. Salinity and the concentrations of Zn and Cr show a moderate tendency of increase.

The Group Median Comparisons Approach

Kruskal-Wallis test indicated lack of clear-cut, consistent temporal trends in river WQ and that the WQI distributions among the 8 monitoring months and ten years are significantly different. The Mann-Whitney U pair-wise tests and Bonferroni group comparisons revealed that the different years and months can be grouped into three distinct groups each.

4.3.4 Culprits of Observed Spatial and Temporal Trends:

Based on the above findings, we ran the Kruskal-Wallis test to identify the WQVs that are simultaneously responsible for the spatial and temporal differences and subsequent groupings. Our results reveal that these vaiables are salinity, SS, TS, pH, DO, BOD, NH_3-N, Cl, PO_4^{3-}, Cd, Cr, Pb, Zn, and E. coli bacteria.

4.4 Multivariate Statistics

Regression Modeling

In this part of our work, we established an OLS, stepwise MLR model predictive of the WQI from the raw WQVs without the need for sub-indexing. Produced prediction model has a very high correlation (ρ_S = 0.901, p = 0.000, α = 0.01) with the WQI formula already in use (the Department of Environment WQI). The model identified the variables most representative to the quality of the river water to decline in significance following the order DO > BOD > NH_3-N > turbidity > pH > COD. The coefficient of determination (R^2) provides a measure of how well the regression model fits the data. This six-predictor model has an R^2 value of 0.812, i.e., it explains 81.2% of the variation in the WQI and implies the model is well-specified.

Pattern Recognition:
Multivariate pattern recognition was investigated in this study following the chemometrics as well as the Artificial Neural Networks approaches.

Pattern Recognition: The Chemometric Approach

The multivariate analysis is widely used to characterize and evaluate river WQ and it is useful for evidencing temporal and seasonal variations caused by natural and anthropogenic processes (Vega *et al.*, 1998; Singh *et al.*, 2004).

Principal Factor Analysis: **Variable/ Parameter Trend Analysis**

Factor analysis is a powerful approach for recognizing patterns. It aims to explain the variance of a large set of inter-correlated variables by transforming them into a smaller set of independent (uncorrelated) ones. The resultant variables can be treated as new sets

called latent factors, which are neither observed nor expressible in terms of the observed variables (Liao *et al.*, 2006).

Dimensionality of the data set was reduced to 26 variables distributed between seven factors that explain 71.63% of variation in the data. Individually, variation in the data that each factor explains are 26%, 10.26%, 9.34%, 8.24%, 8.16%, 4.98%, and 4.70%, respectively.

The first factor has high loadings from TDS, Cl, Na, Mg, salinity, conductivity, K, and Ca. Hence, we combine these items in a scale which might be called "salinity" factor. The second factor has high loadings from SS, TS, turbidity, and COD. This factor might be called "turbidity" factor. The third factor has moderate and above moderate loadings from total coliform bacteria, E. coli bacteria, NH_3-N, and BOD. Hence, it might be called "wastewater" or "microbial scale". The fourth factor has moderate and above moderate loadings from Cd, Pb, and Cr. So, it might be called "Heavy metal" or "industrial pollution" scale. The fifth factor has moderate and above moderate loadings from DO, temperature, and As. The sixth has moderate and above moderate loadings from Fe, Zn, and PO_4^{3-}. Only one factor loads on the seventh factor, so we can drop it and conclude that dimensionality of the data was reduced to 26 variables distributed between six factors explaining 67.0% of variation in the data.

Hierarchical Agglomerate Cluster Analysis (HACA)

We carried out HACA using Ward's method and applying the squared Euclidean distance as the distance measure. Due to the large number of observations, we ran CA first on the 90^{th} percentile, second on the mean, of each WQ parameter for each station and month. Variables introduced to CA correspond to those WQVs with moderate and above moderate loadings on the factors extracted by PFA.

Spatial Trends

Clustering the WQ monitoring stations based on the 90^{th} percentile value for the WQVs produced 3 clusters: (i) cluster I is formed by the stations 2PK25, 2PK33, 2PK34, and 2PK19, (ii) cluster II by the stations 2PK59 and 2PK60, and (iii) cluster III by the stations 2PK22 and 2PK24. The WQV means also clustered the stations into 3 clusters but with different assignations; (i) cluster I is formed by the stations 2PK33, 2PK34, and 2PK19, (ii) cluster II by station 2PK22, and (iii) cluster III by the stations 2PK24, 2PK25, 2PK59, and 2PK60.

Temporal Trends

Clustering the WQ monitoring months based on the 90^{th} percentile value for the WQVs produced 2 clusters; one formed by the months 2, 5, 6, 8, 11, and 12, while the other by the months 3 and 9. However, clustering based on the mean of the WQVs produced 3 clusters: (i) cluster I is formed by the months 2, 5, 8, and 11, (ii) cluster II by the months 6, 9, and 12, and (iii) cluster III only by the month 3. These results imply that for rapid assessment of river WQ, only one station and one month from each cluster is needed to represent a reasonably accurate spatial and temporal assessment of the WQ for the whole river.

Discriminant Function Analysis (DFA)

Our aim in this part of the project was firstly to verify if the various spatial and temporal clusters can, or can not, be used to represent the differences among the CA-defined station and month groups and secondly to determine which variables are of most significance to the classification. Finally, the DFA was performed to relate the WQ clusters to different WQVs and generate predicting equations. New objects can then be classified and the learning objects reclassified by means of the non-elementary discriminant functions.

Spatial Class Prediction

Station "90th percentile" clusters produced two discriminant functions. A 95.3% of the variance in these clusters is explained by the first discriminant function while the second explains the remaining variance. The classification power of the generated function is 82.7%. The expected hit ratio is 50.16%. Since the actual predictive accuracy is 82.7%, we conclude that our discriminant function works appreciably well.
On the other hand, the station "Mean" clusters produced two discriminant functions. The first function accounts for 85.1% of the discriminating ability of the discriminating variables while the second accounts for the rest. The classification power of the function is 65.1% and the expected hit ratio is 48.15%. Hence, the discriminant function works appreciably well.
As a conclusion, monitoring station clustering based on the 90th percentile, rather than mean, values of the WQVs produced a discriminant function of a higher discriminating power.

Temporal Class Prediction

Since the month "90th percentile" clusters have only two categories, one discriminant function explains all the variance in the dependent variable. This discriminant function correctly classifies 81.6% of the cases. Its classification power is 80.4%. The expected hit ratio is 79.0%. Since the actual predictive accuracy is 80.4%, we conclude that our discriminant function works well. On the other hand, the DA identifies two discriminant functions for the month "Mean" clusters. The first of which explains 83.7% of the variance in the month "mean" clusters. Cross validation indicates that the classification power of the generated discriminant function is 72.5%. The expected hit ratio is 58.04%. Since the actual predictive accuracy is 72.5%, we conclude that the discriminant function works well.

4.5 Artificial Neural Networks

ANN Model for Prediction of the WQI from the WQVs

This part of the study aimed at establishing a NN model for prediction of the WQI from the various WQVs. We used a fully-connected, feed-forward, multi-layer perceptron network comprised of a single, 28-neuron input layer, one hidden layer, and the WQI as the output layer. Training was carried out by the standard, supervised, quick-propagation algorithm and activation was triggered by the logistic sigmoid activation function.

The optimum network architecture was obtained by trial and error based on estimation of several models. We found that the optimum network performance is obtained with 10 neurons. The network with the architecture (1:28, 1:10, 1) yielded a correlation coefficient with the experimental WQI of 0.994. That's, this model cares for 98.8% of the variation in the experimental WQI values. This implies that the network architecture is well-specified.
The next step was identifying the optimum learning rate. We tested net work performance with learning rates within the range 0.01-0.90. Selection of the optimum training rate was based on the mean absolute error (MAE). Tests indicated that the optimum network architecture (1:28, 1:10, 1) performs best with a learning rate of 0.90.
Linear regression analysis of calculated results and measured data can be used to evaluate the results of a validation in a more objective and quantitative manner (Flavelle, 1992). Accordingly, the OLS regression was the measure we followed in validating the model and evaluating its final performance. With a learning rate of 0.9, the WQI predictions of this network have a correlation coefficient of 0.990. This means that these predictions have a very high positive correlation with the actual WQI values and that they explain around 97.9% of the variation in the experimentally calculated WQI.

Pattern Recognition: The ANN Approach

Clustering: The Self-Organizing Feature Maps (SOFM).
This part of the project is now ongoing.

Classification: Radial Basis Function Neural Networks (RBF NN)
This part of the project will be started after the former part is finished.

4.6 Land Use Data Acquisition and Analysis:
The LU maps have already been georeferenced, Kina River basin has already been delineated, and LU map digitization is in progress.

(5) Significance

This study will contribute to our understanding of the cumulative contributions of different land uses as they change downstream of the river and hence will contribute to development of water quality regulations and sustainable development practices in the study area.
Study results can be applied in selective chemical monitoring of river water quality. We identify the water quality parameters most significant in describing the spatial and temporal variations in the water quality, identify and propose the optimum locations for future sampling stations, and specify the optimum number of monitoring stations such that highly reliable and cost-effective river water quality data is still secured without compromising reliability and credibility of conclusions. This will reduce monitoring time and costs through reducing the number of the WQVs to test, sampling stations to include in future monitoring programs, number of samples to collect, and the frequency of sampling.
Study will develop a quantitative model and then use this model to estimate future LU changes and their impacts on the water quality of Kinta River. Model predictions will

allow preventive measures/actions to be taken early during development planning processes to cope with any foreseen changes in river WQ in conjunction with planned development.

Information on the hydrologic effects of LUs can provide guidelines not only for resource managers in restoring our aquatic ecosystems, but also for local planners in devising viable and ecologically-sound watershed development plans, as well as for policy makers in evaluating alternate land management decisions. Special care will be paid to making the NN model available for local planners and regulatory authorities involved in restoring or improving the desired WQ in light of predicted or planned changes in LU patterns.

References:

Flavelle, P. 1992. Advances in Water Resources 15 (1): 5-13.

Helsel, D. R. 1987. Hydrological Sciences Journal 32 (2), 179–190.

Helsel, D. R. and Hirsch, R. M. 2002. Statistical Methods in Water Resources. Chapter A3 in Book 4, Hydrologic Analysis and Interpretation. Techniques of Water-Resources Investigations of the United States Geological Survey. U.S. Geological Survey, USA.

Liao, S., Lai, W., Chen, J., Sheu, J. and Lee, C. 2006. Environmental Monitoring and Assessment 122 (1-3): 81–100.

Singh, K. P., Malik, A., Mohan, D. and Sinha, S. 2004. Water Research 38 (18): 3980–3992.

Tsirkunov, V. V., Nikanorov, A. M., Laznik, M. M. and Dongwei, Z. 1992. Water Research WATRAG 26 (9): 1203–1216.

Vega, M., Pardo, R., Barrado, E. and Deban, L. 1998. Water Research 32 (12): 3581–3592.