We are IntechOpen, the world’s leading publisher of Open Access books
Built by scientists, for scientists

3,900
Open access books available

116,000
International authors and editors

120M
Downloads

154
Countries delivered to

TOP 1%
Our authors are among the most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™
Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com
Using Principal Component Scores and Artificial Neural Networks in Predicting Water Quality Index

Rashid Atta Khan\textsuperscript{2}, Sharifuddin M. Zain\textsuperscript{2}, Hafizan Juahir\textsuperscript{1}, Mohd Kamil Yusoff\textsuperscript{1} and Tg Hanidza T.I.\textsuperscript{1}

\textsuperscript{1}Department of Environmental Science, Faculty of Environmental Study, University Putra Malaysia, Serdang
\textsuperscript{2}Chemistry Department, Faculty of Science, University of Malaya, Kuala Lumpur

1. Introduction

The management of river water quality is a major environmental challenge. One of the major challenges is in determining point and non-point sources of pollutants. Industrial and municipal wastewater discharges can be considered as constant polluting sources, unlike surface water runoff which is seasonal and highly affected by climate. According to Aiken et al. (1982), 42 tributaries in Peninsular Malaysia are categorized as very polluted including the Langat River. Until 1999, there were about 13 polluted tributaries and 36 polluted rivers due to human activities such as, industry, construction and agriculture (Department of Environment, Malaysia (DOE), 1999). In 1990, there were 48 clean rivers classified as clean but the number is reduced to 32 rivers in 1999 (Rosnani Ibrahim, 2001).

Surface water pollution is identified as the major problem affecting the Langat River Basin in Malaysia. Increase in developing areas within the river basin has in turn increased pollution loading into the Langat River. To avoid further degradation, the DOE have installed telemetric stations along the river basin to continuously monitor the water quality. As a result, abundant data were collected since 1988. There are 927 monitoring stations located within 120 river basins throughout Malaysia. Water quality data were used to determine the water quality status and to classify the rivers based on water quality index (WQI) and Interim National Water Quality Standards for Malaysia (INWQS). WQI provides a useful way to predict changes and trends in the water quality by considering multiple parameters. WQI is calculated from six selected water quality variables, namely dissolved oxygen (DO), biochemical oxygen demand (BOD), chemical oxygen demand (COD), suspended solid (SS), ammonical nitrogen (AN) and pH (DOE, 1997). It is a well-known phenomenon that the contribution of pollution loading into river systems from the environment involves a complex interaction of many factors (e.g. chemical, physical and meteorological interaction). These primary pollutants are emitted from land use activities surrounding the river basin (e.g. agriculture, forest, urban, industrial and others) Rapid urbanization along the Langat River plays an important role in the increase of point source pollution.
(PS) and non-point source (NPS). In view of this complex interaction, use of modelling techniques to solve this problem, is needed. However, the problem of obtaining models that adequately represent the dynamic behaviour of field data is not easy. Lack of good understanding and description of the phenomena involved, the availability of reliable and complete field data set and the estimation of the numerous parameters involved are the major factors contributing to this problem. Beck (1986) noted that, increase in model complexity will undoubtedly increase the number of parameters, leading to the problems of identification.

Applications of ANN (Artificial Neural Networks) to environmental problems are becoming more common (Silverman and Dracup, 2000; Scardi, 2001; Recknagel et al., 2002; Bowden et al., 2005; Muttil and Chau, 2007). The applications of ANN, which are computing systems that were originally designed to simulate the structure and function of the brain (Rumelhart et al, 1986) is a relatively new concept in environmental modeling. If trained properly, a neural network model is capable of ‘learning’ linear as well as the nonlinear features in the data (Elsner and Tronis, 1992).

ANN consists of a set of simple processing units (neurons) arranged in a defined architecture and connected by weighted channels which act to transform remotely-sensed data into a classification. The classification techniques of ANN are unlike the conventional ones. It is distribution-free, may sometimes use small training sets (Hepner et al., 1990) and, once trained; it is rapid computationally, which will be of value in processing large data sets (Gershon and Miller, 1993). Furthermore, ANNs have been shown to be able to map land cover more accurately compared to many widely used statistical classification techniques (Benediktsson et al., 1990; Foody et al., 1995) and alternatives such as evidential reasoning (Peddle et al., 1994).

It has been proposed that the best tool to model non-linear environmental relationship is ANN (Zhang and Stanley, 1997; Jain and Indurthy, 2003). Research have been undertaken at Imperial College, London which attempts to investigate the capability of ANN approach in modelling spatial and temporal variations in river water quality (Clarici, 1995). ANNs were used as a predictive model to predict cyanobacteria Anabaena spp. in the River Murray, South Australia (Maier et al., 1998). DeSilets et al. (1992), have also used ANN to predict salinity. Ha and Stenstrom (2003), proposed a neural network approach to examine the relationship between storm water quality and various types of land use.

ANN has been successfully applied on the study of river water quality in Malaysia (Zarita Zainudin, 2001; Mohd Ekhwan Toriman and Hafizan Juahir, 2003; Hafizan Juahir et al., 2003a,b; Hafizan et al., 2004a,b; 2005; Ruslan Rainis et al., 2004). An approach for identifying possibilities of water quality improvement could be developed by using this concept. Such information could provide opportunities for better river basin management to control river water pollution in Malaysia. In the Malaysian context, Hafizan Juahir et al. (2003a) showed that the ANN model gives a better performance compared to the autoregressive integrated moving average (ARIMA) model in forecasting DO. The use of ANN for river regulation (Mohd. Ekhwan Toriman and Hafizan Juahir, 2003) and the application of the second order back propagation method (Hafizan Juahir et al., 2004a) on water quality of the Langat River have also been demonstrated.
In natural environment, water quality is a multivariate phenomenon, at least as reflected in the multitude of constituents which are used to characterize the quality of water body. Water quality is very difficult to model because of the different interactions between pollutants and meteorological variables. The principal component analysis (PCA) is one of the approaches to avoid this problem and has received increasing attention as an accepted method in environmental pattern recognition (Simeonov et al., 2003; Wunderline et al., 2001; Helena et al., 2000; Loska and Wiechula, 2003).

The objective of this study is to use the PCA method to classify predictor variables according to their interrelation, and to obtain parsimonious prediction model (i.e., model that depend on as few variables as necessary) for WQI with other physico-chemical and biological data as predictor variables to model the water quality of the Langat river. For this purpose, principal component scores of 23 physico-chemical and biological water quality parameters were generated and selected appropriately as input variables in ANN models for predicting WQI.

2. Methodology

2.1 The data and monitoring sites

The water quality data in this study were obtained from seven stations along the main Langat River (Fig. 1).

![Fig. 1. Data from seven water quality stations (Sb) were selected in this study along the main river.](www.intechopen.com)
The water quality monitoring stations are manned by the DOE and Ministry of Natural Resource and Environment of Malaysia. The selected stations are illustrated in Table 1. The data used in the study is from September 1995 to May 2002. Seven sites were chosen, namely, Teluk Panglima Garang (site 7), Teluk Datok (site 6), Putrajaya (site 5), Kajang (site 4), Cheras (site 3), Hulu Langat (site 2), Pangsoo and Ulu Lui (site 1). Sites 3 to site 7 are located in the region of high pollution load as there are several wastewater drains situated in the middle and downstream of the Langat River basin. Site 2 is partly situated in the middle stream region, designated as moderately polluted. Site 1 and a part of site 2 are located upstream of the Langat River, in an area of relatively low river pollution. It is worth mentioning here that some stations have missing data and not all stations were consistently sampled.

Although there are 30 water quality parameters available, only 23 completely monitored parameters were selected. A total of 254 samples were used for the analysis. The 23 water quality parameters were dissolved oxygen (DO), biological oxygen demand (BOD), electrical conductivity (EC), chemical oxygen demand (COD), ammoniacal nitrogen (AN), pH, suspended solids (SS), temperature (T), salinity (Sal), turbidity (Tur), dissolved solid (DS), total solid (TS), nitrate (NO), chlorine (Cl), phosphate (PO), zinc (Zn), calcium (Ca), iron (Fe), potassium (K), magnesium (Mg), sodium (Na), E.coli and coliform.

<table>
<thead>
<tr>
<th>DOE Station No.</th>
<th>Study Code</th>
<th>Distance From Estuary (km)</th>
<th>Grid Reference</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>2814602</td>
<td>Sb07</td>
<td>4.19</td>
<td>2° 52.027’N 101° 26.241’E</td>
<td>Kampung Air Tawar (penghujung jalan)</td>
</tr>
<tr>
<td>2815603</td>
<td>Sb06</td>
<td>33.49</td>
<td>2° 48.952’N 101° 30.780’E</td>
<td>Telok Datuk, near Banting Town</td>
</tr>
<tr>
<td>2817641</td>
<td>Sb05</td>
<td>63.43</td>
<td>2° 51.311’N 101° 40.882’E</td>
<td>Bridge at Kampung Dengkil</td>
</tr>
<tr>
<td>2918606</td>
<td>Sb04</td>
<td>81.14</td>
<td>2° 57.835’N 101° 47.030’E</td>
<td>Near West Country Estate</td>
</tr>
<tr>
<td>2917642</td>
<td>Sb03</td>
<td>86.94</td>
<td>2° 59.533’N 101° 47.219’E</td>
<td>Kajang bridge</td>
</tr>
<tr>
<td>3017612</td>
<td>Sb02</td>
<td>93.38</td>
<td>3° 02.459’N 101° 46.387’E</td>
<td>Junction to Serdang, Cheras at Batu 11</td>
</tr>
<tr>
<td>3118647</td>
<td>Sb01</td>
<td>113.99</td>
<td>3° 09.953’N 101° 50.926’E</td>
<td>Bridge at Batu 18</td>
</tr>
</tbody>
</table>

Table 1. DOE sampling station at study area.

2.2 Principal component analysis

In this work, PCA was performed on the above mentioned water quality parameters to rank their relative significance and to describe their interrelation patterns. Chosen PC scores of
the 23 water quality parameters were used as input variables in ANN model to predict the WQI. The principal components (PCs) can be expressed as

$$z_j = a_{i1}x_{ij} + a_{i2}x_{2j} + ... + a_{im}x_{mj}$$

Where $z$ is the component score, $a$ is the component loading, $x$ the measured value of variable, $i$ is the component number, $j$ is the sample number and $m$ is the total number of variables.

The PCs generated by PCA are sometimes not readily interpreted; therefore, it is advisable to rotate the PCs by varimax rotation. Varimax rotation ensures that each variable is maximally correlated with only one PC and a near zero association with the other components (Abdul-Wahab et al., 2005; Sousa et al., 2007). Varimax rotations applied on the PCs with eigenvalues more than 1 are considered significant (Kim and Mueller, 1987) where the typical criteria are 75-95% of total variance (Chen and Mynett, 2003). The rotations were carried out, in order to obtain new groups of variables. Variables with communality greater than 0.7 are considered, having significant factor loadings (Stevens, 1986).

### 2.3 Artificial neural networks for WQI prediction

In this work, the back propagation (BP) ANN was used in the development of all the prediction models. The Activation Transfer Function of a back-propagation network is usually a differentiable Sigmoid (S-shape) function, which helps to apply the non-linear mapping from inputs to outputs. A three layer back-propagation ANN is used in this study. The number of input and output neurons is determined by the nature of the problem under study. In this study, the networks were trained, tested and validated with one hidden layer and 1 to 10 hidden neurons. This choice was based on the work of Jiang et al. (2004), who found that the results with one hidden layer was better than that of two hidden layers, and the best performance was obtained using a structure with 3 to 6 neurons in the hidden layer. The output neuron (layer) gives the predicted WQI value.

Two different types of ANN models were developed. In the first type, prediction was performed based on the original PCs. In the second type of ANNs developed, scores of rotated (varimax rotation) PCs (ANN-RPCs) with eigenvalues greater than 1 were selected as input. For this model, prediction of WQI was performed using two to six rotated principal components separately.

The original PCs and rotated PCs (RPCs) data sets consist of 305 observations (305 rows) and are divided into training, testing and validating phases for WQI prediction. The ANN predicted WQI values are compared to the WQI values calculated using the DOE-WQI formula which is based on 6 water quality parameters, namely the DO, COD, BOD, AN, SS and pH (DOE, 1997). The input data matrix consists of 23 water quality variables (column) and 305 observations (rows) [23×305]. The observed data for each station is arranged according to time of observation from September 13, 1995 to June 7, 2002. Table 2 describes the data structure. The validation data is at least 10% of the whole data set, with 75% training set and 25% testing set data (Kuo et al., 2007).
Table 2. The data structure for ANN prediction model.

2.4 Determination of model performance

The model’s behaviour in both learning (training and testing) and validating phase, is evaluated using the following statistical methods; the correlation coefficient (R) at 95% confidence limit, given by equations;

\[
\text{Coefficient of correlation (R), } R = \frac{\left[ \sum x_i \hat{x}_i - \frac{1}{n} \left( \sum x_i \right) \left( \sum \hat{x}_i \right) \right]^2}{\left[ \sum x_i^2 - \frac{1}{n} \left( \sum x_i \right)^2 \right] \left[ \sum \hat{x}_i^2 - \frac{1}{n} \left( \sum \hat{x}_i \right)^2 \right]^\frac{1}{2}} \tag{2}
\]

and the mean bias error or residual error given by;

\[
\text{Mean bias error (MBE), } MBE = \frac{1}{n} \sum_{i=1}^{n} (\hat{x}_i - x_i) \tag{3}
\]

Where \( \hat{x}_i \) and \( x_i \) represent observed values and the corresponding forecast values for \( i = 1, 2, \ldots, n \).

The prediction performance evaluated using these two methods are used to evaluate the accuracy of the forecast and for comparing the forecasting ability of each approach.

The 95% confidence limit is used to determine that the predicted output lie within the confidence range. It is assumed that a predicted value fall into an interval within which there is an associated uncertainty. According to Wackerly et al. (1996), this uncertainty is derived from the residual errors that have already been calculated within that range of values. If the residual errors are randomly distributed, there is a general rule of thumb which states that they will lie within two standard deviations of their mean with a probability of 0.95. This method was used in the measurements of the ANN prediction performance conducted by some researchers (Bishop, 1995; Tibshirani, 1996; Shao et al., 1997; Zhang et al., 1998; Lowe and Zapart, 1999; Townsend and Tarassenko, 1999).

ANN models and statistical analyses were carried out using MATLAB 7.0 and XLSTAT2008 (Excel2003 add-in) for Windows.
3. Results and discussion

Post PCA, out of the 23 principal components generated, only six PCs with eigenvalues higher than 1 (Table 3) were selected for the ANN input parameters. Selected PCs explained 75.1% of the total variation. Furthermore, communality values were high for the selected PCs, for example, the values are 93% for Cond., 95% for Sal, 98% for DS and TS (Table 4). These results further confirm the choice of the selected number of PCs (Stevens, 1986).

For the first six rotated PCs (RPCs), the loadings from PCA are given in Table 4. The highest correlations between variables are noted in bold. For instance, Cond., Sal, DS, TS, Cl, Ca, K, Mg and Na, have high correlations with RPC1. Eighteen variables with strong loadings were included in the six selected RPCs. Significant variables in RPC1 are Cond., Sal., DS, TS, Cl, Ca, K, Mg and Na; in RPC2 they are DO, BOD and AN; in RPC3 they are SS and Tur and in RPC4, NO$_3$ and PO$_4$$^3$-$.$ The only meaningful loads in RPC5 and RPC6 are pH and Zn.

<table>
<thead>
<tr>
<th>Variables</th>
<th>RPC1</th>
<th>RPC2</th>
<th>RPC3</th>
<th>RPC4</th>
<th>RPC5</th>
<th>RPC6</th>
<th>Communalities</th>
</tr>
</thead>
<tbody>
<tr>
<td>DO</td>
<td>-0.205</td>
<td>-0.722</td>
<td>-0.121</td>
<td>0.046</td>
<td>0.485</td>
<td>-0.066</td>
<td>0.82</td>
</tr>
<tr>
<td>BOD</td>
<td>0.035</td>
<td>0.740</td>
<td>0.071</td>
<td>0.110</td>
<td>0.110</td>
<td>0.022</td>
<td>0.58</td>
</tr>
<tr>
<td>COD</td>
<td>0.340</td>
<td>0.103</td>
<td>0.081</td>
<td>-0.166</td>
<td>0.268</td>
<td>0.326</td>
<td>0.34</td>
</tr>
<tr>
<td>SS</td>
<td>-0.042</td>
<td>-0.009</td>
<td>0.920</td>
<td>0.010</td>
<td>-0.025</td>
<td>0.017</td>
<td>0.85</td>
</tr>
<tr>
<td>pH</td>
<td>0.189</td>
<td>-0.109</td>
<td>-0.204</td>
<td>0.020</td>
<td>0.792</td>
<td>-0.083</td>
<td>0.72</td>
</tr>
<tr>
<td>AN</td>
<td>-0.092</td>
<td>0.797</td>
<td>-0.151</td>
<td>0.161</td>
<td>0.023</td>
<td>-0.032</td>
<td>0.69</td>
</tr>
<tr>
<td>T</td>
<td>0.337</td>
<td>0.368</td>
<td>-0.242</td>
<td>-0.298</td>
<td>-0.317</td>
<td>0.208</td>
<td>0.54</td>
</tr>
<tr>
<td>Cond.</td>
<td>0.963</td>
<td>0.022</td>
<td>-0.043</td>
<td>0.035</td>
<td>0.013</td>
<td>-0.022</td>
<td>0.93</td>
</tr>
<tr>
<td>Sal.</td>
<td>0.974</td>
<td>0.023</td>
<td>-0.038</td>
<td>0.030</td>
<td>0.008</td>
<td>-0.004</td>
<td>0.95</td>
</tr>
<tr>
<td>Tur.</td>
<td>-0.031</td>
<td>-0.007</td>
<td>0.863</td>
<td>0.011</td>
<td>-0.140</td>
<td>-0.035</td>
<td>0.77</td>
</tr>
<tr>
<td>DS</td>
<td>0.988</td>
<td>0.017</td>
<td>-0.034</td>
<td>0.013</td>
<td>0.009</td>
<td>-0.005</td>
<td>0.98</td>
</tr>
<tr>
<td>TS</td>
<td>0.985</td>
<td>0.017</td>
<td>0.069</td>
<td>0.014</td>
<td>0.007</td>
<td>-0.003</td>
<td>0.98</td>
</tr>
<tr>
<td>NO$_3$$^-$</td>
<td>0.018</td>
<td>0.033</td>
<td>0.107</td>
<td>0.688</td>
<td>-0.126</td>
<td>0.300</td>
<td>0.59</td>
</tr>
<tr>
<td>Cl</td>
<td>0.986</td>
<td>0.010</td>
<td>-0.029</td>
<td>-0.004</td>
<td>0.020</td>
<td>0.005</td>
<td>0.97</td>
</tr>
<tr>
<td>PO$_4$$^3-$</td>
<td>0.023</td>
<td>0.312</td>
<td>-0.106</td>
<td>0.700</td>
<td>0.112</td>
<td>-0.073</td>
<td>0.62</td>
</tr>
<tr>
<td>Zn</td>
<td>-0.019</td>
<td>0.044</td>
<td>-0.011</td>
<td>0.186</td>
<td>-0.128</td>
<td>0.767</td>
<td>0.64</td>
</tr>
<tr>
<td>Ca</td>
<td>0.980</td>
<td>0.028</td>
<td>-0.026</td>
<td>-0.043</td>
<td>-0.024</td>
<td>0.039</td>
<td>0.97</td>
</tr>
<tr>
<td>Fe</td>
<td>-0.080</td>
<td>0.043</td>
<td>0.475</td>
<td>0.540</td>
<td>0.066</td>
<td>0.192</td>
<td>0.57</td>
</tr>
<tr>
<td>K</td>
<td>0.984</td>
<td>0.004</td>
<td>-0.031</td>
<td>-0.004</td>
<td>0.004</td>
<td>0.010</td>
<td>0.97</td>
</tr>
<tr>
<td>Mg</td>
<td>0.974</td>
<td>0.000</td>
<td>-0.022</td>
<td>-0.028</td>
<td>-0.002</td>
<td>0.037</td>
<td>0.95</td>
</tr>
<tr>
<td>Na</td>
<td>0.986</td>
<td>0.002</td>
<td>-0.025</td>
<td>-0.020</td>
<td>0.005</td>
<td>0.017</td>
<td>0.97</td>
</tr>
<tr>
<td>COLI</td>
<td>-0.254</td>
<td>0.361</td>
<td>0.097</td>
<td>-0.424</td>
<td>0.457</td>
<td>0.056</td>
<td>0.60</td>
</tr>
<tr>
<td>COLIFORM</td>
<td>-0.032</td>
<td>0.049</td>
<td>-0.025</td>
<td>0.042</td>
<td>-0.077</td>
<td>-0.517</td>
<td>0.28</td>
</tr>
</tbody>
</table>

Table 3. Descriptive statistics of selected original PCs with eigenvalues more than 1.

Table 4. Rotated factor loadings using six PCs.
Using the original principal component scores as inputs, the best architecture consists of a three layer network with 23 input neurons, 10 neurons in the hidden layer and one neuron in the output layer. Considering RPC scores as inputs, the best architectures were achieved with almost the same number of hidden neurons. The hidden neurons consist of 9 and 10 neurons respectively. Training was carried out for a maximum 10000 iterations. Selection of the network was performed at maximum correlation coefficient (R) and 95% confidence limit.

Table 5 and Figure 2 illustrate the prediction performances of ANN models using different combinations of PC scores as input variables. ANN using the first 2 PCs (PC1 and PC2) does not perform very well as far as accuracy is concerned for all the training, testing and validation phases. It is observed that the prediction performance of the validation phase is slightly worse compared to the training and testing phases. It is important to point out that for this model, the cumulative percentage in explaining the variance given by these two RPCs is only 49.8%. None of the strong loading variables contains the variables forming the WQI equation. DO, BOD and pH loadings in PC2 explain only 10.4% of the total variance.

Based on the results, it is apparent that the WQI prediction performance increases with the increase in number of input variables. The highest accuracy in predicting WQI is given by model ANN-RPC6, which contains six RPCs with 75.1% variation explained, giving an $R^2$ value of 0.64 (training), 0.87 (testing), and 0.72 (validation) respectively.

<table>
<thead>
<tr>
<th>Model</th>
<th>No.of PC</th>
<th>$R^2$</th>
<th>MBE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Training</td>
<td>Testing</td>
</tr>
<tr>
<td>ANN-RPC2 (2 inputs)</td>
<td>2</td>
<td>0.43</td>
<td>0.70</td>
</tr>
<tr>
<td>ANN-RPC3 (3 inputs)</td>
<td>3</td>
<td>0.60</td>
<td>0.78</td>
</tr>
<tr>
<td>ANN-RPC4 (4 inputs)</td>
<td>4</td>
<td>0.53</td>
<td>0.79</td>
</tr>
<tr>
<td>ANN-RPC5 (5 inputs)</td>
<td>5</td>
<td>0.53</td>
<td>0.79</td>
</tr>
<tr>
<td>ANN-RPC6 (6 inputs)</td>
<td>6</td>
<td>0.64</td>
<td>0.87</td>
</tr>
<tr>
<td>ANN-PC23</td>
<td>23</td>
<td>0.60</td>
<td>0.85</td>
</tr>
</tbody>
</table>

Table 5. The prediction performances of the different ANN models.

From table 5, it can be observed that the prediction performance of the ANN model using original PCs (23 input PC scores) is not significantly different from the RPC models. However, as RPC models use fewer variables and is far less complex, the advantage over the ANN-PC23 model is obvious. Comparing the MBE values, it is generally observed that the signs for the validation phases are negative for both the un-rotated and rotated PC models. This is an indication that the predicted WQI values are consistently underestimated in both approaches.
Using Principal Component Scores and Artificial Neural Networks in Predicting Water Quality Index

Part I

- **Training**
  - $R^2 = 0.4268$
  - [Graph showing scatter plot with line of best fit]

- **Testing**
  - $R^2 = 0.6951$
  - [Graph showing scatter plot with line of best fit]

- **Validation**
  - $R^2 = 0.3183$
  - [Graph showing scatter plot with line of best fit]

- **ii)**
  - $R^2 = 0.6021$
  - [Graph showing scatter plot with line of best fit]

- $R^2 = 0.7842$
  - [Graph showing scatter plot with line of best fit]

- $R^2 = 0.6067$
  - [Graph showing scatter plot with line of best fit]
Fig. 2. The prediction performances for different combination of PC scores during training, testing and validation phases: (i) 2 RPCs, (ii) 3 RPCs, (iii) 4 RPCs, (iv) 5 RPCs, (v) 6 RPCs and, (vi) 23 original PCs.
This study also attempts to allocate 95% confidence interval on the WQI prediction produced by the best ANN model. Figure 3, 4 and 5 show the comparison between predicted values and the upper (UL) and lower limits (LL) lying within 95% confidence interval. This was carried out for ANN-RPC6 and ANN-PC23 models. It can be seen that only 4.3% out of the 305 predicted values were identified beyond the 95% confidence limit (1% fall below the LL and 3.3% fall beyond the UL) for ANN-RPC6. For ANN-PC23, 25% of the 305 observations fall beyond the upper and lower of 95% confidence interval limit (14% fall below the LL and 11.8% fall beyond the UL). This basically shows that by using reduced rotated PC scores as input, better results can be obtained without losing information. It is thus apparent that ANN prediction using scores of varimax rotated PCs result in a more accurate WQI prediction.

Fig. 3. Predicted WQI within the 95% confidence interval during training phase using (a) six rotated PCs, and (b) 23 original PCs.
Fig. 4. Predicted WQI within the 95% confidence interval during testing phase using (a) six rotated PCs, and (b) 23 original PCs.
4. Conclusion

In this work, a combination of PCA and ANN is used to predict WQI based on 23 historical water quality parameters. The original predictors were selected based on the available Malaysian DOE data. To obtain the latent variables as inputs into the ANN, two different approaches were used; one based on un-rotated original PCs and the other based on varimax rotated PCs.

Using six PCs, significant loadings are observed for Cond, Sal, DS, TS, Cl, Ca, K, Mg and Na in PC1, DO, BOD and AN in PC2, SS and Tur in PC3, NO3- and PO43- in PC4, pH in PC5 and Zn in PC6. ANN models based on these 6 PC scores can predict WQI with acceptable...
Using Principal Component Scores and Artificial Neural Networks in Predicting Water Quality Index

accuracy (within 95% confidence limit). Moreover, the ANN model using the 23 original PCs as input, do not render the prediction more accurate, even with a complex network structure. The use of rotated PC scores based models is clearly more effective and efficient due to the elimination of collinearity and reduction of predictor variables without losing important information.

5. Acknowledgment
The authors acknowledge the financial and technical support for this project provided by the Ministry of Science, Technology and Innovation and Universiti Putra Malaysia under the Science Fund Project no. 01-01-04-SF0733. The authors wish to thank the Department of Environment, and Department of Irrigation and Drainage, Ministry of Natural Resources and Environment of Malaysia, Institute for Development and Environment (LESTARI), Universiti Kebangsaan Malaysia, Universiti Malaya Consultancy Unit (UPUM) and Chemistry Department of Universiti Malaya, who have provided us with secondary data and valuable advice.

6. References


www.intechopen.com


In the book "Chemometrics in practical applications", various practical applications of chemometric methods in chemistry, biochemistry and chemical technology are presented, and selected chemometric methods are described in tutorial style. The book contains 14 independent chapters and is devoted to filling the gap between textbooks on multivariate data analysis and research journals on chemometrics and chemoinformatics.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:
