Data mining and multivariate statistical analysis for ecological system in coastal waters

Kwok-wing Chau and Nitin Muttil

ABSTRACT

In this study, data mining using box plots and multivariate statistical analysis using factor analysis are employed for a spatio-temporal analysis of coastal water quality data from Tolo Harbour, Hong Kong. The analysis of box plots reveals pronounced spatial heterogeneity of the parameters studied. The spatial analysis clearly shows monitoring station TM2 in the Harbour Subzone to be most susceptible to eutrophication with the highest nutrient and algal biomass concentrations. The factor analysis brings to light dominant parameters to the ecological system under the coastal marine environment. The temporal analysis confirms the considerable decline in nutrient levels in recent years. In spite of this decline, the factor analysis indicates that nutrient processes play an important role even in recent years, suggesting an adequate supply of nutrients. It seems that they are being released from sources other than known point sources, possibly from nutrients accumulated in the sediments, necessitating steps to be undertaken for their control also. This study demonstrates the use of data mining techniques in the ecological system in Tolo Harbour.

Key words | box plots, data mining, factor analysis, harmful algal blooms, multivariate statistical analysis

Kwok-wing Chau (corresponding author)
Nitin Muttil
Department of Civil and Structural Engineering,
Hong Kong Polytechnic University,
Hung Hom, Hong Kong,
China
Tel: +852 2766 6014
E-mail: *cekwchau@polyu.edu.hk*

INTRODUCTION

A major impact of eutrophication is the stimulation of algal growth and the production of harmful algal blooms (HABs). HAB incidents can have a significantly devastating economic impact on the local fishing industry and on tourism. Over the past two decades, massive fish kills due to HABs and/or hypoxia (low dissolved oxygen levels) have occurred in some of the marine fish culture zones in Hong Kong (Sin & Chau 1992). Thus, better understanding of the complex ecological processes and HAB dynamics is of the utmost importance. Research on HABs have been done for more than 20 years and the general ecological response of phytoplankton to environmental conditions has been extensively studied and incorporated in process-based mathematical models of eutrophication (e.g. Thomann & Mueller 1987; Chau 2004). Nevertheless, the prediction of algal blooms remains a very difficult problem, owing to the extremely complicated ecological dynamics.

In recent years, with the general availability of computing systems with ever-expanding capabilities, there is a growing tendency to use data mining (DM) techniques to complement or even replace process-based models. There are two primary goals of data mining in practice, namely “prediction” and “description” (Fayyad *et al.* 1996). Prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest, whereas description focuses on finding interpretable patterns describing the data. With the aim of prediction, extensive use of machine learning techniques has been reported in ecological modelling (Recknagel 2001). These include artificial neural networks (Recknagel *et al.* 1997, 2002; Yabunaka *et al.* 1997; Maier *et al.* 1998; Scardi & Harding 1999; Karul *et al.* 2000; Jeong *et al.* 2001; Scardi 2001; Wei *et al.* 2001; Lee *et al.* 2003), evolutionary based techniques (Bobbin & Recknagel 2001; Recknagel *et al.* 2002; Jeong *et al.* 2003; Chau 2005; Muttil & Lee 2005), fuzzy and neuro-fuzzy techniques (Maier *et al.* 2001; Chen & Mynett 2003), and so on. DM techniques with the goal of “description” have also been used, but to a lesser extent. These include

doi: 10.2166/hydro.2007.003

principal component analysis (Petersen *et al.* 2001; Chen & Mynett 2003), cluster analysis (Brosse *et al.* 2001), machine learning techniques (Muttil & Chau 2007), etc.

In this study, we use descriptive DM techniques to reveal the spatial and temporal ecological dynamics of the coastal waters of Hong Kong. Data mining using box plots and multivariate statistical analysis using factor analysis are employed in this study. In the following sections, we first present brief descriptions of the data mining techniques used, which are followed by the analysis of box plots and factor analysis of the ecological and related water quality data.

DATA MINING METHODOLOGIES

DM and knowledge discovery in databases (KDD) is concerned with extracting useful information from databases (Fayyad 1997). The process involves discovering useful (hidden) patterns in data: knowledge extraction, exploratory data analysis, information harvesting, data dredging, etc. Various steps of the DM and/or KDD process include data warehousing, target data selection, cleaning, preprocessing, transformation and reduction, DM, model selection (or combination), evaluation and interpretation, and finally consolidation and use of the extracted "knowledge". It can be said to be on the interface between computer science and statistics, integrating multiple technologies from both disciplines, which include database management and warehousing, machine learning, decision support, and others such as visualisation and parallel computing (Figure 1, adapted from Thuraisingham (1999)). Statistics is a major area contributing to DM and various statistical analysis software packages are now being marketed as data mining tools. Thus, the main part of DM is concerned with the analysis of data and the use of software techniques for finding patterns and regularities in sets of data. The choice of a particular combination of techniques to apply in a particular situation depends on both the nature of the data mining task to be accomplished and the nature of the available data.

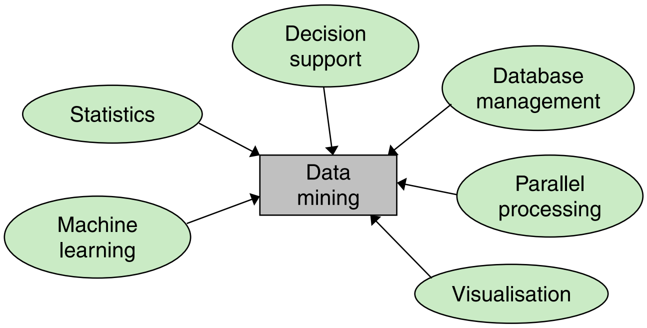

Figure 1 | Knowledge discovery in databases and data mining: integration of multiple technologies.

Different DM techniques have been applied to ecological engineering field. Boogaard (1998) presented self-organization feature maps for analysis of hydrological and ecological data sets. Hall & Anderson (1999) used a deterministic ecological risk assessment for copper in European saltwater environments with a hazard quotient approach. Russom (2002) performed data mining on environmental toxicology information with the focus on currently available web resources. Su *et al.* (2004) applied a spatiotemporal assignment mining model to determine the spatio-temporal relationship between environmental factors and fish distribution. Chen & Mynett (2005) employed self organization feature maps to analyse eutrophication in Tahu Lake. Tadesse *et al.* (2005) employed rule-based regression tree models for predicting drought-related vegetation stress, integrating satellite, climate, and biophysical data over the US central plains. Bui *et al.* (2006) presented knowledge discovery from models of soil properties developed through data mining with piecewise linear tree models. Stockwell (2006) developed improved ecological niche models by data mining large environmental datasets for surrogate models.

The techniques employed in this study are described briefly in the following subsections.

Visual data mining techniques

Visual data mining refers to the visual presentation of data to extract useful information. The use of visualisation techniques allows users to summarise, extract and grasp more complex patterns and results than mathematical or text type descriptions of the same. Box plots are used in this study for a spatial and temporal analysis of time series water quality data.

Box plots, or box and whisker diagrams, provide an excellent visual summary of a set of data. They show a measure of central location (the median), two measures of spread or variation (the range and inter-quartile range), the skewness (from the orientation of the median relative to the quartiles) and potential outliers (marked individually). More specifically, the line across the box represents the median, which is the point where 50% of the data is above it and 50% below it. The bottom of the box is at the first quartile, Q_1 (where at most 25% of the data fall below it) and the top is at the third quartile, Q_3 (where at most 25% of the data is above it). The box itself represents the middle 50% of the data. The whiskers are the lines that extend from the bottom and top of the box to the lowest and highest observations inside the range defined by a lower limit of $Q_1 - 1.5(Q_3 - Q_1)$ to an upper limit of $Q_3 + 1.5(Q_3 - Q_1)$, where $(Q_3 - Q_1)$ is the inter-quartile range. Box plots are especially useful when comparing two or more sets of data (Chambers *et al.* 1983).

Factor analysis

Multivariate data often includes a large number of measured variables, and sometimes these variables "overlap" in the sense that groups of them may be dependent or correlated. Factor analysis (FA) is a way to fit a model to multivariate data to estimate just this sort of interdependence. In a FA model, the measured variables depend on a smaller number of unobserved (or latent) "factors". In other words, this statistical approach involves finding a way of condensing the information contained in a number of original measured variables into a smaller set of factors with a minimum loss of information. Because each factor may affect several variables in common, they are known as "common factors". Each variable is assumed to be dependent on a linear combination of the common factors, and the coefficients are known as "loadings". FA can also be used to generate hypotheses regarding causal mechanisms or to screen variables for subsequent analysis. There are four basic steps in FA: data collection and generation of the correlation matrix; extraction of initial factor solution; rotation and interpretation; construction of scales or factor scores to use in further analyses.

FA and principal component analysis (PCA) use the same set of mathematical tools (spectral decomposition, projection, etc.), but there are substantial differences between the two data analysis techniques. Both are dimension-reduction techniques, in the sense that they can be used to replace a large set of measured variables with a smaller set of new variables. However, the two methods are different in their goals and underlying models. The biggest difference between FA and PCA comes from the model philosophy. FA imposes a strict structure of a fixed number of common factors whereas the PCA determines p factors in decreasing order of importance. The most important factor in PCA is the one that maximises the projected variance (the first principal component). On the other hand, the most important factor in FA is the one that (after rotation) gives the maximal interpretation. Often this is different from the direction of the first principal component (Hardle & Simar 2003). Thus, PCA can be used only when the goal is to simply summarise or approximate the data with fewer dimensions, whereas FA can be used to give an interpretation model for the correlation among the data.

In this study, the adopted factor extraction method is the most commonly used principal component extraction method. The eigenvalues of the correlation matrix measure the amount of the variation explained by each factor and will be the largest for the first factor and become smaller for the subsequent factors. The goal of factor rotation is to find a parametrisation in which each variable has only a small number of large loadings, i.e. is affected by a small number of factors, preferably only one. This can often facilitate the interpretation of the representation of the factors. The varimax method of orthogonal rotation using the Kaiser normalization method (Kaiser 1958) is used in this study. Rotated factors are most widely called "varifactors". The higher the loading of a variable (either positive or negative), the more that variable contributes to the variation accounted for by the particular varifactor. In practice, only loadings with absolute values greater than 60% are selected for the factor interpretation (Jolliffe 1986). A factor with an eigenvalue greater than or equal to one is usually considered as being of statistical significance (the Kaiser criterion).

Before applying FA, it is necessary to test the validity of applying it using the Kaiser–Meyer–Olkin measure of sampling adequacy. This statistical measure indicates the proportion of variance in the variables that might be caused by the underlying factors. High values (close to 1.0) generally indicate that a FA may be useful with the data. If the value is less than 0.50, the results of the FA will probably not be very useful.

DATA AND MODELLING APPROACH

The selected DM techniques are applied to water quality data from Tolo Harbour (Figure 2), which are measured under the

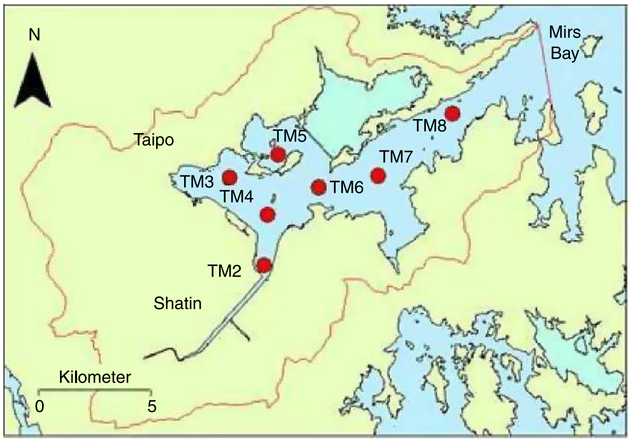

Figure 2 | The study site: Tolo Harbour indicating the seven monitoring stations.

water quality monitoring program of the Environmental Protection Department (EPD) of the Hong Kong government. The EPD has seven water quality monitoring stations in Tolo Harbour, named as TM2, TM3, … , TM8. Based on the spatial variation in water quality, the harbour is divided into three subzones: the inner Harbour Subzone (with stations TM2, TM3 and TM4), the intermediate Buffer Subzone (with stations TM5 and TM6) and the outer Channel Subzone (consisting of stations TM7 and TM8). The following subsections provide further details of the study site and the data used.

The study site

Tolo Harbour is a semi-enclosed bay connected to the open sea at Mirs Bay (Figure 2) with a gradient of improving water quality from the more enclosed and densely populated inner Harbour Subzone to the outer "better flushed" Channel Subzone. The nutrient enrichment in the weakly flushed harbour due to municipal and livestock waste discharges has been a major environmental concern and eutrophication has resulted in frequent algal blooms (Chau & Jin 1998). Consequently, occasional massive fish kills were recorded as a result of severe dissolved oxygen depletion or toxic algal blooms. Morton (1988) reported that the inner Tolo Harbour was effectively dead as a marine disaster in the late 1980s (in Figure 3, the increase in frequency of HABs in the late 1980s can be observed). At that time, a critical stage had been reached, which prompted the Hong Kong government to implement an integrated Tolo Harbour Action Plan (THAP). The measures implemented include: controlling livestock pollution, restoring old landfills, enforcing the Water Pollution Control Ordinance and building sewer networks in rural areas (EPD 2003). THAP resulted in a significant reduction of pollutant loading, which in turn improved the water quality. Further improvement in the water quality took place after the implementation of the Tolo Harbour Effluent Export Scheme (THEES), which became fully operational in early 1998. Under the THEES, fully treated effluent from the two sewage treatment plants in Shatin and Taipo (see Figure 2) are transported to a new pumping station at Shatin, and are exported to Victoria Harbour for discharge through a series of sewer pipes and tunnel. Earlier, before the THEES was introduced and implemented, the treated effluent used to be discharged into Tolo Harbour.

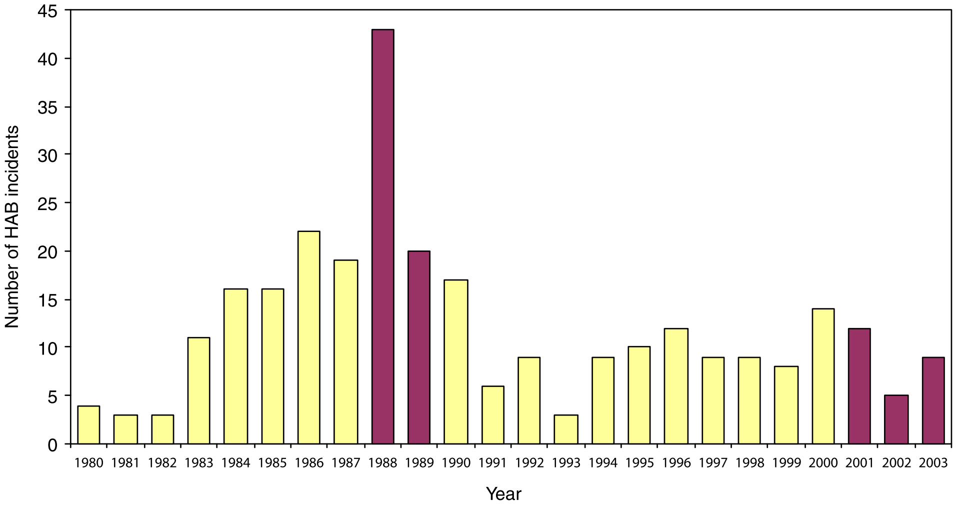

Figure 3 | Frequency of HABs in Tolo harbour.

The dataset used

Depth-averaged water quality data provided by the EPD are used in this study. The data are measured either biweekly or monthly. Fourteen parameters are used in this study, which are presented in Table 1. The available data covered a period from 1983 to 2003 for all seven water quality monitoring stations in Tolo Harbour.

Datasets from any two time periods can be selected for temporal analysis. In this study, the two most representative datasets are adopted. The first dataset is from a period when the water quality in Tolo Harbour was at its worst, which is the late 1980s. During this period, the HAB incidents increased significantly and reached a peak in 1988, when a total of 43 incidents were reported (Figure 3). To represent this time period, two years of data from 1988–1989 are selected. The second dataset is from a period when the water quality had improved significantly after the implementation of the THAP and the THEES. The data from 2001–2003 is selected as the second dataset. Both the selected time periods of data are shaded dark in Figure 3.

Table 1 | List of water quality variables

Variable name	Symbol	Units
Nutrients		
Ammonia nitrogen	NH_4	mg/L
Nitrate nitrogen	NO_3	mg/L
Total nitrogen	TN	mg/L
Orthophosphate	PO_4	mg/L
Total phosphorus	TP	mg/L
Physical properties		
Suspended solids	SS	mg/L
pH	pH	–
Turbidity	TURB	NTU
Water temperature	TEMP	°C
Dissolved oxygen	DO	mg/L
Secchi disc depth	SD	m
Salinity	SAL	PSU
Organic constituents		
5 d biochemical oxygen demand	BOD5	mg/L
Biological indicator		
Chlorophyll-a	CHL	μg/L

MINING OF WATER QUALITY DATA FROM TOLO HARBOUR

Spatio-temporal analysis using box plots

In this section, we present a spatial and temporal analysis of the 14 selected water quality parameters using their box plots. The spatial analysis is undertaken by using the data from the seven marine water monitoring stations, whereas the temporal analysis is performed using two sets of data, namely from 1988–1989 and from 2001–2003.

Box plot analysis of CHL and DO

CHL and DO are generally taken as primary parameters for water quality monitoring and algal biomass estimation, box plots for which are presented in Figure 4. From the box plot for CHL, it is observed that the spread of the box plots gradually decreased from the TM2 station to TM8 for both the periods of data. During the period 1988–1989, the CHL values were much higher at TM2 than at the other stations, indicated by the larger value of Q_3 and a longer top whisker. It can also be observed that, for this box plot, the median is at about 20 μg/L, indicating that 50% of the CHL values in 1988–1989 are above 20 μg/L. Typically, CHL concentrations exceeding 20 μg/L would be considered to constitute an algal bloom (Chau & Jin 1998). The box plots for 2001–2003 indicate lower CHL values, as compared to those for 1988–1989, indicating a reduction in algal biomass concentration. For the box plots of DO, the period 1988–1989 have much longer bottom whiskers, indicating much lower DO values. During the 1988–1989 period, DO values reached minimum values of about 1 mg/L at stations TM2, TM3, TM7 and TM8, whereas the minimum values during 2001–2003 were around 4 mg/L, again indicating an

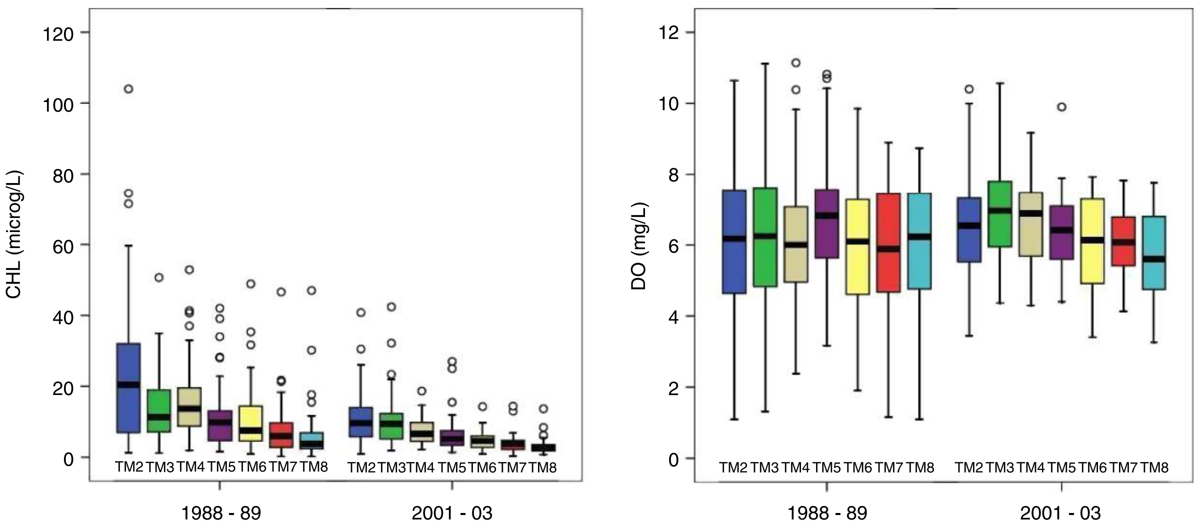

Figure 4 | Box plots for chlorophyll-a and dissolved oxygen.

improvement in water quality. The Water Quality Objective set by the EPD for the Tolo Harbour is that depth-averaged DO should be greater than or equal to 4 mg/L in all the samples (EPD 2003). DO levels less than 4 mg/L are a cause for concern, as it may cause fish kills due to hypoxia. As far as the spatial variation is concerned, DO does not seem to indicate any significant variation amongst the seven monitoring stations. This indicates that there are no severe single pollution sources within the entire water body and that the circulation within the water body is good, which agrees well with observation data.

Box plot analysis of nutrients and BOD5

The box plots for the nutrients and BOD5 are presented in Figure 5. With the exception of NO_3 and PO_4, all the remaining nutrient variables showed a gradual decrease in concentration from stations TM2 to TM8. For all the nutrients, and especially for NO_3 and PO_4, the values at TM2 during 1988–1989 were exceptionally high. This high concentration of nutrients very clearly indicates the poor water quality in Tolo Harbour during the period 1988–1989. Thus, the drastic increase in the number of HABs in Tolo Harbour in 1988 is not surprising (see Figure 3). As indicated by the box plots for 2001–2003, the water quality after about 10–12 years has shown significant improvement. The BOD5 also showed high values at TM2 during 1988–1989, whereas for both time periods the BOD5 at TM8 was clearly lower than at the other monitoring stations, indicating much better water quality.

Box plot analysis of physical properties

The box plots for the physical properties are presented in Figure 6. As expected, SAL showed a gradual increase in concentration as we move closer to the sea, from station TM2 to TM8. As far as the temporal behavior of SAL is concerned, it showed reduced values in 2001–2003, indicating the improved oceanographic conditions of marine water at that time (as salinity reduces, solubility of oxygen in water increases). SD also showed increasing values from TM2 to TM8, indicating increasing transparency and light penetration from TM2 to TM8. With the improvement in water quality in 2001–2003, the SD values also showed a slight increase. The SS and TURB did not indicate any spatial pattern from TM2 to TM8, but their values at TM2 were clearly higher than at the other stations. TURB values showed a certain increase in the period 2001–2003, as compared to the values in 1988–1989. This phenomenon might be explained by annual variations. As with TURB, TEMP values also showed a significant increase in 2001–2003, shown by an upward shift in the boxes and the median by about 3°C. The minimum water temperature in 1988–1989 can be seen in the box plots to be around 13°C, whereas in 2001–2003, it was about 16°C. This is undesirable because increase in water temperature cause

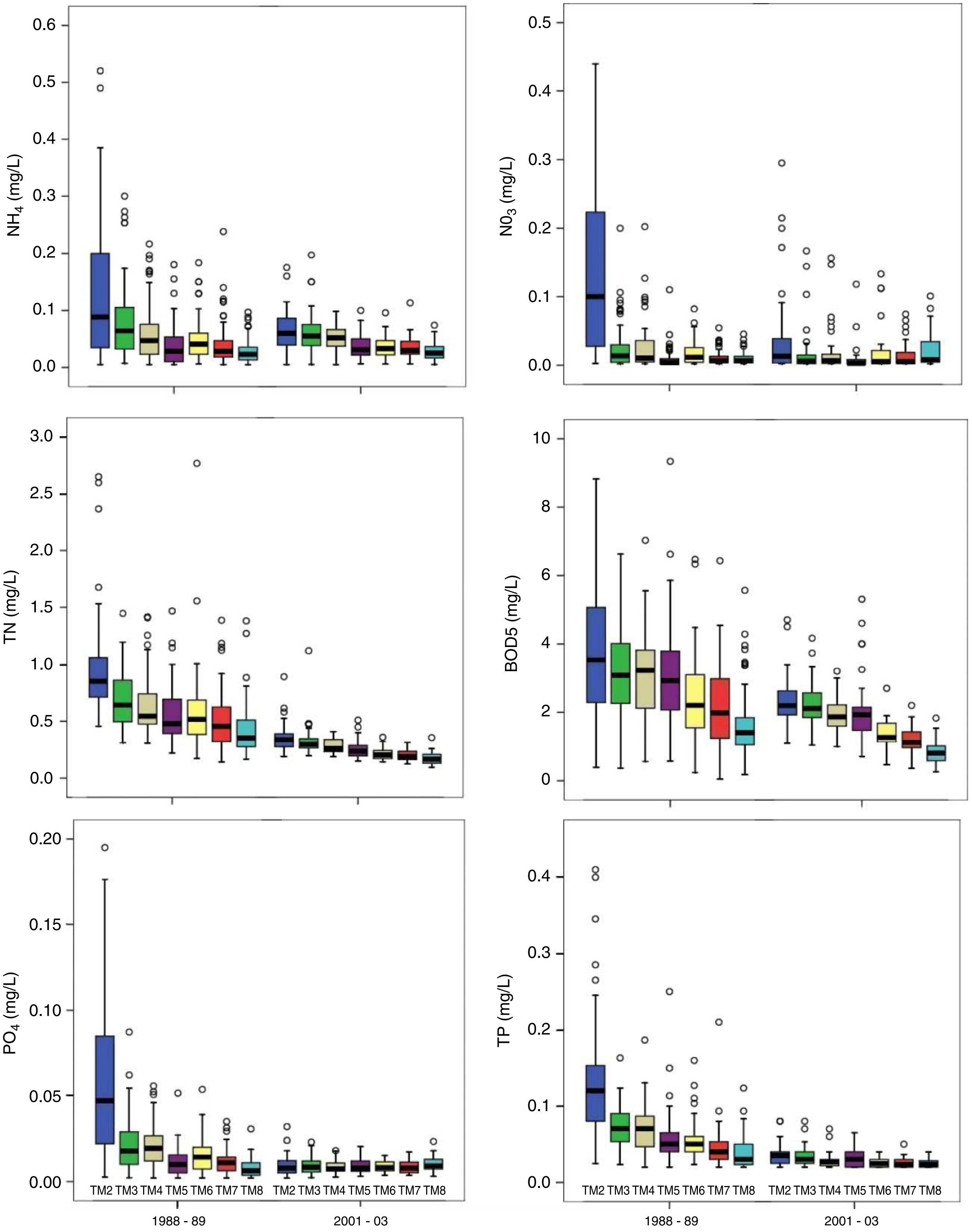

Figure 5 | Box plots for nutrients and 5 d biochemical oxygen demand.

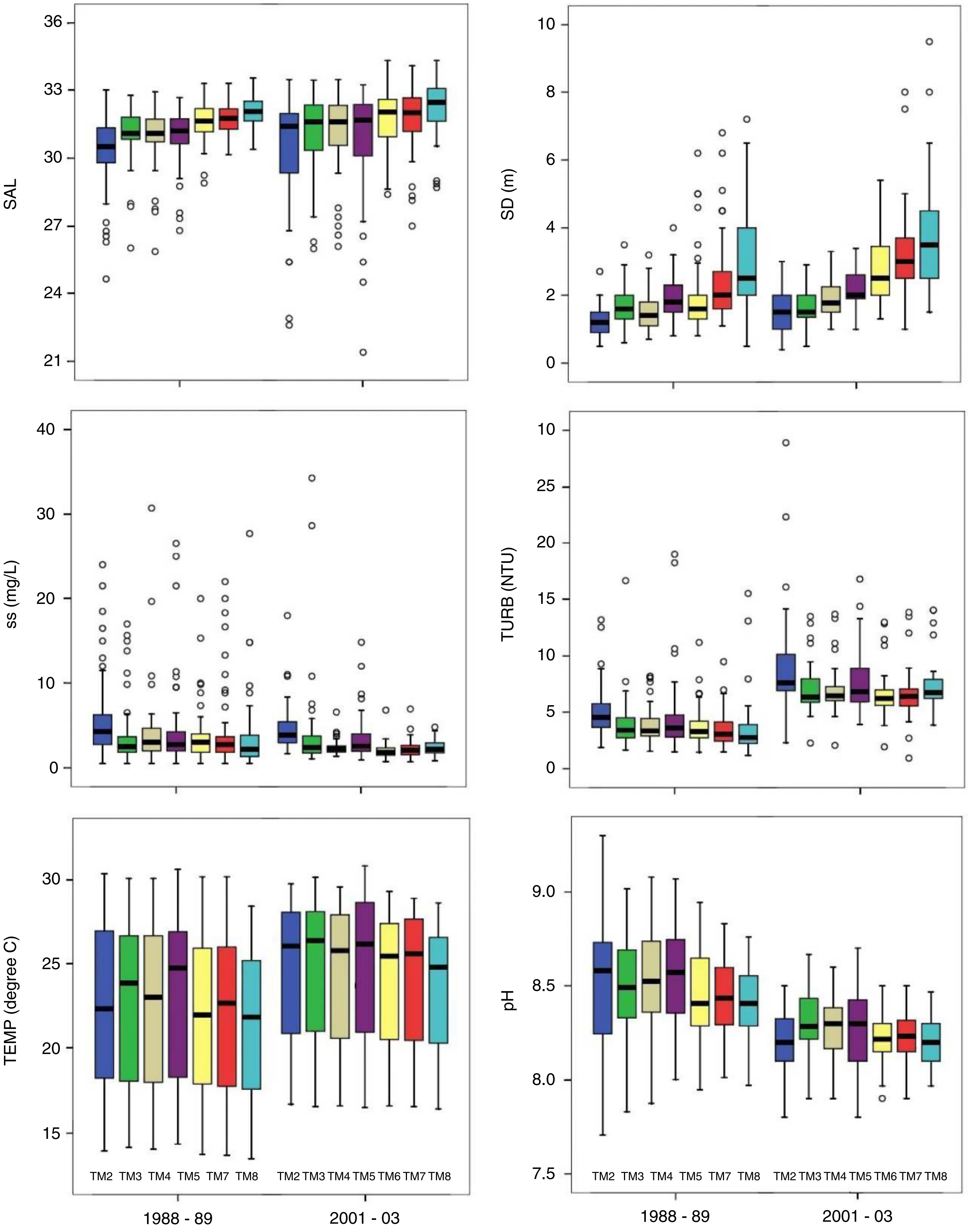

Figure 6 | Box plots for physical properties.

oxygen solubility to decline. Finally, the pH values did not show any clear spatial change from TM2 to TM8. But the pH values clearly showed smaller values in 2001–2003, indicating that the water was more alkaline during the period 1988–1989.

Thus, as far as the spatial variation is concerned, the water quality in the Channel Subzone was clearly much better than in the Harbour Subzone. Further, out of the three monitoring stations in the Harbour Subzone, the water quality in TM2 has consistently been the worst. This suggests TM2 to be the most weakly flushed monitoring station and with consequently the highest concentration of nutrients. The analysis of the temporal variation between the two time periods clearly indicated a significant improvement in water quality in 2001–2003, as compared to that about 10 years earlier.

Factor analysis

The FA was applied to the two datasets considered in this study, namely the 1988–1989 and the 2001–2003 data. In order to isolate the ecological processes from the hydrodynamic effects as much as possible, using the data from monitoring station TM2, which was found to have the worst water quality, for FA seems to be most appropriate. But, since for the 2001–2003 dataset, only monthly values of the water quality variables are available, we only have 36 data records in this dataset. Thus, in order to increase the number of data records, data from the three monitoring stations in the Harbour Subzone (TM2, TM3 and TM4) were combined for both datasets from the different time periods.

Factor analysis for 1988–1989 dataset

Prior to the factor analysis, the data first underwent some preprocessing. The Kaiser–Meyer–Olkin measure of sampling adequacy for this dataset was found to be 0.706, and since it was greater than 0.5, applying the FA seems to be reliable. Table 2 shows the factor loadings obtained from the principal components factor analysis with varimax rotation. The first four varifactors, accounting for 68.66% of the total variation, are retained on the basis of the "eigenvalue greater than one" rule. Factor loadings with values greater than 0.60 are shown in bold font and only these are used for the factor interpretation.

Table 2 | Factor loadings from a principal component factor analysis for the 1988–1989 dataset

	Varifactors			
Variable	**1**	**2**	**3**	**4**
PO_4	0.871	0.050	0.080	0.054
NO_3	0.828	0.004	− 0.018	− 0.028
NH_4	0.783	− 0.364	0.048	0.063
TP	0.725	0.239	0.371	0.368
TN	0.527	0.074	0.466	0.426
DO	− 0.070	0.760	0.421	− 0.188
BOD5	0.052	0.729	0.325	0.353
CHL	0.031	0.716	0.059	0.339
pH	− 0.323	0.670	− 0.377	0.082
SAL	− 0.370	− 0.542	0.409	− 0.112
TEMP	− 0.212	− 0.202	− 0.767	0.169
SS	0.024	− 0.018	0.663	0.193
TURB	0.081	0.080	0.121	0.781
SD	− 0.063	− 0.183	0.094	− 0.708
Eigenvalues	3.994	2.826	1.614	1.180
% of variance	28.529	20.187	11.527	8.427
Cumulative %	28.529	48.715	60.243	68.669

Extraction method: principal component analysis. Rotation method: varimax with Kaiser normalisation.

From Table 2, it is observed that the first two varifactors were dominant and together accounted for 48.71% of the total variance, whereas the remaining two varifactors were secondary and accounted for 11.52% and 8.42% of the variance, respectively. The first varifactor, with an eigenvalue of 3.99, explained 28.52% of the total variance. It is clearly dominated by the nutrients, PO_4, NO_3, NH_4, TP and TN, which exhibited significant positive loadings. The Harbour Subzone, being a largely enclosed and consequent weakly flushed bay, was highly vulnerable to nutrient enrichment during the late 1980s, to which period this dataset belongs. Thus, this varifactor can be clearly interpreted as representing the nutrient processes, namely the nitrogen and phosphorus cycles – the hydrolysis of organic nitrogen to ammonia nitrogen and its oxidation to nitrate nitrogen and also the decay of phosphorus.

The second varifactor, with an eigenvalue of 2.82 and accounting for 20.18% of the total variance, was also significant. The variables DO, BOD5 and CHL exhibited high positive loadings, pH showed low to moderate positive loading and SAL exhibited a low to moderate negative loading. This varifactor seems to represent the hydrobiological processes like phytoplankton primary production, microbial degradation and dissolved oxygen budget. Phytoplankton has two opposite effects on the oxygen level: oxygen production by photosynthesis on the one hand and oxygen consumption due to its respiration and death (microbial degradation) on the other hand (Lee *et al.* 1991a,b). Thus, this varifactor seems to indicate the connectivity between phytoplankton or algal biomass (represented by CHL) and the oxygen production (DO) and the consumption of oxygen (BOD5).

The remaining two varifactors explain relatively lesser variance, as compared to the first two. They seem to represent the physical properties, as they include TEMP, SS, TURB and SD. Since there has been an abundance of nutrients in the Harbour Subzone, the critical limiting factor for phytoplankton growth dynamics was clearly not the nutrients. The formation of algal blooms is not just dependent on the availability of the nutrients, although they are essential. The third and fourth varifactors seem to indicate the importance of several external environmental factors, i.e. water temperature (TEMP), suspended solid (SS), the degree of penetration of sunlight into the water column (SD and TURB).

Factor analysis for 2001–2003 dataset

The Kaiser–Meyer–Olkin measure of sampling adequacy for this dataset was found to be 0.706, and for this dataset also, applying the FA seems useful. The factor loadings for this dataset are presented in Table 3, in which loading values greater than 0.60 are shown in bold font. For this dataset, the first five varifactors are retained, which account for 69.64% of the total variance. It is observed that the first three varifactors together accounted for 54.03% of the total variance, whereas the remaining two varifactors were less dominant and together accounted for 15.60%.

Table 3 | Factor loadings from a principal component factor analysis for the 2001–2003 dataset

	Varifactors				
Variable	**1**	**2**	**3**	**4**	**5**
NO_3	0.827	0.254	−0.218	0.109	−0.047
TN	0.659	0.551	0.254	0.196	0.110
SD	−0.589	−0.177	−0.124	0.301	−0.212
SAL	−0.570	−0.133	−0.227	−0.539	0.034
NH_4	0.327	0.762	−0.117	−0.092	0.068
PO_4	0.144	0.760	0.075	0.106	−0.067
TP	0.008	0.729	0.437	0.271	0.123
BOD5	0.049	0.027	0.792	−0.044	0.140
CHL	0.184	0.210	0.641	0.423	0.252
pH	−0.123	0.143	0.606	−0.116	−0.253
DO	0.158	−0.149	0.561	−0.545	−0.178
TEMP	0.010	0.080	−0.076	0.837	−0.174
SS	−0.043	0.106	−0.016	−0.078	0.849
TURB	0.512	−0.109	0.044	−0.062	0.643
Eigenvalues	3.866	1.941	1.759	1.165	1.020
% of variance	27.612	13.863	12.564	8.320	7.288
Cumulative %	27.612	41.475	54.038	62.359	69.646

Extraction method: principal component analysis. Rotation method: varimax with Kaiser normalisation.

Similar to the pattern observed in the FA for the 1988–1989 dataset, nutrient processes, phytoplankton primary production, microbial degradation and the external environmental factor patterns clearly appeared in this dataset also. But, as shown in Table 3, only TN and NO_3

dominated the first varifactor, whereas the other nutrient variables have much lower loadings in this varifactor. This could be because of the fact that, in recent years, nitrogen and phosphorus nutrients in Tolo Harbour displayed a gradual decline and almost reached their lowest levels in ten years (EPD 2003). Thus, for this dataset, the effect of nutrient processes is relatively less, being subdivided into the first two varifactors. The third varifactor, consisting of BOD5, CHL, pH and DO represented the hydro-biological processes, similar to the pattern observed in the second varifactor for the 1988–1989 dataset, but accounting for a much lesser percentage of variance (12.56%), as compared to 20.18% in the first dataset, indicating a lesser presence of algal biomass. In spite of the nutrients exhibiting a gradual decrease in the harbour in recent years, this FA reveals that the nutrient processes are still dominant. Perhaps the nitrogen and phosphorus nutrient concentrations are still at adequate levels to meet their demand for phytoplankton growth. This could explain why, despite the decrease of nutrient concentrations, the chlorophyll-a in Tolo Harbour has remained relatively stable in recent years, as reported by EPD (2003). The last two varifactors seem to represent the external environmental factors, as did the last two varifactors for the 1988–1989 dataset.

The nutrient levels in Tolo Harbour, because of the vigorous pollution control and nutrient removal measures (through the implementation of the THAP and the THEES) have shown a gradual decline in the recent years. But, in spite of this decline, this study indicates an adequate supply of nutrients. Significant reduction of pollutant loadings from point sources has been achieved; still, the presence of nutrients in the Tolo Harbour indicates that the nutrients necessary for algal blooms are not just from external sources, but also from internal sources, as observed by Chau (2002). Investigators have found that a large amount of nutrients discharged into natural aquatic ecosystems can accumulate in sediments in organic and inorganic forms, and they can be released into the water under some environmental conditions (Evans 2001). Thus, as suggested by Xu *et al.* (2004), steps for eliminating internal pollutant loadings from sediments have also to be undertaken, along with the efforts to control the pollutant loadings from various point sources.

CONCLUSIONS

In this study, ecological and related water quality data taken over different time periods from seven monitoring stations in Tolo Harbour are analysed by descriptive DM techniques. The results from the analysis of box plots reveal pronounced spatial and temporal patterns and the heterogeneity of the parameters studied. The studies of spatial heterogeneity showed that, out of the three monitoring stations in the Harbour Subzone, TM2 is the most susceptible to eutrophication. Its nearly landlocked location leads to higher nutrient concentrations, weaker flushing and consequent higher algal biomass. The factor analysis indicates nutrient and hydro-biological processes to be most dominant and the external environmental factors seem to be relatively less dominant. The temporal analysis using box plots confirmed the fact that the level of nutrients in Tolo Harbour has shown a significant decline in recent years. But, in spite of this decline, it is revealed in the factor analysis that the nutrient processes play an important role even in recent years, suggesting an adequate supply of nutrients for phytoplankton growth. Since nutrients from external sources like pollutant loadings from point sources have been significantly reduced, it seems that they are being released from nutrients accumulated in the sediments. It is therefore proposed that, along with steps to control pollution loadings from external sources, it is necessary to undertake steps to control pollutant loadings from internal sources also.

ACKNOWLEDGEMENTS

This research was supported by the Research Grants Council of Hong Kong (PolyU5132/04E). The authors also wish to thank the Environmental Protection Department for providing the water quality data used in this study.

REFERENCES

Bobbin, J. & Recknagel, F. 2001 Knowledge discovery for prediction and explanation of blue–green algal dynamics in lakes by evolutionary algorithms. *Ecol. Model.* **146**, 253–262.

Boogaard, venden H. F. P. 1998 Self organization feature maps for analysis of hydrological and ecological data sets. In *Hydroinformatics 98'*. A.A. Balkema, Rotterdam.

Brosse, S., Giraudel, J. L. & Lek, S. 2001 Utilisation of non-supervised neural networks and principal component analysis to study fish assemblages. *Ecol. Model.* **146**, 159–166.

Bui, E. N., Henderson, B. L. & Viergever, K. 2006 Knowledge discovery from models of soil properties developed through data mining. *Ecol. Model.* **191** (3–4), 431–446.

Chambers, J., Cleveland, W., Kleiner, B. & Tukey, P. 1983 *Graphical Methods for Data Analysis*. Wadsworth.

Chau, K. W. 2002 Field measurements of SOD and sediment nutrient fluxes in a land-locked embayment in Hong Kong. *Adv. Environ. Res.* **6** (2), 135–142.

Chau, K. W. 2004 A three-dimensional eutrophication modeling in Tolo Harbour. *Appl. Math. Model.* **28** (9), 849–861.

Chau, K. W. 2005 A split-step PSO algorithm in prediction of water quality pollution. *Lect. Notes Comput. Sci.* **3498**, 1034–1039.

Chau, K. W. & Jin, H. S. 1998 Eutrophication model for a coastal bay in Hong Kong. *J. Environ. Eng. ASCE* **124** (7), 628–638.

Chen, Q. & Mynett, A. E. 2003 Integration of data mining techniques and heuristic knowledge in fuzzy logic modelling of eutrophication in Taihu Lake. *Ecol. Model.* **162**, 55–67.

Chen, Q. & Mynett, A. E. 2005 Self organization feature maps to analyse eutrophication in Taihu Lake. In *IAHR, Seoul, Korea – Spatial Dynamics*.

EPD 1986–2003 *Marine Water Quality in Hong Kong*. Annual reports published by Environmental Protection Department, Government of Hong Kong Special Administrative Region.

Evans, R. D. 2001 Interactions between sediments and water: summary of the Eighth International Symposium. *Sci. Tot. Environ.* **266** (1–3), 1–5.

Fayyad, U. 1997 Editorial. *Data Min. Knowl. Discov.* **1**, 5–10.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. 1996 From data mining to knowledge discovery: an overview. In *Advances in Knowledge Discovery and Data Mining*, (ed. U. Fayyad *et al.*), MIT Press, Boston, MA. pp. 1–34.

Hall, L. W. & Anderson, R. D. 1999 A deterministic ecological risk assessment for copper in European saltwater environments. *Mar. Pollut. Bull.* **38** (3), 207–218.

Hardle, W. & Simar, L. 2003 *Applied Multivariate Statistical Analysis*. Springer, Berlin.

Jeong, K. S., Joo, G. J., Kim, H. W., Ha, K. & Recknagel, F. 2001 Prediction and elucidation of algal dynamics in the Nakdong River (Korea) by means of a recurrent artificial neural network. *Ecol. Model.* **146**, 115–129.

Jeong, K. S., Kim, D. K., Whigham, P. & Joo, G. J. 2003 Modelling Microcystis aeruginosa bloom dynamics in the Nakdong River by means of evolutionary computation and statistical approach. *Ecol. Model.* **161**, 67–78.

Jolliffe, I. T. 1986 *Principal Component Analysis*. Springer. Berlin.

Kaiser, H. F. 1958 The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **23**, 187–200.

Karul, C., Soyupak, S., Cilesiz, A. F., Akbay, N. & Germen, E. 2000 Case studies on the use of neural networks in eutrophication modelling. *Ecol. Model.* **134**, 145–152.

Lee, J. H. W., Huang, Y., Dickman, M. & Jayawardena, A. W. 2003 Neural network modelling of coastal algal blooms. *Ecol. Model.* **159**, 179–201.

Lee, J. H. W., Wu, R. S. S., Cheung, Y. K. & Wong, P. P. S. 1991a Dissolved oxygen variations in marine fish culture zone. *J. Environ. Eng.* **117** (6), 799–815.

Lee, J. H. W., Wu, R. S. S. & Cheung, Y. K. 1991b Forecasting of dissolved oxygen in marine fish culture zone. *J. Environ. Eng.* **117** (6), 816–833.

Maier, H. R., Dandy, G. C. & Burch, M. D. 1998 Use of artificial neural networks for modelling cyanobacteria *Anabaena* spp. in the River Murray, South Australia. *Ecol. Model.* **105**, 257–272.

Maier, H. R., Sayed, T. & Lence, B. J. 2001 Forecasting cyanobacterium Anabaena spp. in the River Murray, South Australia, using B-spline neurofuzzy models. *Ecol. Model.* **146**, 85–96.

Morton, B. 1988 Editorial: Hong Kong's first marine disaster. *Mar. Pollut. Bull.* **19**, 299–300.

Muttil, N. & Chau, K. W. 2007 Machine learning paradigms for selecting ecologically significant input variables. *Eng. Appl. Artif. Intell.* **20** (6), 735–744.

Muttil, N. & Lee, J. H. W. 2005 Genetic programming for analysis and real-time prediction of coastal algal blooms. *Ecol. Model.* **189** (3–4), 363–376.

Petersen, W., Bertino, L., Callies, U. & Zorita, E. 2001 Process identification by principal component analysis of river water-quality data. *Ecol. Model.* **138**, 193–213.

Recknagel, F. 2001 Applications of machine learning to ecological modelling. *Ecol. Model.* **146**, 303–310.

Recknagel, F., Bobbin, J., Whigham, P. & Wilson, H. 2002 Comparative application of artificial neural networks and genetic algorithms for multivariate time-series modelling of algal blooms in freshwater lakes. *J. Hydroinf.* **4** (2), 125–134.

Recknagel, F., French, M., Harkonen, P. & Yabunaka, K. 1997 Artificial neural network approach for modelling and prediction of algal blooms. *Ecol. Model.* **96**, 11–28.

Russom, C. L. 2002 Mining environmental toxicology information: web resources. *Toxicology* **173** (1–2), 75–88.

Scardi, M. 2001 Advances in neural network modelling of phytoplankton primary production. *Ecol. Model.* **146**, 33–45.

Scardi, M. & Harding, L. W. 1999 Developing an empirical model of phytoplankton primary production: a neural network case study. *Ecol. Model.* **120**, 213–223.

Sin, Y. S. & Chau, K. W. 1992 Eutrophication studies on Tolo Harbour, Hong Kong. *Water Sci. Technol.* **26** (9–11), 2551–2554.

Stockwell, D. R. B. 2006 Improving ecological niche models by data mining large environmental datasets for surrogate models. *Ecol. Model.* **192** (1–2), 188–196.

Su, F., Zhou, C., Lyne, V., Du, Y. & Shi, W. 2004 A data-mining approach to determine the spatio-temporal relationship between environmental factors and fish distribution. *Ecol. Model.* **174** (4), 421–431.

Tadesse, T., Brown, J. F. & Hayes, M. J. 2005 A new approach for predicting drought-related vegetation stress: integrating satellite, climate, and biophysical data over the US central plains. *ISPRS J. Photogramm. Remote Sensing* **59** (4), 244–253.

Thomann, R. V. & Mueller, J. A. 1987 *Principles of Surface Water Quality Modeling and Control*. Harper and Row, New York.

Thuraisingham, B. 1999 *Data Mining: Technologies, Techniques, Tools, and Trends*. CRC Press, Boca Raton, FL.

Wei, B., Sugiura, N. & Maekawa, T. 2001 Use of artificial neural network in the prediction of algal blooms. *Water Res.* **35** (8), 2022–2028.

Xu, F. L., Lam, K. C., Zhao, Z. Y., Zhan, W., David Chen, Y. & Tao, S. 2004 Marine coastal ecosystem health assessment: a case study of the Tolo Harbour, Hong Kong, China. *Ecol. Model.* **173** (4), 355–370.

Yabunaka, K., Hosomi, M. & Murakami, A. 1997 Novel application of a back-propagation artificial neural network model formulated to predict algal bloom. *Water Sci. Technol.* **36** (5), 89–97.