Revealing Patterns in Coastal Water Quality Data using Statistical Analysis

Muttil, N. [1] and K.W. Chau [2]

[1] School of Architectural, Civil and Mechanical Engineering and Institute for Sustainability & Innovation, Victoria University, P.O. Box 14428, Melbourne 8001, Vic., Australia
[2] Department of Civil and Structural Engineering, Hong Kong Polytechnic University, Hung Hom, Hong Kong

Keywords: *Harmful algal blooms, data mining; box plots, multivariate statistical analysis, factor analysis, water quality modelling, Hong Kong*

EXTENDED ABSTRACT

A major impact of eutrophication is the stimulation of algal growth and the production of harmful algal blooms (HABs). HABs can have profound negative effects on the environment, which include severe dissolved oxygen depletion, fish kills, discoloration of marine water, beach closures, etc. Owing to the extremely complicated ecological processes, previous research efforts indicate that it is far from accurately unravelling the causality and dynamics of HABs and predicting their occurrence with acceptable accuracy and lead-time. Thus, techniques to better understand the complex interrelated processes in the ecosystem are needed.

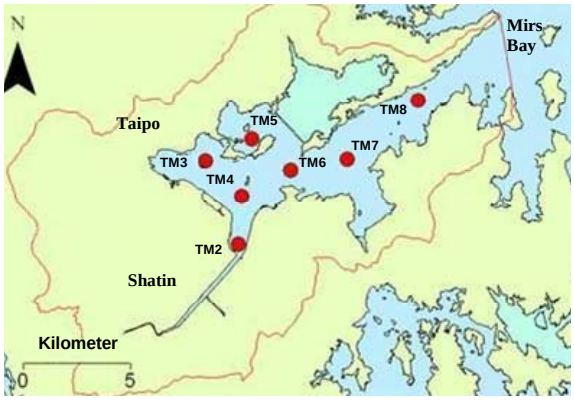

Figure 1. The study site: Tolo Harbour indicating the seven monitoring stations

In this study, visual 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 (Figure 1). To study spatial patterns, data from the seven monitoring stations (shown in Figure 1) are used. To study temporal patterns, two datasets from different periods were used: one when the water quality in Tolo Harbour was at its worst (in the late eighties) and the second dataset being more recent, when the water quality had improved significantly.

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 brings to light the dominant ecological processes in the coastal marine environment. It indicates nutrient processes and the hydro-biological processes to be most dominant and the external environmental factors are indicated to be relatively less dominant.

The temporal analysis using box plots confirms 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 the recent years, suggesting 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.

This study demonstrates the use of data mining techniques to exhibit prominent spatial and temporal patterns and to reveal dominant variables governing key ecological processes and thus provide further insight into the HAB dynamics.

1. INTRODUCTION

A major impact of eutrophication is the stimulation of algal growth and the production of harmful algal blooms (HABs). Algal bloom is the phenomenon of the rapid increase (or blooming) in the number of microscopic algae (tiny marine plants of the class of phytoplankton) to an abnormally high concentration. HABs can have profound negative effects on the environment, which include severe dissolved oxygen depletion, fish kills, discoloration of marine water, beach closures, etc. Thus, better understanding of the complex ecological processes and HAB dynamics is of utmost importance.

Research on HABs has been conducted over 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. In spite of this, the causality and dynamics of algal blooms are not well-understood and the prediction of algal blooms remains a very difficult problem, owing to the extremely complicated ecological dynamics.

In recent years, with the availability of sophisticated personal computers with ever-expanding capabilities, there is a growing tendency to use data mining (DM) techniques to complement or even replace process-based models. These include artificial neural networks (Recknagel et al., 2002; Lee et al., 2003) and evolutionary based techniques (Recknagel et al., 2002; Muttil and Lee, 2005).

In this study, we use descriptive DM techniques for revealing the spatial and temporal ecological dynamics of the coastal waters of Hong Kong. Visual 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 is followed by the analysis of box plots and factor analysis of the ecological and related water quality data.

2. DATA MINING METHODOLOGIES

DM is concerned with extracting useful information from databases. Statistics is a major area contributing to DM and various statistical analysis softwares are now being marketed as data mining tools. Thus, the main part of data mining 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. The techniques employed in this study are described briefly in the following sub-sections.

2.1. Visual data mining using box plots

Visual data mining refers to the visual presentation of data to extract useful information. The use of visualisation techniques allows users to summarize, 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. In a box plot, the line across the box represents the median, the bottom of the box is at the first quartile (Q_1) and the top is at the third quartile (Q_3). 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 provide an excellent visual summary of a set of data and are especially useful when comparing two or more sets of data.

2.2. 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". Each variable is assumed to be dependent on a linear combination of the 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.

In this study, the adopted factor extraction method is the most commonly used principal component extraction method. The eigenvalues of 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 parameterization 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. A factor with an eigenvalue greater than or equal to one is usually considered as being of statistical significance (the Kaiser criterion).

3. DATA AND MODELLING APPROACH

The selected DM techniques are applied to water quality data from Tolo Harbour (see Figure 1), which are measured under the 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 sub-sections provide further details of the study site and the data used.

3.1. The study site

Tolo Harbour is a semi-enclosed bay connected to the open sea at Mirs Bay (shown in Figure 1) 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. 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. 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 landfill, enforcing the Water Pollution Control Ordinance and building sewer networks in rural areas (EPD, 2001). 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 1) 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.

3.2. 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 the seven water quality monitoring stations in Tolo Harbour.

Table 1. List of water quality variables

Variable Name	Symbol	Units
Nutrients:		
Ammonia Nitrogen	NH4	mg/L
Nitrate Nitrogen	NO3	mg/L
Total Nitrogen	TN	mg/L
Orthophosphate	PO4	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-day Biochemical Oxygen Demand	BOD5	mg/L
Biological indicator:		
Chlorophyll-a	CHL	μg/L

In order to study the temporal patterns, datasets from 2 different time periods were selected. The first dataset is from a period when the water quality in Tolo Harbour was at its worst, which is the late eighties. During this period, the HAB incidents increased significantly and reached a peak in 1988, when a total of 43 incidents were reported. To represent this time period, 2 years of data from 1988-89 is 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 2001-03 is selected as the second dataset.

4. MINING WATER QUALITY DATA

4.1. 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. Please note that due to a limit on number of pages, box plots for only 8 variables are presented here. 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 1988-89 and from 2001-03.

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 2.

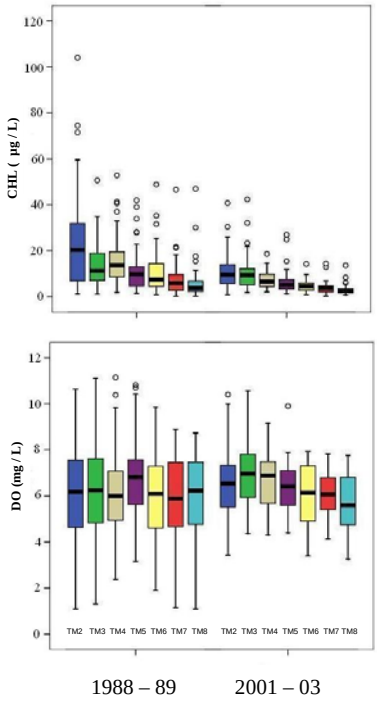

Figure 2. Box plots for CHL and DO

From the box plot for CHL, it is observed that the spread of the box plots gradually decreased from the TM2 station to the TM8 for both the periods. During 1988-89, the CHL values were much higher at TM2 than the other stations. 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 1988-89 are above 20 µg/L. Typically, CHL concentrations exceeding 20 µg/L would be considered to constitute an algal bloom. The box plots for 2001-03 indicate lower CHL values, indicating a reduction in algal biomass concentration. The box plots of DO over the period 1988-89 have much longer bottom whiskers, indicating much lower DO values. Higher DO values during 2001-03 indicate an improvement in water quality. As far as the spatial variation is concerned, DO does not seem to indicate any significant variation amongst the seven stations.

Box plot of nutrients and BOD5

The box plots for the nutrients and BOD5 are presented in Figure 3. With the exception of NO3 and PO4, all the remaining nutrients showed gradual decrease in concentration from TM2 to TM8. For all the nutrients and especially for NO3 and PO4, the values at TM2 during 1988-89 were exceptionally high. This high concentration of nutrients clearly indicates the poor water quality in Tolo harbour during the period of 1988-89.

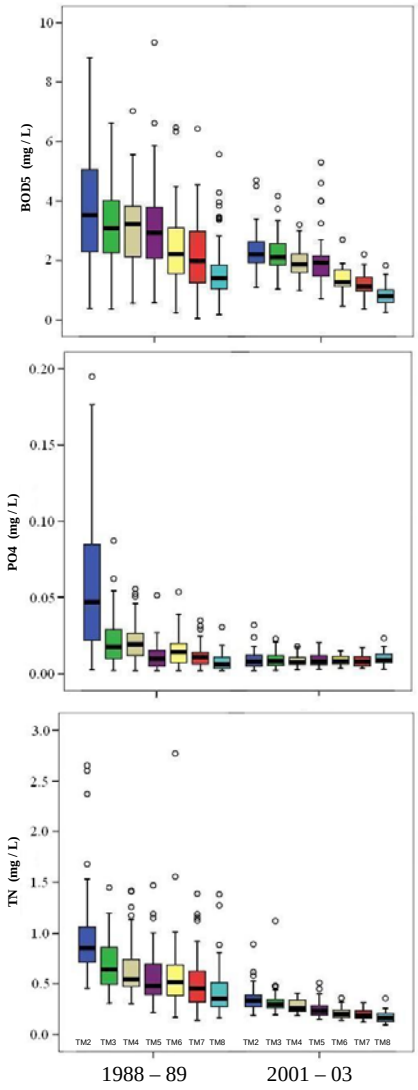

Figure 3. Box plots for nutrients and BOD5

As indicated by the box plots for 2001-03, the water quality during this period has shown significant improvement. The BOD5 also showed high values at TM2 during 1988-89, whereas for both time periods, the demand 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 4.

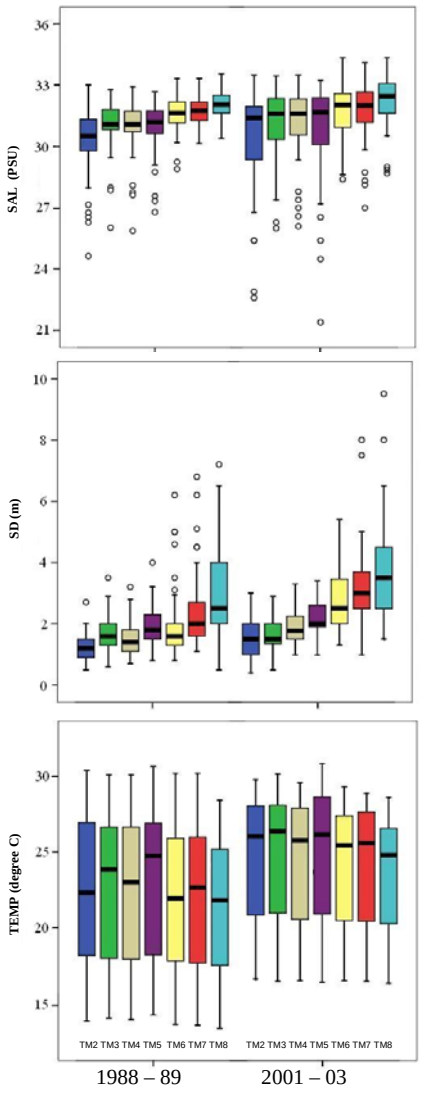

Figure 4. Box plots for the physical properties

As expected, SAL showed gradual increase in concentration as we move closer to the sea, from TM2 to TM8 station. As far as the temporal behaviour of SAL is concerned, it showed reduced values in 2001-03, 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 improvement in water quality in 2001-03, the SD values also showed 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 significant increase in the period 2001-2003, as compared to the values in 1988-89. As with TURB, TEMP values also showed significant increase in 2001-03, shown by an upward shift in the boxes and the median by about 3 ˚C. The minimum water temperature in 1988-89 can be seen in the box plots to be around 13 ˚C, whereas in 2001-03, it was about 16 ˚C. This is undesirable because increases in water temperature cause 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 lesser values in 2001-03, indicating that the water was more alkaline during the period 1988-89.

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 3 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 consequent highest concentration of nutrients. The analysis of the temporal variation between the 2 time periods clearly indicated a significant improvement in water quality in 2001-03, as compared to that about 10 years earlier.

4.2. Factor analysis

The FA was applied to the 2 datasets considered in this study, namely the 1988-89 and the 2001-03 data. In order to isolate the ecological processes from the hydrodynamic effects as much as possible, using the data from TM2 monitoring station (which was found to have the worst water quality) for FA seems to be most appropriate. But, for the 2001-03 dataset, only monthly values of the water quality variables are available, resulting in only 36 data records in this dataset. Thus, to increase the number of data records, data from the 3 monitoring stations in the Harbour Subzone (TM2, TM3 and TM4) were combined, resulting in (36 x 3) 108 data records for this dataset.

Factor analysis for 1988-89 dataset

Table 2 shows the factor loadings obtained from the principal components factor analysis with varimax rotation. The first four varifactors accounting for 68.67% 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.

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 11.53% and 8.43% of the variance respectively. The first varifactor, with an eigenvalue of 3.99, explained 28.53% of the total variance. It is clearly dominated by the nutrients, PO4, NO3, NH4, 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 eighties, 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.

Table 2. Factor loadings from a principle component factor analysis for the 1988-89 data

Variable	Varifactors			
	1	2	3	4
PO4	**0.871**	0.050	0.080	0.054
NO3	**0.828**	0.004	-0.018	-0.028
NH4	**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**
Eigenvalue	3.994	2.826	1.614	1.180
% Variance	28.529	20.187	11.527	8.427
Cumulative	28.529	48.715	60.243	68.669

The second varifactor, with an eigenvalue of 2.83 and accounting for 20.19% of the total variance was also significant. The variables DO, BOD5 and CHL exhibited high positive loadings, pH showed low-moderate positive loading and SAL exhibited a low moderate negative loading. This varifactor seems to represent the hydro-biological 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 one hand and oxygen consumption due to its respiration and death (microbial degradation) on the other hand. 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. Simultaneous favourable presence of other external environmental factors like the water temperature (TEMP), the degree of penetration of sunlight into the water column (SD and TURB), etc., lead to the optimal conditions for the algae to bloom. The third and fourth varifactors seem to indicate the importance of these environmental factors.

Factor analysis for 2001-03 dataset

The factor loadings for this dataset are presented in Table 3. For this dataset, the first five varifactors are retained, which account for 69.65% of the total variance. It is observed that the first three varifactors together accounted for 54.04% of the total variance, whereas the remaining two varifactors were less dominant and together accounted for 15.61%.

Table 3. Factor loadings from a principle component factor analysis for the 2001-03 dataset

Variable	Varifactors				
	1	2	3	4	5
NO3	**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
NH4	0.327	**0.762**	-0.117	-0.092	0.068
PO4	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
% Variance	27.612	13.86	12.564	8.320	7.288
Cumulative	27.612	41.47	54.038	62.359	69.646

Similar to the pattern observed in the FA for the 1988-89 dataset, nutrient processes, phytoplankton primary production, microbial degradation and the external environmental factor patterns clearly appeared in this dataset also. But, only TN and NO3 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 several years, nitrogen and phosphorus nutrients in Tolo Harbour displayed a gradual decline and almost reached their lowest levels in ten years (EPD, 2001). 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 1988-89 dataset, but accounting for a much lesser percentage of variance (12.56%), 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 in adequate levels to meet their demand for phytoplankton growth. This could explain why, despite the decrease of nutrient concentrations, the chlorophyll-a in the Tolo Harbour has remained relatively stable in recent years, as reported by EPD (2001). Again, the last 2 varifactors seem to represent the external environmental factors.

The THAP and THEES resulted in significant reduction of pollutant loadings from point sources; 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, 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.

5. CONCLUSIONS

In this study, ecological and related water quality data taken over different time periods from seven monitoring stations in Tolo Harbour are analysed using DM techniques. 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. The factor analysis indicates nutrient and hydro-biological processes to be most dominant and the external environmental factors seem to be relatively less dominant. In spite of the decline of nutrients in recent years, it is revealed in the factor analysis that the nutrient processes play an important role even in the recent years, suggesting adequate supply of nutrients for phytoplankton growth. 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.

6. REFERENCES

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

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, *Science of the Total Environment*, 266 (1-3), 1-5.

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

Lee, J. H. W., Y. Huang, M. Dickman and A.W. Jayawardena (2003), Neural network modelling of coastal algal blooms. *Ecological Modelling*, 159, 179-201.

Morton, B. (1988), Editorial: Hong Kong's first marine disaster. *Marine Pollution Bulletin*, 19, 299-300.

Muttil, N., and J.H.W. Lee (2005), Genetic programming for analysis and real-time prediction of coastal algal blooms. *Ecological Modelling*, 189, 363-376.

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