Water quality assessment and apportionment source of pollution from neighbouring rivers in the Tapeng Lagoon (Taiwan) using multivariate analysis: a case study

S.-W. Liao, J.-Y. Sheu, J.-J. Chen and C.-G. Lee

Department of Environmental Engineering and Science, Tajen University, Pingtung, 10907, Chinese Taiwan (E-mail: *swliao@mail.tajen.edu.tw*)

Abstract Factor analysis was conducted to explain the characteristics and variation in the quality of water during the disassembly of oyster frames and fishery boxes. The result shows that the most important latent factors in the Tapeng Lagoon are the ocean factor, the primary productivity factor, and the fishery pollution factor. Canonical discriminant analysis is applied to identify the source of pollution in neighbouring rivers outside the Tapeng Lagoon. The two constructed discriminant functions (CDFs) showed a marked contribution to all the discriminant variables, and that total nitrogen, algae, dissolved oxygen, and total phosphate combined in the nutrient effect factor. The recognition capacities in these two CDFs were 95.6% and 4.4%, respectively. The water quality in the Kaoping river most strongly affected the water quality in the Tapeng Lagoon. Disassembling the oyster frames and fishery boxes improved the water quality markedly. However, environmental topographic conditions indicate that strengthening stream pollution prevention and constructing another entrance to the ocean are the best approaches for improving the quality of water in the Tapeng Lagoon by reducing eutrophication. These approaches and results yield useful information concerning habitat recovery and water resource management.

Keywords Canonical discriminant analysis; factor analysis; Tapeng Lagoon; water quality

Water Science & Technology **Vol 54 No 11–12 pp 47–55** © IWA Publishing 2006

Introduction

In this work, the data in the large database obtained during the development of disassembling the oyster frames and fishery boxes in the Tapeng Lagoon and in the neighbouring rivers were analysed using various multivariate statistical approaches to extract information on the similarities or dissimilarities among the sampling sites, the identification of the water quality variables that cause spatial and temporal variations in the quality of lagoon water, the hidden factors that govern the structure of the database, the effect of the neighbouring rivers on the water quality parameters, and the identification of sources of pollution from neighbouring rivers. In conclusion, the quality of water in the rivers and lagoon was analysed. Secondly, factor scores were determined from a factor analysis of water variables, which were used to determine spatial and temporal variations of water quality in the Tapeng Lagoon. Thirdly, canonical discriminant analysis was conducted to construct the functions of monitoring parameters and to estimate the contribution of possible sources to the concentration of the parameters that are associated with neighbouring rivers. The aim of this work is to perform multivariate analysis and demonstrate its applicability and effectiveness in environment research. This is the first study in Taiwan that takes such an approach, and the results may be useful in developing a methodology for use by the government in refining its management programmes.

doi: 10.2166/wst.2006.835

Materials and methods

Site description

The Tapeng Lagoon is a semi-enclosed coastal lagoon, which has only one tidal inlet through which lagoon water is exchanged with the coastal currents along the Kaoping coast on the narrow shelf in the southwest of Taiwan. It has a total area of around 532 hectares. The area is enclosed by the Taiwan Strait to the west, the central mountains to the east, the Kaoping river and the Tungkang river to the north, and the Lingbeng river to the south. Therefore, seasonal variations affect the bay. Much aquaculture, including oyster farming and fish cages, is conducted in the bay area, because the water is calm and rich. However, all of the fishery facilities were disassembled when the Tapeng Lagoon National Scenic Area was established in 2003. Therefore, the balance between nature and anthropogenic disturbance was broken.

Sampling and analysing water

Water samples were collected in July and August in 2002 and 2004 at all 13 sites to compare the changes caused by deconstructing the fishery facilities in the Tapeng Lagoon. Grid sampling was designed to represent the quality of the lagoon water. A total of 21sites (seven in each river) were sampled outside the lagoon in July and August 2004 in the three neighbouring rivers to identify the source that most worsened the quality of water in the Tapeng Lagoon (Figure 1). Sampling, preservation, and transportation of the water samples to the laboratory followed *Standard Methods* (1995). The temperature, electrical conductivity, and pH of the water were measured on-site using a mercury thermometer. All other parameters were determined in a laboratory, following standard protocols (*Standard Methods*, 1995). The samples were analysed to evaluate the 14 parameters - T (temperature), T-Alk (total alkalinity), Cl^- (concentration of chloride), EC (electrical conductivity), TKN (total concentration of kjeldahl nitrogen), PO_4^{2-} (concentration of phosphate), pH, DO (concentration of dissolved oxygen), HPC (heterotrophic plate count), UV-254, T. coli (total concentration of coliform), algae (number of algae), chloro-a (concentration of chlorophyl-a), SO_4^{2-} (concentration of sulfate) and TS (total solids).

Statistical methods

Environmental monitoring typically generates many data that are difficult to analyse and interpret because their relationships among the variables are complex. Multivariate approaches have been used successfully to support the interpretation of complex field measurements and extract meaningful information from such databases (Ruiz and Blasco, 1990). Factor analysis (FA) can determine the most important factors that contribute to the structure of data (Jenerette *et al.*, 2002). Canonical discriminant analysis (CDA) is used to interpret the spatial distribution of a bioassemblage given various environmental parameters (Huberty, 1994; Momen and Zehr, 1998; Shin and Fong, 1999).

Factor analysis is an approach that explains the observed relations among many variables in terms of simpler relationships to offer insight into the structure that underlies the variables. It is a powerful approach for recognizing pattern, which aims to explain the variance of a large set of inter-correlated variables and transform them into a smaller set of independent (uncorrelated) variables (Sharma, 1996). CDA determines how a set of quantitative variables may differentiate among many known classes. CDA yields linear functions of quantitative variables that maximally separate two or more groups of individuals, while minimizing variation within groups. This approach distinguishes uncorrelated canonical discriminant functions (CDFs), which are the linear combinations of the original variables that most strongly separate the averages of the groups of observations (Rencher, 1992).

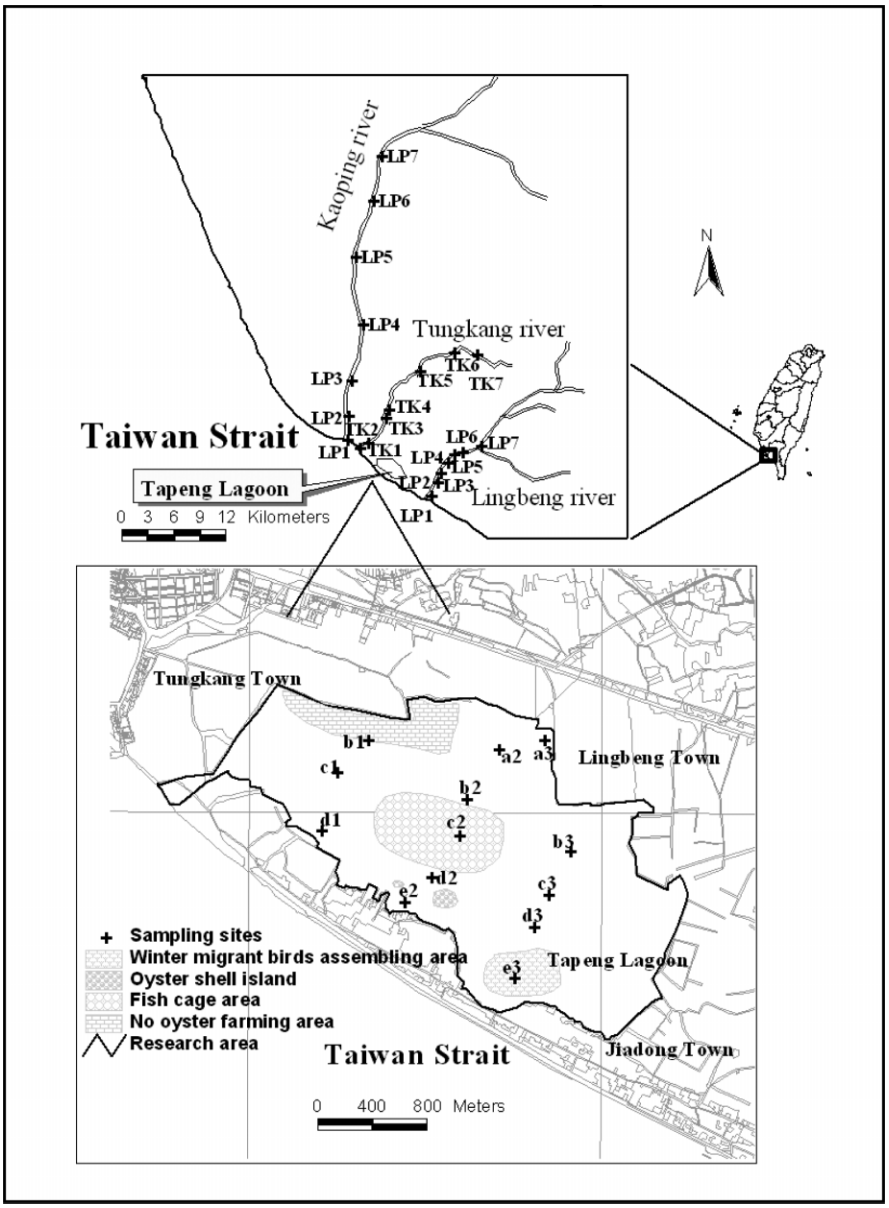


Figure 1 Water sampling sites in the Tapeng Lagoon and neighbouring rivers

Results and discussion

Factor analysis

The first latent factor explained 38.21% of the total variation of the water variables (Table 1), indicating that T, Cl, EC, and TKN represent the underlying dimensions that account for the correlation among the variables. The temperature of the seawater was always steady. Total nitrogen represents nutrition, and can influence the production of phytoplankton. Therefore, the first latent factor was called 'the ocean current and primary production factor'. Meanwhile, Cl and EC had high positive factor loadings with the first latent factor (not shown in the text), indicating that it is strongly positively correlated with the first factor. T and TKN exhibited high negative factor loadings with the first latent factor. Factor scores are calculated to determine the level of pollution. In Figure 2(a),

Table 1 Factor loadings of varimax-rotated factor matrix for 2002 and 2004 samples

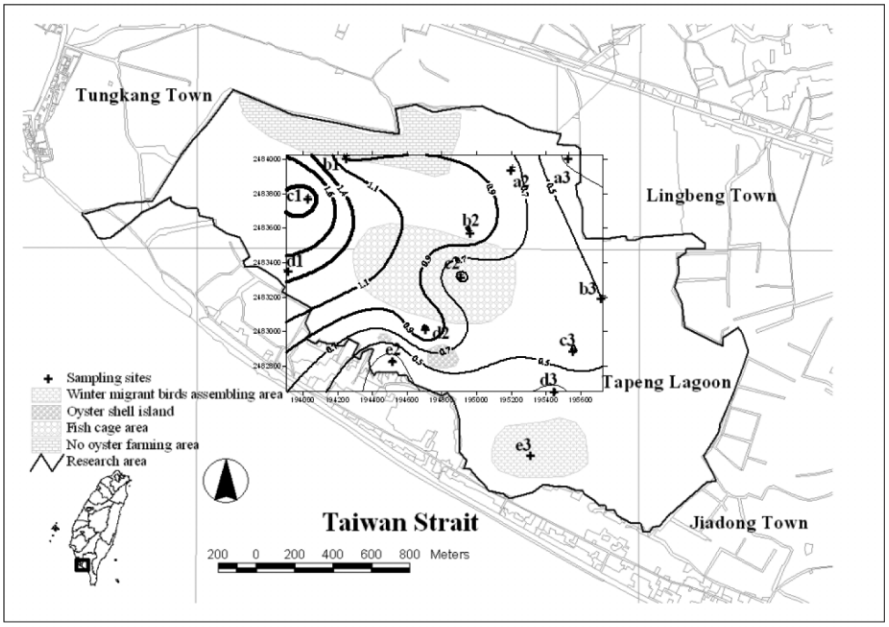
Variables	2002			2004	
	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2
T	−0.8130	−0.1707	−0.3650	−0.8064	0.5282
Alk	0.1455	0.1432	0.6125	0.9505	0.1089
Cl	0.8184	0.1589	0.0219	−0.9285	−0.1084
EC	0.7942	0.2719	0.3140	0.0129	−0.9546
TKN	−0.6681	0.3921	−0.0546	−0.5264	0.8035
TP	−0.2092	0.0504	0.7440	−0.2230	0.9388
UV-254	−0.0912	−0.8657	−0.0310	0.9160	−0.3336
pH	0.5118	0.1877	0.5452	−0.8337	0.1500
HPC	−0.1163	−0.9209	−0.0991	−0.6917	0.6193
DO	0.2604	0.0086	0.7159	0.6991	−0.1325
Algae	0.4898	−0.1603	0.7063	0.2653	−0.0720
Chloro-a	−0.3278	0.0017	−0.4578	0.0682	0.7464
Eigenvalue	4.59	1.84	1.4	6.51	2.66
Total variance	0.3821	0.1532	0.117	0.5427	0.2204

the bold line and the fine line represent the highest and lowest concentrations of contaminants, respectively. Figure 2(a) shows that the Cl content and EC were highest at the entrance of the lagoon and were associated with the highest ocean exchange capacity. The TKN content was highest at the sites in the southeast of the lagoon, indicating that this location may have suffered from the low ocean exchange capacity.

The second latent factor was responsible for 15.32% of the total variation of water variables and exhibited high negative factor loadings for UV-254 and HPC (Table 1). UV-254 represents the water organic content. HPC represents the aerobic and anaerobic microbiological content, denoting the level of water pollution; many fishery activities exist in the samples. Therefore, the second latent factor was called the 'fishery pollution causes factor'. Figure 2(b) shows that the areas with the highest concentrations of UV-254 and HPC were at the lakeside, which finding coincided with the laboratory data because ocean exchange capacity was low. The third factor explained 11.7% of the total variation in the water variables and exhibited high positive factor loadings for TP, DO, and algae. TP is nutritious compound and a food for phytoplankton and algae. The amount of algae strongly influenced the DO content. Therefore, the third latent factor was called the 'primary productivity influence factor'. The concentrations of TP and algae were highest to the northeast and southeast of the lake. Weak ocean exchange capacity caused nutritious materials to accumulate and many algae to breed. More oxygen was dissolved during the day because of photosynthesis.

The water data (26×12) sampled in 2004 showed that the first two factors explained around 76.31% of the total variance of the samples. The first latent factor explained 54.27% of the total variation in the water variables (Table 1). Therefore, T, Alk, Cl, UV-254, pH, HPC, and DO explained the correlation among the variables. By analogy, the first latent factor was 'the ocean factor and the fishery pollution factor'. It shows that the pollution level was highest at the southern, southeast, and eastern sides of the lake. The second factor explained 22% of the total variation in the water variables and exhibited high positive factor loadings on TKN, TP, and chlorophyl-a, and had a high negative factor loading for EC. The second latent factor was called, 'the ocean factor and the fishery pollution factor'. It shows that nutrition content was highest at the central and southern sides of the lagoon. The drop in nutrition caused the TKN and TP content in 2004 to be lower than that in 2002 after the fishery facilities were disassembled. This showed that water quality had improved in 2004, but eutrophication still occurred at the end of the bag-shaped lagoon.

(a) Distribution of scores for factor 1 in 2002



(b) Distribution of scores for factor 2 in 2002

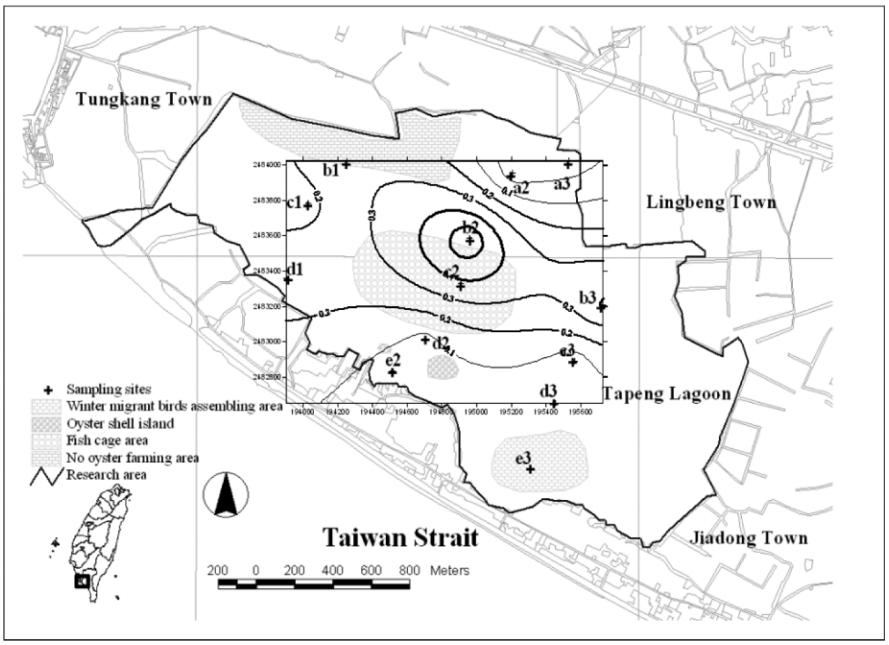


Figure 2 Distribution of factor scores for 2002

Discriminant analysis

Total standardized canonical coefficients (TSCC) specify the joint effects of independent variables of a given CDF, so are more informative than the total canonical structure coefficients (TCSC) (Rencher, 1992). TSCCs can be misleading when independent variables are related (Cruz-Castillo *et al.*, 1994). In this work, the TCSCs were used to interpret the CDFs because significant correlations were obtained among some of the independent variables (not shown in the text), and the TSCCs were used to yield the CDFs.

A forward stepwise approach was applied to determine which variables could be incorporated in the model (StatSoft, 1996). Therefore, an *F*-test was conducted to identify the most discriminating variables. The process was terminated when the differences ceased to be significant. Table 2 shows that all of these discriminant variables were significant according to Wilk's Lambda test. The order of inclusion in the model, according to the *F*-test, was – TKN, chlorophyl-a, algae, DO, HPC, EC, UV-254, TP, T, pH, Cl, TS, SO_4^{2-}, and Alk. In these models, the main factors were the ocean factor, the nutrient factor, the dissolved oxygen factor, the primary productivity factor, and the environmental pollution factor. Accordingly, these rivers vary greatly. Some parts of these rivers were polluted with domestic and industrial wastewaters but only one river was polluted with agricultural wastewater. The discriminatory capacity followed the order TKN, chlorophyl-a, algae, DO, HPC, EC, UV-254, and TP. These variables were the most important in these models. TKN and TP are nutrients that influence the large-scale formation of algae, and thus the measurements of chlorophyll-a, algae, and DO. The HPC content shows the level of pollution.

Canonical correlation coefficients exceeded 0.8 for both CDFs (Table 3). Eigenvalues also exceeded 1.0 for both CDFs. These two CDFs together explained 100% (95.6%, 4.4%, respectively) of the variance at the 21 sampling sites. The value of every discriminant variable was standardized to determine the relationship between the discriminant variables and functions. The standardized CDFs were obtained as follows.

$$\begin{aligned}\mathrm{CDF}_1 = {} & 1.59\,\mathrm{T} + 0.91\,\mathrm{Alk} - 8.37\,\mathrm{Cl} + 2.81\,\mathrm{EC} - 6.80\,\mathrm{TKN} - 1.38\,\mathrm{TP} - 1.09\,\mathrm{UV\text{-}254} \\ & + 1.30\,\mathrm{pH} - 3.42\,\mathrm{HPC} + 2.05\,\mathrm{DO} + 2.25\,\mathrm{Algae} + 9.15\,\mathrm{Chlorophyl\text{-}a} \\ & + 3.88\,\mathrm{SO_4^{2-}} - 0.56\,\mathrm{TS}\end{aligned}$$

$$\begin{aligned}\mathrm{CDF}_2 = {} & -0.40\,\mathrm{T} - 1.06\,\mathrm{Alk} - 7.83\,\mathrm{Cl} + 1.05\,\mathrm{EC} - 0.18\,\mathrm{TKN} - 2.53\,\mathrm{TP} + 1.02\,\mathrm{UV\text{-}254} \\ & + 0.53\,\mathrm{pH} - 3.33\,\mathrm{HPC} + 0.12\,\mathrm{DO} + 0.06\,\mathrm{Algae} + 6.21\,\mathrm{Chlorophyl\text{-}a} \\ & + 7.46\,\mathrm{SO_4^{2-}} - 0.88\,\mathrm{TS}\end{aligned}$$

Three river classes were distinguished using two CDFs. CDF_1 had the highest canonical coefficient, 0.99, and was defined by eight discriminant variables whose canonical coefficients had high absolute values. These were chlorophyl-a, Cl, TKN, SO_4^{2-}, HPC, EC, algae, and DO. Accordingly, CDF_1 comprised the nutrient factor, the domestic wastewater factor, and the ocean factor. The canonical structure coefficients indicate that the variance of variables in CDF_1 followed the ascending order TKN, algae, DO, and TP (Table 3). TKN and TP are nutrients that can affect the production of algae, and thus influence the DO content. Table 2 shows that the measured TKN in the Tungkang river and the Lingbeng river exceeded that in the Kaoping river. Visual reconnaissance in the field established that some segments of these two rivers suffered from livestock breeding and tillage wastewater. The DO was highest in the Lingbeng river and lowest in the Tungkang river. Sampling sites in the Lingbeng river were concentrated in the middle–lower reaches, where duck breeding farms were densely located and nutrients and algae were prevalent. TP is also an important factor in determining algae life, but it is not prevalent in these three rivers. In summary, CDF_1 is called the 'nutrient factor', based on canonical structure coefficients.

CDF_2 had a higher canonical correlation of 0.92 and was defined by five discriminant variables, which had canonical coefficients with high absolute values: Cl, SO_4^{2-}, chlorophyll-a, HPC, and TP (Table 3). CDF_2 comprised the nutrient factor, the domestic

Table 2 Outcomes of CDA determined by forward stepwise method for the three neighbouring rivers of the Tapeng Lagoon

Discriminant variables	Wilks' Lambda	F-remove	*p*-level	Mean values and standard derivation		
				Kaoping river	Tungkang river	Lingbeng river
T (°C)	0.003	2.099	0.218	28.64 (0.41)	28.78 (0.95)	30.79 (1.94)
Alk (mg/L)	0.002	0.714	0.534	206.70 (130.09)	148.27 (53.71)	188.50 (74.21)
Cl (mg/L)	0.002	1.235	0.366	516.00 (1108.07)	1817.44 (4622.29)	266.36 (461.09)
EC (μs/cm)	0.003	2.822	0.151	1786.71 (2726.10)	1157.86 (1385.78)	1206.00 (1440.98)
TKN (mg/L)	0.018	30.127	0.002	0.18 (0.08)	0.83 (0.25)	0.46 (0.48)
TP (mg/L)	0.003	2.405	0.185	0.00 (0.00)	0.01 (0.01)	0.01 (0.02)
UV-254 (cm^{-1})	0.003	2.727	0.158	0.62 (0.04)	98.37 (258.24)	0.84 (0.09)
pH	0.002	1.527	0.304	7.95 (0.30)	7.76 (0.30)	7.91 (0.19)
HPC (CFU/mL)	0.003	3.209	0.127	5.42×10^7 (7.27×10^7)	2.76×10^7 (3.87×10^7)	7.48×10^7 (1.19×10^8)
DO (mg/L)	0.004	4.950	0.065	5.74 (1.97)	4.46 (1.58)	7.03 (1.13)
Algae (unit/mL)	0.007	9.679	0.019	191.29 (101.55)	81.71 (35.07)	113.00 (47.64)
Chloro-a (mg/m^3)	0.008	11.541	0.013	22.86 (5.71)	22.86 (5.71)	20.17 (34.23)
SO_4^{2-} (mg/L)	0.002	0.740	0.523	125.30 (49.72)	239.85 (408.14)	92.64 (17.84)
TS (mg/L)	0.002	0.960	0.444	984.43 (2251.56)	343.57 (595.52)	561.14 (307.38)

Table 3 Outcomes of total standardized canonical coefficients (TSCC) and total canonical structure coefficients (TCSC) between canonical discriminant functions (CDF_1, CDF_2) and discriminant variables

Discriminant variables	CDF_1		CDF_2	
	TSCC	TCSC	TSCC	TCSC
T (°C)	1.59	0.01	−0.40	−0.36
Alk (mg/L)	0.91	0.03	−1.06	−0.00
Cl (mg/L)	−8.37	−0.02	−7.83	0.06
EC (μs/cm)	2.81	0.01	1.05	0.04
TKN (mg/L)	−6.80	−0.08	−0.18	−0.04
TP (mg/L)	−1.38	−0.03	−2.53	−0.13
UV-254 (m^{-1})	−1.09	−0.03	1.02	0.05
pH	1.30	0.03	0.53	−0.02
HPC (CFU/mL)	−3.42	0.02	−3.33	−0.08
DO (mg/L)	2.05	0.04	0.12	−0.24
Algae (unit/mL)	2.25	0.06	0.06	0.13
Chloro-a (mg/m^3)	9.15	−0.00	6.21	0.03
SO_4^{2-} (mg/L)	3.88	−0.02	7.46	0.07
TS (mg/L)	−0.56	0.02	−0.88	0.03
Chi-square test	75.87		21.14	
Canonical coefficient	0.99		0.92	
Eigenvalue	115.65		5.29	
Cumulative variance	0.96		1.00	

wastewater factor, and the ocean factor. The order of variance of variables that contributed to CDF_2 was T, DO, algae, and TP. Although the discriminant capacity was only 4.4%, CDF_2 was called the 'nutrient factor'. Figure 3 presents a dendrogram of the two CDFs to distinguish among the three rivers. Finally, five sampling sites that were significantly influenced by the ocean tide and located at the entrance of Tapeng Lagoon were selected; these were b1, c1, d1, b2, and c2. These two constructed canonical discriminant functions were substituted with standardized water quality parameters, indicating that the

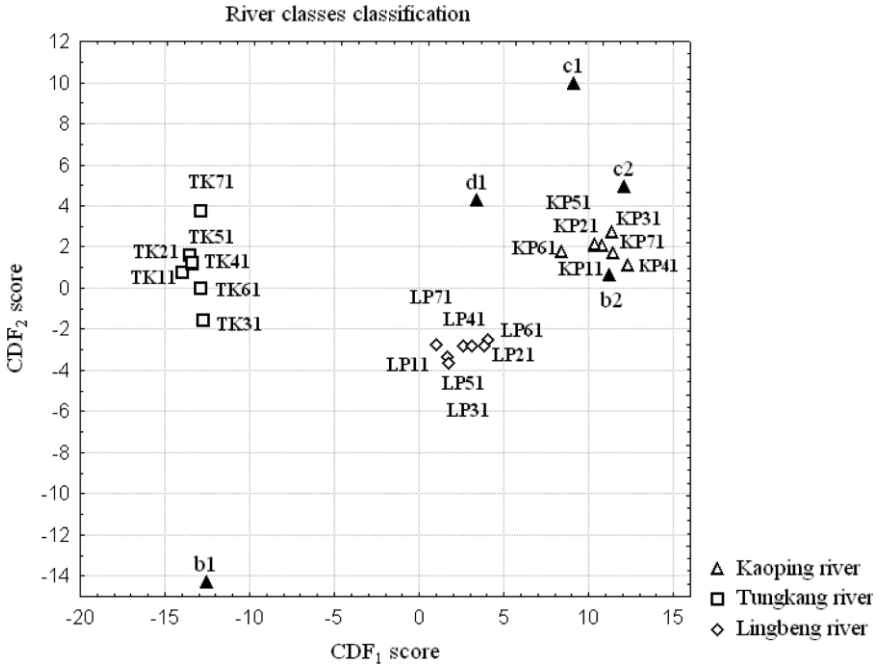


Figure 3 Dendrogram of three rivers neighbouring the Tapeng Lagoon in two discriminant functions

quality, except at b1, was similar to that in the Kaoping river (Figure 3). Therefore, the Kaoping river most strongly affected the water quality in the Tapeng Lagoon.

Conclusions

Analysis of the samples shows that the physical and chemical properties of the Tapeng Lagoon can be explained by three latent factors – the ocean factor, the primary productivity factor, and the fishery pollution factor. Furthermore, in these two constructed linear discriminant functions the main factor was the nutrient factor which recognized that the Kaoping river most affected the water quality of the Tapeng Lagoon. The water quality improved after disassembling the oyster frames and fishery boxes. On the basis of environmental topography characteristics, strengthening stream pollution prevention and making another entrance and exitway through the ocean are the best ways to raise the water quality in the Tapeng Lagoon by reducing eutrophication. These approaches and results will help the government to refine the current monitoring programme by selecting determinants of physical and chemical analyses of river water samples, which also may be applicable to other wetlands with similar properties or similar environmental problems.

References

Cruz-Castillo, J.G., Ganeshanandam, S., Mackay, B.R., Lawes, G.S. and Woolley, D.J. (1994). Applications of canonical discriminant analysis in horticultural research. *HortScience*, **29**, 1115–1119.

Huberty, C.J. (1994). *Applied Discriminant Analysis*. John Wiley and Sons, New York.

Jenerette, G.D., Lee, J.D., Waller, W. and Carlson, R.E. (2002). Multivariate analysis of the ecoregion delineation for aquatic systems. *Environ. Manage.*, **29**, 67–75.

Momen, B. and Zehr, J.P. (1998). Watershed classification by discrimiant analyses of lakewater-chemistry and terrestrial characteristics. *Ecol. Appl.*, **8**, 497–507.

Rencher, A.C. (1992). Interpretation of canonical discriminant functions, canonical varieties, and principal components. *Am. Stat.*, **46**, 217–225.

Ruiz, F.V. and Blasco, G.P. (1990). Application of factor analysis to the hydrogeochemical study of a coastal aquifer. *J. Hydrol.*, **119**, 169–177.

Sharma, S. (1996). *Applied Multivariate Techniques*, John Wiley & Sons, New York.

Shin, P.K.S. and Fong, K.Y.S. (1999). Multiple discriminant analysis of marine sediment data. *Mar. Pollut. Bull.*, **39**, 285–294.

Standard Methods for the Examination of Water and Wastewater (1995). 19th edn, American Public Health Association/American Water Works Association/Water Environment Federation, Washington DC, USA.

StatSoft (1996). *STATISTICA for Windows*, StatSoft, Inc, Tulsa OK, USA.