Statistical analysis of the physico–chemical data on the coastal waters of Cochin

C. S. Padmanabha Iyer,[*a] **Manonmani Sindhu,**[a] **Savita G. Kulkarni,**[b] **Sanjeev S. Tambe**[b] **and Bhaskar D. Kulkarni**[b]

[a]*Centre for Marine Analytical Reference and Standards, Regional Research Laboratory, Thiruvananthapuram-695019, India. E-mail: iyer_csp@yahoo.com; Fax: 91-471-491712/490186*
[b]*National Chemical Laboratory, Pune, India*

Measurements of temperature, salinity, dissolved oxygen, nitrogen as ammonia, nitrate and nitrite, and phosphate along with chlorophyll were carried out at three stations on the coastal waters of Cochin, south west India, at two levels of the water column over a period of five years. The data set has been factorised using principal component analysis (PCA) for extracting linear relationships existing among a set of variables. A graphical display of the scores generated from the PCA was done by means of boxplots and biplots, which helped in the interpretation of the data. The major factors conditioning the system are related to the input of fresh water from the estuary of the Periyar river and the high organic load of the bottom sediment in the coastal area which results in a reducing environment, as reflected in the parameters of dissolved oxygen, ammoniacal-nitrogen and nitrite-nitrogen. Another factor which contributes to the variation in the system is related to the unloading activity in the port area. The present approach presents a logical way to interpret the complex data of the physico–chemical measurements.

Introduction

A major monitoring programme has been initiated to collect data on the coastal water quality of India.[1] This paper analyzes the water quality data collected from the Cochin coast, south west India. The said coastal area is geographically complex due to the fact that it is connected through an estuary to the Periyar River, on the banks of which are a number of industries. This is also the site of major shipping activity. The data comprising measurements of several physico–chemical variables from coastal waters are analyzed to detect the spatial and temporal variations therein. For this purpose, a number of statistical methods such as principal component analysis (PCA),[2] correlation matrix, biplots[3,4] and boxplots[5] have been employed. Earlier, Barbieri *et al.* have carried out a study on the coastal water quality in the Gulf of Trieste.[6]

Experimental

Measurements of temperature (x_1), salinity (x_2), dissolved oxygen (x_3), nitrogen as ammonia (x_4), nitrite (x_5), nitrate (x_6), phosphate (x_7) and chlorophyll (x_8) have been made at three stations, at two levels of the water column, over a period of five years. The stations chosen were near the coast and 5 km and 10 km towards the sea. A total of 54 samples were drawn at two depths, just below the surface and just above the bottom of the sea, from the three stations identified as per the standard procedures. The measurements of the above stated eight variables were made using standard methods: (i) temperature and salinity *in situ* by probes, (ii) dissolved oxygen using the Winkler method, (iii) ammonia, nitrite, nitrate and phosphate spectrophotometrically, and (iv) chlorophyll by spectrofluorimetry after extraction with 90% acetone.[7–9] Quality assurance of the data collected was carried out by periodical standardization.

Principal component analysis (PCA)

Physico–chemical events in the coastal waters and their relationships reveal that these do not occur as independent events. In fact, the events and data emanating there from are correlated. Any technique that captures the important events based on the variability in the data thus provides both a reduction in the data size and a summary of the information contained in the original data set. Principal component analysis (PCA) is one such technique that extracts linear relationships existing among a set of variables. It decomposes a single interdependent data set comprising measurements of multiple variables into a transformed variable defined by eigenvectors of the covariance of the data. In essence, PCA generates a set of pseudo measurements (scores or latent variables) which are linearly independent and each of which captures the maximum amount of variability in the data in the descending order. The PCA procedure transforms an original set of variables into a new set of principal components, which are at right angles (*i.e.* uncorrelated) to each other. Specifically, it decomposes a 2-D matrix $X(I,J)$, consisting of I measurements of J variables into scores and loadings matrices, as given by:

$$X = TP' + E \qquad (1)$$

where each column of the matrix X is mean centered (mean = 0) and variance scaled (standard deviation = 1); $T(I,J)$ represents the matrix of J principal components scores (each column of matrix T refers to a principal component); P' denotes the transpose of original data. In general, first R principal components are sufficient to describe a large amount of variance in the original matrix, X, and thus eqn. (1) can be recast as

$$X = \sum_{R=1}^{R} t_r (p_r)^T + E' \qquad (2)$$

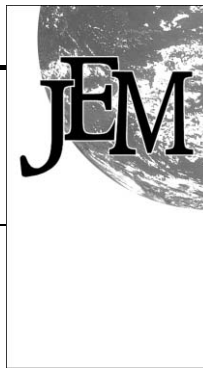Where t_r represents the rth I dimensional score vector; $(p_r)^T$

DOI: 10.1039/b209219k

Table 1 Maximum, minimum, mean and standard deviation of the data

	Temperature/ °C	Salinity/Practical salinity units	DO/ μmol dm^{-3}	NH$_3$-N/ μmol dm^{-3}	NO$_2$-N/ μmol dm^{-3}	NO$_3$-N/ μmol dm^{-3}	PO$_4$-P/ μmol dm^{-3}	Chlorophyll a/ μg dm^{-3}
Maximum	32	35.43	6.64	26	2.23	38.83	3.43	25.35
Minimum	26	24.42	1.77	0	0	0	0.13	0.17
Mean	29.785	33.451	4.386	2.99	0.4403	3.455	0.8385	4.1659
Std. deviation	1.1313	1.5535	0.82	4.3464	0.474	7.286	0.688	4.5288

Table 2 Correlation matrix

	Temperature	Salinity	DO	NH$_3$-N	NO$_2$-N	NO$_3$-N	PO$_4$-P	Chlorophyll a
Temperature	1	−0.0843	0.4403	0.2421	0.1278	0.3654	0.0869	0.1583
Salinity	−0.0843	1	−0.1820	0.0383	−0.4113	−0.2351	−0.3824	−0.4130
DO	0.4403	−0.1820	1	0.0938	−0.1281	0.1197	−0.0744	0.3058
NH$_3$-N	0.2422	0.0383	0.0938	1	−0.1115	0.1175	−0.1117	−0.1147
NO$_2$-N	0.1278	−0.4113	−0.1281	−0.1115	1	0.1287	0.3780	0.3027
NO$_3$-N	0.3654	−0.2351	0.1197	0.1175	0.1287	1	0.2601	0.1177
PO$_4$-P	0.0869	−0.3824	−0.0744	−0.1117	0.3780	0.2600	1	0.4332
Chlorophyll a	0.1583	−0.4130	0.3058	−0.1147	0.3027	0.1177	0.4332	1

Table 3 Principal component loadings

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Temperature	0.2868	0.5348	−0.0968	−0.1551	0.4728	−0.2648	0.2627	0.4854
Salinity	−0.4639	0.1295	−0.0613	−0.2880	0.1624	−0.6729	−0.2971	−0.3374
DO	0.2158	0.5073	0.5632	−0.0097	0.0060	0.1351	0.1384	−0.5843
NH$_3$-N	−0.0248	0.4499	−0.4351	0.7092	−0.2607	−0.1280	−0.0967	−0.1048
NO$_2$-N	0.3875	−0.3060	−0.2316	0.2357	0.6742	0.0417	−0.2102	−0.3828
NO$_3$-N	0.3212	0.2515	−0.4766	−0.5540	−0.2525	0.2803	−0.3714	−0.1390
PO$_4$-P	0.4369	−0.2740	−0.1793	−0.0997	−0.3444	−0.4897	0.5253	−0.2400
Chlorophyll a	0.4596	−0.0742	0.4094	0.1321	−0.2126	−0.3497	−0.5967	0.2733

denotes transpose of rth J dimensional loading vector (p_r) and E', the residual matrix. As can be seen from eqn. (2), the original data matrix X can now be represented in terms of only R score vectors ($R < J$) and thus data reduction is achieved. The sum of squares of elements of a score vector (t_r) relates to the eigenvalue (also termed as "trace") associated with that score vector and is a measure of the variance captured by that principal component. Accordingly, larger the magnitude of the eigenvalue (trace), the more significant is the corresponding principal component.
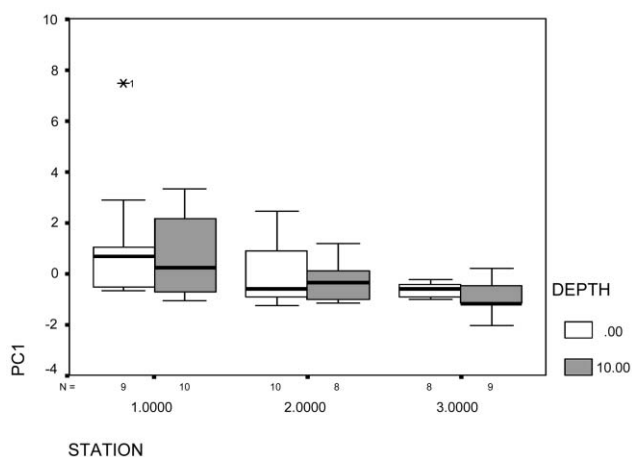
Results and discussion

The data comprising measurements of the eight physico–chemical variables (x_1 to x_8) were organized in a matrix form (X) with variables in columns and samples as rows. Next, elements of each matrix column were mean centered with mean = 0 and standard deviation = 1. The data matrix X pre-processed this way was used to obtain the correlation matrix. In Table 1 are given the maximum, minimum and mean values of the physico–chemical variables along with their standard deviations. From the correlation matrix given in Table 2, it is

Table 4 Traces of principal component[a]

Trace 1	30.18
Trace 2	20.41
Trace 3	13.36
Trace 4	9.95
Trace 5	8.70
Trace 6	8.01
Trace 7	5.42
Trace 8	3.96

[a]The trace gives the percentage variance captured by the principal component. In this case the first 6 components capture nearly 90% of the variance

seen that a strong linear relationship exists between some of the variables. Thus, the principal components analysis was conducted using NIPALS algorithm[10] on the preprocessed matrix X to characterize the linear correlations and the loadings of the eight variables for each of the principal components is obtained. These are entered in Table 3. It may be noted that the loading values expresses the influence of each original variable with a principal component. The traces, which describe the amount of variance captured by each of the eight principal components are given in Table 4. From Table 4, it is seen that the first three principal components (PCs) capture a major portion (63.95%) of the variance present in the data. It is also seen that the principal component (PC1) explaining experimental variance of 30.18%, has high loading values for salinity (−0.460), chlorophyll (0.459) and phosphate (0.436). Boxplots (Fig. 1) of PC1 scores of the two layers in the three

Fig. 1 Boxplot of PC1 against stations at two depths.
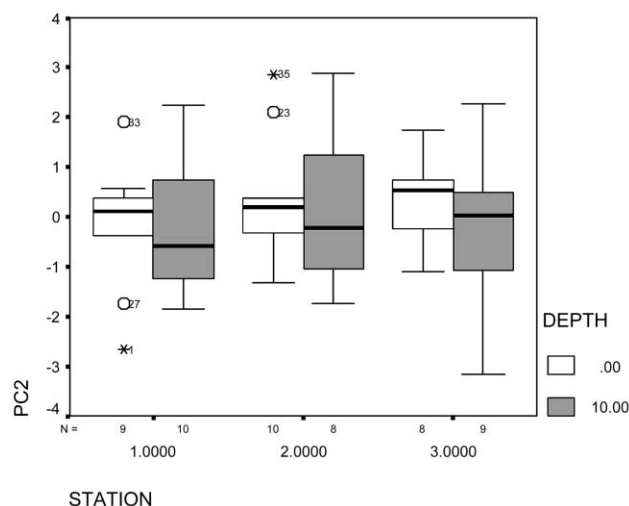
Fig. 2 Boxplot of PC2 against stations at two depths.

Fig. 3 Scores of sample and loadings of variable.

sampling stations indicate that the samples drawn near the shore (Station 1) have higher scores compared to Station 2 and 3. This is clearly due to the fact that Station 1 is influenced by the inflow of water from the Periyar river *via* the estuary, which also brings in the nutrients, especially phosphate. Similarly, temperature is also affected due to the mixing of the water from the estuary with that of oceanic water. The higher score for chlorophyll is also explained, as due to the presence of nutrients.

The examination of the loading vector associated with PC2 shows high values for temperature (0.535), dissolved oxygen (0.507), ammonia-nitrogen (0.449), phosphate-phosphorus (−0.274) and nitrite-nitrogen (−0.306). Boxplot of PC2 scores of the two layers in the sample stations (Fig. 2) shows that there is more variation in the score values at 10 m depth compared to that at the surface in all the three stations. This variation is due to the reaction taking place in the reducing environment at the bottom, which get reflected in the measurement of dissolved oxygen, ammoniacal nitrogen and nitrite-nitrogen. It is reported that organic content of sediments in this area is high of the order of 6–8% C.

The loadings corresponding to PCs three to eight can be interpreted similarly although it is seen that from the corresponding traces that PCs three to eight together explain approximately same amount (50%) of variance as that by the PC1 and PC2. An interesting observation can be made while comparing the loadings of PC1 and PC6. It is seen that loadings of both PCs have high magnitudes for the same variables *viz.* salinity, phosphate-phosphorus, chlorophyll, nitrate-nitrogen, temperature and dissolved oxygen. Also, the correlation analysis of the principal component loadings has revealed that the highest correlation (= 0.416) exists between PC1 and PC6. The next highest correlation existing between PC2 and PC6 has a magnitude equal to 0.285. Hence it is seen that the correlation between PC1 and PC6 is significantly higher than the others (Table 5). Thus, a biplot of PC1 against PC6 was prepared (Fig. 3) wherein scores of samples drawn
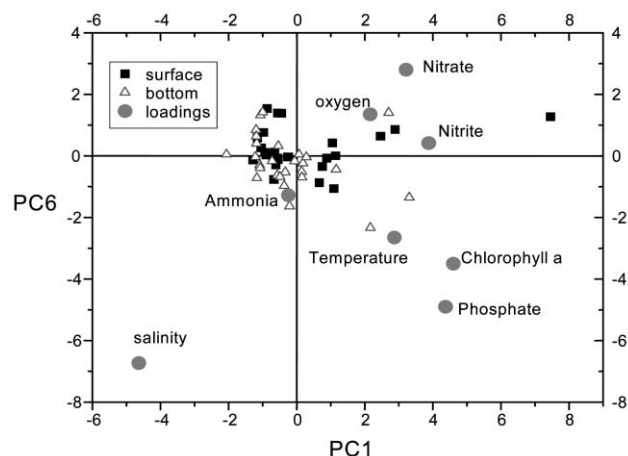
from the sea surface and sea bottom, and the loadings of variables have been plotted. It can been seen in this figure that: (i) nitrite-nitrogen, nitrate-nitrogen and dissolved oxygen are present in the first quadrant, (ii) ammoniacal-nitrogen and salinity are in the third quadrant, and (iii) temperature, chlorophyll a and phosphate lie in the fourth quadrant. The presence of multiple variables in the same quadrant suggests close association between them. The association between dissolved oxygen, nitrite and nitrate is understandable as their equilibrium values are controlled by oxidation–reduction processes. The association of chlorophyll a with temperature and phosphate can be explained on the premise that phosphate helps in the plankton production and thus of chlorophyll; temperature also is a factor in the synthesis of chlorophyll. The presence of ammonia and salinity in the third quadrant indicates influx of ammonia in the coastal waters due to the discharges of Periyar river, carries the waste effluents from the fertilizer unit. The change in salinity is associated with dilution due to the freshwater inflow. Scanning the plot, one can see that among the nutrients, phosphate and ammonia lie in different quadrants, and thus they are separated. It can thus be concluded that the sources for ammonia and phosphate are not identical. It is reported that there exists a major unloading activity of rock-phosphate in the port, which is an additional source, different from that of ammonia. In Fig. 3, two clusters of points are seen; one in the third quadrant where ammonia and salinity lie and another in the fourth quadrant, where temperature, chlorophyll a and phosphate are present. The first cluster appears in the months of February to April and the second cluster during the month of November. It may be noted that the former period represents a part of the summer months while the latter, a winter month.

Conclusion

The present paper analyzes the coastal water quality data collected from the Cochin coast in south west India. Based on the principal component analysis, a statistical model has been constructed, which explains the relationships between the

Table 5 Canonical correlation analysis

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
PC1	1.000	−0.333	0.127	0.007	−0.061	0.416	0.164	0.273
PC2	−0.333	1.000	0.087	0.005	−0.042	0.285	0.112	0.187
PC3	0.127	0.087	1.000	−0.002	0.016	−0.109	−0.043	−0.071
PC4	0.007	0.005	−0.002	1.000	0.001	−0.006	−0.002	−0.004
PC5	−0.061	−0.042	0.016	0.001	1.000	0.052	0.020	0.034
PC6	0.416	0.285	−0.109	−0.006	0.052	1.000	−0.140	−0.233
PC7	0.164	0.112	−0.043	−0.002	0.020	−0.140	1.000	−0.092
PC8	0.273	0.187	−0.071	−0.004	0.034	−0.233	−0.092	1.000

various physico–chemical variables that have been monitored and the environmental conditions. The influence of the discharge of the river water through the Periyar estuary into the coastal water is explained by the first principal component, as it reflects the changes in salinity and nutrient concentrations. The second principal component is related to the changes, which take place in the reducing environment at the sea bottom. The biplot of PC1 *versus* PC6 reveals presence of two clusters one of which is related to the summer months (February to April) and the other is related to November which falls in the midst of winter. It can be concluded that the principal component analysis can be used to advantage to relate the voluminous data generated from any environmental monitoring programme to the physical chemical factors prevailing in the area of study.

Acknowledgements

References

1 *Annual Reports*, Department of Ocean Development, Government of India, 1996–2000.
2 K. V. Mardia, J. T. Kent and J. M. Bibby, *Multivariate Analysis*, Academic Press, London, 1994.
3 K. R. Gabriel, *Biometrika*, 1971, **58**, 453.
4 J. C. Gower and D. J. Hand, *Biplots*, Chapman & Hall, London, 1996.
5 M. J. Norusis, *SPSS for Windows 6.1.3 User's Guide*, SPSS Inc., Chicago, 1996.
6 P. Barbieri, G. Adami, S. Predonzani, E. Reisenhofer and D. L. Massart, *J. Environ. Monit.*, 1999, **1**, 69–74.
7 K. Grasshoff, M. Enrhardt and K. Krenling, *Methods of Seawater Analysis*, Verlag Chemie, Weinheim, 1983.
8 APHA-AWWA-WPCF, *Standard Methods for the Examination of Water and Waste Water*, American Public Health Association, Washington, DC, 19th edn., 1995.
9 *Protocols for the Joint Global Ocean Flux Study (JGOFS) Core Measurements*, UNESCO, 1994.
10 P. Geladi and B. R. Kowalski, *Anal. Chim. Acta*, 1986, **185**, 1–17.