

Evaluation of significant sources influencing the variation of water quality of Kandla creek, Gulf of Katchchh, using PCA

S. G. Dalal · P. V. Shirodkar · T. G. Jagtap ·
B. G. Naik · G. S. Rao

Received: 7 August 2008 / Accepted: 27 January 2009 / Published online: 27 March 2009
© Springer Science + Business Media B.V. 2009

Abstract To evaluate the significant sources contributing to water quality parameters, we used principal component analysis (PCA) for the interpretation of a large complex data matrix obtained from the Kandla creek environmental monitoring program. The data set consists of analytical results of a seasonal sampling survey conducted over 2 years at four stations. PCA indicates five principal components to be responsible for the data structure and explains 76% of the total variance of the data set. The study stresses the need to include new parameters in the analysis in order to make the interpretation of principal components more meaningful. The PCA could be applied as a useful tool to eliminate multi-collinearity problems and to remove the indirect effect of parameters.

Keywords Water quality · PCA · Significant sources · Kandla creek

Introduction

Maintenance of quality of nearshore marine water is a sensitive issue, as a wide variety of anthropogenic waste is discharged into it. Sewage and municipal wastes, landfill drainage, pulp and paper industry, forestry operations, agricultural runoff, aquaculture and fish processing plants, etc. are some major point and non-point sources of pollution. Such wastes, being the most ubiquitous and direct sources, contaminate the coastal zone to a large extent, often leading to eutrophication, which refers to the process of natural or man-made enrichment with inorganic nutrients (Kennish 1992; Schramm and Nienhuis 1996). From a coastal zone management perspective, it is important to understand the relative capacities of waters to absorb such wastes for sustainable development and to adopt proper management strategies. Since aquatic environment shows spatial and temporal variations in water quality, there is a need to devise a monitoring program that will provide the best representation and reliable estimate of a particular water mass. This is necessary to avoid frequent water samplings at many sites.

A 2-year survey (October 2002–September 2003 and June 2004–May 2005) to establish national databases on water quality was undertaken in Kandla creek. The Environmental Guidelines for Ports and Harbours (EGPH 1989), laid down by the Ministry of Environment & Forests,

S. G. Dalal (✉) · P. V. Shirodkar ·
T. G. Jagtap · B. G. Naik
National Institute of Oceanography,
Dona Paula, Goa 403004, India
e-mail: dalal@nio.org

G. S. Rao
Kandla Port Trust, New Kandla,
Gandhidham, Gujarat, India

Department of Environment, Forests & Wildlife, Government of India, New Delhi were followed to assess the pollution levels in marine waters. In the present communication, the large data matrix obtained during the monitoring program (1,224 observations) in Kandla creek was utilized to extract information on: (a) the temporal variations and (b) evaluate the significant sources contributing to water quality parameters.

Material and methods

Study area

Kandla creek lies between latitude $22^{\circ}55'$ to $23^{\circ}05'N$ and longitude $70^{\circ}05'$ to $70^{\circ}02'E$ in the Gulf of Katchchh. Kandla Port is one of the major ports, situated about 90 km from the mouth of the

Gulf of Katchchh (Fig. 1). This port has many terminals for handling oil and oil products, fertilizers, and general cargo. Hectic loading and unloading activities and movement of personnel at the port are bound to have adverse effects on the creek water. Tidal height in the creek ranges from 0.83 to 7.2 m and a surface current varies from 1.5 to 5 kn.

Sampling and analysis

The sampling strategy was designed to cover a wide range of water quality determinants at key sites. The sampling for water was done once each season—premonsoon/summer (February–May), monsoon (June–September), and post monsoon/winter (October–January) at four key stations—mouth of the creek, off cargo jetty, off IOC oil jetty and at the junction, where Sara and Phang

Fig. 1 Location of port industrial units and sampling stations during a monitoring program of Kandla creek

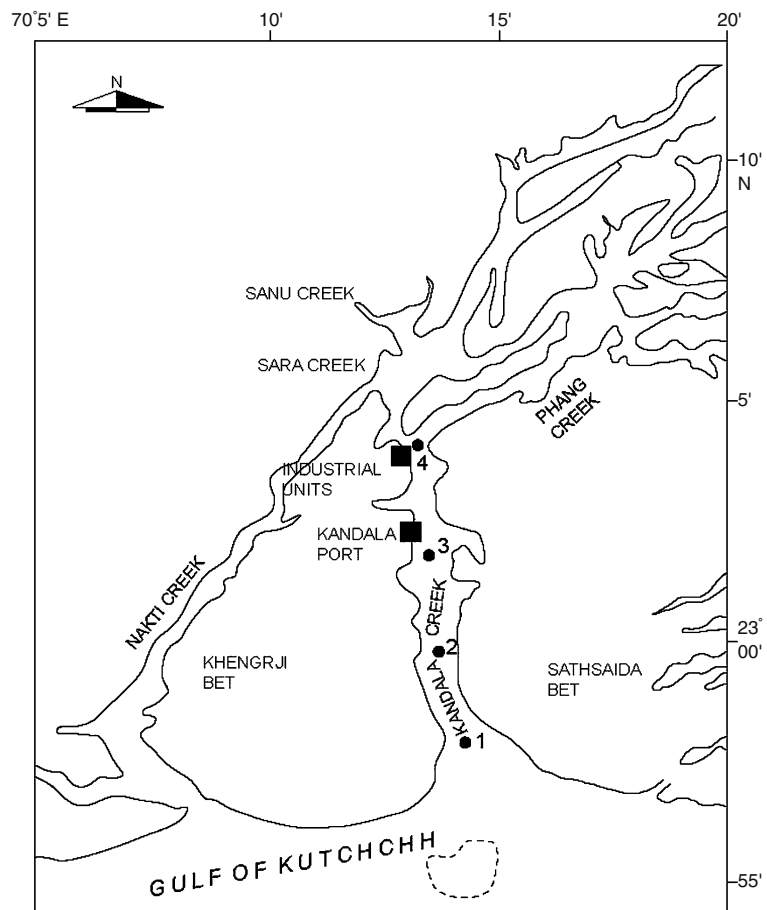
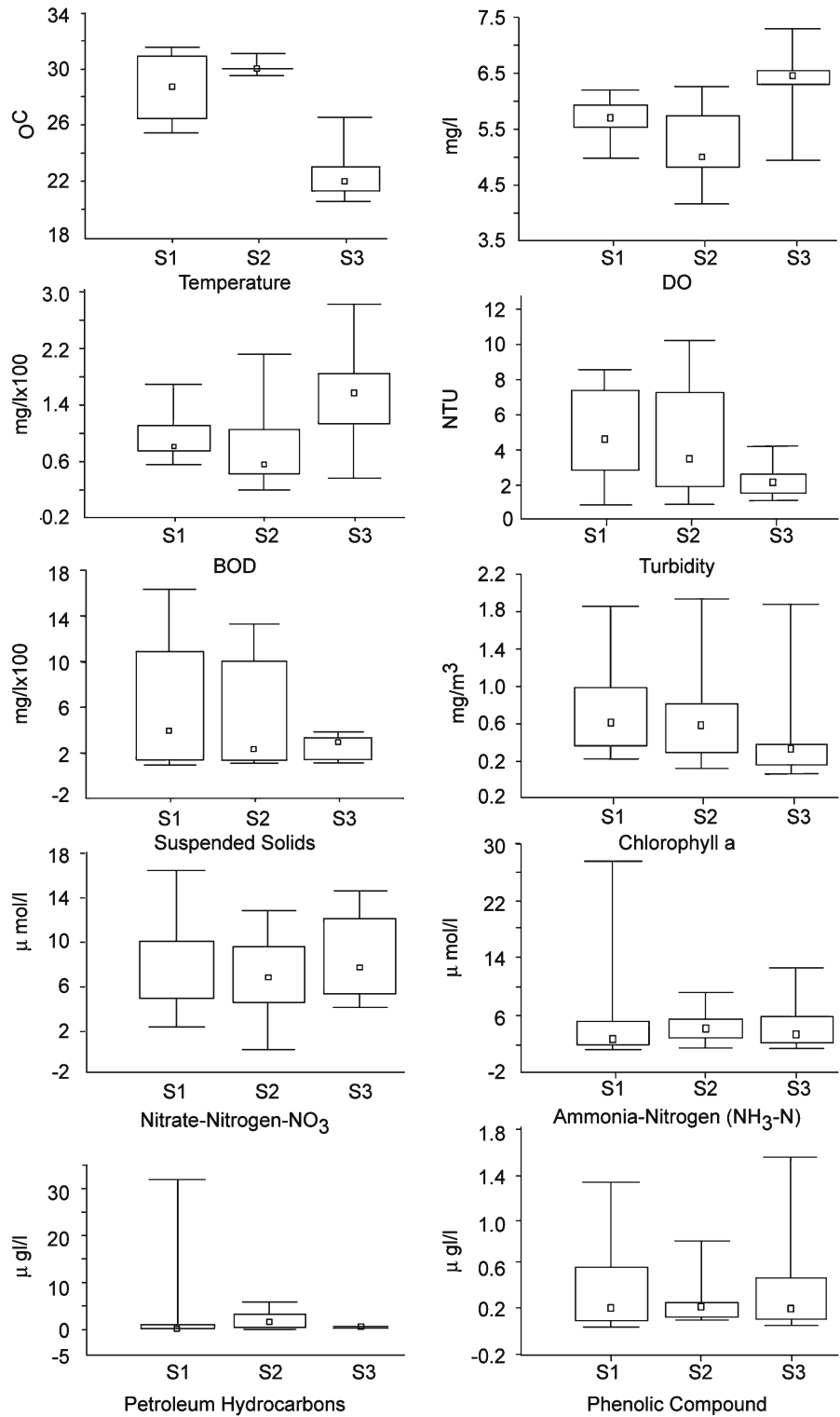


Fig. 2 Temporal variations—box plot for selected parameters for three seasons (S1 premonsoon, S2 monsoon, S3 post monsoon)



creeks meet Kandla creek (Fig. 1), over a 2-year period from October 2002 to September 2003 and June 2004 to May 2005. The water samples were collected from surface, mid-depth, and bottom water layers at each station and on-the-spot analyses were made for few parameters while for other parameters, the samples were analyzed at the shore laboratory following standard methods for water analysis (APHA 1975; Grasshoff et al. 1983). Water quality variables analyzed include dissolved oxygen (DO), water temperature (TEMP), pH, salinity (SAL), suspended solids (SSOL), biochemical oxygen demand (BOD), phosphate (PO₄-P), nitrite (NO₂-N), nitrate (NO₃-N), ammonia (NH₃-N), and silicate (Si₂O₃-Si). Turbidity (TURB) was measured by nephelometric method. Dissolved/dispersed petroleum hydrocarbons (PHC) were extracted from seawater with double distilled hexane and quantified by using Shimadzu RF-1501 fluorescence spectrofluorometer. Phenol (PHE) was extracted with chloroform after complexing with 4-aminoantipyrine, and the color was measured spectrophotometrically. For estimation of chlorophyll *a* (CHL), 500 ml of water sample was filtered through GF/F glass fiber filter paper extracted in 90% acetone overnight. The extracts were used for the estimation of fluorescence before and after acidification using Turner Design Fluorometer. The fluorescence values were converted to chlorophyll and phaeophytin (PHE) using appropriate calibration factor. Primary production (PP) was measured using ¹⁴C technique. The data quality was checked by careful standardization and procedural blank measurements of spiked and duplicate samples.

Multivariate statistical methods

The water quality parameters in three different seasons (winter, summer, and monsoon) were assigned a numerical value in the data file which, as a variable corresponding to the season, was correlated (pair by pair) with all the measured parameters.

In order to avoid misclassification due to wide differences in data dimensionality (Ross 1988), data was standardized through *z*-scale transformation before applying PCA. Standardization

tends to minimize the influence of variance of parameters. It also eliminates the influence of different units of measurement and renders the data dimensionless. The main concern of the PCA is to understand the mode of action or behavior of components of a system and its subsystems (Petersen et al. 2001; Bengraïne and Marhaba 2003). The use of PCA for water quality assessment has increased in the last few years, mainly due to the need to obtain appreciable data reduction for analysis and decision (Morales et al. 1999). Bartlett's sphericity test (χ^2 with degrees of freedom = $1/2[p(p-1)]$) was used to verify the applicability of PCA to raw data (Stevens 1986). The STATISTICA 6.0 software package was employed for data treatment.

Results and discussion

Exploratory data analysis

Box plots of selected parameters during three seasons (Fig. 2) at four sampling stations were examined. By inspecting these plots, it was possible to perceive differences between the seasons. Our first approach to establish the parameter-associated temporal variation was by use of the Spearman *R*. The bivariate results (Table 1) show that the five parameters having significant correlation with the season ($p < 0.05$) are: nitrates, suspended solids, ammonia, water temperature,

Table 1 Spearman non-parametric correlation coefficient (*R*) for selected parameters between seasons (bold figures indicate significance at $p < 0.05$)

Parameters	Summer and monsoon	Summer and winter	Monsoon and winter
TEMP	-0.39	-0.78	0.34
DO	-0.22	0.28	0.08
BOD	0.06	0.30	0.05
TURB	0.71	0.38	0.37
SSOL	0.93	0.62	0.64
CHL	-0.21	0.74	-0.35
NO ₃	0.48	0.84	0.55
NH ₃	0.50	0.38	0.86
PHC	0.24	0.12	0.31
PHE	0.27	0.25	0.22

and turbidity. The season-correlated parameters were taken as representing the major source of temporal variations in water quality. In view of the source types in the creek, these correlations can be explained on the basis of seasonal features in the monitoring region. The correlation matrix suggests that the temperature, SSOL, NO₃, and NH₃ are the most significant parameters to discriminate between the seasons, which also means that these parameters account for most of the expected temporal variations in the water quality; this also suggests that the anthropogenic input, which was the major pollution source mainly derived from the discharge of wastewater into the system, was independent of the season as it was present throughout the year. Some of these correlations can be explained by climatic changes associated with the three seasons. Land drainage and strong tidal currents in the creek bring in a large amount of colloidal particles into suspension during monsoon, whereas, during summer, the strong tidal currents in low water level of the creek disturb the settled sediments bringing in large amount of colloidal particles in water, thereby increasing the turbidity. This was also true for suspended solids, which were significantly higher during the summer and winter periods as compared to the monsoon. Although instances of waste releases due to port activities were evident in Kandla creek, the influence of seasonal changes appears to be fairly large.

Data treatment

In the application of PCA to water quality data from Kandla Port monitoring stations, correlation matrix of variables ($R_{p \times p}$) was used to obtain eigenvalues and weights of parameters. Since the four sampling stations were combined to calculate the correlation matrix, the correlation coefficients should be interpreted with caution as they are simultaneously affected both by spatial and temporal variations. The Scree plot was used here to identify the number of PCs to be retained in order to comprehend the underlying data structure (Jackson 1993). Eigenvector λ was used to obtain unrotated factor loadings. V_{ia} indicates the values of rotated factor loadings, which were obtained by varimax rotation. Rotated loadings in PCs

indicate the percent contribution of corresponding variable to the PC and are called the loading of i th variable in k th PC (correlation between variable i th and k th PC). These loading values were used to group variables in PCs.

Score value (s_{kj}) for j th observation in k th PC was obtained from the weight of variables in PCs and standardized variables by using the following equation:

$$s_{kj} = a_{1j}z_{1j} + a_{2k}z_{2j} + \dots + a_{pk}z_{pj}$$

where $j = 1, 2, \dots, n$ is the number of observations; $k = 1, 2, \dots, q$ the number of selected PC numbers, and p the number of independent variables.

By applying Bartlett’s sphericity test, a value of 844.764 for the Bartlett chi-square statistics was found ($df = 136, p < 0.01$), confirming that the parameters are not orthogonal but correlated, therefore explaining the data variability with a lesser number of parameters (called principal components).

PCA on 17 parameters yielded five principal components explaining sample variance of about 76% (Table 2). The varimax rotation was then performed to secure increased principal components of environmental significance; a similar approach based on PCA has been used to identify the main components in water quality (Vega et al. 1998; Helena et al. 2000; Wunderlin et al. 2001; Simeonov et al. 2003; Singh et al. 2004).

In the present study, the first PC that explains 29% of total variance has significant loadings (> 0.70) on salinity, suspended solids, turbidity, and petroleum hydrocarbons (Fig. 3). High loadings on suspended solids, turbidity, and salinity were due to the natural effects of strong tidal currents (tidal range 7 m) and intrusion of saline

Table 2 Eigenvalues and percentage of explained variance by first five components by PCA

Component	Eigenvalue	% total variance	Cumulative %
1	4.922	28.95	28.95
2	4.282	25.18	54.13
3	1.588	9.34	63.47
4	1.107	6.51	69.98
5	1.018	5.99	75.97

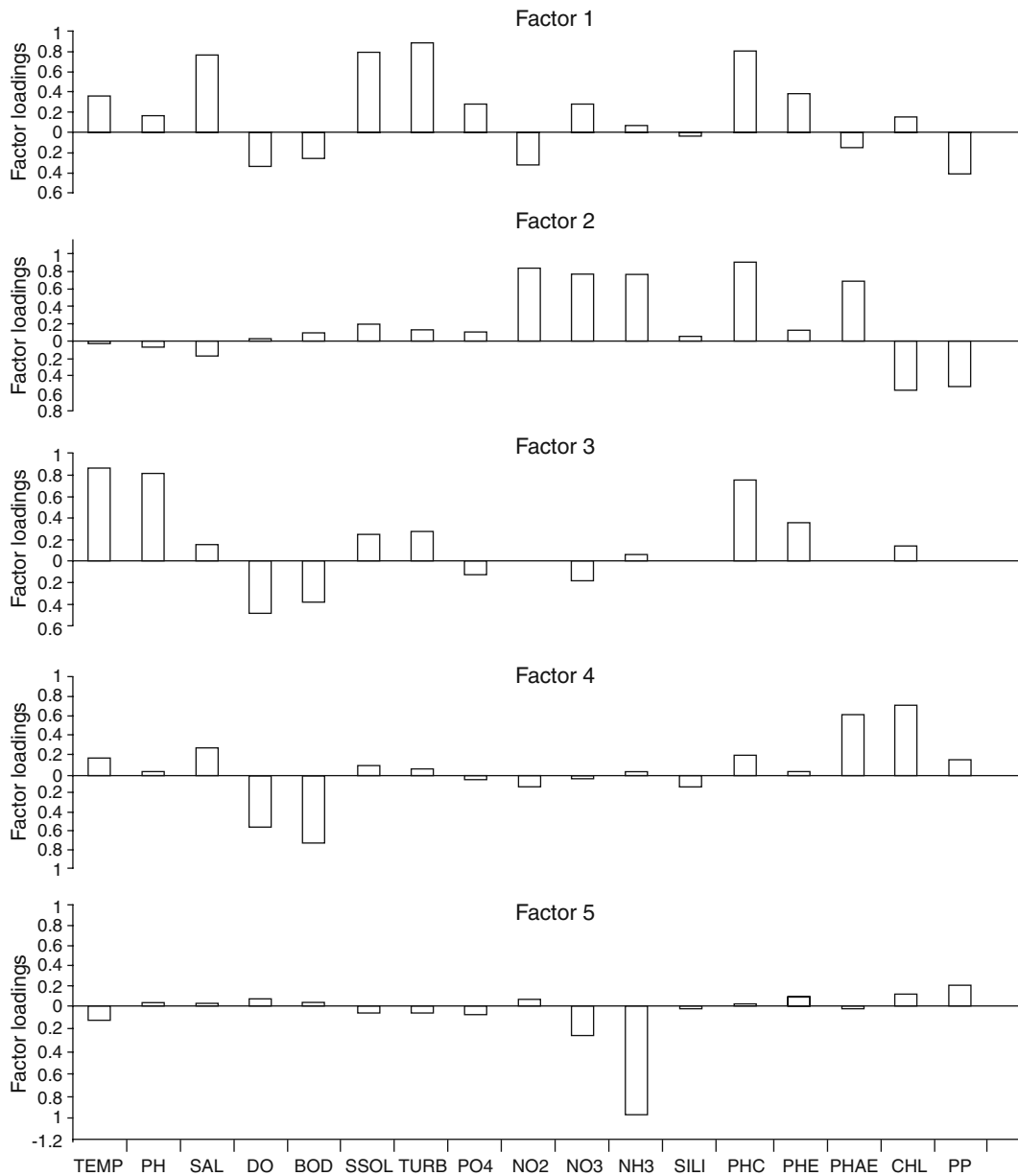


Fig. 3 Factor loadings of PCA on 17 water quality parameters

waters from the salt works. The high loading of petroleum hydrocarbons was due to the spillage from loading and unloading activities of oil and other petroleum products. The second PC that explains 25% of the total variance correlates to water-soluble nitrogenous species, i.e., $\text{NO}_2\text{-N}$, $\text{NO}_3\text{-N}$, $\text{NH}_3\text{-N}$, and petroleum hydrocarbons (Fig. 3). The main sources of nitrate were due to agricultural activities and the increased drainage

from fields after precipitation events. Within the part of variance described by the third PC, temperature and pH have an opposite sign in comparison to the DO and BOD. The contribution of salinity was negligible; phosphate and nutrients were only weakly involved (Fig. 3). The pattern can be interpreted in terms of biological activity as either primary production by algae or their subsequent microbial decomposition. The fourth

PC significantly negatively correlated with DO and BOD with positive loadings on chlorophyll *a*, and phaeophytin (Fig. 3) represents the ‘organic source’ of the creek water. In the fifth PC, one can derive that the high loadings of ammonia (Fig. 3) were mainly due to the discharge from the nearby fertilizer plant and the oil jetty.

Given that using a cut-off criteria $\lambda_0 \leq 0.70$ retains too many parameters, and that a cut-off of $\lambda_0 > 0.70$ leads to retention of too many components (Jackson 1993), we do not suggest the eigenvalues (λ) criteria. Our results also suggest that a cumulative proportion of variance (α_0) criterion is inappropriate. Solutions based on proportionately more parameters will be less stable and the resulting eigenvector coefficients will be less reliable. As a consequence, the interpretability of the analyses will be compromised.

From an ecologist’s view, parameter selection is useful for reducing the number of parameters required for statistical analyses since it can improve the reliability and stability of final results (Williams and Titus 1988; Grossman et al. 1991). In this study, PCA did not result in much data reduction, as one still needs 14 parameters (about 80% of the 17 parameters) to explain 76% of the data variance. Therefore, it becomes necessary to include new parameters such as toxic trace metals (viz. lead, cadmium, and mercury) in water and sediments in the monitoring program to make the interpretation of PCA more meaningful. However, PCA serves as a means to identify parameters that have the highest contribution in the creek water quality.

Conclusion

The study demonstrates the value of PCA of large and complex databases in deriving better information about the water quality and analytical protocols. PCA is also a powerful pattern recognition technique that attempts to explain the variance of a large dataset of intercorrelated parameters with a smaller set of independent parameters. Five components explain 76% of the total variation. It may be necessary to include certain new parameters such as toxic trace metals and eliminate some observed parameters for future

monitoring studies. If resources become limited, the selected parameters may provide a suggestion for future data collection in environment monitoring studies. Our results suggest that a cumulative proportion of variance (α_0) criterion is inappropriate for PCA.

Acknowledgements The authors wish to thank the Director, NIO, India for facilities and Kandla Port Trust Authorities for sponsoring the environmental monitoring program in Kandla creek. This is NIO (CSIR) contribution Number 4515.

References

- APHA (1975). *Standard methods for the examination of water and waste water* (14th ed.). APHA-AWWA-WPCE, American Public Health, Washington DC20036.
- Bengraïne, K., & Marhaba, T. F. (2003). Using principal component analysis to monitor spatial and temporal changes in water quality. *Journal of Hazardous Materials, B100*, 179–195. doi:10.1016/S0304-3894(03)00104-3.
- EGPH (1989). *Environmental guidelines for ports and harbour projects*. New Delhi: Govt. of India.
- Grasshoff, K., Ehrhardt, M., & Krimling, K. (1983). *Methods of seawater analyses* (Second revised and extended edition, 419 pp.). Weinheim: Verlag Chemie.
- Grossman, G. D., Nicckerso, D. M., & Freeman, M. C. (1991). Principal component analyses of assemblages structure data: Utility of tests based on eigenvalues. *Ecology, 72*(1), 341–347. doi:10.2307/1938927.
- Helena, B., Pardo, R., Vega, M., Barrado, E., Fernandez, J. M., & Fernandez, L. (2000). Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis. *Water Research, 34*, 807–816. doi:10.1016/S0043-1354(99)00225-0.
- Jackson, D. A. (1993). Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches. *Ecology, 74*, 2201–2214.
- Kennish, M. J. (1992). *Ecology of estuaries: Anthropogenic effects* (494 pp.). Florida: CRC.
- Morales, M. M., Mart, P., Llopis, A., Campos, L., & Sagrado, J. (1999). An environmental study by factor analysis of surface seawater in the Gulf of Valencia (western Mediterranean). *Analytica Chimica Acta, 394*, 109–117. doi:10.1016/S0003-2670(99)00198-1.
- Petersen, W., Bertino, L., Callies, U., & Zorita, E. (2001). Process identification by principal component analysis of river water-quality data. *Ecological Modelling, 138*, 193–213. doi:10.1016/S0304-3800(00)00402-6.
- Ross, P. J. (1988). *Taguchi techniques for quality engineering*. New York: McGraw-Hill.
- Schramm, W., & Nienhuis, P. H. (1996). *Marine benthic vegetation: Recent changes and the effects of eutrophication* (470 pp.). Berlin: Springer.

- Simeonov, V., Stratis, J. A., Samara, C., Zachariadis, G., Voutsas, D., Anthemidis, A., et al. (2003). Assessment of the surface water quality in Northern Greece. *Water Research*, *37*, 4119–4124. doi:[10.1016/S0043-1354\(03\)00398-1](https://doi.org/10.1016/S0043-1354(03)00398-1).
- Singh, K. P., Malik, A., Mohan, D., & Sinha, S. (2004). Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India): A case study. *Water Research*, *38*, 3980–3992. doi:[10.1016/j.watres.2004.06.011](https://doi.org/10.1016/j.watres.2004.06.011).
- Stevens, J. (1986). *Applied multivariate statistics for the social science* (515 pp.). Hillsdale: Erlbaum.
- Vega, M., Pardo, R., Barrado, E., & Deban, L. (1998). Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. *Water Research*, *32*, 3581–3592. doi:[10.1016/S0043-1354\(98\)00138-9](https://doi.org/10.1016/S0043-1354(98)00138-9).
- Williams, K. K., & Titus, K. (1988). Assessment and sampling stability in ecological applications of discriminate analysis. *Ecology*, *69*(4), 1275–1285. doi:[10.2307/1941283](https://doi.org/10.2307/1941283).
- Wunderlin, D. A., Diaz, M. P., Ame, M. V., Pesce, S. F., Hued, A. C., & Bistoni, M. (2001). Pattern recognition techniques for the evaluation of spatial and temporal variation in water quality. A case study: Suquia river basin (Cordoba Argentina). *Water Research*, *35*, 2881–2894. doi:[10.1016/S0043-1354\(00\)00592-3](https://doi.org/10.1016/S0043-1354(00)00592-3).