Journal of Mathematics and Statistics, 1 (4): 277-281, 2005
ISSN 1549-3644

Modelling the Effect of Salinity on the Multivariate Distribution of a Water Quality Index

Daniela Cocchi and Michele Scagliarini
Dipartimento di Scienze Statistiche "Paolo Fortunati", Via Belle Arti 41, 40126 Bologna Italy

Abstract: Salinity is a tracer of water and nutrient flows and is employed in studies on coastal environment to establish the coastal or remote origins of pollutants and nutrients. The present article analyses the effects of salinity on the multivariate distribution of the Trophic Index TRIX. The approach taken is based on a multivariate nested model with salinity as an explicative variable. We believe that our results are of use in discerning those areas of the North-West Adriatic Sea influenced by the Po River.

Key words: Multivariate normal distribution, trophic index, maximum likelihood estimation, variance-covariance structure, multivariate nested model

INTRODUCTION

The use of indicators to characterize the status of a specific environment is of considerable importance since these indicators play a fundamentally important role in linking policy objectives and targets, in communicating data priorities to different nations and in reporting complexity in simple ways that both policy makers and the public can understand.

In this study we are going to analyze the index used to evaluate the trophic state of the coastal environment: the trophic index TRIX. This topic is of considerable importance since the eutrophication due to overloading with nitrogen and phosphorus nutrients has produced changes in the structure and functioning of marine ecosystems, together with a reduced degree of biodiversity and falling income from fishery, mariculture and tourism.

The TRIX index is routinely used by Italian authorities to monitor the trophic condition of the Adriatic Sea. In particular, upper limits have been established for the yearly average value of the index for the whole coastline. In order to draw up suitable environmental policies, one has to remember that the trophic conditions in the Northern Adriatic Sea are mainly due to the contribution of the Po River, even though in such a complex system, local contributions are also of some importance.

It is important that regional authorities recognize the existence of the two sources of contributions, since they wield the power to intervene at the local level, whereas the effects due to the Po River are beyond their control. Detailed studies of nutrient flows are useful in determining the sources of those nutrients that contribute towards eutrophication and the ways in which nutrients are transported to the ecosystem and they may also suggest methods of combating this problem. Such studies require a wide range of approaches, technology and technical skills and are best addressed by the cooperation between multidisciplinary teams of scientists.

In this context, studies of the coastal water environment[1,2] indicate that salinity is a very peculiar tracer that can be used to interpret the coastal or remote origins of pollutants and nutrients. In the said works, "tracers" are used on the basis of deterministic models.

The aim of the present work is to set out the problem from a statistical point of view involving: a multivariate approach, which enables us to study the entire area using a single model; statistical methodology, which allows us to verify whether salinity has diverse effects on the trophic condition in the studied area.

THE INDEX AND THE DATA

The index TRIX was proposed in[3] as a means of evaluating the trophic state of coastal waters:

$$TRIX = (k/n)\sum_{i=1}^{n}\frac{\log X_i - \log L_i}{\log U_i - \log L_i} \qquad (1)$$

In (1), n is the number of considered components, X_i is the measured value of component i, while U_i and L_i are the upper and lower limits of each component. The n=4 components are: chlorophyll-a (X_1), oxygen saturation (X_2), mineral nitrogen (X_3) and total phosphorus (X_4). Their ranges are currently standardized to 3 log units, *i.e.* $\log U_i - \log L_i = 3$ and k=10 is an expansion factor for ranging the index out from the interval [0,1] to [0,10]. Variables X_1 and X_2 are direct expressions of productivity, while variables X_3 and X_4 are nutritional factors.

The data set consists of the weekly point measurements, taken from 1998 to 2002, of the four

Corresponding Author: Michele Scagliarini, Dipartimento di Scienze Statistiche, Università di Bologna, Via Belle Arti 41, 40126 Bologna, Italy, Tel. +39 051 2098253, Fax +39 051 232153

variables contributing towards the TRIX index and of salinity. Measurements come from thirteen monitoring sites along the Emilia-Romagna coast. This information has been kindly provided by the Emilia-Romagna Region and the Regional Agency for Environmental Protection (ARPA).

Five monitoring points are situated at a distance of 500 metres from the coastline: Lido di Volano (identification number: 1), Porto Garibaldi (id.no. 2), Cesenatico (id.no. 3), Rimini (id.no. 4) and Cattolica (id.no. 5); together they constitute the inshore subgroup. The other eight monitoring sites, which constitute the offshore subgroup, are situated at three kilometres out to sea from Lido di Volano (id.no. 6), Porto Garibaldi (id.no. 7), Cesenatico (id.no. 10) and Cattolica (id.no. 11) and at ten kilometres out to the sea from Lido di Volano (id.no. 8), Porto Garibaldi (id.no. 9), Cesenatico (id.no. 12), Cattolica (id.no. 13). The map of the monitoring network is shown in Fig. 1.

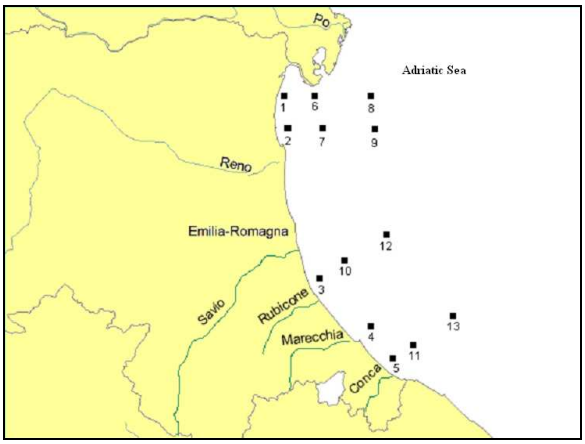

Fig. 1: Monitoring sites

Marine trophic conditions are evaluated by the public authorities according to the yearly average of the TRIX index. Since at present data sets are only available for the last few years, our work is based on monthly averages of the index.

The monthly averages at each site are well approximated by the normal distribution. Means, variances and the results of the Kolmogorov-Smirnov normality tests of the monthly data are given in Table 1.

In this coastal zone the water circulation pattern is influenced by several factors such as bottom bathymetry, surface winds, inputs from the Po River, etc.. This is also evident from the correlation matrix corresponding to the monthly TRIX average, reported in Table 2, featuring groups of highly correlated sites such as the four sites out to sea along the southern stretch of coastline (sites 10÷13) and the four sites out to sea along the northern part of the coast (sites 6÷9). Therefore, we may reasonably assume correlation among sites and thus adopt a multivariate modelling approach.

Table1: Sample means, variances and normality tests

Site	$\hat{\mu}_j$	$\hat{\sigma}_j^2$	$Pr(K \geq K_c)$
1	6.116	0.277	0.200
2	5.991	0.351	0.200
3	5.836	0.221	0.200
4	5.487	0.382	0.200
5	5.254	0.451	0.186
6	6.087	0.332	0.200
7	5.892	0.430	0.168
8	5.988	0.488	0.200
9	5.724	0.640	0.200
10	5.224	0.633	0.200
11	4.900	0.800	0.200
12	4.907	0.719	0.200
13	4.670	0.727	0.200

MULTINORMALITY TEST

Let Y_{vi} be the TRIX monthly average at site v (v=1,2,..,13) during month i (i=1,2,...,60). The multivariate process for the entire area under study may thus be considered to be as follows:

$$\mathbf{y} = [Y_1, Y_2, ..., Y_{13}]^T \quad (2)$$

with mean vector $\boldsymbol{\mu}_y$ and covariance matrix $\boldsymbol{\Sigma}_y$ where each component is normally distributed.

To verify whether $\mathbf{y}$ may be treated as a multinormal process, we adopt the test procedure proposed by[4].

Let $\mathbf{x}$ be a p-variate normal

$$\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (3)$$

Let $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ be a random sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We call

$$\mathbf{X}_{n \times p} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)^T \quad (4)$$

a data matrix from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, or simply a "normal data matrix".

The normal data matrix $\mathbf{X}$ can be written also as

$$\mathbf{X} = (\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, ..., \mathbf{x}_{(n)}) \quad (5)$$

Where $\mathbf{x}_{(j)}$ (j=1,2,...,p) is the n-vector whose elements denote the observations regarding the j-variable. Measures of multivariate skewness and kurtosis are, respectively,

$$\beta_{1,p} = \mathrm{E}\left\{(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{z}-\boldsymbol{\mu})\right\}^3 \quad (6)$$

$$\beta_{2,p} = \mathrm{E}\left\{(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right\}^2 \quad (7)$$

Where $\mathbf{x}$ and $\mathbf{z}$ are identically and independently distributed.

If $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then,

$$\beta_{1,p} = 0 \quad (8)$$

$$\beta_{2,p} = p(p+2) \quad (9)$$

Table 2: Correlation matrix

sites	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1.000	0.756	0.313	0.377	0.301	0.836	0.682	0.752	0.621	0.409	0.319	0.372	0.347
2		1.000	0.472	0.526	0.482	0.857	0.828	0.778	0.750	0.578	0.493	0.524	0.476
3			1.000	0.733	0.715	0.344	0.599	0.351	0.459	0.789	0.721	0.744	0.729
4				1.000	0.919	0.481	0.653	0.529	0.621	0.820	0.905	0.820	0.797
5					1.000	0.406	0.591	0.480	0.568	0.839	0.924	0.800	0.810
6						1.000	0.853	0.864	0.794	0.510	0.437	0.494	0.404
7							1.000	0.826	0.893	0.730	0.648	0.677	0.608
8								1.000	0.860	0.579	0.543	0.533	0.497
9									1.000	0.729	0.659	0.660	0.616
10										1.000	0.883	0.922	0.874
11											1.000	0.842	0.874
12												1.000	0.873
13													1.000

Let

$$\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}^T\mathbf{1} \tag{10}$$

be the sample mean vector and

$$\mathbf{S} = \frac{1}{n}\left(\mathbf{X}^T\mathbf{Z} - \frac{1}{n}\mathbf{X}^T\mathbf{1}\mathbf{1}^T\mathbf{X}\right) \tag{11}$$

be the sample covariance matrix, where **1** is a column vector of *n* ones.

The sample counterparts of the measures $\beta_{1,p}$ and $\beta_{2,p}$ are:

$$b_{1,p} = \frac{1}{n^2}\sum_{r,s=1}^{n} g_{rs}^3 \tag{12}$$

and

$$b_{2,p} = \frac{1}{n}\sum_{r=1}^{n} g_{rr}^2 \tag{13}$$

where

$$g_{r,s} = (\mathbf{x}_r - \bar{\mathbf{x}})^T\mathbf{S}^{-1}(\mathbf{x}_s - \bar{\mathbf{x}}) \tag{14}$$

and $\mathbf{x}_r$ denotes the *r*-th rows of **X**.

In[5] has been shown that $b_{1,p}$ and $b_{2,p}$ have the following asymptotic distributions

$$\frac{1}{6}nb_{1,p} \sim \chi_f^2, \text{ where } f=p(p+1)(p+2)/6 \tag{15}$$

and

$$\left\{b_{2,p} - p(p+2)\right\}\Big/\left\{8p(p+2)/n\right\}^{1/2} \sim N(0,1) \tag{16}$$

These statistics can be used to test the null hypothesis of multinormality:

$$H_0: \beta_{1,p} = 0 \tag{17}$$

$$H_0: \beta_{2,p} = p(p+2) \tag{18}$$

In our case $\mathbf{y}_i$ (i=1,2,...,n) is the row vector of the TRIX monthly average in month i (p=13 and n=60). The data gives us

$$b_{1,p} = 49.559 \tag{19}$$

$$\frac{1}{6}nb_{1,p} = 495.58 \tag{20}$$

$$b_{2,p} = 196.694 \tag{21}$$

and

$$\left\{b_{2,p} - p(p+2)\right\}\Big/\left\{8p(p+2)/n\right\}^{1/2} = 0.332 \tag{22}$$

The degrees of freedom for the chi-square are 455 and fixed $\alpha = 0.05$, the critical value is 506, therefore $(1/6)nb_{1,p}$= 495.58 is not significant. We may also note that the observed value 0.332 is not significant at the 5% level. The hypothesis of multinormality is therefore not rejected for the entire region under study; the multivariate distribution of the TRIX is suitably approximated by a multivariate Gaussian process.

THE PROPOSED MODEL

Salinity in coastal ecosystems is an important indication of inputs from coastal drainage basins and it can be used as a tracer of river plumes, nutrient and sediment loading.

In order to relate the multivariate distribution of the index ($\mathbf{y}\sim N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$) to salinity, we adopt a nested link structure; that is, we assume that the mean $\boldsymbol{\mu}_y$ of the multivariate Gaussian process can be expressed as a linear function of the tracer

$$\boldsymbol{\mu}_y = \mathbf{Z}\boldsymbol{\beta} \tag{23}$$

where

$$\mathbf{Z} = \begin{bmatrix} 1 & 1 & \ldots & 1 & 0 & 0 & \ldots & 0 \\ Z_1 & Z_2 & \ldots & Z_5 & & & \ldots & \\ 0 & 0 & \ldots & 0 & 1 & 1 & \ldots & 1 \\ 0 & 0 & \ldots & 0 & Z_6 & Z_7 & \ldots & Z_{13} \end{bmatrix}^T \tag{24}$$

Z_j is the salinity average for all data at site j (j=1,2,...,13) and

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{0A} & \beta_{1A} & \beta_{0B} & \beta_{1B} \end{bmatrix}^T \quad (25)$$

is a vector of parameters. The structure of $\boldsymbol{\beta}$ reflects our working hypothesis of a different effect of salinity in the two groups: *A)* inshore group (sites 1÷5); *B)* offshore group (sites 6÷13).

The spatial dependence of the observations is governed by bathymetry and by currents that control water movement. This leads to a complex pattern which we have incorporated in the covariance matrix $\boldsymbol{\Sigma}_y$, where we have assumed the existence of groups of dependent sites characterized by the homogenous circulation of sea water:

* the five sites (sites 1÷5) situated at 500 metres from the coast line with constant variance σ^2_A and constant correlation ρ_A
* the eight sites (sites 6÷13) of the offshore group are split into two sub-group
* sites 6÷9 (group B1) out to sea along the northern part of the coast line and directly influenced by the Po River;
* sites 10÷13 (group B2) out to sea along the southern part of the coast line.

Moreover, we assume the following: a constant correlation, ρ_B, among sites of group B and different variances in the two sub-groups σ^2_{B1} for sites 6÷9 and σ^2_{B2} for sites 10÷13; constant correlation between group A and group B: ρ_{AB}.

According to the above, we get

$$\boldsymbol{\Sigma}_y = \begin{bmatrix} \boldsymbol{\Sigma}_A & & \boldsymbol{\Sigma}_{AB} & \\ & & & \\ \boldsymbol{\Sigma}_{AB}^T & \boldsymbol{\Sigma}_{B1} & & \boldsymbol{\Sigma}_{BB} \\ & \boldsymbol{\Sigma}_{BB}^T & & \boldsymbol{\Sigma}_{B2} \end{bmatrix} \quad (26)$$

where

$$_{4\times4}\boldsymbol{\Sigma}_A = \begin{bmatrix} \sigma^2_A & \rho_A\sigma^2_A & \rho_A\sigma^2_A & \rho_A\sigma^2_A \\ \cdots & \sigma^2_A & \rho_A\sigma^2_A & \rho_A\sigma^2_A \\ \cdots & \cdots & \sigma^2_A & \rho_A\sigma^2_A \\ \cdots & \cdots & \cdots & \sigma^2_A \end{bmatrix} \quad (27)$$

$$_{4\times4}\boldsymbol{\Sigma}_{B1} = \begin{bmatrix} \sigma^2_{B1} & \rho_B\sigma^2_{B1} & \rho_B\sigma^2_{B1} & \rho_B\sigma^2_{B1} \\ \cdots & \sigma^2_{B1} & \rho_B\sigma^2_{B1} & \rho_B\sigma^2_{B1} \\ \cdots & \cdots & \sigma^2_{B1} & \rho_B\sigma^2_{B1} \\ \cdots & \cdots & \cdots & \sigma^2_{B1} \end{bmatrix} \quad (28)$$

$$_{4\times4}\boldsymbol{\Sigma}_{B2} = \begin{bmatrix} \sigma^2_{B2} & \rho_B\sigma^2_{B2} & \rho_B\sigma^2_{B2} & \rho_B\sigma^2_{B2} \\ \cdots & \sigma^2_{B2} & \rho_B\sigma^2_{B2} & \rho_B\sigma^2_{B2} \\ \cdots & \cdots & \sigma^2_{B2} & \rho_B\sigma^2_{B2} \\ \cdots & \cdots & \cdots & \sigma^2_{B2} \end{bmatrix} \quad (29)$$

$$_{4\times4}\boldsymbol{\Sigma}_{BB} = \begin{bmatrix} \rho_B\sigma_{B1}\sigma_{B2} & \rho_B\sigma_{B1}\sigma_{B2} & \rho_B\sigma_{B1}\sigma_{B2} & \rho_B\sigma_{B1}\sigma_{B2} \\ \cdots & \rho_B\sigma_{B1}\sigma_{B2} & \rho_B\sigma_{B1}\sigma_{B2} & \rho_B\sigma_{B1}\sigma_{B2} \\ \cdots & \cdots & \rho_B\sigma_{B1}\sigma_{B2} & \rho_B\sigma_{B1}\sigma_{B2} \\ \cdots & \cdots & \cdots & \rho_B\sigma_{B1}\sigma_{B2} \end{bmatrix} \quad (30)$$

$$_{5\times8}\boldsymbol{\Sigma}_{AB} = \begin{bmatrix} \rho_{AB}\sigma_A\sigma_{B1} & \cdots & \rho_{AB}\sigma_A\sigma_{B1} & \rho_{AB}\sigma_A\sigma_{B2} & \cdots & \rho_{AB}\sigma_A\sigma_{B2} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{AB}\sigma_A\sigma_{B1} & \cdots & \rho_{AB}\sigma_A\sigma_{B1} & \rho_{AB}\sigma_A\sigma_{B2} & \cdots & \rho_{AB}\sigma_A\sigma_{B2} \end{bmatrix} \quad (31)$$

The likelihood of the model is

$$L\left(\boldsymbol{\beta}, \sigma^2_A, \sigma^2_{B1}, \sigma^2_{B2}, \rho_A, \rho_B, \rho_{AB}; Y, Z\right) = (2\pi)^{\frac{-np}{2}} \left|\boldsymbol{\Sigma}_y\right|^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{y}_i - \mathbf{Z}\boldsymbol{\beta}\right)^T \boldsymbol{\Sigma}_y^{-1}\left(\mathbf{y}_i - \mathbf{Z}\boldsymbol{\beta}\right)\right) \quad (32)$$

and the maximum likelihood estimates of parameters are given in Table 3.

Table 3: Parameters estimates

Parameter	Estimates	Std. err.	p-value
β_{0A}	9.3164	0.3076	0.000
β_{1A}	-0.1167	0.0098	0.000
β_{0B}	11.2243	0.2104	0.000
β_{1B}	-0.1909	0.0075	0.000
σ^2_A	0.3566	0.0416	0.000
σ^2_{B1}	0.4782	0.0686	0.000
σ^2_{B2}	0.6935	0.0990	0.000
ρ_A	0.5097	0.0608	0.000
ρ_B	0.6974	0.0434	0.000
ρ_{AB}	0.6126	0.0474	0.000

RESULTS

For the mean of the process the results confirm a common linear dependence pattern. Table 4 shows, for each site, the evaluation $\tilde{\mu}_j$ obtained by $\tilde{\boldsymbol{\mu}}_y = \mathbf{Z}\hat{\boldsymbol{\beta}}$.

The fit of the estimated model is good (Fig. 2) and the mean of the relative absolute differences between the sample means $\hat{\mu}_j$ (Table 1) and $\tilde{\mu}_j$ is 1.07%.

Table 4: Values of $\tilde{\mu}_j$ obtained by $\tilde{\boldsymbol{\mu}} = \mathbf{Z}\hat{\beta}$

Site	Z_j	$\tilde{\mu}_j = \hat{\beta}_{0A} + \hat{\beta}_{1A} Z_j$ Inshore	$\tilde{\mu}_j = \hat{\beta}_{0B} + \hat{\beta}_{1B} Z_j$ offshore
1	27.228	6.139	
2	28.317	6.012	
3	31.994	5.583	
4	32.729	5.497	
5	33.283	5.432	
6	27.225	-	6.027
7	27.861	-	5.906
8	27.295	-	6.014
9	28.784	-	5.729
10	32.025	-	5.111
11	33.216	-	4.883
12	32.920	-	4.940
13	34.240	-	4.688

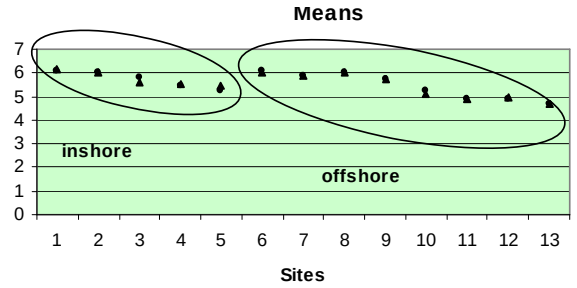

Fig. 2: Fit for the mean level •= $\hat{\mu}_j$ and ▲= $\tilde{\mu}_j$

The 95% confidence intervals for β_{0A} [8.708, 9.925] and for β_{0B} [10.808, 11.640] are disjointed. The same holds for β_{1A}, for which the 95% confidence interval is [-0.136, -0.097] and for β_{1B} with a 95% confidence interval of [-0.206, -0.176].

This means that the effect of salinity on the mean of the index distribution follows a common pattern (linear), albeit with a significant difference in intensity.

As far as the variance and correlation parameters are concerned, the mean of the relative absolute differences between the sample variances $\hat{\sigma}_j^2$, Table 1 and the estimate parameters in Table 3 is 17.9%. The fit is also displayed in Fig. 3. While for the correlations, the overall difference is 27.4%.

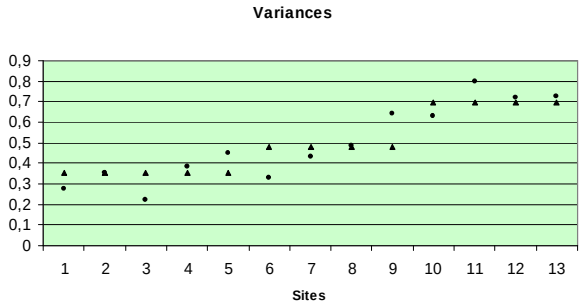

Fig. 3: Fit for the variances •= $\hat{\sigma}_j^2$, ▲= $\hat{\sigma}_A^2$ (sites 1÷5), ▲= $\hat{\sigma}_{B1}^2$ (sites 6÷9), ▲= $\hat{\sigma}_{B2}^2$ (sites 10÷13)

CONCLUSION

An understanding of the connection between water quality in coastal areas and river catchments is important if we are to remedy current environmental effects and be in a position to plan future options so as to improve the coastal ecosystem.

The problem of how to define the areas influenced by fluvial discard is a substantial one for which there is no one single solution. The literature includes several contributions in which salinity is used, in deterministic models, as a tracer in order to distinguish the origin of nutrients and pollutants.

In this study, a multivariate model has been used to identify the effects of salinity on the multivariate distribution of the index TRIX. In particular, the effects of the tracer are relevant on the mean of the TRIX distribution, while the covariance matrix is substantially given by the sea current pattern.

Results have revealed two groups of monitoring sites with significant differences in the characteristics of the relationship. The differences can be seen as evidence of the existence of two different situations: the offshore monitoring points subject to the influence of the Po river and the inshore monitoring sites mainly influenced by local factors such as urban discharges or nutrients carried in to the sea by local rivers.

ACKNOWLEDGMENTS

This work was partly funded by a 2004 grant (Sector 13: Economics and Statistics, Project protocol no. 2004137478 for Research Projects of National Interest by MIUR).

REFERENCES

1. Pitts, A.P. and N.P. Smith, 1998. Tracking Florida bay water across hawk channel using salinity as a natural tracer. Harbor Branch Oceanographic Institution, Ft. Pierce, Florida
2. Burrage, D., J. Miller, D. Johnson, J. Wesson and J. Johnson, 2002. Observing sea surface salinity in coastal domains using an airbone surface salinity mapper. Proc. Marine Technol. Soc., Oceans 2002 MTS/IEEE Conf., pp: 2014-2015.
3. Vollenweider, R.A., F. Giovanardi, G. Montanari and A. Rinaldi, 1998. Characterization of the trophic conditions of marine coastal waters with special reference to the nw adriatic sea: proposal for a trophic scale, turbidity and generalized water quality index. Environmetrics, 9: 329-357.
4. Mardia, K.V., J. T. Kent and J.M. Bibby, 1979. Multivariate Analysis, Academic Press: London.
5. Mardia, K.V., 1970. Measures of multivariate skewness and kurtosis with applications. Biometrika, 57: 519-530.