

A statistical model for evaluating a water quality index¹

Un modello statistico per valutare un indice di qualità dell'acqua

Daniela Cocchi, Michele Scagliarini

Dipartimento di Scienze Statistiche “Paolo Fortunati”, Università di Bologna

e-mail: cocchi@stat.unibo.it, scagliar@stat.unibo.it

Riassunto: Il lavoro ha l'obiettivo di studiare come la distribuzione dell'indice di qualità dell'acqua marina adottato nella legislazione italiana dipenda dalla salinità. Si propone un modello Gaussiano multivariato in cui si ipotizza che i parametri dipendano dalla salinità tramite iperparametri. I risultati mettono in luce l'effetto non omogeneo della salinità sulla struttura parametrica della distribuzione multivariata dell'indice.

Keywords: multivariate normal distribution, trophic index, maximum likelihood estimation, variance-covariance structure, linear link function.

1. Introduction

The TRIX index has been proposed by Vollenweider *et al.* (1998) for evaluating the trophic state of coastal waters:

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

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 for 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. Index (1) is regularly used by Italian authorities for monitoring the trophic condition of the Adriatic Sea. In particular, upper limits for the yearly average value of the index have been given for the entire coastline (DLGS 152/99). In order to design 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, although in such a complex system, local factors also have to be taken into consideration. It is important that regional authorities acknowledge the existence of the two sources of contributions: while these authorities have the power to intervene at the local level, the effects of the Po River are beyond their control. Studies of coastal water pollution (Pitts and Smith, 1998) have shown that salinity is a very

¹ The work is partially funded by a 2002 grant (Sector 13: Economics and Statistics, Project n. 7, protocol n. 2002134337_01 for Research Projects of National Interest by MIUR). We thank R. Pastres for the helpful discussion about the argument of this paper.

special tracer that can be used to establish the coastal or remote origin of pollutants and nutrients. In this study we are going to use salinity as a covariate in a regional, *i.e.* multivariate, model for the TRIX index.

2. The data

The data set consists of weekly point measurements, taken from 1992 to 2002, of the four variables contributing towards the TRIX index, and of salinity. Measurements come from 13 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). The following five monitoring points are situated at 500 metres from the coastline, and as such constitute the inshore subgroup: Lido di Volano (number 1), Porto Garibaldi (no. 2) Cesenatico (no. 3), Rimini (no. 4) and Cattolica (no. 5). The other eight monitoring sites, comprising the offshore subgroup, are situated 3 kilometres out to sea from Lido di Volano (no. 6), Porto Garibaldi (no. 7), Cesenatico (no. 10) and Cattolica (no. 11), and at 10 kilometres out to sea from Lido di Volano (no. 8), Porto Garibaldi (no. 9), Cesenatico (no. 12), Cattolica (no. 13). The sea's trophic state is evaluated by the public authorities according to the yearly average of the TRIX index. Since only 11 years of data are presently available, our work is based on monthly averages of the index. The distributions of the monthly average in each site are well approximated by a normal distribution. Means and standard deviations of the monthly data are shown in Table 1 below.

Table1: Means and standard deviations of TRIX monthly averages (1992-2002)

Site	1	2	3	4	5	6	7	8	9	10	11	12	13
$\bar{Y}_s$	6.146	6.053	5.783	5.399	5.168	6.068	5.865	5.989	5.639	5.157	4.847	4.870	4.640
$s(Y)_s$	0.469	0.518	0.522	0.623	0.658	0.524	0.576	0.603	0.767	0.824	0.869	0.861	0.850

3. A regional model for the TRIX index as a function of salinity

Let Y_{si} be the monthly TRIX average at site s ($s=1, 2, \dots, 13$; $i=1, 2, \dots, 132$). Exploiting the normality of the Y_s components, we assume, for the entire region in question, a multivariate Gaussian process $\mathbf{Y}|\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}=[\mu_1, \mu_2, \dots, \mu_{13}]^T$ and $\boldsymbol{\Sigma}$ is the (13×13) covariance matrix of the process. The local parameters of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be expressed as functions of a suitable synthesis of salinity, Z . We have examined several possible types of link function using theoretical knowledge and empirical evidence, and consequently suggest that covariate Z_s consists of the average salinity at site s ($s=1, \dots, 13$). We have built a stratified model whereby sites are assigned to two groups: A) inshore group (sites 1÷5); B) offshore group (sites 6÷13). The mean of the process is linearly dependent upon salinity, with different parameters for each sites group:

$$\boldsymbol{\mu} = \mathbf{Z}\boldsymbol{\beta} \quad (3)$$

$$\text{where } \mathbf{Z} = \begin{bmatrix} 1 & 1 & \dots & 1 & 0 & 0 & \dots & 0 \\ Z_1 & Z_2 & \dots & Z_5 & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 1 & 1 & \dots & 1 \\ 0 & 0 & \dots & 0 & Z_6 & Z_7 & \dots & Z_{13} \end{bmatrix}^T \text{ and } \boldsymbol{\beta} = [\beta_{0A} \quad \beta_{1A} \quad \beta_{0B} \quad \beta_{1B}]^T.$$

Guided by an explorative study, not reported here, together with *a priori* knowledge we suggest that the covariance matrix consists of a block diagonal structure:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_A & 0 \\ 0 & \boldsymbol{\Sigma}_B \end{bmatrix}, \quad \text{where} \quad \boldsymbol{\Sigma}_B = \begin{bmatrix} \boldsymbol{\Sigma}_{B1} & 0 \\ 0 & \boldsymbol{\Sigma}_{B2} \end{bmatrix}. \quad (4)$$

In (4), $\boldsymbol{\Sigma}_A = (\mathbf{Z}_A \boldsymbol{\tau}_A)^2 \mathbf{I}_5$ with $\mathbf{Z}_A = [Z_1 \quad Z_2 \quad \dots \quad Z_5]^T$; it is also $\boldsymbol{\Sigma}_B$ (8×8) a block diagonal, with different covariance structures for the northern and southern part of the coast: $\sigma_{ii}^2 = (\tau_{B1})^2$, for $i=6, \dots, 9$ and $\sigma_{ij} = \tau_{B11}$ for $i,j=6, \dots, 9 \quad i \neq j$; $\sigma_{ii}^2 = (\tau_{B2})^2$ for $i=10, \dots, 13$ and $\sigma_{ij} = \tau_{B22}$ for $i,j=10, \dots, 13 \quad i \neq j$. Given this model, the variability of inshore sites directly depends on salinity, and sites are independent of each other. In the case of the offshore sites, two sub-groups have been identified: B1) the four sites situated along the northern part of the coastline; and B2) the four sites along the southern section of the coastline. For both sub-groups, constant variances and covariances have been assumed. The likelihood of the model is ($N=132$):

$$L(\boldsymbol{\beta}, \tau_A, \tau_{B1}, \tau_{B11}, \tau_{B2}, \tau_{B22}; \mathbf{Y}, \mathbf{Z}) = (2\pi)^{-N/2} |\boldsymbol{\Sigma}|^{-N/2} \exp\left(-\frac{1}{2} \sum_{i=1}^N (\mathbf{Y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_i - \boldsymbol{\mu})\right) \quad (5)$$

The maximum likelihood estimates of hyperparameters are reported in Table 2.

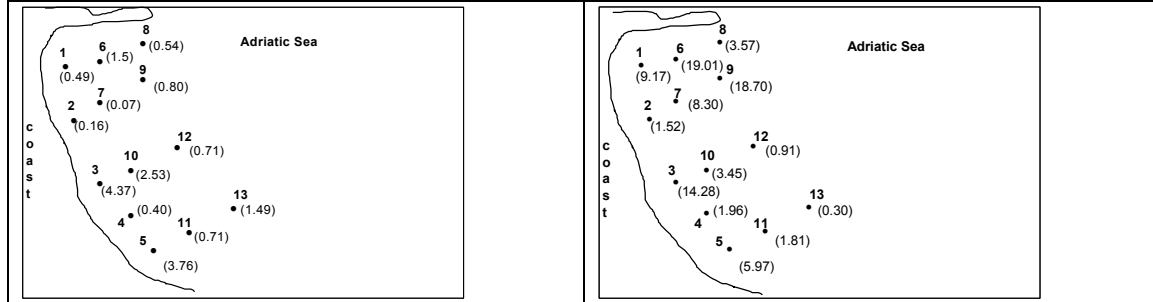
Table 2: Estimates of hyperparameters

Par.	Estimates	Std. err.	p-value	Par.	Estimates	Std. err.	p-value
β_{0A}	10.0768	0.2766	0.0000	τ_A	0.0187	0.0005	0.0000
β_{1A}	-0.1426	0.0091	0.0000	τ_{B1}	0.6241	0.0321	0.0000
β_{0B}	12.2951	0.3053	0.0000	τ_{B2}	0.8528	0.0488	0.0000
β_{1B}	-0.2274	0.0102	0.0000	τ_{B11}	0.2979	0.0398	0.0000
				τ_{B22}	0.6500	0.0831	0.0000

In the case of the process mean, results confirm a pattern of common linear dependency, with a different influence exercised by salinity in the two strata: the differences between the values of parameters β_{0A} , β_{0B} and β_{1A} , β_{1B} are all statistically significant, since their confidence intervals, not shown here, are not overlapping. With regard to variability, the differences between the pairs of parameter estimates - τ_{B1} and τ_{B2} , and τ_{B11} and τ_{B22} - are once again statistically significant. The performance of the model is evaluated by comparing the site overall sample means and variances shown in Table (1) with the estimates of the $\boldsymbol{\mu}$ vector obtained by (3) and the diagonal elements of $\boldsymbol{\Sigma}$ obtained by (4), shown in Table 3 below. For the mean level, the fit of the estimated model is good. Figure 1-left shows the respective percentage difference between the two quantities (in brackets) for each site. Figure 1-right shows the respective percentage difference between the variances in Table 1 and the estimates of the diagonal in $\boldsymbol{\Sigma}$ (in brackets) for each site.

Table 3: Evaluations of μ_s and σ_s by (3) and (4)

Site	1	2	3	4	5	6	7	8	9	10	11	12	13
$\tilde{\mu}_s$	6.176	6.063	5.529	5.420	5.362	6.066	5.898	6.078	5.685	5.026	4.750	4.834	4.568
$\tilde{\sigma}_s$	0.512	0.526	0.596	0.611	0.618	0.624	0.624	0.624	0.624	0.853	0.853	0.853	0.853

Figure 1: Relative percent differences between overall site means and variances and their estimates by a regional model

4 Conclusions

The present work develops the investigation of the influence of salinity on the TRIX index computed at different sites, using an overall regional model. Salinity is a fundamental tracer of the index's components. The analysis does not take dynamic features into consideration since it aims to evaluate the impact of salinity on the distribution of the index. The main result is the evident effect of salinity on the average level of the index's distribution. A clear-cut distinction between inshore and offshore sites is revealed not only by intuition, but also by statistical tools. The strong differences between the means of the estimated parameters is a sign of the diverse effects salinity has on the TRIX distribution. The model performs less well in variance estimation, and further investigation is needed to improve it. However, the proposal of further stratification of the offshore group, based on proximity to the Po River delta, would seem promising. On the basis of the proposed model, we thus conclude that there is a difference between local inshore effects, which may be dealt with by local authorities, and the more general effects of human settlement of the entire Po River valley, which are beyond their control.

References

- DLGS 11 maggio 1999, n. 152 and subsequent modifications.
- Pitts A. P., Smith, N.P. (1998) *Tracking Florida Bay Water Across Hawk Channel Using Salinity As A Natural Tracer*. Harbor Branch Oceanographic Institution, Ft. Pierce, Florida
- Vollenweider, R.A., Giovanardi, F., Montanari, G., Rinaldi, A. (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.