



EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre



Institute for Environment and Sustainability
Inland and Marine Waters Unit
Soil and Waste Unit
21020 Ispra (VA) – Italy

TN No. I.04.xxx

PRELIMINARY STUDY OF SPATIO-TEMPORAL DATA ANALYSIS TOOLS FOR THE MANAGEMENT OF COASTAL LAGOONS

DITTY PROJECT (Development of an information technology tool for the management of Southern European lagoons under the influence of river-basin runoff)

(EESD Project EVK3-CT-2002-00084)

J. M. Zaldívar, F. Somma, F. Bouraoui

European Commission, Joint Research Centre, Institute for Environment and Sustainability, Ispra, Italy

M. Austoni, G. Giordani, P. Viaroli

Department of Environmental Sciences, University of Parma, Italy

M. Plus

Station d'Arcachon, Institut Français de Recherche pour l'Exploitation de la Mer (Ifremer), France

April 2004

DELIVERABLE D4

DISTRIBUTION LIST

JRC Internal

Austoni, M. (3 copies)
Bouraoui, F. (3 copies)
Somma, F. (3 copies)
Zaldívar, J.M. (3 copies)

External Members

MUMM, Belgium

P. Luyten, A. Norro

ECOLAG, France

C. Aliaume, T. Do Chi, M. Trousellier

Hydrosciences, France

J.F. Boyer, B. Picot, M.G. Tournoud
S. Payraudeau

LAMETTA, France

F. Valette

IPIMAT/CRIP, Portugal

M. Falção

Aegean University, Greece

G. Tsirtsis, D. Kitsiou, T. Nitis

University of Siena, Italy

C. Mocceni, L. Torsello

Municipality of Gera, Greece

P. Vogiatzis

Cepralmar, France

N. Mazouni

Ifremer, France

A. Chapelle, A. Fiandrino, L. Loubersac,
M. Plus

University of Parma, Italy

A. Bodini, A. Carletti, G. Di Leo
G. Giordani, P. Viaroli

Fernando Pessoa University, Portugal

P. Duarte

Murcia University, Spain

J. Martinez, M.A. Esteve

Provincia di Ferrara, Italy

S. Bencivelli, P. Magri

ICN, Portugal

J.C. Barros

DIREN, France

D. Crepin

For Endorsement

Head of Unit:

Competent Director: (if the distribution list includes people or organizations external to the JRC)

Name : S. EISENREICH

Name : M. GRASSERBAUER

Date :

Date :

Signature :

Signature :

Legal Notice

The information contained in this document may not be disseminated, copied or utilized without the written authorization of the Commission. The Commission reserves specifically its rights to apply for patents or to obtain other protection for the matters open to intellectual or industrial protection.

The distribution of this document is limited to the persons given in the distribution list.

Neither the Commission of the European Communities nor any person acting on behalf of the Commission is responsible for the use, which might be made of the following information.

CONTENTS

| | |
|---|----|
| 1. INTRODUCTION | 5 |
| 2. TEMPORAL ANALYSIS TOOLS | 5 |
| 2.1. Simple descriptive techniques | 7 |
| 2.2. Trend analysis | 8 |
| 2.3. Spectral analysis/Filtering | 9 |
| 2.4. Hurst exponent of a time series | 11 |
| 2.5. Non-linear time series analysis techniques | 13 |
| 2.6. Problems in environmental time series | 20 |
| 2.7. Multiple time series analysis | 23 |
| 3. SPATIAL ANALYSIS TOOLS | 24 |
| 3.1. Exploratory spatial data analysis | 25 |
| 3.2. Structural analysis: the intrinsic model | 28 |
| 3.3. Residuals in kriging | 30 |
| 3.4. Best linear unbiased estimation (BLUE) | 31 |
| 3.5. Model validation | 31 |
| 3.6. Anisotropy | 32 |
| 3.7. Variable mean models | 33 |
| 3.8. Multivariate analysis | 33 |
| 4. EXISTING SOFTWARE | 39 |
| 5. CONCLUSIONS | 40 |
| REFERENCES | 41 |
| APPENDIX 1. QUESTIONNAIRE FOR END-USERS | 44 |

DITTY PROJECT (Development of an information technology tool for the management of Southern European lagoons under the influence of river-basin runoff)
(EESD Project EVK3-CT-2002-00084)

Deliverable n° 4: Data analysis tools

Preliminary study on analysis tools to be implemented to facilitate the data analysis in coastal lagoons. (**WP1: Data Compilation** *Task 1.4. Study and implementation of data analysis tools to be implemented*).

Objective: The **WP1: Data Compilation** has as its main objective to collect all the relevant information concerning data, modeling approaches, relevant projects, and studies, and to decide in collaboration with the end-users on selected priority problems to be analyzed at each test site. Also during this phase data analysis techniques (linear and non-linear time series analysis, statistical analysis) will be developed. The best suited approaches for distinguishing important aspects of the historical time series, i.e. trends, seasonal variations, etc. will afterwards be implemented into several components of the DSS. Data analysis techniques to be implemented will be decided in collaboration with end users and stakeholders

Summary

The objective of this preliminary work is to present the state of the art of data analysis techniques to assess temporal and spatial trends, seasonal variations, spatial distribution, etc. in the data routinely collected for the management of coastal lagoons. Furthermore, a questionnaire has been developed to select between these techniques, which are the most suitable for coastal lagoons and at which level they should be implemented. Specific software that could be implemented in the monitoring of the lagoon has also been developed. Finally, existing open source software able to carry out several of these analysis techniques has also been reviewed.

1. INTRODUCTION

A time series is a set of observations on a variable, x , made sequentially in time, $\{x(t_1), x(t_2), \dots, x(t_n)\}$. If in addition to possessing temporal reference, the set of observations have a spatial reference, then let us assume that the location may be specified as a vector \mathbf{s} that contains all the necessary information to identifying the spatial location of the value $x(t_i)$; in this case the data might be expressed in the form: $\{x(i, t_1), x(i, t_2), \dots, x(i, t_n) / \mathbf{s}(i)\}_{i=1, \dots, m}$.

Whereas time series analysis is concerned with: the description of the salient features of the series, the explanation of the variation of another time series, and the prediction of its future values, spatial data analysis represents a collection of techniques and models that explicitly use the spatial referencing associated with each data value that is specified within the system under study (Haining, 2003). Also in this case spatial data analysis includes amongst other activities the identification of data properties, formulating hypothesis and drawing out useful information from data.

The objective of this preliminary work is to present the state of the art of data analysis techniques to assess trends, seasonal variations, cyclical fluctuations, random component, spatial distribution, etc. in the data routinely collected for the management of coastal lagoons. Examples of spatio-temporal time series have been taken from several lagoons from the DITTY project. It should be noted that whereas the main body of time series analysis has been developed for usually equally spaced observations, for the case of environmental time series this is not always the standard situation. Furthermore, another characteristic of these series is the existence of intervals with no data points due to problems with the instrumentation or lack of funding to continue the sampling campaign. For these reasons, most of the existing analysis tools can not always be applied directly to our series of interest.

Specific software has been developed for the case of early detection of anoxic conditions, whereas the use of existing open source software able to carry out several of these analysis techniques has also been reviewed. Finally, a questionnaire has been developed for selecting, between these techniques, the ones that are most suitable for coastal lagoons and the level at which they should be implemented, i.e. inside the database, stand-alone software routines, inside the GIS, in the monitoring devices (Appendix 1).

2. TEMPORAL ANALYSIS TOOLS

As mentioned before, a time series is a collection of observations made sequentially in time. Many types of time series occur in environmental studies. For example, the values of water

temperature and salinity on a sampling station in Sacca di Goro from 1989 to 1998 are presented in figures 1 and 2. Other examples include rainfall, air temperature measured in successive hours, days or months, nutrient concentrations in inland and marine waters, etc.

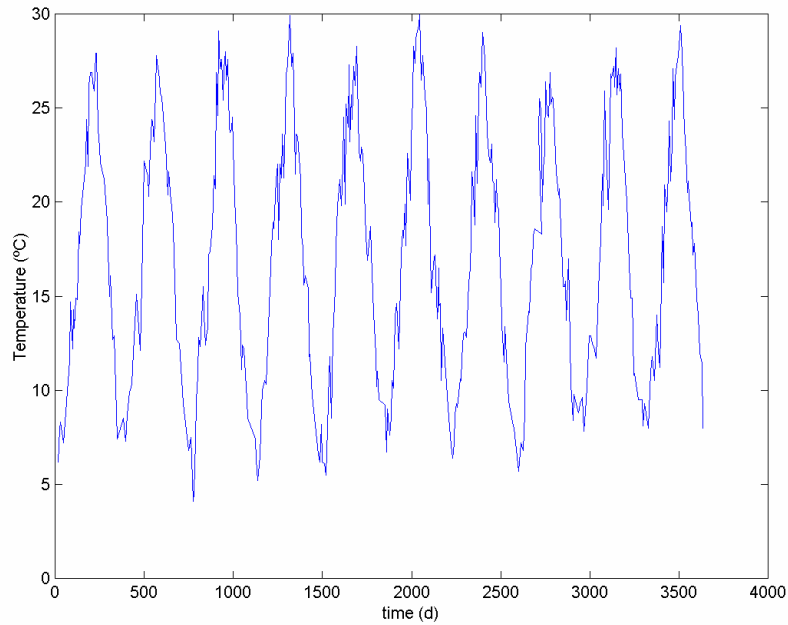


Figure 1. Temperature in Sacca di Goro (central buoy) from 1989 to 1998.

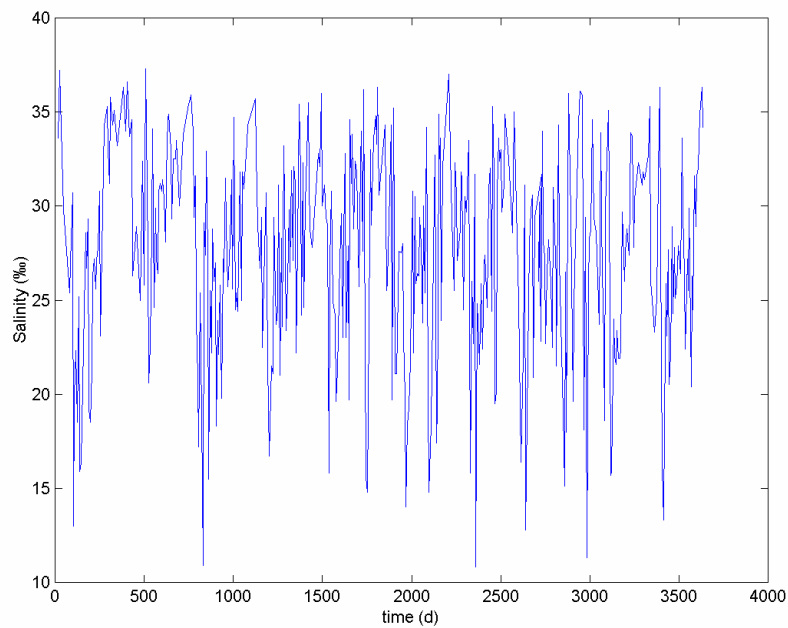


Figure 2. Salinity in Sacca di Goro (central buoy) from 1989 to 1998.

Methods of analyzing time series constitute an important area of statistics. There are several possible objectives in analyzing a time series. These objectives may be classified as (Chatfield, 1997): description, explanation, prediction and control. Most of the time-series literature is concerned with linear methods and models. Despite their simplicity, linear methods often work

well and may provide an adequate approximation for the task at hand. However, non-linear time series are more the normality than the exception and, hence, non-linear time series analysis tools have been recently developed (Abarbanel, 1996; Kantz and Shreiber 1997; Diks, 1999; amongst others) in statistics as well as in physics.

2.1. Simple descriptive techniques

Traditional methods of time series analysis are mainly concerned with decomposing the variation in a series into trend, seasonal variation, other cyclic changes and the remaining irregular fluctuations (Chatfield, 1997). By trend we mean a long-term change in the mean level; by seasonal effect we assume variations that are annual in period, as for example the temperatures of Sacca di Goro in figure 1; by other cyclic changes we consider other types of variations that do not have an annual period, for example daily variations of temperature; by irregular fluctuations we consider the residuals left after removing the other variations. These residuals may or may not be random. Even though this decomposition approach is sometimes not adequate, it is suitable when trend and/or seasonality dominate the variation in the time series.

Transformations

One evident concept is that the relationship between the quantities in the variable and the information is designed to present should be commensurate. If, for example, there are data which naturally falls into clusters or extreme events that condition the maximum of a variable then transformation of the data may be necessary. For example, data in clusters could be better represent by histograms; log transformations are good compressors, square root is relatively mild; higher roots also may be used. A typical problem with environmental data is the discontinuities in the time series. Several techniques have been developed but all of them present some drawbacks and, if possible, insertion of “estimated” data should be avoided. Chatfield (1997) considers several cases in which logarithmic or square root transformation may be necessary:

- To stabilize the variance
- To make the seasonal effect additive
- To make the data normally distributed

Scaling

Linear time series analysis models are immune to scaling difficulties. However, for non-linear models, e.g. neural networks, scaling may be a problem mainly when more than one variable is used. Furthermore, multivariate non-linear time series analysis implicitly or explicitly assumes that variables having a large variation are more important than variables having a small variation. Normally, the recommended (Masters, 1995) procedure for each variable separately should consist on the following steps:

- Compute the mean (in case of outliers use the median)

- Subtract it to the series
- Compute the standard deviation
- Divide each observation by the standard deviation
- Use this centred and scaled data

2.2. Trend analysis

Detection of temporal trends is one important objective in environmental monitoring. Therefore, a variety of statistical procedures have been developed to distinguish between random fluctuations and temporal changes in environmental variables. The easiest, but not always correct, procedure consists in computing a straight line that minimizes the mean squared difference between the line and the observed data. More adequate statistical procedures have been developed for univariate and multivariate systems- to take into account spatially distributed systems. These tests which are divided into parametric and non-parametric test, have been recently reviewed by El-Shaarawi (1993) and Esterby (1996). Here we only report in a test that is widely used when dealing with environmental data.

Non parametric Mann-Kendall test

The non-parametric test of Mann-Kendall (Hirsh *et al.*, 1982) does not make any assumption regarding the data distribution. Furthermore, it deals with incomplete, seasonal data with serial dependence, and any type of trend (linear and non linear). The basic principle of this test is to examine the sign of all pair wise differences of observed values. The first step of the test is to determine the sign of the $n(n-1)/2$ differences between the pairs $(x_j;x_k)$ with $j>k$ and compute the Mann-Kendall S with the following convention:

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^n \text{sgn}(x_j - x_k) \quad \text{where} \quad \begin{aligned} \text{sgn}(x_j - x_k) &= 1 \quad \text{if } x_j - x_k > 0; \\ \text{sgn}(x_j - x_k) &= 0 \quad \text{if } x_j - x_k = 0; \\ \text{sgn}(x_j - x_k) &= -1 \quad \text{if } x_j - x_k < 0. \end{aligned} \quad (1)$$

where x_1, \dots, x_n are the variables ordered in a chronological way, and n is the number of points to be analyzed. For small data sets ($n \leq 40$), the null hypothesis (H_0) of no trend versus an increasing trend hypothesis (H_1) is tested by comparing the probability of obtaining a value of S , $p(S)$, found in Kendall's table (1975) and the significance level of the test, α . The null hypothesis is rejected if $p(S)$ is larger than α . For larger data sets, it is necessary to determine the variance of S (Gilbert, 1987) to compute the Z test statistics is then computed as follows:

$$\begin{aligned} Z &= \frac{S-1}{\text{VAR}(S)^{0.5}} \quad \text{if } S > 0; \\ Z &= 0 \quad \text{if } S = 0; \\ Z &= \frac{S+1}{\text{VAR}(S)^{0.5}} \quad \text{if } S < 0. \end{aligned} \quad (2)$$

If the null hypothesis of no trend (H_0) is true, Z follows a standard normal distribution. The value of Z is thus compared to the standard normal probability distribution associated with the significance level, which was assumed to be 0.05 for the rest of the analysis. Additional details about the test can be found in Hirsh *et al.* (1982) and Gilbert (1987). Existing open-source software to perform the test is reviewed in Section 4. As an example, the results of the application of this technique to data from Thau lagoon on MPN (Most Probable Number) measurements of *E. coli* are presented in figure 3.

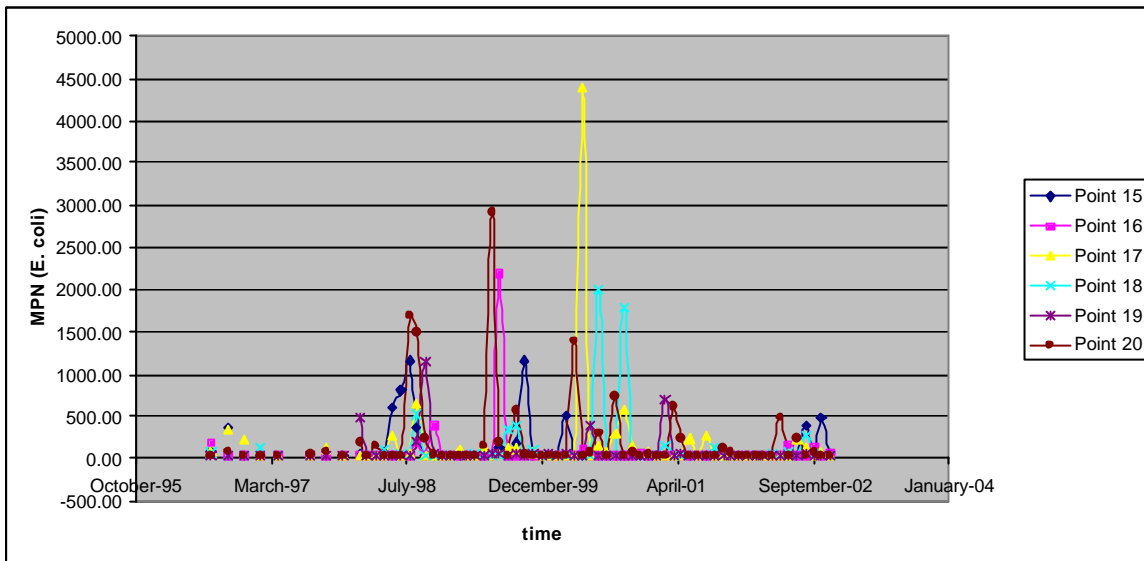


Figure 3. MPN (*E. Coli*) measured at several stations in Etang de Thau.

Using data from Ifremer REMI monitoring network and taking monthly averages for stations 15- 20, by applying Mann-Kendall test one obtains that there is no global tendency whereas exist for station 15 a decreasing tendency for July, for station 16 a decreasing tendency for August and for stations 17 and 19 a decreasing tendency for September.

2.3. Spectral analysis/Filtering

The Fourier transform establishes a one-to-one correspondence between the signal in the time domain and in the frequency domain, ie. how certain frequencies contribute to the signal. The power spectrum of a measured signal $s(t)$ helps in distinguishing among periodic, quasiperiodic and complex motions (chaotic or random) and is typically used to study stationary signals. It can be easily computed using standard software available¹, i.e. using the Fast Fourier Transform (FFT) algorithm.

¹ There are many software packages available to determine the FFT of a given signal, for example in Numerical Recipes, MATLAB, etc.

In principle, when analyzing the power spectrum of $s(t)$ we can obtain different type of information. More specifically, the power spectrum of a periodic motion with frequency f_1 has peaks at f_1 and its harmonics $2f_1, 3f_1, \dots$. A quasiperiodic motion with m rationally independent frequencies f_1, \dots, f_m has peaks at all the linear combinations of the basic frequencies with integer coefficients. In case of noise or chaotic signals we will obtain a similar power spectrum characterized by a high broadband distribution and a continuous spectra.

In practice we always have a finite discrete time series with a limited precision and therefore all peaks become finite in both height and width and there will be noise contamination. Furthermore, as said before if the time series has a trend, it should be removed from the data before carrying out a spectral analysis. In case of seasonality it would also be convenient to remove it, since any other effects will be relatively small and are unlikely to be visible in the spectrum of the raw data.

There are various features to look for once the spectrum of a given time series has been estimated (Chatfield, 1997): are there any peaks in the spectrum? is the spectrum large at low frequency, indicating possible non-stationarity in the mean? what is the general shape of the spectrum? Spectral analysis may be useful for some series in which there is no obvious trend or seasonal variation. As an example, in figure 4 we have analyzed water level data for 1992 at the Ravenna station, provided by APAT (Servizio Mareografico Nazionale). These data has been used as forcing function for the 3-D model of Sacca di Goro (Marinov *et al.*, 2004).

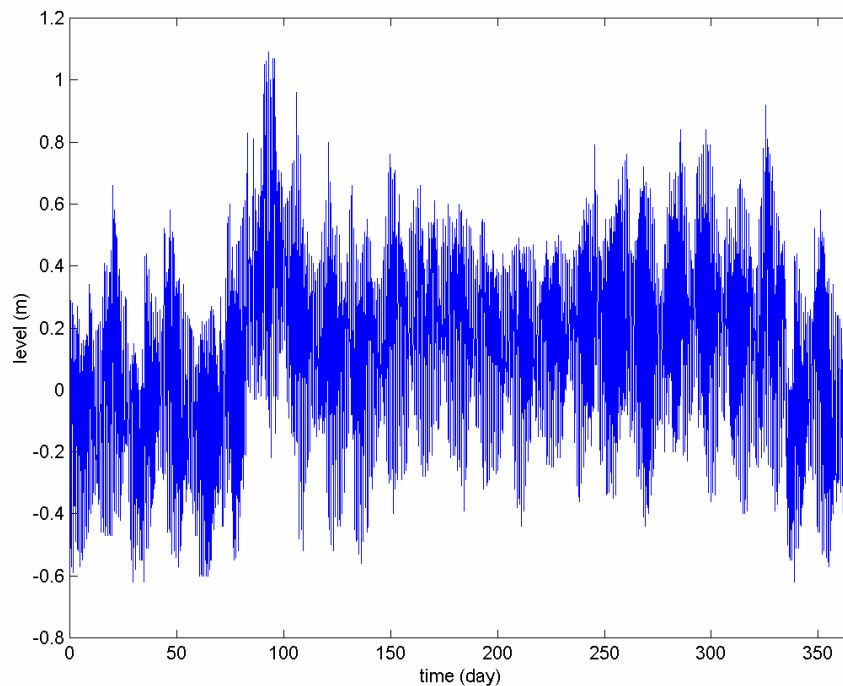


Figure 4. Water level at Ravenna station during 1992.

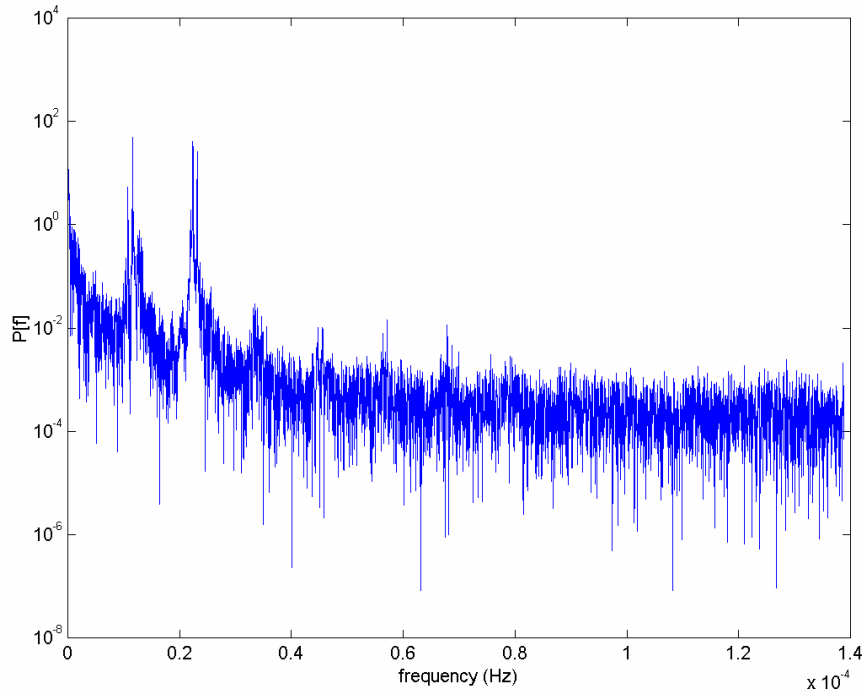


Figure 5. The Fourier power spectrum for the data from fig. 4.

The analysis of the power spectrum of these data (figure 5) indicates the existence of periodicities related to the diurnal and semidiurnal tides with a period of 12 and 24 hours, respectively and a broadband spectrum.

2.4. Hurst exponent of a time series

The Hurst exponent is a measure of the long-time correlations in a time series and was originally used to characterize flow in rivers and dams (Hurst, 1951). The Hurst exponent allows the classifying of a time series, being able to distinguish the existence of long-range correlations from random noise. In this method, also known as rescaled range analysis (R/S analysis), the span of a random process is divided by its variance, resulting in a new variable that depends on the length of the data record. Let us define the time average of the time series $s(t)$ over the interval of time τ :

$$\langle s \rangle_t = \frac{1}{\tau} \sum_{t=1}^{\tau} s(t) \quad (3)$$

Let us also define $A(t, \tau)$, the accumulated departure of $s(t)$ from the mean as:

$$A(t, \tau) = \sum_{u=1}^{\tau} (s(u) - \langle s \rangle_t) \quad (4)$$

so that the span of the process is defined by:

$$S(\tau) = \max_{1 \leq t \leq \tau} A(t, \tau) - \min_{1 \leq t \leq \tau} A(t, \tau) \quad (5)$$

Let us also introduce the standard expression for the variance:

$$V(\mathbf{t}) = \sqrt{\frac{1}{\mathbf{t}} \sum_{t=1}^{\mathbf{t}} (s(t) - \langle s \rangle_{\mathbf{t}})^2} \quad (6)$$

The rescaled Hurst analysis consists in studying the properties of the ratio:

$$R(\mathbf{t}) = \frac{S(\mathbf{t})}{V(\mathbf{t})} \quad (7)$$

The dependence of $R(\mathbf{t})$ on the number of data points follows an empirical power law described as $R(\mathbf{t}) = R_0 \mathbf{t}^H$ obtained over a wide length of time lengths \mathbf{t} , where H is the Hurst exponent. The Hurst exponent, $0 \leq H \leq 1$, is equal to 0.5 for random, white noise series, < 0.5 for rough anticorrelated series, and > 0.5 for positively correlated series.

The deficiencies in time-series analysis for identifying, describing and modeling long-range correlations was pointed out by Mandelbrot and Van Ness (1968). Mandelbrot (1983), using the theory of fractional Brownian motion (*fBm*), showed that fractional Brownian motion could provide an explicit statistical realization of the power law scaling, supporting the interpretation of natural phenomena in terms of fractal functions.

Estimating the Hurst coefficient of time series

Several methods are available for estimating the Hurst coefficient of a one-dimensional time series: scaled windowed variance, dispersional analysis, Hurst rescaled range analysis, autocorrelation measures, and power spectral analysis. Bassingthwaite and Raymond (1994) have demonstrated that the last three methods for estimating H are highly biased and variable. For example, for a series of 512 points, a 95 % confidence interval for H , based upon a rescaled range estimate of $H=0.5$ will include every H from 0.2 to 0.9. Autocorrelation analysis estimates are highly biased towards $H=0.5$. Fourier spectral analysis based on the periodogram has also a high variance in its estimates.

In this work we have used the scaled windowed variance method (Cannon *et al.*, 1997) to estimate H . In this method the signal is repeatedly divided into windows, but instead of computing the standard deviation of the means within the windows, the means of the standard deviations within the windows are used to obtain an estimate of H (Cannon *et al.*, 1997). Short data sets are difficult to characterize. Noise present in a real time series can mask long-range correlations among the signal elements. In the case of white noise, it would induce a bias toward $H = 0.5$. In this case it is possible to detect this by excluding small window sizes when computing the linear regression of $\log(R(\mathbf{t}))$ versus $\log(\mathbf{t})$.

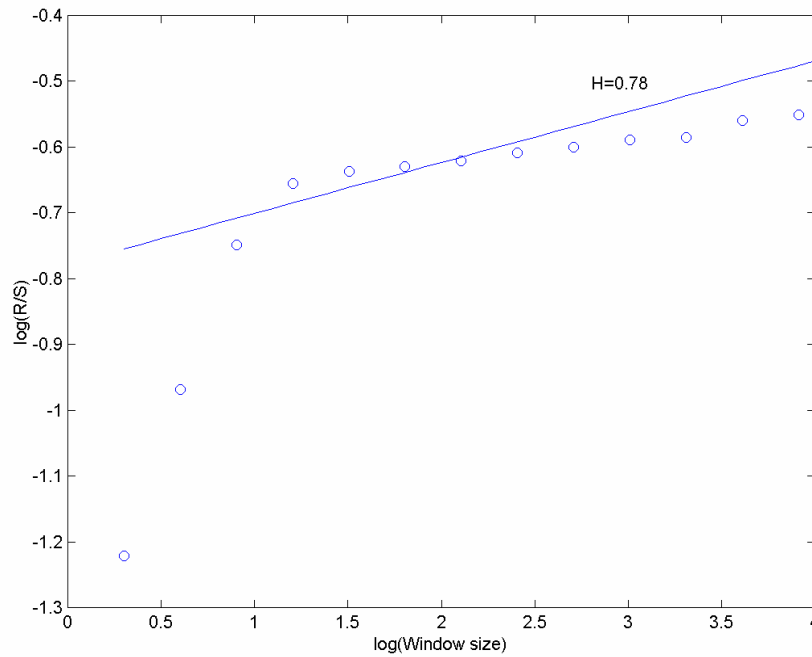


Figure 6. Hurst exponent calculated using the scaled window variance method and the recommended points in Table 2 (3-8) of Cannon *et al.* (1997).

2.5. Non-linear time series analysis techniques

The theory of dynamical systems is based on the concept of state space and a flow that governs the temporal evolution of the states and leads, for dissipative systems, to different types of attractors for sufficiently large times. Experimentally, it is not always possible to measure the complete state of a system and, normally, when analyzing a dynamical system, we have access to few observable quantities which, in the absence of noise, are related to the state space coordinates by:

$$s(t)=h(x(t)) \tag{8}$$

where h is an unknown non-linear function.

Non-linear time series analysis is based on the embedding theory (Takens, 1981). The theory of embedding is a way to pass from the space of measurements (Eq. 8) to a state space similar to that of the underlying dynamical system we are interested in analyzing. Techniques of state space reconstruction were introduced by Packard *et al.* (1980) and Takens (1981), which showed that it is, in theory, possible to address this problem using measurements of a sufficient long time series of the dynamical system of interest. Takens (1981) proved that, under certain conditions, points on the dynamical attractor in the underlying original system state space have a one-to-one correspondence with measurements of a limited number of variables. This fact opened a new field of research. In fact if the equations defining the underlying dynamical system are not known and we are not able to measure all the state space variables, the state space of the original system is not directly accessible to us. However if by measuring few variables we are able to reconstruct a one-

to-one correspondence between the reconstructed state space and the original, this means that it is possible to identify unambiguously our original state space from the measurements. In order to understand the relationship that occurs between the space of measurements and the state space, let us consider the following dynamical system:

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(\mathbf{x}) \quad \mathbf{x} = (x_1, x_2, x_3) \quad (9)$$

We can define $\mathbf{y} = (y_1, y_2, y_3)$ as follows: $\mathbf{y} = (x_1, dx_1/dt, d^2x_1/dt^2)$, then the equations of motion take the form:

$$\begin{aligned} \frac{dy_1}{dt} &= y_2 \\ \frac{dy_2}{dt} &= y_3 \\ \frac{dy_3}{dt} &= G(y_1, y_2, y_3) \end{aligned} \quad (10)$$

for some function $G [(\mathbf{F} \cdot \nabla)^2 F_1]$. In this coordinate system, modeling the dynamics reduces to constructing the single function G of three variables, rather than three separate functions, each of three variables. In this way we may pass from the state space (x_1, x_2, x_3) to that of the space of derivatives $(x_1, dx_1/dt, d^2x_1/dt^2)$. The dynamics in this new space will be related to the dynamic of the original space, in general, by a non-linear transformation, which is called the observability map. The extension of this approach to higher-dimensional dynamical systems is straightforward by considering higher derivatives. The advantage in considering the space of derivatives is that we can approximate them from measurements of x_1 . But which kind of information about the original space is preserved in the new one? There are two types of preserved information: qualitative and quantitative:

- Qualitative information is that which allows a qualitative description of the dynamics, they are, for example singularity of the field, the closeness of an orbit, the stability of a fixed point, etc.
- Quantitative information can be of three different types, which involve metric, dynamical and topological invariants. Metric methods (Grassberger and Procaccia, 1983) depend on the computation of various fractal dimensions or scaling functions. Dynamical methods (Wolf *et al.*, 1985) rely on the estimation of local and global Lyapunov exponents and Lyapunov dimensions as well as on entropy. Topological methods (Gilmore, 1998) involve determination of specific topological invariants of the attractor as relative rotation rates for the unstable periodic orbits embedded in the attractor, etc.

Takens (1981) showed that instead of the space of derivatives, $\{s(t), \dot{s}(t), \ddot{s}(t), \dots\}$, one can use that of delay coordinates, $\{s(t), s(t-\Delta t), s(t-2\Delta t), \dots\}$, where Δt is a time delay opportunely chosen. Looking at the following approximation of the derivative of $s(t)$:

$$\frac{ds(t)}{dt} \equiv \frac{s(t + \Delta t) - s(t)}{\Delta t}$$

$$\frac{d^2s(t)}{dt^2} \equiv \frac{s(t + 2\Delta t) - 2s(t + \Delta t) + s(t)}{2\Delta t^2} \quad (11)$$

it is clear that the new information brought from every new derivative is contained in the series of the delay coordinates. In fact, in case of high dimensions, the use of high order derivatives is not wise because they tend to amplify the noise in the measurements considerably.

Another method in common use is principal value decomposition, also called principal component analysis, factor analysis, or Karhunen-Loève decomposition. Their use for state space reconstruction was proposed by Broomhead and King (1986). The simplest way to implement this procedure is to compute the covariance matrix of the signal with itself and then to compute the eigenvalues, i.e. if $s(t)$ is the signal at time t , the elements of the covariance matrix C are:

$$c_{ij} = \langle s(t)s(t + (i - j)t) \rangle^T \quad (12)$$

where i and j go from 1 to n where n is bigger or equal to the dimension of the system in this new space. The eigenvectors of C define a new coordinate system. Typically, one calculates the dimension of the reconstructed phase space by considering only eigenvectors whose eigenvalues are “large”. This method allows to consider a time delay of one step and calculate an embedding dimension which is the rank of the covariance matrix C . The problem of this method is that the noise on the data tends to smear out the deterministic behavior, and, in the directions associated with small or vanishing singular values of C , the noise will dominate. In some sense we can say that this method tends to amplify the error due to the noise as for the method of derivatives.

Then, from the space of derivatives, time lags or eigenvectors, it is possible to extract information about the underlying system, which was generating the measured data. To obtain the system (Eq. 10) we assume implicitly that we know the dimension of the original system (Eq. 9), but this is not usually the case. Furthermore, we have also assumed that we are able to measure a state space variable, for example x_l , which is also not the case (see Eq. 8). In most of the situations we are measuring an observable quantity which is nonlinearly related, by an unknown function h , to the state space variables.

Different dissipative dynamical systems in state spaces of widely different dimensions may be equivalent, i.e. they have the same qualitative dynamics, if their asymptotic dynamics are confined in a state space with the same dimensionality. The question is now how to find the minimum dimension in order to describe the qualitative dynamics of a system?

A general existence theorem for embedding in Euclidean spaces was given by Whitney (1936) who proved that a smooth (C^2) n -dimensional manifold may be embedded in \mathbf{R}^{2n+1} . This theorem is the basis of the time delay reconstruction (or embedding) techniques for phase space portraits from time series measurements proposed by Takens (1981), who proved that, under certain

circumstances, if d_E (the dimension of our reconstructed vector or the embedding dimension) is greater or equal than $2n+1$, where n is the dimension of the original state space, then the reconstructed points fill out a reconstructed attractor which is diffeomorphic (i.e. a one-to-one differentiable mapping with a differentiable inverse exists) to the original attractor. Sauer *et al.* (1991) generalized this result by replacing the $2n+1$ condition to $d_E > 2D_C$, where D_C is the box-counting dimension. Furthermore, Casdagli *et al.* (1991) extended Taken's results to the case of low level observational noise, i.e. $s(t) = h(\mathbf{x}(t)) + \mathbf{s}(t)$.

Apart from the previous three methods mentioned above, there are several other methods of reconstructing state space from the observed quantity $s(t)$ that have appeared in literature (for a critical review see Breeden and Packard, 1994). The method of reconstruction can make a large difference in the quality of the resulting coordinates, but in general it is not clear which method is the best. The lack of a unique solution for all purposes is due in part to the presence of noise and the finite length of the data set. Even different algorithms for achieving the same goal often have different optimal representations. No single representation will be optimal for all possible objectives (Breeden and Packard, 1994).

After looking at the different time series, their frequency and properties, it has been concluded that the only time series where this type of analysis may be carried out are high frequency series, which means hourly values. In the DITTY project there are only two series: oxygen concentration and water level. Concerning the last time series an analysis for predicting "high waters" at Venice lagoon was already published by Zaldívar *et al.* (2000). There, the idea was to predict in advance the occurrence of high water levels that is a source of problems in Venice. However, for the coastal lagoons analyzed this not seems to be a relevant issue at the moment but may become important, depending on sea level rise in the following years. For example, the Burana-Volano watershed is under sea level and high tides could provoke an intrusion of the sea in the agricultural fields as happened in the past. The second type of time series is the hourly oxygen concentration. In this case, anoxic crises are an important issues to the studied lagoons and, hence, it was decided to carry out a preliminary test to see if non-linear time series techniques could be applied to this problem.

Early detection of anoxic crises

An important and recurrent problem in coastal lagoons are anoxic crises. Anoxic crises occur mainly in summer when temperatures are high and are triggered off mainly by organic matter decomposition. This organic matter in the case of Sacca di Goro is the product of macroalgal decomposition (Viaroli *et al.*, 2004), whereas in Etang de Thau is due to organic matter that accumulates in the oyster farms or where the oysters are processed before commercialization (Chapelle *et al.*, 2001).

The early detection of anoxic crises, with sufficient time in advance for countermeasures to be taken, is therefore an important aspect in the management of coastal lagoons. The main problems associated with this detection are that anoxia has a spatial localized initiation, and, hence, normally oxygen measurement is not always available; that the frequency of the measurement is not high enough, in the majority of the cases, to predict in advance the occurrence of an anoxic crises; and that the oscillations in the signal made the calculation of the derivatives quite difficult.

An early warning method that has been used in several fields is the calculation of the derivatives (Hub and Jones, 1986). In this case, when the first derivative of the oxygen concentration is negative, i.e. there is a decrease of oxygen in the system, and the second derivative is negative, i.e. there is an acceleration of this decrease then the probability that an anoxic crises would occur is high. This associated with a limit value for the oxygen concentration could be used as a warning. The problem in this case is the noise amplification when calculating the derivatives. Another approach also used in other fields (Zaldívar *et al.*, 2004) based on non-linear dynamical system theory is the calculation of the divergence of the system which tell us how fast two close points in state space are approaching or separating. This quantity may be calculated on-line (Bosch *et al.*, 2004) using state space reconstruction techniques. The measurement of the divergence has proved to be less sensitive to noise and with earlier detection in the case of chemical reactors.

As a first attempt to show the feasibility of these approaches we have decided to use 1990 data (from 18 May to 31 December) from a buoy at the center of Sacca di Goro, for which hourly measurements are available. Figure 7 shows the time series of oxygen concentration. As can be seen the time series shows high frequency fluctuations in accordance with model predictions (Zaldívar *et al.*, 2003). Furthermore, due to several reasons, in some periods the measurements are not available or constant. If we consider hypoxic conditions concentrations of O₂ lower than 2 g/m³ and anoxic conditions when concentration is 0 g/m³ (<0.2) then it is possible to see that these conditions were meet several times in Sacca di Goro during 1990. In fact figure 8 shows in yellow the period of hypoxic conditions and in red the anoxic ones.

Using this data set, we have developed two algorithms for early detection of anoxic conditions. The first algorithm is based on the first and second derivative of the oxygen concentrations and gives an alarm when:

$$\frac{dO_2}{dt} < -0.5 \quad \& \quad \frac{d^2O_2}{dt^2} < -0.2 \quad (13)$$

The second criterion states that when the variation in state space volume is lower than 0.1, $\Delta V_{ps}(t) < -0.1$, then an alarm is triggered off. This quantity is related to the divergence of the system (the trace of the Jacobian) by the following relationship (Bosch *et al.*, 2004):

$$div = \frac{\dot{V}_{ps}(t)}{V_{ps}(t)} \quad (14)$$

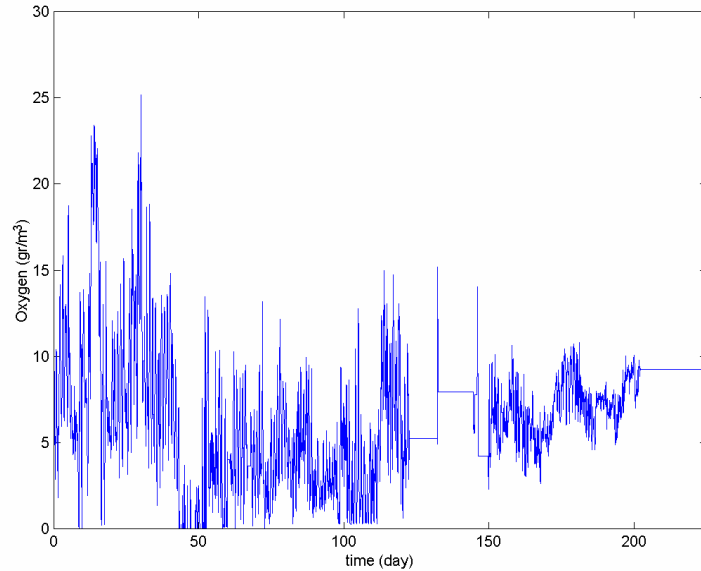


Figure 7. Experimental oxygen concentrations in the water column from the buoy of Consorzio Sacca di Goro in 1990.

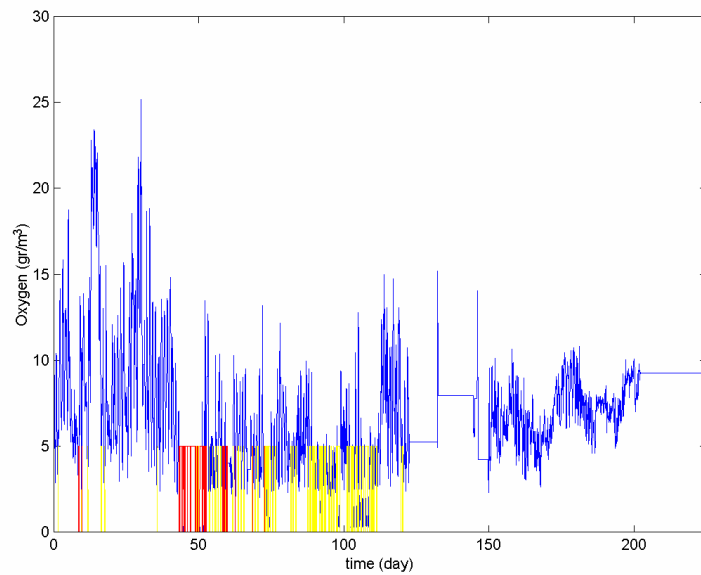


Figure 8. Hypoxic (yellow) and anoxic (red) conditions in the water column from the buoy of Consorzio Sacca di Goro in 1990.

and may be easily obtained from the time series using state space reconstruction techniques.

In both cases, the limits were chosen after visual inspection of the values, requiring the detection of the anoxic periods, but a proper selection, using an oxygen mathematical model would be necessary. The results of both criteria are shown in figures 9 and 10, respectively. As can be

seen the number of alarms provided by both algorithms is high in comparison with the real anoxic conditions, i.e. $O_2 < 0.2 \text{ g/m}^3$ but consistent with hypoxic conditions, i.e. $O_2 < 2 \text{ g/m}^3$.

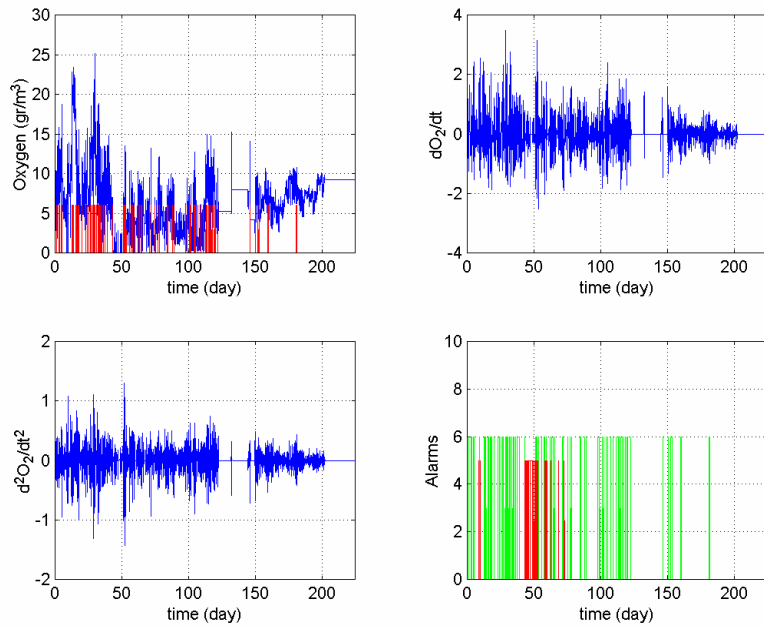


Figure 9. a/ Oxygen concentration and alarms (red) given by the first criterion; b/ First derivative; c/ Second derivative; d/ Anoxic conditions (red) and alarms (green).

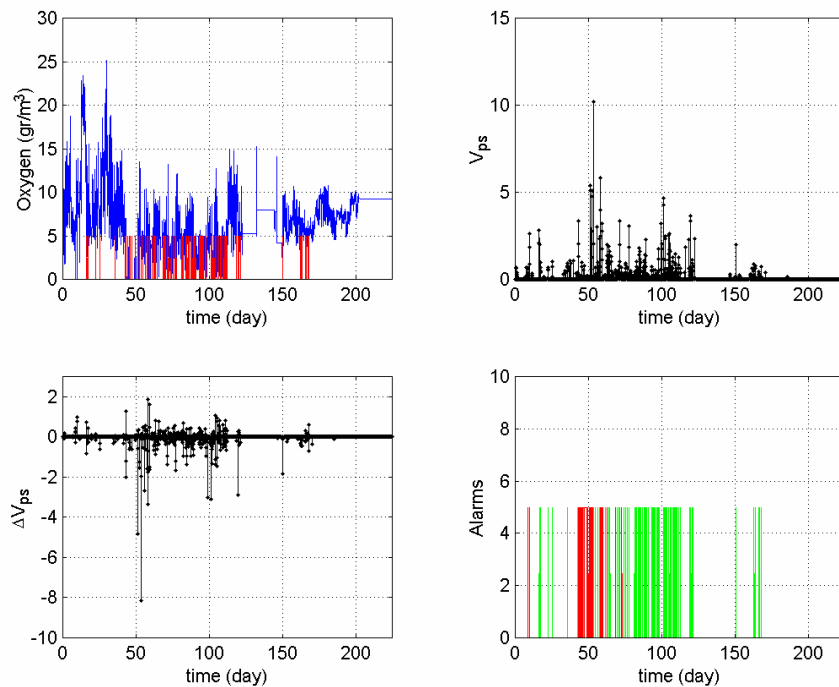


Figure 10. a/ Oxygen concentration and alarms (red) given by the second criterion; b/ State space volume; c/ state space variation; d/ Anoxic conditions (red) and alarms (green).

Concerning early detection, the results are summarized for two cases using both algorithms in figures 11 and 12. As can be seen both algorithms are able to detect the anoxic crises in advance.

However, as this is a process that takes only few hours it is clear that the time in advance is of the order of hours. The first algorithm does not detect the first anoxic crises in figure 11, whereas the second algorithm performs in the opposite way. For the case of the strong anoxic crises that occurred in July 1990, both algorithms detect in advance the initiation whereas the second algorithm also gives several alarms when there is a reduction of oxygen during the crisis. In this case the first algorithm would give the alarm approximately 4 hour in advance with respect to the second algorithm.

Such preliminary results show the possibility to apply techniques developed from non-linear dynamical systems theory to environmental time series. However, only few of them have the adequate frequency to be able to capture the real dynamics being most of them under sampled. Furthermore, these results are only provisional and further development is necessary before developing an operational device for early anoxic detection. Of course, this will depend on the interest of end-users in this approach which will be discussed during DITTY meetings.

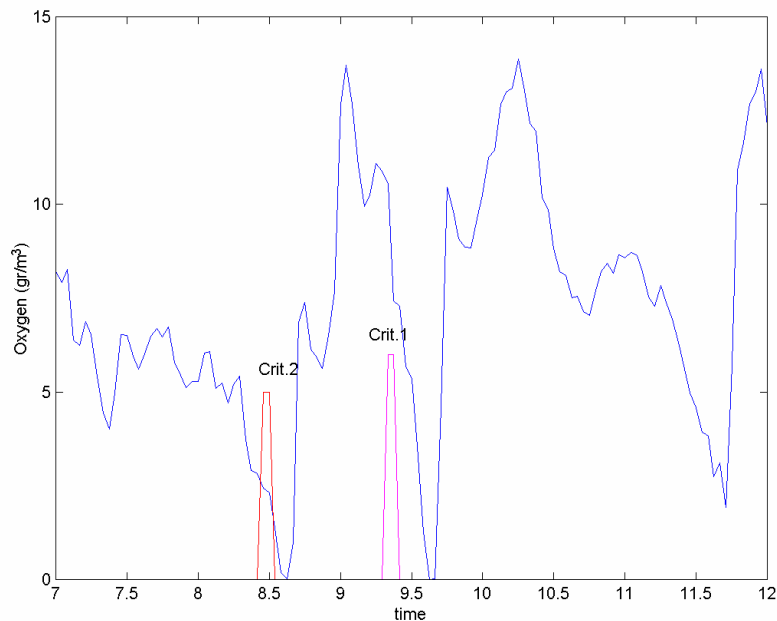


Figure 11. Results obtained using both algorithms for detection in advance of anoxic conditions for two cases.

2.6. Problems in environmental time series

There are different aspects that should be carefully studied before attempting to go further using time series analysis methods (linear and non-linear). A long and exhaustive discussion can be found in Schreiber (1999). The main problems one should be aware of can be summarized as: has the phenomenon been sufficiently sampled? Is the data set stationary or can one remove the nonstationary part? Is the level of noise sufficiently low so that one can obtain useful information using time series techniques?

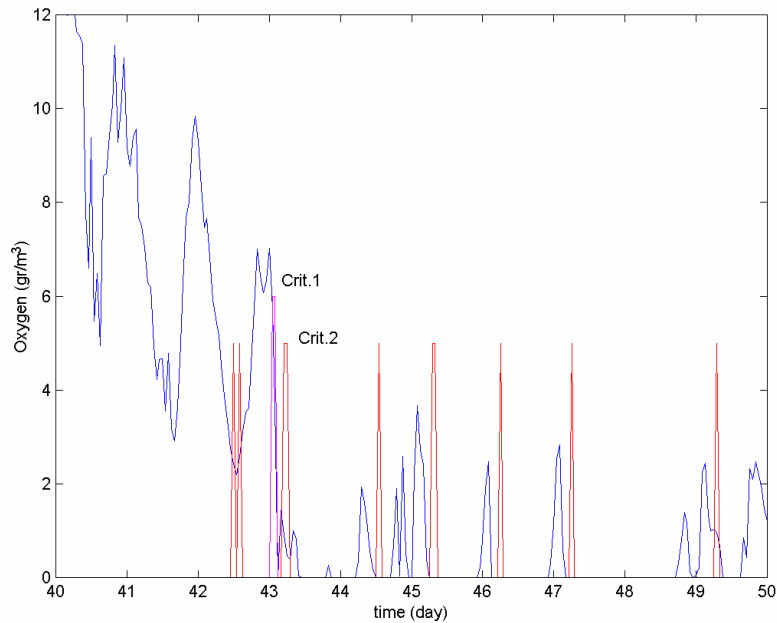


Figure 12. Results obtained using both algorithms for detection in advance of anoxic conditions for the severe anoxic crises at the beginning of July.

Stationarity

A time series is said to be strictly stationary if its statistical distribution does not change across time. More specifically, suppose we have a set of m samples of the series $s(t)$ made at times t_1 through t_m . These need not to be contiguous times. Strict stationarity implies that the joint probability density function of those m samples is identical to the joint probability distribution of another m samples taken at times t_1+k through t_m+k . This must be true for all the choices of m and k , as well as choices of the m relative sample times.

Why is stationarity so important? Because almost all methods developed by linear and nonlinear time series analysis assume that the time series we are analyzing is stationary, which implies that the parameters of the system, which has generated the time series, remain constant. For this reason time-series analysis often requires one to turn a nonstationary series into a stationary one so as to use these theories. Unfortunately, nonstationary signals are very common in particular when observing natural or economical phenomena, and in some cases the nonstationary components, such as the trend, may sometimes be of more interest from that of the stationary part obtained by removing the trend or the seasonal variation from the signal.

Even though a precise definition of stationarity exists, there is no magic formula for deciding whether a series is nonstationary. Only in the case of ARMA models whose weights are exactly known it is possible to assess the stationarity. However, strong violations of the basic requirements that the dynamical properties of the system must not change, beyond their statistical fluctuations, can be checked simply by measuring such properties, i.e. mean, variance, spectral

components, correlations, etc., for several segments of the data set. Another useful tool for detecting nonstationarity behavior is by studying the autocorrelation function of the series which is equal to the autocovariance divided by the variance:

$$c_k = \frac{\frac{1}{n} \sum_{i=k+1}^n (s_i - \bar{s})(s_{i-k} - \bar{s})}{c_0} \quad (15)$$

where $\bar{s} = \frac{1}{n} \sum_{i=1}^n s_i$ and $c_0 = \frac{1}{n} \sum_{i=1}^n (s_i - \bar{s})^2$.

As general indications, if a time series contains a trend, then the values of c_k will not come down to zero when the lag increases except for very large values. If a time series contains a seasonal fluctuation, then the plot of c_k versus k will also exhibit an oscillation at the same frequency. If a time series is random then 95% of the values of c_k will lie between $\pm 2/\sqrt{n}$ (Chatfield, 1997). Nonlinear time series analysis has also developed its own techniques to study nonstationarity (Kantz and Shreiber, 1997).

Once nonstationarity has been detected, there are different possibilities. The obvious one is detrending. However, the choice of detrending is still an unsolved issue in testing environmental time series. The traditional time series approach for removing nonstationarity consists on differencing. There are two different types. The first is adjacent-point differencing. In this type of differencing, a new series is defined as the relative change of the original time series at each time step. The second type of differencing is seasonal differencing. In this type a new series is computed as the change relative to the sample at some fixed distance in the past, for example a year, a month, etc. Differencing is only effective if the series exhibit a homogeneous nonstationarity. Related to this technique the logarithmic first differences have been widely used in fitting stochastic environmental models. In this case the original time series is transformed as,

$$u(t) = \ln s(t+1) - \ln s(t) = \ln \left(\frac{s(t+1)}{s(t)} \right) \quad (16)$$

Unfortunately, these techniques will increase the signal to noise ratio in the original time series and in some cases after the transformation we will find only noise.

Another technique that has been used by environmentalists is the log linear detrending. In this case we have a new resulting time series, $u(t)$, given by:

$$u(t) = \ln s(t) - (k_0 + k_1 t) \quad (17)$$

where k_0 is the intercept and k_1 is the constant growth rate, and $s(0) = \exp(k_0)$. Log linear detrending retains the long-term correlations in environmental fluctuations, since its time scale represent the whole period of the available time series.

Many time series, especially those found in the environment, are dominated by a strong periodic component. In some cases, a cyclic component overwhelms all other aspects of the data and, hence, in these cases it is necessary to isolate it from the time series in order to process effectively all the information contained in the series. In case of a strong seasonal component a method that has been widely used consists in assuming that the series can be represented as the sum of two components. One of these is slow global variation and the other is rapid local oscillation. A low-pass filter is then used to isolate these two components and they are treated separately (Masters, 1995).

Observations at unequal intervals

Normally time series analysis techniques (linear and non-linear) have been developed for time series measured at equal intervals of time. When observations are taken at unequal intervals, there are additional complications to time series analysis. Two general cases may be considered: the situation where observations are taken at random time points is different from the situation where some observations are missing from a time series measured at equal intervals. In the latter case it is also important to assess whether points are missing at random or there is some systematic pattern. In any case this aspect reduces the methods of analysis available and in some cases only few tools may be applied. Unfortunately, this is the case for most of the time series existing so far in the DITTY project.

2.7. Multiple time series analysis

So far only one time series has been considered. However, when time series have been obtained simultaneously at multiple locations, the Mann-Kendall test may be computed for each of the locations. An estimate of the magnitude of trend at each station can be obtained using Sen's procedure. If monotonic trends at all stations point to the same direction, hypotheses can be made about the existence of a regional-wide trend. Such homogeneity of trend direction at multiple stations can be tested, again using the Mann-Kendall test. The homogeneity *chi-square* statistic χ_h^2 is computed first:

$$\mathbf{c}_h^2 = \mathbf{c}_{total}^2 - \mathbf{c}_{trend}^2 = \sum_{j=1}^M Z_j^2 - M \bar{Z}^2 \quad (18)$$

where m is the number of stations, $Z^2 = \frac{S_j}{[\text{VAR}(S_j)]^{1/2}}$, S_j is the Mann-Kendall trend statistic for

the j -th stations and $\bar{Z} = \frac{1}{M} \sum_{j=1}^M Z_j$. If the trend at each station is in the same direction, then χ_h^2

has a *chi-square* distribution of with $M-1$ degrees of freedom. To test for trend homogeneity between stations at the α significance level, then χ_h^2 needs to be less than the α critical value (selected from test table for the $M-1$ degrees of freedom). If $\chi_h^2 > \alpha$ then the *null hypothesis* is

rejected, and no regional-wide statements can be made. However, the validity of the test requires the individual series to be uncorrelated and that no seasonal cycles are present.

If seasonal trends are present, van Belles and Hughes (1984) show that nonparametric aligned rank test, used by Farrel (1980) as proposed by Sen (1968), is more likely to detect monotonic trends. This is especially true when only a few years of data are available.

3. SPATIAL ANALYSIS TOOLS

In analyzing data related to coastal lagoons, data are often ordered with respect to spatial coordinates other than with respect to time. Methods for analysis of spatial data have been described in great detail, for example, by Ripley (1981), Cressie (1993), Kitanidis (1997) and Haining (2003), and some similarities can be drawn between time series and spatial data analysis. Among the main reasons for analyzing the behavior of a time series is to discover trends and to derive a model that can be used to predict the value of the variable at those points in time where no measurements are available. By the same token, we analyze spatial data to derive models that allow us to extrapolate values at unmeasured locations. However, major differences arise from the different features of the time (one-dimensional and unidirectional) and space domains (three-dimensional, no constrains in direction), mainly in trend analysis. It is obvious at this point that filling the spatial gaps in a set of measurements means to derive a best estimate for the value at that point, or, in other word, to minimize the error, calculated as difference between the actual value (where measurement is not available) and its estimate. The models used to “fill the gaps” are probabilistic models, i.e. a set of mathematical equations that summarize information drawn from the data set (e.g. explaining variability) and derive a prediction. A series of steps can be outlined when analyzing and processing spatial data. The following paragraphs will give some indications about each of the phases involved, starting from univariate spatial data analysis and later expanding to multivariate (the latter one describing the relationship between several variables in space) and the software available to perform such steps. In particular multivariate analysis could be of interest in the case of watershed and lagoon ecosystems, where spatial trends in one variable field can be related to variability in other variable fields (e.g. lagoon salinity concentrations influenced by tides, inputs from watershed and precipitations; nutrient concentration related to chlorophyll a concentration).

3.1. Exploratory spatial data analysis

It is generally helpful to “look” at the data before any models for spatial interpolation are fitted or hypothesis formally tested. Exploratory analysis be can of help in familiarize with the data, detect pattern of regularity, point out some very obvious feature of the data set, and uncover

aberrations (unusually high or low values, i.e. outliers), as well as direct further analysis. A review of basic ideas for data organization and graphical display can be found in Chambers *et al.* (1983). Exploratory analysis is particularly important in the case of the lagoons of the Ditty project, where time intersection between available time series at multiple locations is very low. Some very simple techniques can be used at this stage, such as *summary statistics*, *histograms*, *box plots*, *density estimates*, *maps* and *experimental variograms*.

Summary statistics is the set of numbers that gives us basic information about the characteristics of the data set: central value of the data set (*mean*, *media*, *mode*), spread (variance, *standard deviation*, *interquartile range*, *kurtosis coefficient*) and symmetry (*skewness coefficient*). *Summary statistics* can give an indication as to whether data are normally distributed (*mean*, *media*, *mode* coinciding, *kurtosis* and *skewness* equal to 3 and 0 respectively). The *histogram* is the most common way to portray data distribution, and can give a first indication about data distribution (e.g. unimodal); however, the visual impression derived may depend critically on the choice of intervals. For very small data sets (e.g. less than 50 elements) *histograms* can be misleading, and *box plots* come in as a more useful exercise. An example of such analysis is presented in figure 13.

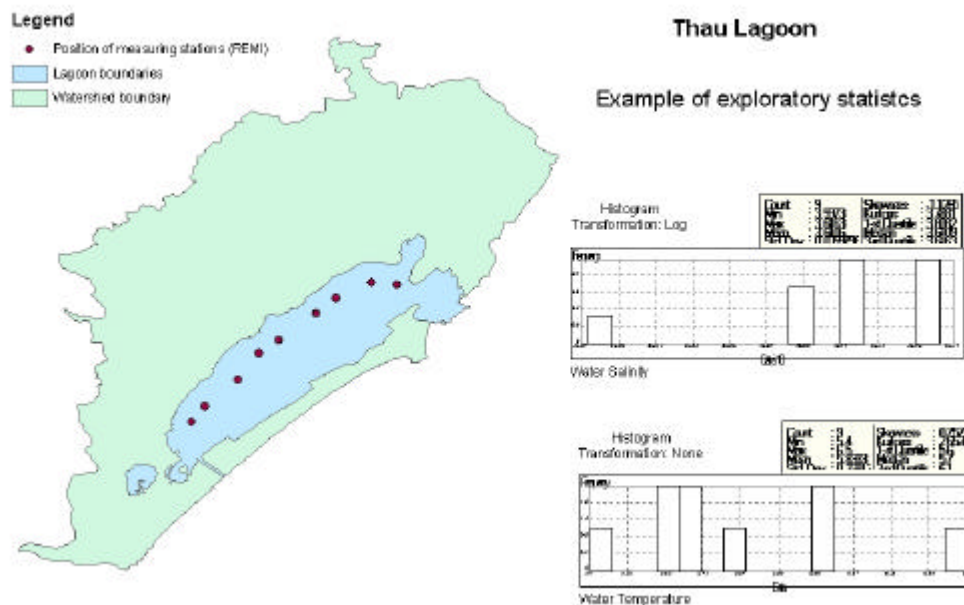


Figure 13. Summary statistics and histograms for temperature and salinity data recorded at the Thau Lagoon.

Box plots describe the data set using a five number statistics (minimum value, first quartile, median, third quartile, maximum value, “*crude*” and “*refined*” formulation), so that skewness of the distribution (and outliers, if any) becomes immediately apparent. *Kernel density estimates* can

be considered as a slightly more sophisticated approach than histograms, and attempt to provide a smooth estimate of probability density. *Maps* are also a very useful and easy way to look at data, as long as care is taken to avoid spurious details that could mask important information. In the case of univariate analysis, scatter plots of the variable against spatial location in one, or two dimensions can yield useful information about spatial patterns and unusual values. A very appealing exercise is the creation of surfaces (figure 14), using *inverse distance weighting*, *global* or *local polynomial interpolation*, *kriging* (each of these methods comes with a number of options,

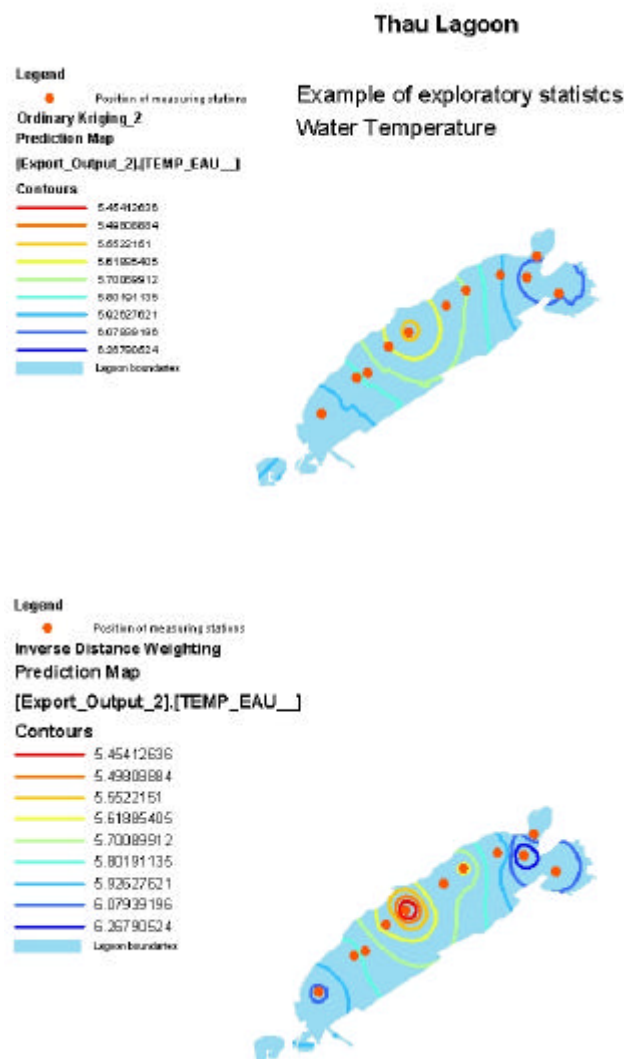


Figure 14. Kriging (top) and Inverse Distance interpolation (bottom) using ArcGIS for temperature data recorded at the Thau Lagoon (note that data from different sampling dates were joined into one series to overcome the 9-data minimum set in ArcGIS).

depending on the software used). However, note that most of the spatial series available for the lagoons in the Ditty projects do not have enough sampling points to build valid statistics. In

particular, some programs will not even allow the simplest interpolations if the number of sampling points is below the program-set minimum (e.g., ArcGIS requires a minimum of 10 sampling points for kriging).

Spatial structure can also be investigated using the *experimental variogram* (or *semivariogram*), where spatial distance between all possible couple of points in the series is plotted against half the square difference between the respective values (*raw variogram*). Points in this scatter plot known as *raw variogram* have coordinates:

$$h = |x_i - x_i'| \quad (19)$$

and

$$g(h) = \frac{1}{2} [z(x_i) - z(x_i')]^2 \quad (20)$$

where x is the variable spatial coordinate ($i=1$, or $i=1,2$, or $i=1,2,3$ for one, two and three dimensions respectively) and $z(x)$ the value of the variable. The *experimental variogram* (figure 15) is derived simply by dividing the h axis into N intervals (bins) and plotting the points:

$$h_k = \frac{1}{N_k} \sum_{i=1}^{N_k} |x_i - x_i'| \quad (21)$$

and

$$g(h_k) = \frac{1}{2N_k} \sum_{i=1}^{N_k} [z(x_i) - z(x_i')]^2 \quad (22)$$

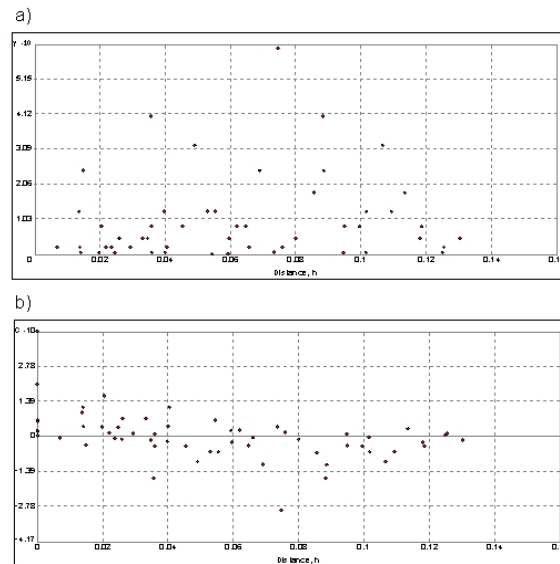


Figure 15. Experimental variogram (a) (lag size = 0.136; number of lags: 10) and covariance plot (b) for temperature data recorded at the Thau Lagoon.

Next, the points $[h_k, g(h_k)]$ are connected to form the experimental variogram. The shape of the graph obtained will depend on the selected intervals; however, there is no optimal number of bins to be selected, and the rationale behind the choice of classes is similar to the one adopted for the histograms. Through its characteristics (*nugget*, *range* and *sill*) the *experimental variogram* allows to detect two important structural characteristics of the data: variability at the scale of the sampling span, depending on the behavior of the *variogram* near the origin; variability at the scale of the sampling domain, depending on the behavior of the *variogram* at large distances. Continuity, smoothness and stationarity of the data series can also be inferred from the shape of the *experimental variogram*.

As a general approach, the experimental variogram and other tools so far presented should not monopolize the analysis, although some of the choices made at this stage will influence the conclusions drawn from the exploratory exercise (tentative selection of a model that fits our data series).

As a final note, unimodal near-symmetric (i.e. *normal*) distributions are the easiest to analyze, needing only two numbers to be sufficiently described. Sometimes data themselves are not normally distributed, but some power transform of the data (e.g. logarithmic transformation) might fit the Gaussian distribution. Tests can be performed, to determine whether data depart sufficiently from the null hypothesis that the observations were sampled independently from a normal distribution. General tests are the *chi-square* and the *Kolmogorov-Smirnov*; the *Shapiro-Wilks* (1965) test has been developed specifically for the normal distribution.

All the tools presented are available in a number of software packages; spreadsheet software can be used at this stage (e.g. EXCEL software package) or GIS software (e.g. ArcGIS with Spatial Analyst extension).

3.2. Structural analysis: the intrinsic model

Structural analysis is the process through which the family of functions that includes the solutions to the problem is determined, namely the family of functions $z(x)$, all derived from the recorded measurements of the variable $z(x_1), z(x_2), \dots, z(x_n)$, at locations x_1, x_2, \dots, x_n , that describe the variable field (x is a vector with one, two or three components). The process entails fitting equations that describe the first two moments of the family of functions. The form of the expression describing the first two moments is called *the model*. Many expressions can be used to represent these moments. One of the most used (general purpose) models is the *intrinsic model*, postulating that the mean is constant (and its exact value non needed), and the mean square difference is a function of the separation distance only. Such model is also called *isotropic* because it uses only the length and not the orientation of the linear segment connecting two points. The function to be used is the *theoretical variogram* (different from the *experimental* one, which is computed from the data), and describes the distribution of variability among scales. Other more

specific models can be selected, depending on the data (experience at similar sites) and site information available. It is an iterative procedure, comprising the steps: i) exploratory analysis, on the basis of which a model is selected; ii) parameter estimation and iii) parameter validation (evaluation model performance on test case).

Among the most common stationary models are:

- the *Gaussian model*, where:

$$R(h) = \mathbf{s}^2 \exp\left(-\frac{h^2}{L^2}\right) \quad (23)$$

$$\mathbf{g}(h) = \mathbf{s}^2 \left(1 - \exp\left(-\frac{h^2}{L^2}\right)\right) \quad (24)$$

with parameters $\mathbf{s}^2 > 0$ and $L > 0$. The model has covariance with parabolic behavior at the origin (i.e. $\mathbf{g}(h) \propto h^2$), and is appropriate for a differentiable regionalized variable.

- the exponential model, where:

$$R(h) = \mathbf{s}^2 \exp\left(-\frac{h}{l}\right) \quad (25)$$

$$\mathbf{g}(h) = \mathbf{s}^2 \left(1 - \exp\left(-\frac{h}{l}\right)\right) \quad (26)$$

with parameters $\mathbf{s}^2 > 0$ and $l > 0$.

- the *spherical model*, where:

$$R(h) = \begin{cases} \left(1 - \frac{3h}{2a} + \frac{1}{2} \frac{h^3}{a^3}\right) \mathbf{s}^2, & \text{for } 0 \leq h \leq a \\ 0, & \text{for } h > a \end{cases} \quad (27)$$

$$\mathbf{g}(h) = \begin{cases} \left(\frac{3h}{2a} - \frac{1}{2} \frac{h^3}{a^3}\right) \mathbf{s}^2, & \text{for } 0 \leq h \leq a \\ \mathbf{s}^2, & \text{for } h > a \end{cases} \quad (28)$$

with parameters $\mathbf{s}^2 > 0$ and $\mathbf{a} > 0$.

Other stationary models are available in literature (e.g., going back to stationary models, the *hole-effect* model, appropriate for one-dimensional problems, the *nugget-effect* model, appropriate to represent microvariability and random measurement error); many more can be custom built by superposition of existing models. Among the most common intrinsic non-stationary models are the *power model*, with variogram: $\mathbf{g}(h) = \mathbf{q}h^s$, with $\mathbf{q} > 0$ and $0 < s < 2$ (for $s = 1$ we obtain the *linear model*); the logarithmic model, with variogram: $\mathbf{g}(h) = A \log(h)$, with $A > 0$ (appropriate for integrals over regionalized volumes, e.g. solute mass over a finite volume when mass data, and not point measurements, are available).

3.3. Residuals in kriging

Residuals are differences between observations and model predictions. In statistical modeling (regression, time series, analysis of variance, and geostatistics), parameter estimation and model validation depend heavily on the examination of residuals (Belsley *et al.*, 1980; Borgman, 1988; Dubrule, 1983a; Box and Jenkins, 1976). In kriging residual are built following the procedure here forth described. Given the model where the unknown spatial function $z(x)$ is a realization of an intrinsic function with variogram $\mathbf{g}(h)$, the kriging estimate of z at a second point x_2 given only the first measurement x_1 is $\check{z}_2 = z(x)$ and $\mathbf{s}_2^2 = 2\mathbf{g}(x_1 - x_2)$. In this case the actual error and the residual are, respectively:

$$\mathbf{d}_2 = z(x_2) - \check{z}_2 \quad (29)$$

$$\text{and: } \varepsilon_2 = \mathbf{d}_2/\mathbf{s}_2 \quad (30)$$

The general terms for residual and normalized residual are:

$$\mathbf{d}_k = z(x_k) - \check{z}_k, \quad \text{for } k = 2, \dots, n \quad (31)$$

$$\text{and: } \varepsilon_k = \mathbf{d}_k/\mathbf{s}_k, \quad \text{for } k = 2, \dots, n \quad (32)$$

To the ensemble of all possible realizations of a random function the ensemble of residuals (one for each realizations) can be associated, which are treated like random variables for which ensemble statistics or probability distributions are computed. If the residuals are uncorrelated from each other and normalized to have a unit variance (i.e. they are orthonormal), then the selected model is correct and kriging is a minimum variance unbiased estimator.

3.4. Best linear unbiased estimation (BLUE)

Once the variogram is chosen (in the case of the intrinsic model), conditioning can be performed, thus eliminating from consideration members of the family of functions that are inconsistent with the available data (set of measurements). Conditioning is performed through kriging, applying the best linear unbiased estimation (*BLUE*) to the intrinsic functions. In this method, the estimate is a linear function of the data with coefficients that are selected to minimize the mean square error and to yield an expected error of zero. The objective function:

$$E\left[\left(z_0 - \hat{z}(x_0)\right)^2\right] = -\sum_{i=1}^n \sum_{j=1}^n I_i I_j \mathbf{g}(\|x_i - x_j\|) + 2 \sum_{i=1}^n I_i \mathbf{g}(\|x_i - x_0\|) \quad (33)$$

must be minimized, with the constrain:

$$\sum_{i=1}^n I_i = 1 \quad (34)$$

The corresponding set of equations is solved resorting to Lagrange multipliers, thus obtaining $\lambda_1, \lambda_2, \dots, \lambda_n$ (linear estimators), v (Lagrange multiplier). At times (although not often), it can be convenient to use *kriging with moving neighbourhood*, where only observations close to the point of estimate are used; however, such computation-saving technique is adopted only in special cases. Several commercial algorithms can perform the operation, from MatLab, to the already mentioned ArcGIS, with full kriging functionality. Kriging is an *exact interpolator* in the case of continuous functions and can be so also in the case of a variogram that exhibits a *nugget effect*, and assures uniqueness of solution (assuming that there are no redundant measurements).

3.5. Model validation

Testing the model selected implies the use of statistical tests. Under the assumption that our random variable is normally distributed, the orthonormal residuals of the selected model are computed, and then their average:

$$Q_1 = \frac{1}{n-1} \sum_{k=2}^n \mathbf{e}_k \quad (35)$$

being \mathbf{e} the standard error, along with mean and variance. Under the null hypothesis Q_1 is normally distributed with mean 0 and variance $1/(n-1)$, with probability density function equal to:

$$f(Q_1) = \frac{1}{\sqrt{2\mathbf{p}/(n-1)}} \exp\left(-\frac{x^2}{2/(n-1)}\right) \quad (36)$$

The model is rejected (a 95% confidence level is usually adopted) if:

$$|Q_1| > \frac{2}{\sqrt{n-1}} \quad (37)$$

The same procedure can be applied to higher order statistics:

$$Q_2 = \frac{1}{n-1} \sum_{k=2}^n \mathbf{e}_k^2 \quad (38)$$

with the same considerations presented above. Residuals \mathbf{e} should also be tested for normality, either by simply plotting them on normality paper or performing *goodness of fit* tests (Shapiro-Wilk, 1965; Filliben, 1975). Again, variogram estimation is an iterative procedure, where several models are fitted and tested. As a general rule, assuming that minimum criteria (i.e.: Q_1 near 0, Q_2 about 1, orthonormal residuals) are met by more than one model, then the simplest model available that satisfactorily fits the data should be adopted (Occam's razor principle). In fact, assuming that minimum criteria are met, in most cases the results from kriging will not be sensitive to the exact shape of the variogram (Brooker, 1986). Parameter uncertainty (Kitanidis, 1986) and additional information (Omre, 1987) can be incorporated into the prediction.

3.6. Anisotropy

In case of anisotropic variable fields, the variogram will depend not only on the separation distance, but also on direction (for example as in the case of flow in a porous medium). Anisotropic models require more parameters than the isotropic ones, and choice of an appropriate system of spatial coordinates becomes crucial. For example, in the case of anisotropy in the z (vertical) direction, the variogram is:

$$\mathbf{g}(h_x, h_y, h_z) = \mathbf{s}^2 \left[1 - \exp \left(- \sqrt{ \left(\frac{h_x}{20} \right)^2 + \left(\frac{h_y}{20} \right)^2 + \left(\frac{h_z}{2} \right)^2 } \right) \right] \quad (39)$$

conveying the information that correlation length in the horizontal plane is 10 times greater than in the vertical direction. Linear estimation (kriging) in this case is the same as in the case of isotropic case. However, structural analysis is more complicated because anisotropic variogram have more parameters than isotropic ones. Experimental variograms will be used to explore the degree of anisotropy. Other than separation distance, direction angle need to be accounted for. Data pairs are grouped with respect to orientation. In absence of information pointing to another model (from stochastic theory), the geoanisotropic model may be used: in this case, correlation may be transformed to isotropic by rotation and stretching of the axes. Calculations involved may be carried out using algorithms available in most computer packages (e.g. MATLAB). Anisotropic

fields can also be analyzed using GIS software packages (e.g. ArcGIS with Geostatistical Analyst extension).

3.7. Variable mean models

Linear minimum-variance unbiased estimation can be readily extended to the linear model with variable mean. In this case *universal kriging* can be applied. The spatial function is given by the linear model:

$$z(x) = m(x) + \hat{\mathbf{I}}(x) \quad (40)$$

$$\text{with: } m(x) = \sum_{k=1}^p f_k(x) \mathbf{b}_k \quad (41)$$

where f is the base function and \mathbf{b} the drift coefficient (*linear model*). One way to interpret the expression for $z(x)$ is to think of it as consisting of a deterministic part (*trend*), involving exact determination, and a stochastic part, involving approximate correlations. Polynomial, trigonometric, orthogonal functions can be used as base functions. Restrictions imposed to eliminate the drift coefficients to satisfy the unbiasedness condition allow to confine the analysis to the *generalized covariance function (GCF)*. A preliminary estimate of the *GCF* can be obtained fitting an equation to the experimental variogram of the detrended data. Such estimate can be later improved using the method of orthonormal residuals, as in the constant-mean case. A particular form of *GCF* is the *polynomial (PGCF; Matheron, 1973)*, with intrinsic functions of higher order. Interpolation through splines can also be related to kriging or Bayesian estimation (Cressie, 1986; Dubrule, 1983b; Kimeldorf and Wahba, 1970).

3.8. Multivariate analysis

Cokriging (Kitanidis, 1997, and references therein) is the equivalent of *kriging* in the case of multiple variable. The variables are thus processed separately, e.g. the most appropriate model selected, the variogram determined, and the model validated. In other words, related functions are modeled as realizations of jointly distributed random fields; the random fields are described through their joint first two moments; crosscovariance is thus determined (unlike autocovariance, it does not need to be symmetric) and generalized cross-covariance; following second moment characterization, linear minimum-variance unbiased estimation is performed. The method of orthonormal residuals can be used for parameter estimation and model validation.

Multivariate Analysis (Hair et al., 1998) techniques are normally applied to decrease the dimensionality of the input variables space and homogenize the information distribution for input and output variables (data samples selection). Two typical representatives are principal

components analysis (PCA) and clustering techniques. In this short note an exhaustive enumeration will not be possible. For a comprehensive review the reader is referred to Hair et al., (1998). Here we briefly describe Correspondence Analysis due to its application to Thau lagoon.

Correspondence Analysis (CA): A statistical ordination technique

Sampling campaigns performed with the aim of studying the spatial typology of an ecosystem often lead to the construction of two-dimensional tables, sometimes very big in size, showing the measured variable (species, biomasses, size-classes, ...) in function of the sampled stations. The objective is then to identify the spatial structure of the variable. Among the statistical multivariate analysis, Correspondence Analysis (CA) allows finding the best graphical simultaneous representation for the measured variables and the sampled stations, by partitioning the total data scattering amongst several axes (or factors), that express and grade the “organised dispersion” of the scatter plot (Gros and Hamon, 1988).

The following description is taken from Gros and Hamon (1988) and Lefebvre (1983): Let X be a matrix with n species in row and p stations in column, x_{ij} represent the measured variable (species for example) for the i^{th} ($i=1, \dots, n$) species at the j^{th} ($j=1, \dots, p$) station. We will only consider hereafter the species-plots in the space of stations-plots \mathfrak{R}^p , but everything is also valid for the stations-plots in the space of species-plots \mathfrak{R}^n . In fact, in their respective spaces, these two scatter plots are part of a r -dimensional sub-space, with $r=\text{rank}(X)$.

First step consists in representing the i^{th} species by its “profile”, that is the point with the following coordinates :

$$\left(\frac{x_{i1}}{x_{i.}}, \dots, \frac{x_{ij}}{x_{i.}}, \dots, \frac{x_{ip}}{x_{i.}} \right) \quad \text{where} \quad x_{i.} = \sum_{j=1}^p x_{ij} \quad (42)$$

Then, two species with different effectives but showing the same profile will be represented by the same point. Nonetheless, the memory of their respective effectives is not lost because a weight will be attributed to each line, thus composing a $p+1$ new column in the table. Then the point with coordinates $\frac{x_{ij}}{x_{i.}}$ ($j=1, \dots, p$) will have the following weight (m_i):

$$m_i = \frac{x_{i.}}{x_{..}}, \quad \text{where} \quad x_{..} = \sum_{i=1}^n \sum_{j=1}^p x_{ij} \quad (43)$$

This weight will act on the final repartition of points, giving more influence to more abundant species (if abundance is the measured variable).

The origin is the barycentre G which coordinates are defined by the last column (weights) :

$$\left(\frac{x_{.1}}{x_{..}}, \dots, \frac{x_{.j}}{x_{..}}, \dots, \frac{x_{.p}}{x_{..}} \right) \quad \text{where} \quad x_{.j} = \sum_{i=1}^n x_{ij} \quad (44)$$

Let $d^2(i, G)$ be the distance between the point i and the barycentre G , it is calculated as following:

$$d^2(i,G) = \sum_j \left(\frac{x_{.i}}{x_j} \right) \left(\frac{x_{ij} - x_j}{x_i - x_{.i}} \right)^2 \quad (45)$$

where $\frac{x_{.i}}{x_j}$ is a stabilizing factor for the squared distance.

This system thus gives more importance to the species with high abundance (or biomass or whatever is the measured variable), but without totally neglecting the stations where the different species are weakly represented.

Next step, as mentioned before, is to split the total data scattering amongst several axes (or factors). The optimization of this partitioning is based on the calculation of each profile inertia compared to G ($IN(i,G)$):

$$IN(i,G) = m_i \times d^2(i,G) \quad (46)$$

And the total scattering of points is quantified as $\sum_{i=1}^n IN(i,G)$, that can be decomposed into two terms :

$$\sum_{i=1}^n IN(i,G) = \sum_i IN(i \perp D) + \sum_i IN(i // D) \quad \text{where } D \text{ is any straight line passing through } G,$$

$IN(i \perp D)$ is the point i inertia transverse to D and $IN(i // D)$, point i inertia located on the line D .

The optimization criteria define the best line $D1$ (first factor) as following:

$$IN(i // D1) = I_1 = \max_D \left\{ \sum_i IN(i // D) \right\} \quad (47)$$

The first factor is thus calculated with the objective of having the bigger dispersion on the axis. For the second axis (2^{nd} factor), the points are projected in the sub-space orthogonal to $D1$ in G , and so on in order to determine the $r-1$ axes with inertia $?_1, \dots, ?_{(r-1)}$ where $?_{(a-1)} = ?_a$.

Total inertia is given by :

$$\sum_{i=1}^n IN(i,G) = \sum_{a=1}^{r-1} I_a \quad (48)$$

Each factor is characterized by its relative inertia t_a , expressing the total scatter plot distribution along axis a :

$$t_a = \frac{I_a}{\left(\sum_a I_a \right)}, \quad a=1, \dots, r-1 \quad (49)$$

It is possible, using matrix algebra, to demonstrate that the coordinate of a point-species on axis a in \mathfrak{R}^p is the gravity centre of the p points-stations coordinates on axis a in \mathfrak{R}^n , given the weights that represent the species profile considered. Likewise, the coordinates of a point-station is the homothetic of the point-species barycentre, given the weights of the station profile. This

property, called “barycentric principle”, allows to superimpose the stations and the species graphs, and taking into account a corrective factor $\sqrt{I_a}$, to put one species at the barycentre of the stations where the species is present (if the species is located at one unique station, then it will be superimposed on the station).

Example : Spatial analysis of macrophyte distribution in the Thau lagoon (Plus, 2001).

During May and June 1998, within the framework of the French National Program for Coastal Environment (PNEC), a re-evaluation of the macrophyte community (distribution and biomass) was carried out in the Thau lagoon. Sampled stations are presented in figure 16.

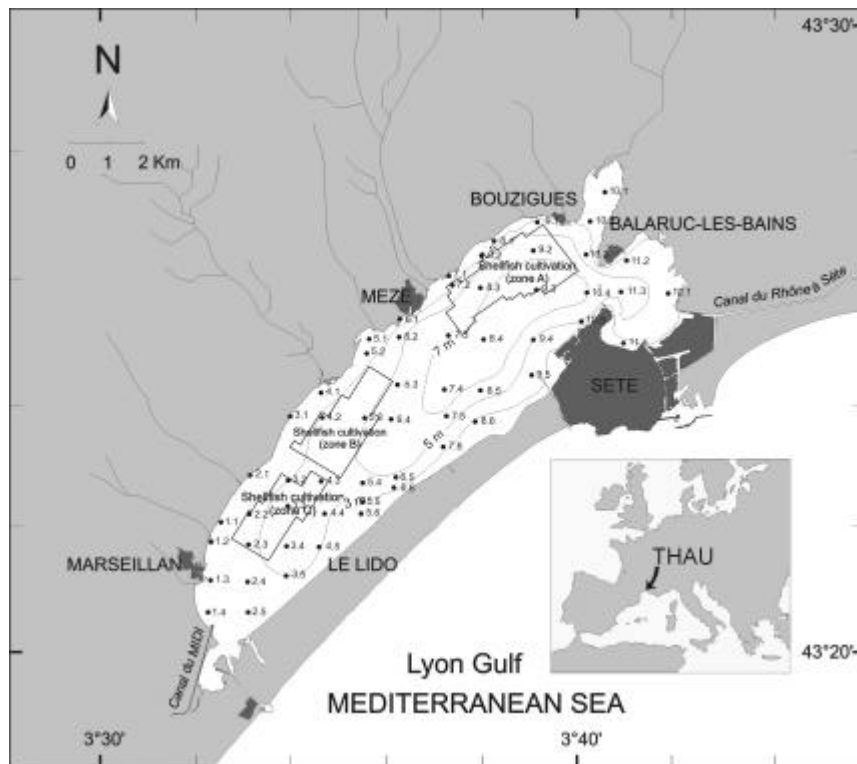


Figure 16. The Thau lagoon and the evenly distributed stations sampled during the PNEC campaign (black points).

The biomass data (table 1) were processed with Correspondence Analysis (CA) in order to represent the spatial structure of the macrophyte community and identify areas with a homogeneous macrophyte association. The results for the CA are presented in figure 17. The analysis allowed to extract for axis 1, 2 and 3, respectively 18, 17 and 13 % of the total variance.

Table 1. Macrophyte biomasses measured at the different stations. Hi: *Halopity incurvus*; Zm: *Zostera marina*; Zn: *Zostera noltii*; Gc: *Gracilaria compressa*; Cl: *Chaetomorpha linum*; Gd: *Gracilaria dura*; Rt: *Rytiphloea tinctoria*; Gl: *Gracilariopsis longissima*; Ac: *Alsidium corallinum*; Ur: *Ulva rigida*; Mo: *Monostroma obscurum*.

| Station | Hi | Zm | Zn | Gc | Cl | Gd | Rt | Gl | Ac | Ur | Mo |
|---------|--------|--------|--------|--------|--------|--------|--------|--------|--------|-------|--------|
| 1,1 | 117,2 | 0,0 | 3280,0 | 0,0 | 0,0 | 0,0 | 3,6 | 0,0 | 0,0 | 0,0 | 0,0 |
| 1,2 | 4160,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 1000,0 | 0,0 | 27,2 | 0,0 | 0,0 |
| 1,3 | 1600,0 | 1600,0 | 12,0 | 0,0 | 0,0 | 0,0 | 6,4 | 0,0 | 4,4 | 0,0 | 0,0 |
| 1,4 | 5480,0 | 0,0 | 0,0 | 135,2 | 0,0 | 162,0 | 162,4 | 0,0 | 10,4 | 22,0 | 0,0 |
| 2,1 | 26,4 | 640,0 | 1440,0 | 0,0 | 136,6 | 58,8 | 0,8 | 0,0 | 0,0 | 0,0 | 0,0 |
| 2,2 | 5880,0 | 720,0 | 0,0 | 0,0 | 0,0 | 349,2 | 12,0 | 0,0 | 193,0 | 0,0 | 0,0 |
| 2,3 | 3655,2 | 0,0 | 0,0 | 0,0 | 0,0 | 173,6 | 1360,0 | 0,0 | 197,2 | 0,0 | 0,0 |
| 2,4 | 8520,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 202,4 | 0,0 | 147,6 | 0,0 | 0,0 |
| 2,5 | 4000,0 | 1200,0 | 960,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 3,1 | 0,0 | 0,0 | 2976,0 | 0,0 | 422,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 3,2 | 1320,0 | 260,0 | 0,0 | 0,0 | 0,0 | 2140,0 | 0,0 | 0,0 | 400,0 | 0,0 | 0,0 |
| 3,3 | 4800,0 | 222,0 | 0,0 | 0,0 | 15,2 | 4,4 | 0,0 | 0,0 | 38,8 | 0,0 | 0,0 |
| 3,4 | 6560,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 2040,0 | 0,0 | 0,0 | 10,8 | 0,0 |
| 3,5 | 2800,0 | 1040,0 | 480,0 | 0,0 | 0,0 | 0,0 | 90,4 | 0,0 | 0,0 | 0,0 | 0,0 |
| 4,1 | 0,0 | 146,8 | 0,0 | 88,8 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 4,2 | 0,0 | 1680,0 | 0,0 | 0,0 | 0,0 | 7,6 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 4,3 | 47,4 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 1,8 | 0,0 | 0,0 | 0,0 | 0,0 |
| 4,4 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 4,5 | 0,0 | 8,0 | 720,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 2,4 | 0,0 |
| 5,1 | 0,0 | 4172,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 5,2 | 0,0 | 184,0 | 0,0 | 9,2 | 0,0 | 0,0 | 0,0 | 45,2 | 0,0 | 0,0 | 0,0 |
| 5,3 | 0,0 | 2,0 | 0,0 | 38,0 | 0,0 | 15,6 | 0,0 | 102,8 | 37,0 | 0,0 | 0,0 |
| 5,4 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 5,5 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 2,0 | 0,0 |
| 5,6 | 0,0 | 0,0 | 1840,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 6,1 | 0,0 | 1000,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 6,2 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 6,3 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 6,4 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 6,5 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 6,6 | 0,0 | 440,0 | 2280,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 7,1 | 0,0 | 0,0 | 0,0 | 2266,7 | 773,3 | 0,0 | 0,0 | 0,0 | 1,9 | 746,7 | 0,0 |
| 7,2 | 0,0 | 0,0 | 0,0 | 1040,0 | 0,0 | 6,0 | 0,0 | 320,0 | 2000,0 | 0,0 | 0,0 |
| 7,3 | 0,0 | 0,0 | 0,0 | 1,6 | 0,0 | 0,0 | 0,0 | 13,7 | 0,0 | 0,0 | 0,0 |
| 7,4 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 7,5 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 7,6 | 0,0 | 0,0 | 1680,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 8,1 | 0,0 | 0,0 | 0,0 | 56,4 | 391,2 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 8,2 | 0,0 | 520,0 | 0,0 | 1200,0 | 380,0 | 1320,0 | 0,0 | 85,6 | 280,0 | 0,0 | 0,0 |
| 8,3 | 0,0 | 0,0 | 0,0 | 780,0 | 160,0 | 120,0 | 0,0 | 820,0 | 0,0 | 0,0 | 0,0 |
| 8,4 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 8,5 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 8,6 | 0,0 | 2120,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 557,6 | 0,0 | 0,0 | 0,0 |
| 9,1 | 0,0 | 306,6 | 0,0 | 826,6 | 560,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 373,3 |
| 9,2 | 0,0 | 100,0 | 0,0 | 119,3 | 300,0 | 0,9 | 0,0 | 26,0 | 0,0 | 0,0 | 0,0 |
| 9,3 | 0,0 | 0,0 | 0,0 | 11,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 9,4 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 120,0 | 0,0 | 0,0 | 0,0 |
| 9,5 | 0,0 | 96,0 | 1600,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 10,1 | 0,0 | 0,0 | 0,0 | 669,6 | 6720,0 | 40,8 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 10,2 | 0,0 | 866,7 | 0,0 | 800,0 | 1013,3 | 2706,7 | 0,0 | 13,7 | 60,5 | 0,0 | 0,0 |
| 10,3 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 10,4 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 10,5 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 11,2 | 0,0 | 0,0 | 0,0 | 800,0 | 34,4 | 0,0 | 0,0 | 50,0 | 0,0 | 800,0 | 0,0 |
| 11,3 | 0,0 | 0,0 | 0,0 | 360,0 | 0,0 | 0,0 | 0,0 | 2060,0 | 0,0 | 0,0 | 0,0 |
| 11,4 | 0,0 | 0,0 | 0,0 | 200,0 | 0,0 | 0,0 | 0,0 | 80,0 | 0,0 | 260,0 | 1200,0 |
| 12,1 | 0,0 | 0,0 | 0,0 | 5720,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |

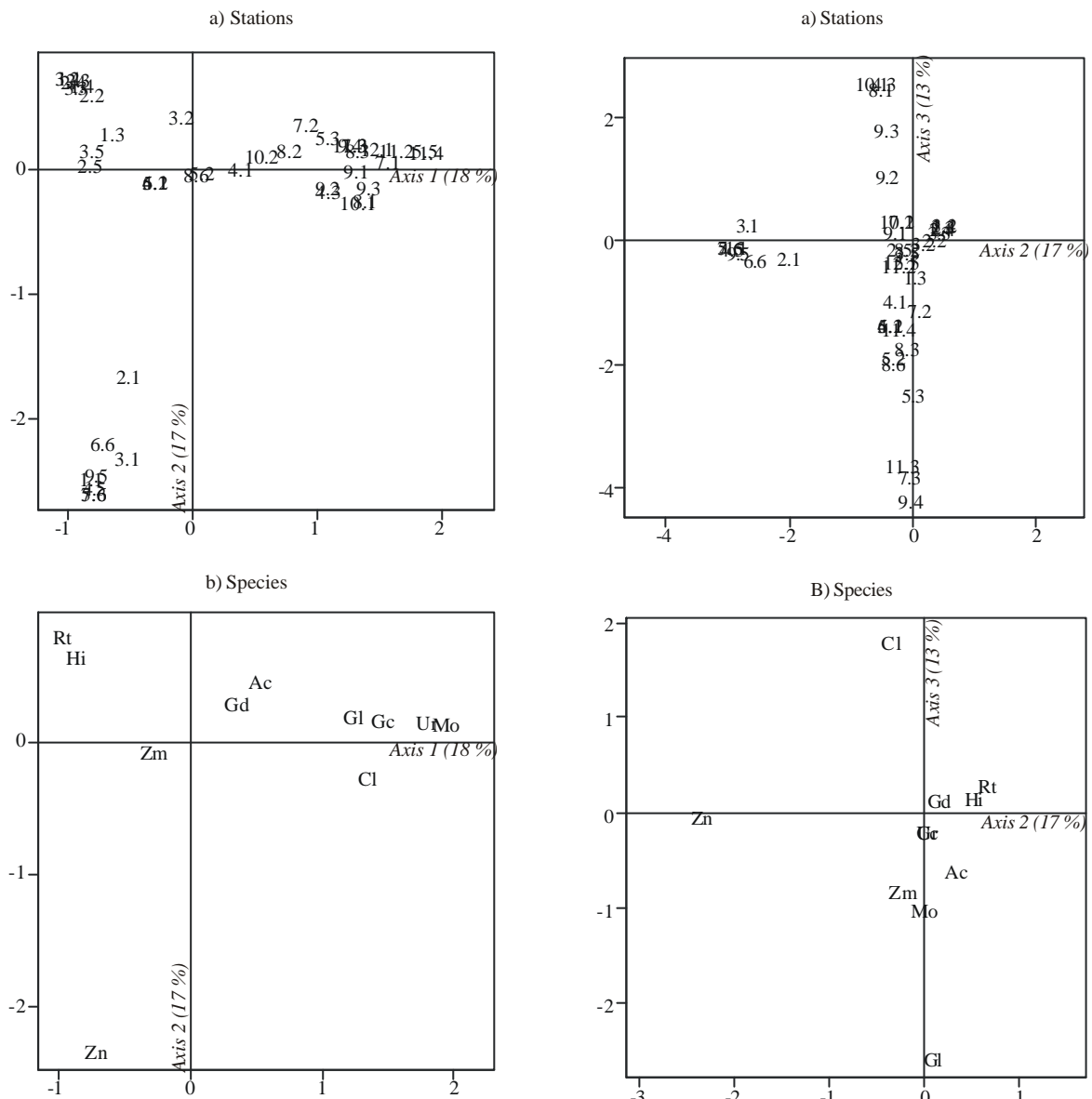


Figure 17. CA result: distribution of point-stations (up) and points-species (bottom) along axes 1 and 2 (left side) and axes 2 and 3 (right side).

The superposition of stations and species graphs then led to the definition of seven different zones in the Thau lagoon, each zone being characterized by a typical species or an association of species. Finally, a simplified macrophyte cartography for the Thau lagoon could be drawn (figure 18).

4. EXISTING SOFTWARE

A general package for multivariable analysis is ADE-4(Thioulouse *et al.* , 1997; <http://pbil.univ-lyon1.fr/ADE-4/ADE-4.html>), which include usual one-table methods like principal component analysis (PCA) and correspondence analysis (CA), spatial data analysis methods (using a total

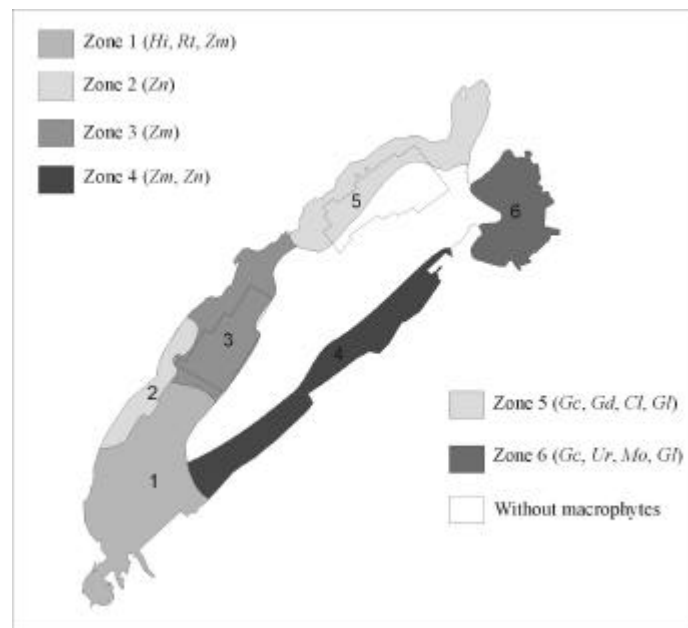


Figure 18. Simplified macrophyte mapping which resulted from the CA analysis. For each zone characteristic species are given in brackets. Labels as in Table 2.

variance decomposition into local and global components, analogous to Moran and Geary indices), discriminant analysis and within/between groups analyses, many linear regression methods including lowess and polynomial regression, multiple and PLS (partial least squares) regression and orthogonal regression (principal component regression), projection methods like principal component analysis on instrumental variables, canonical correspondence analysis and many other variants, coinertia analysis and the RLQ method, and several three-way table (k-table) analysis methods. A dynamic graphic module allows interactive operations like searching, zooming, selection of points, and display of data values on factor maps.

Multivariate and partial Mann-Kendall test for trend analysis can be found at: <http://www.mai.liu.se/~cllib/welcome/PMKtest.html>. The program developed within the EU-Project IMPACT (Estimation of human impact in the presence of natural fluctuation) is embedded in an excel file that can be download there. There is also a manual (Libiseller, 2002) and a paper describing the methodology (Libiseller and Grimbball, 2002).

Non-linear time series analysis algorithms have been implemented in the TISEAN software package (<http://www.mpiyks-dresden.mpg.de/~tisean>) (Kantz and Schreiber, 1997).

Some routines can be performed on MatLab:

- RAND and RANDN for generation of random variable;
- EXPM and LOGM for computation of matrix exponential and matrix logarithm (for anisotropy transformations).

Additionally, MatLab can be programmed for kriging (Kitanidis, 1997).

VARIOWIN (Pannatier, 1996) is a software for spatial data analysis in 2D.

Geo-EAS (Geostatistical Environment Assessment Software, Englund, and Sparks, 1991) is a package developed by the US Environmental Protection Agency. It contains a collection of interactive software tools for performing two-dimensional geostatistical analyses of spatially distributed data. Outputs are: data maps, univariate statistics, scatter plots/linear regression, and variogram computation and model fitting.

ArcView[®] (by ESRI) is popular geographic information system package. The Spatial Analyst extension provides an extensive suite of interpolation options, including inverse distance weighted, splining, trend surfacing and kriging.

ArcGIS[®] (also by ESRI) is a further development of ArcView[®]; its Geostatistical Analyst extension allows for more complete spatial analysis of data.

IDRISI has the most extensive suite of statistical and geostatistical operations in addition to standard neighborhood statistical operations, including inverse distance weighting, contour-based, kriging and trend surface interpolation. The software features logistic regression, multiple regression, trend surface, principal components and autocorrelation operations. In addition, IDRISI includes menu links to the GSTATS geostatistical software package.

Some libraries have been developed for S-PLUS[®] software (Insightful Corp.) and allow performing easily CA, extract the coordinates and axis contributions, etc.

Statbox[®] (Alysd) is another toolbox that allows the calculation of multidimensional statistics without leaving Excel[®]. More info and free demo : <http://www.alsyd.com/FP/Statbox.html>

5. CONCLUSIONS

Our aim in this exercise has been to give an idea of several tools available for data analysis with selected examples. We have not pretended to be exhaustive since a big amount of literature on analyzing biological and environmental field data exists already. The main idea was to bring to the reader (end-user) analysis techniques we are familiar with or we have been using in our research work, to see if these techniques may be of use for the management of coastal lagoons and at which level in the development of the information technology platform they should be included. For this reason a questionnaire has been prepared in APPENDIX 1.

Concerning software development, some programs have been developed, mainly the non-linear time series analysis for carrying out Hurst analysis and early detection of anoxic crises, but actually it is possible to find open source software for a big amount of the above mentioned techniques and statistical analysis package able to carry out multivariate analysis (PCA, CA, variance partitioning, clustering, etc.), time series analysis, trend analysis and statistical modeling.

REFERENCES

- Abarbanel, H. D. I., *Analysis of Observed Chaotic data*, 1996, Springer-Verlag, New-York.
- Bassingthwaighte, J. B. and Raymond, D. C., 1994, Evaluating rescaled range analysis for time series. *Ann. Biomed. Eng.* **22**, 432-444.
- Belsley, D.A., Kuh, E., and Wesch, R.E., 1980, *Regression Diagnostics*, Wiley, New York.
- Borgman, L.E., 1988, New advances in methodology for statistical tests useful in geostatistics studies, *Math. Geology*, 20(4), 383-403.
- Bosch, J., Strozzi, F., Zbilut, J. P., Zaldívar, J. M., 2004, On line application of the divergence criterion for runaway detection in isoperibolic batch reactors: Simulated and experimental results. *Computers and Chemical Engineering* **28**, 527-544.
- Box, G.E.P., and Jenkins, G.M., 1976, *Time Series Analysis*, Holden-Day, San Francisco.
- Breeden, J. L and N. H. Packard, 1994, A learning algorithm for optimal representation of experimental data, *Int. J. of Bifurcations and Chaos* **4**, 311- 326.
- Brooker, P.I., 1986, A parametric study of robustness of kriging variance as a function of range and relative nugget effect for a spherical variogram, *Math. Geology*, 18(5), 477-488.
- Broomhead, D. S. and G. P. King, 1986, Extracting qualitative dynamics from experimental data, *Physica D* **20**, 217-236.
- Burden, R. L. and Faires, J. D., *Numerical Analysis* , 3rd Ed., PWS, Boston.
- Cannon, M. J. Percival, D. B., Caccia, D. C., Raymond, G. M. and Bassingthwaighte, J. B., 1997, Evaluating scaled windowed variance methods for estimating the Hurst coefficient of time series. *Physica A* **241**, 606-626.
- Casdagli, M., Eubank, S., Farmer, J. D., Gibson, J., 1991, State space reconstruction in the presence of noise. *Physica D* **51**, 52-98.
- Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tuckey, P.A., 1983, *Graphical methods for data analysis*, Wadsworth, Belmont, CA.
- Chapelle, A., Lazure, P. and Souchu, P. 2001. Modélisation numérique des crises anoxiques (malaigues) dans la lagune de Thau (France). *Oceanologica Acta* **24**, S87-S97.
- Chatfield, C., 1997, *The analysis of time series*, 5th edition, Chapman & Hall, London.
- Chatfield, C., 2001. *Time-series forecasting*. Chapman & Hall/CRC. Boca Raton. Florida.
- Cressie, N., 1993. *Statistics for spatial data*. Wiley. New York.
- Diks, C., 1999. *Nonlinear time series analysis: Methods and applications*. World Scientific, Singapore.
- Dubrule, O., Cross validation of kriging in a unique neighbourhood, 1983a, *Math. Geology*, 15(6), 687-699.
- Dubrule, O., Two methods with different objectives, kriging and splines, 1983b, *Math. Geology*, 15(2), 245-257.
- El-Shaarawi. A. H., 1995. Trend detection and estimation with environmental applications. *Mathematics and Computers in Simulation* **39**, 441-447.
- Englund, E., and Sparks, A., 1991, "Geo-EAS 1.2.1 User's Guide," EPA Report #600/8-91/008 EPA-EMSL, Las Vegas, NV.

- Esterby S., 1996. Review of methods for the detection and estimation of trends with emphasis on water quality application. *Hydrological Processes* **10**, 127-149.
- Filliben, J.J., 1975, The probability plot correlation test for normality, *Technometrics*, 17(1), 111-117.
- Gilbert, O.R. 1987. *Statistical methods for environmental pollution monitoring*. Van Nostrand Reinhold Company Inc., New York, N.Y.
- Gilmore, R., 1998, Topological analysis of chaotic dynamical systems. *Rev. Mod. Phys.* **70**, 1455-1529.
- Grassberger, P. and Procaccia, I., 1983, Characterization of strange attractors, *Phys.Rev. Lett.* **50**, 346-349
- Gros P., Hamon D., 1988. Typologie biosédimentaire de la Baie de Saint-Brieuc (Manche-Ouest), et estimation de la biomasse des catégories trophiques macrozoobenthiques. Ifremer report DERO/88.27 EL, 66 p. + annexes 86 p.
- Haining, R., 2003, *Spatial data analysis: Theory and practice*. Cambridge University Press. Cambridge.
- Hair, J., Anderson, R., Tatham, R. and Black, W., 1998. *Multivariate Data Analysis*. Prentice-Hall, International.
- Hirsh, R. M., J. R. Slack, and R. A. Smith. 1982. Techniques of trend analysis for monthly water quality data. *Water Resour. Res.* **18**, 107-121.
- Hub L. and Jones, J. D., 1986. Early on-line detection of exothermic reactions, *Plant/Operation Progress* **5**, 121-131.
- Hurst, H. E., 1951, Long-term storage capacity of reservoirs, *Trans. Am. Soc. Civ. Eng.* **116**, 770-779.
- Kantz H., Shreiber T., *Nonlinear time series analysis*, 1997, Cambridge University Press
- Kendall, M. G., 1975. *Rank Correlation Methods*, Charles Griffin, London
- Kimeldorf, G., and Wahba, G., 1970, A correspondence between Bayesian estimation of stochastic processes and smoothing by splines, *Ann. Math. Statistics*, 41, 495-502.
- Kitanidis, P.K., 1986, Parameter uncertainty in estimation of spatial functions: Bayesian analysis, *Water Resources Res.*, 22(4), 499-507.
- Kitanidis, P.K., 1997, *Introduction to Geostatistics: Applications in Hydrogeology*, Cambridge University Press, Cambridge.
- Kolmogorov, A.N., 1958, A new invariant for transitive dynamical systems, *Dokl. Akad. Nauk SSSR* **119**, 861-864.
- Lefebvre J., 1983. Introduction aux analyses statistiques multidimensionnelles. In : Masson (ed.), 3rd edition, 275 p.
- Libiseller, C. and Grimwall A., 2002. Performance of partial Mann-Kendall tests for trend detection in the presence of covariates. *Environmetrics* **13**, 71-84.
- Libiseller, C., 2002, A program for the computation of multivariate and partial Mann-Kendall tests. <http://www.mai.liu.se/~cllib/welcome/PMKtest.html>
- Mandelbrot, B. B. and Van Ness, J. W., 1968, Fractional Brownian motions, fractional noises and applications. *SIAM Rev* **10**, 422-437.
- Mandelbrot, B. B., 1983, *The fractal geometry of Nature*, W. H. Freeman. New York.
- Marinov, D., Norro, A. and Zaldivar, J. M., 2004.
- Masters, T., 1995, *Neural, novel & hybrid algorithms for time series prediction*. Wiley, New York.

- Matheron, G., 1973, The intrinsic random functions and their applications, *Adv. Appl. Prob.*, 5, 439-468.
- More, E., The variogram and its estimation, 1984. In: *Geostatistics for Natural Resources Characterization*, Verly et al. (eds.), Vol. 1, D. Reidel, Dordrecht, The Netherlands.
- Packard, N., Crutchfield, J., Farmer, D. and Shaw, R., 1980, Geometry from a time series. *Phys. Rev. Lett.* **45**, 712-715.
- Pannatier, Y., 1996, "VARIOWIN: Software for Spatial Data Analysis in 2D," Springer-Verlag, New York, NY.
- Plus M., 2001. Etude et modélisation des populations de macrophytes dans la lagune de Thau (Hérault, France). PhD thesis, University of Paris 6, 354 p. + annexes 15 p.
- Ripley, B. D., 1981, *Spatial Statistics*. Wiley. Chichester.
- Sauer, T., Yorke, Y and Casdagli, M., 1991, Embedology, *J. Stat. Phys.* **65**, 579-616.
- Schreiber, T., 1999, Interdisciplinary application of nonlinear time series methods. *Physics Reports*.
- Shapiro, S.S., and Wilks, M.B., 1965, An analysis of variance tests for normality, *Biometrika*, 52, 691-710.
- Takens, F., 1981, in *Dynamical Systems and Turbulence*, Warwick 1980, vol. 898 of Lecture Notes in Mathematics, edited by A. Rand and L.S Young, Springer, Berlin, pp. 366-381.
- Thioulouse, J., Chgessel, D., Dolédec, S. and Olivier J. M. ,1997. ADE-4: A multivariate analysis and graphical display software. *Statistics and Computing* **7**, 75-83.
- Viaroli P., Giordani G. , Bartoli M., Naldi M., Azioni R., Zizzoli D., Ferrari I., Zaldívar J. M., Bencivelli, S., Castaldelli G., Fano E. A., 2004. The Sacca di Goro and an arm of the Po river. The handbook of Environmental Chemistry (Ed. in Chief: O. Hutzinger) Volume 5. Water pollution: estuaries. Volume editor: P.J. Wangersky Springer-Verlag, Berlin (Accepted).
- Whitney, H., 1936, Differentiable manifolds. *Ann. Math.* **37**, 645-655.
- Zaldívar, J. M., Bosch, J., Strozzi, F., and Zbilut, J. P., 2004, Early warning detection of runaway initiation using chaos-like features. *Communications in Nonlinear Science and Numerical Simulation* (Accepted).
- Zaldívar, J. M., Cattaneo, E., Plus, M., Murray, C. N., Giordani. G. and Viaroli, P., 2003, Long-term simulation of main biogeochemical events in a coastal lagoon: Sacca di Goro(Northern Adriatic Coast, Italy). *Continental Shelf Research* **23**, 1847-1876.
- Zaldívar, J.M., Galván, I. M., Strozzi, F., Gutiérrez, E., and Tomasin, A., 2000, Forecasting high waters at Venice Lagoon using chaotic time series analysis and non-linear neural networks. *J. of Hydroinformatics* **2**, 61-84.

APPENDIX 1. QUESTIONNAIRE FOR END-USERS

Please indicate the importance (on a scale 0 to 5) of having any of the listed tools implemented in the DITTY DSS prototype and the level at which you wish the tool be implemented (DB = data base; GIS = Geographical Information System; ST = stand-alone tool; DSS = embedded in the decision support system).

| Statistical Techniques | 0 | 1 | 2 | 3 | 4 | 5 | DB | GIS | ST | DSS |
|--|----------|----------|----------|----------|----------|----------|-----------|------------|-----------|------------|
| Time-series analysis | | | | | | | | | | |
| Pre-processing tools (outliers,transformations,scaling, etc) | | | | | | | | | | |
| Trend analysis | | | | | | | | | | |
| Spectral analysis | | | | | | | | | | |
| Non-linear time series analysis | | | | | | | | | | |
| Others (please indicate) | | | | | | | | | | |
| Spatial Analysis | | | | | | | | | | |
| Exploratory analysis (summary stats, histograms, etc.) | | | | | | | | | | |
| Trend analysis | | | | | | | | | | |
| Model fitting and validation through kriging | | | | | | | | | | |
| Spline interpolation | | | | | | | | | | |
| Anisotropy analysis | | | | | | | | | | |
| Others (please indicate) | | | | | | | | | | |
| Multivariate analysis | | | | | | | | | | |
| Cokringing | | | | | | | | | | |
| Principal Component Analysis (PCA) | | | | | | | | | | |
| Correspondence Analysis (CA) | | | | | | | | | | |
| Clustering | | | | | | | | | | |
| Common trend analysis | | | | | | | | | | |
| Others (please indicate) | | | | | | | | | | |

Specific techniques: (Please, could you please indicate data analysis techniques not included above that you consider necessary in your work?)
