Mediterranean Marine Science
Vol. 2/2, 2001, 67-85

Spatial analysis of hydrological and phytoplanktonic data of the Bay of Tunis. Multivariate cartography

R. B. HAMADOU [1], F. IBANEZ [1], S. SOUISSI [2] and A.C. CATHELINEAU [1]

[1] Laboratoire d'Oceanographie de Villefranche. LOV – UMR 7093
Station Zoologique, BP 28, F-06230 Villefranche-sur-Mer, France
e-mail: hamadou@obs-vlfr.fr

[2] Ecosystem Complexity Research Group, Station Marine de Wimereux
Université des Sciences et Technologies de Lille, CNRS - UMR 8013 ELICO,
28 avenue Foch, B.P. 80, F-62930 Wimereux, France

Abstract

A method of cartography originally used in geology was adapted to generate regionalization and to obtain 2-D maps of multivariate marine data. The ecological purpose of the method is to divide the studied area through homogeneous regions presenting common multivariate characteristics. Firstly, transformation was applied to the original matrix of hydrological parameters in order to satisfy the condition of multinormality. Then, associative analysis was used in order to produce an easy to interpret partition of sites. The level of heterogeneity between each station and the properties of each group was assessed by measuring the Bayesian probabilities. These conditional probabilities measure the chance that each site has of belonging to a predefined group of sites. Based on the geographical positions of the stations, the probability values for each group of stations were mapped using kriging interpolation algorithm. The obtained maps of iso-probabilities for the different groups of stations were used to define homogenous zones on a single map. Including the phytoplanktonic dataset afterwards, the indicator species were identified for each zone.

This multivariate analysis was applied to a hydrological and phytoplanktonic dataset of the Bay of Tunis. Measures at surface were made at 17 stations, sampled monthly over 2 years. The results illustrated a partition of the bay considering four groups, two coastal and two central groups of stations. The importance of the inshore influence was demonstrated in the setting up of such a regionalization through the inflow of alluvium and other products of coastal activities. The significant presence of the toxic phytoplanktonic community in the bay suggests the need to institute a monitoring program.

Keywords: Mediterranean Sea, Bay of Tunis, Multivariate classification, Multivariate mapping, Regionalization, Probability, Kriging.

Introduction

In all scientific fields, research has been conducted to improve classification and discretization on space-time records. Many ecological tasks can be regarded as a procedure of regionalization, clustering individual sampling sites and as stated by LEGENDRE

& LEGENDRE (1998) the analysis of spatial patterns is of prime interest to ecologists. As in climate regionalization (WHITE *et al.,* 1991) or geostatistics (HARFF & DAVIS, 1990; GILL, 1993; HARFF *et al.,* 1993), the process of marine mapping involves identification of homogeneous regions. Unlike the regionalization of land sites, where locations close to one another on the ground present generally similar properties (OLIVER & WEBSTER, 1989), marine stations can show a high level of spatial heterogeneity under important conditions of wind, currents, stratification, freshwater inflow, and so on.

Several methods were used to compute regionalization using multivariate classification (OLIVER & WEBSTER, 1989; BOURGAULT *et al.,* 1992), principal components analysis (RICHMAN & LAMB, 1985; BOYER *et al.,* 1997) or others with the aim of decreasing the prejudiced element of classification and making possible the handling of increasingly large datasets. The results of such methods were groups of stations assimilated by visual appreciation to regions of the ecosystem. The goal of this study is to define an algorithm combining both multivariate classification and the spatial models of regionalized variables, permitting the switch from group of single stations to region of the area.

The problem thus no longer consists of delineating groups, as in cluster analysis, but in interpreting them (LEGENDRE & LEGENDRE, 1998). In fact, there is clearly a need for the identification of regional characteristics in the fields of nature monitoring, conservation, and management. In this way, the present algorithm incorporating association analysis (GILL, 1993), provides a simple way to characterize groups regarding the used descriptors. Association analysis is designed to handle binary-coded multivariate data (WILLIAMS & LAMBERT, 1959, 1960, 1961) and was originally performed for species presence-absence data. On the other hand, biological characterization was made regarding the index values method proposed by DUFRÊNE & LEGENDRE (1997) that allows an identification of the most representative species for each considered group of stations.

This algorithm will be applied to a set of data (hydrology and phytoplankton abundance) obtained in coastal water in the Bay of Tunis.

This paper presents the different theoretical steps of the developed method that will try to link the hydrological characterization to phytoplankton dynamics. The obtained result is a partition of the bay into homogeneous regions regarding their hydrological attributes coupled with a biological description.

The advantages and possible extension of the method are discussed.

Materials and Methods

Description of the site

The Bay of Tunis is located between 10°17' and 11°37' longitude East and 36°42' and 36°53' latitude North (Fig. 1), covers 361.5 km^2 and presents an average depth of 15 m.

The bay communicates with the Gulf of Tunis in the north and with 2 lagoons in the south-west. The hydrodynamic of the surface layer depends on the meteorological conditions, wind direction and speed (KOUKI, 1984) affecting the water exchange with the gulf (BEN CHARRADA, 1997) and homogeneity in the hydrological and planktonic characteristics of the bay. Several anthropic sources are present around the bay (e.g. wadies, a thermal power station, and industrial activities) involving local perturbations along the bay coasts.

Tunis Bay data set

Hydrological measures and phytoplankton catches using 2 litre Ruttner bottles (DALY YAHIA-KÉFI, 1998; DALY YAHIA, 1998) were made at 17 stations sampled monthly

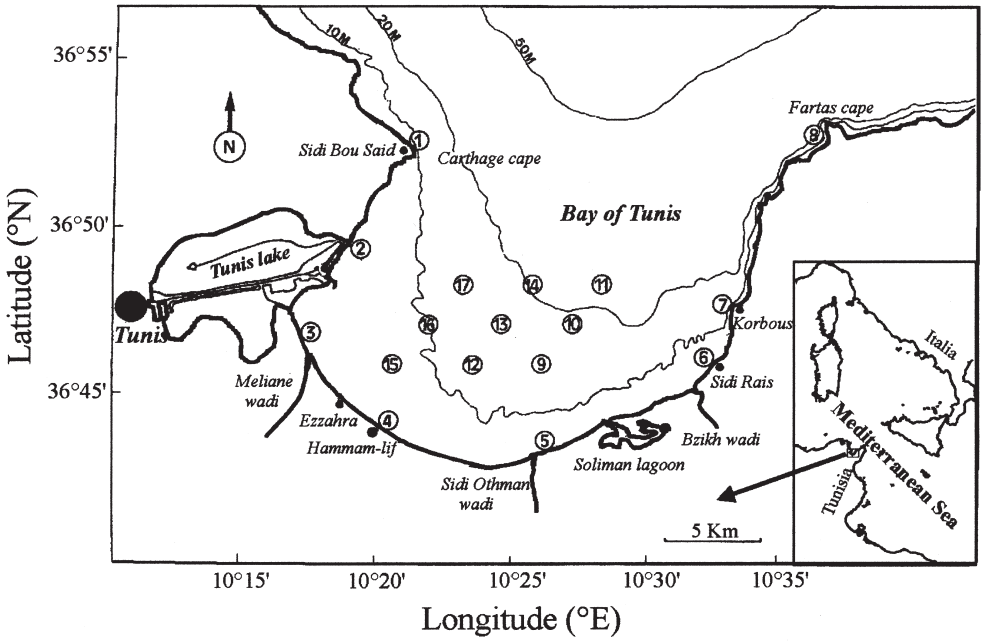


Fig. 1: Location of the study stasions on the Bay of Tunis with illustration of the different sources of water in Flow: wadies and lakes for the fresh water and the Gulf of Tunis for the inflow from offshore.

from December 93 to November 95. Sampling was conducted on the surface (-0.5m).

12 hydrological variables were measured: sea surface temperature (°C), salinity (p.p.m), dissolved oxygen (mg l^{-1}), pH, turbidity (NTU), ammonium (µmol l^{-1}), nitrites (µmol l^{-1}), nitrates (µmol l^{-1}), phosphates (µmol l^{-1}), silicates (µmol l^{-1}), chlorophyll-a (mg m^{-3}) and pheo-pigments (mg m^{-3}). Methods used for these measures are presented in SOUISSI *et al.* (2000).

188 species including 95 diatoms and 93 taxons of dinoflagellates were identified and counted according to THRONDSEN's method (1995) (DALY YAHIA-KÉFI, 1998).

Numerical analyses

The algorithm being described here corresponds to a succession of numerical analyses that allow a suitable multivariate classification of stations regarding their abiotic attributes. The clusters are then used to compute a regionalized variable allocated to all locations of the bay using a method of interpolation. A biological rendering is used to identify the relationships between biotic and abiotic characteristics for each region.

The analysis starts with two matrices (Fig. 2) corresponding to hydrological (H) and phytoplanktonic species abundance (S) data; a short period of time was to be chosen within the two years allowing the best ecological interpretation and an appropriate sense of regionalization. To illustrate the usefulness of this method, the spring season was considered in this analysis due to the significant dynamics and phenomena observed in this period of the year (i.e. phytoplankton efflorescence, high wind speed, and first stratifications). The period from March 1994 to May 1994 was used.

The matrix H relative to hydrological data is a three-dimensional one with 17 rows for the 17 stations of the bay, 12 columns relative to the 12 hydrological parameters and the third

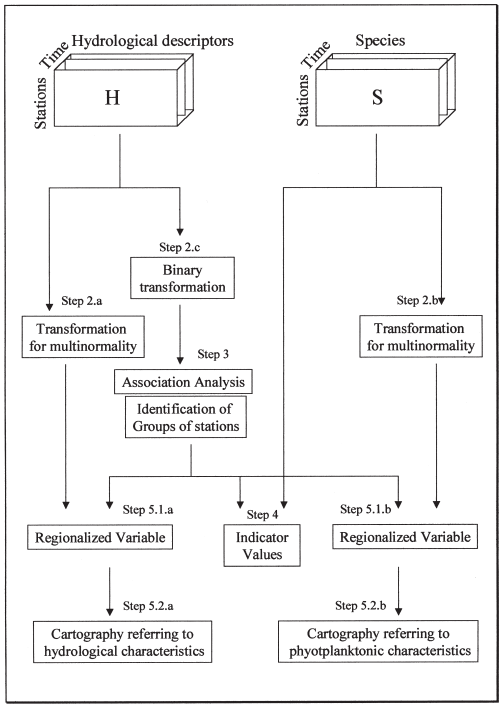


Fig. 2: Organogram presenting the different steps of the numerical analysis.

dimension is for the time space with the three suggested months.

The matrix S relative to the species abundances presents the same dimensions as the matrix H except for the number of columns, 188 species were arranged in the second space.

The following stages of numerical analyses are represented in Figure 2.

Step 1.: Descriptor selection

For the hydrological descriptors the selection was made using the equivalent vectors method, "EVM" (ESCOUFIER, 1970) (see Appendix 1). For a defined period, the EVM algorithm allows the detection of descriptors best correlated with the first principal components, already calculated for the original hydrological data. The new subset could explain a high percentage of the total variance of the original dataset (Fig. 3). The percentages of variances explained by the descriptors, ranked in decreasing magnitude are cumulated and then a threshold corresponding by eye to a sill for this function allows the most representative parameters to be detected; in this case study, five parameters were retained at the cumulative variance of 0,93.

In order to eliminate scarce species from future numerical treatments, a selection was made. Species presenting a high percentage of zeros (more than 80%) were eliminated. Also a criterion of abundance 'CA' taking into account the mean and standard deviation of abundance observed during the defined period (3 months) and for all stations was used for the previously selected species.

$$CA_s = M_S + 2 \cdot STD_S \qquad (1)$$

where M_S and STD_S are respectively the mean and the standard deviation of species s

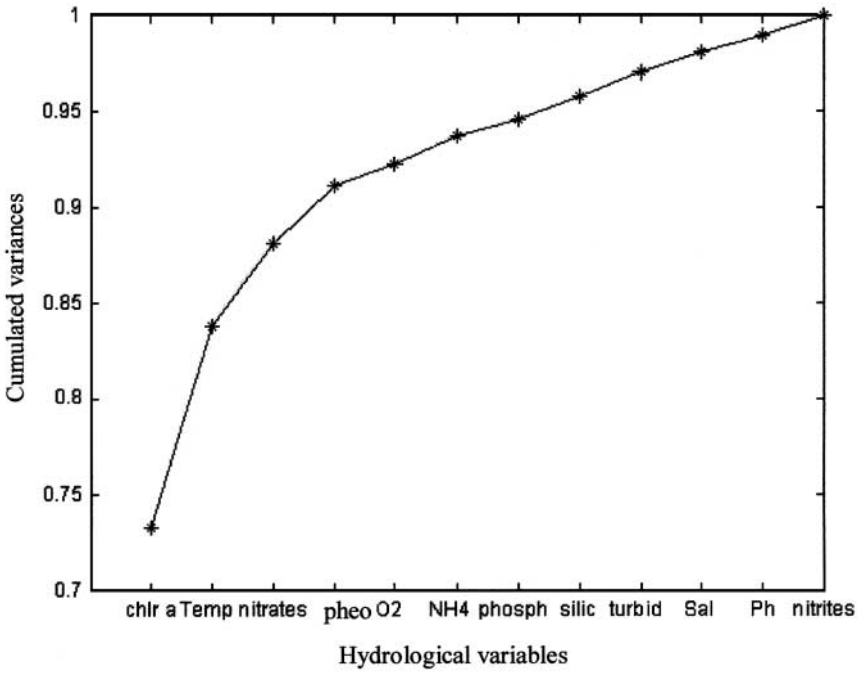


Fig. 3: Percentage of variance of each variable cumulated according to the equivalent vectors method (ESCOUFIER, 1970).

The same number of species as the first selection was held and the species common to both selection results were finally retained.

The goal was to select species that presented high abundance during the whole period and not only for restricted time intervals. Table 1 presents the retained 29 species supplied by the selection procedure from the previous 188.

Step 2.: Normalization of the data

The numerical analyses used require multinormality of the data or transformation to a binary form. Therefore, transformations yielding the best multinormality were searched for (SNEATH & SOKAL, 1973; BOURGAULT *et al.,* 1992; HERNANDEZ ENCINAS, 1994; SOKAL & ROHLF, 1995; LEGENDRE & LEGENDRE, 1998).

The selection of the suitable transformation procedure depends on the multinormality of the delivered data. In order to detect the best multinormality among the transformed data sets, the Mahalanobis generalized distance D^2 (MAHALANOBIS, 1936) was computed between each observation and the centroid of the multivariate distribution.

$$D_i^2 = d_i \cdot \Sigma^{-1} \cdot d_i' \qquad (2)$$

where d_i corresponds to the vector of centered data of the descriptors at observation i;
Σ^{-1} is the inverse of the variance--covariance matrix between descriptors.

In case of multinormality, D^2 should be normally distributed. The ordered distances are represented on a normal probability plot, and then a linear regression is computed to test how well it matches multinormality (DAGNELIE, 1975).

The search for multinormality of the data is to permit the application of the regionalization method, the transformed data are to be used only in its computing.

Table 1
The 29 retained phytoplankton species with 13 dinoflagellates and 16 diatoms. Classification in descending order of total abundance of the species. Diat: Diatom; Dino: dinoflagellate

Class	Species		Class	Species
Dino	*Scrippsiella spp.*		Dino	*Gymnodinium spp.*
Diat	*Asterionella japonica*		Diat	*Lauderia borealis*
Diat	*Nitzschia seriata*		Dino	*Alexandrium spp.*
Diat	*Chaetoceros fragilis*		Diat	*Lithodesmioides polymorphum*
Dino	*Protoperidinium quinquecorne*		Dino	*Gyrodinium spirale*
Dino	*Prorocentrum micans*		Diat	*Bellerochea malleus*
Diat	*Thalassionema nitzschioides*		Diat	*Rhizosolenia delicatula*
Diat	*Nitzschia closterium*		Diat	*Bacillaria paxillifera*
Dino	*Protoperidinium spp.*		Diat	*Bellerochea horologicallis*
Diat	*Rhizosolenia stolterfothii*		Dino	*Diplopsalis spp.*
Dino	*Ceratium furca*		Dino	*Gonyaulax spp.*
Diat	*Rhizosolenia fragilissima*		Diat	*Leptocylindrus minimus*
Diat	*Thalassiosira levanderi*		Dino	*Ceratium fusus seta*
Dino	*Prorocentrum triestinum*		Diat	*Leptocylindrus danicus*
			Dino	*Cerataulina pelagica*

Step 2.a.: Transformation of hydrological data

For the hydrological data, Arcsinus transformation was chosen because of its good match with the multinormality of the transformed data (Fig. 4A.).

$$Y_{ij} = \arcsin\sqrt{\frac{X_{ij}}{100}} \qquad (3)$$

$i = 1,\ldots, k.$ and $j = 1,\ldots, n.$

where Y_{ij} is the transformed value for the observation i and the variable j
X_{ij} is the original value
k is the number of observations.
n is the number of variables.

Step 2.b.: Transformation of phytoplanktonic data

The logarithmic transformation was retained for the same reasons for the phytoplanktonic data (Fig. 4B).

Step 2.c.: Coding hydrological data to binary form

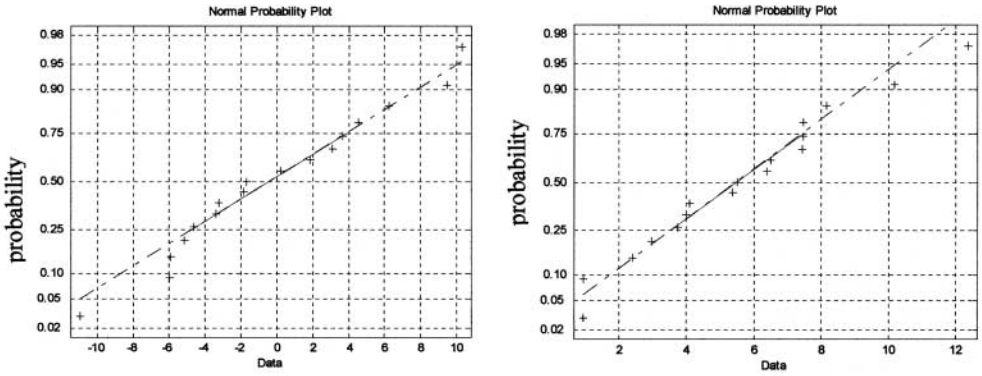
A third transformation is computed on the non-transformed original hydrological dataset which is converted into a binary form: stations presenting values higher than a particular threshold are coded 1 and 0 inversely; the median threshold was chosen, as it is independent of the values of outliers. The codification can classify the observations into more than two classes; in this work, using the median as the partition to separate the observations that present either high or low values of the considered variable.

The original dataset used for this transformation consists of a matrix of values of the five selected hydrological variables at each station. The temporal dimension was the three considered months.

This transformation is necessary to the computation of the classification method; association analysis.

Step 3. : Classification: Association analysis

The aim of this study is to obtain homogeneous groups of stations of the bay defined by their similar temporal fluctuations of hydrological parameters. So a clustering procedure was used.



A: Hydrological data **B: phytoplanktonic data**

Fig. 4: Normal probability plots for both hydrological (A) and phytoplanktonic (B) transformed data. Arcsinus and logarithmic transformations are used successively for the hydrological and the phytolplanktonic datasets.

The exploited dataset consists of the matrix of binary values obtained after transformation in step 2.c.

Association analysis is preferred for this study due to its advantages as a simple means of partition computation and the easy interpretation of the partition results (GILL, 1993).

In contrast to the agglomerative methods, association analysis, a hierarchical divisive classification, uses the whole set of stations as the starting point. The principle is to divide it into two subgroups or clusters, after which each subgroup is divided once again; the procedure is repeated until a criterion chosen to stop the partition is reached (LANCE & WILLIAMS, 1967).

The criterion for selecting the best dividing attribute is based on the discriminatory power between variables regarding the considered group of stations. At each step of the classification, it is necessary to detect the descriptor most strongly associated with all the others. The measure of association between descriptors (i) and (j) is the chi-square statistic χ^2, given by

$$\chi^2_{ij} = \frac{\left[\left(NN_{ij} - N_i N_j\right)^2 N\right]}{\left[N_i\left(N - N_i\right) N_j\left(N - N_j\right)\right]} \qquad (4)$$

where N = total number of stations; N_i, N_j, the number of stations in which the variable *(i)* and *(j)* are coded 1; and N_{ij} = number of stations in which both variables *(i)* and *(j)* are coded 1.

The χ^2 values relative to each descriptor *(j)* are summed up:

$$A_j = \sum_{i=1,\, i \neq j}^{m} \chi^2_{ij} \qquad (5)$$

where m = number of variables

This value provides a measure of its "overall associativity" in the considered group of stations (GILL, 1993). The variable possessing the highest overall associativity is selected as the best divisive attribute, the same measure serves also as termination coordinate of the branches in the dendrogram. Division stops when a desired number of clusters is reached or when the last group is considered to be homogeneous enough to make further partition unusable.

Step 4.: Biological characterisation: Indicator species

The identification of species representative of each group of stations, detected taking into account association analysis, is made according to a flexible asymmetrical approach proposed by DUFRÊNE & LEGENDRE (1997) and annotated in LEGENDRE & LEGENDRE (1998). The Indicator Value (IndVal) consists of identifying indicator species defined to be mostly found only in the considered group and homogeneously present in the majority of stations belonging to that group.

The IndVal for each species j in each cluster of sites k; is estimated by the product of two values A_{kj} and B_{kj}:

$$IndVal_{kj} = 100\, A_{kj}\, B_{kj} \qquad (6)$$

A_{kj} is a measure of *specificity:*

$$A_{kj} = Nindividuals_{kj} / Nindividuals_{+k} \qquad (7)$$

where $Nindividuals_{kj}$ is the mean abundance of species j across the sites pertaining to cluster k, $Nindividuals_{+k}$ is the sum of the mean abundances of species j within the various clusters.
Whereas B_{kj} is a measure of *fidelity:*

$$B_{kj} = Nstations_{kj} / Nstations_{k+} \qquad (8)$$

For B_{kj}, $Nstations_{kj}$ is the number of sites in cluster k where species j is present, $Nstations_{k+}$ is the total number of sites in that cluster.

The indicator value of species j for a partition of sites is the largest value of $IndValkj$ observed over all clusters k of that partition (LEGENDRE & LEGENDRE, 1998)

$$IndVal_j = \max [IndVal_{kj}] \qquad (9)$$

Indicator species are identified as characteristic of each cluster, and their assemblage represents the biological characterization of the groups of stations.

Step 5. : Regionalization

In this step, the bay is partitioned into as many homogeneous zones as clusters obtained by the hydrological classification. The mapping representation shows whether the structure is rather smoothly continuous or marked by sharp discontinuities

The used spatial analysis deals with multivariate regionalization, a characterization made through interpolated maps taking into account several ecological descriptors (hydrological in this case study) at the same time. The aim is to visualize the probabilities of grid nodes in the bay pertaining to each group of stations obtained by the preceding classification.

First, probabilities of stations belonging to each predefined group of stations, based on hydrological parameter classification, are computed (see Step 5.1); they correspond to the estimation of the likeness of hydrological characteristics of each station to those of each group. Second the probability values were estimated at all nodes and finally interpolated by the kriging method (MATHERON, 1965). Thus a contour map of iso-probabilities for each group is obtained (see Step 5.2).

To search for suggested relationships between the hydrological and phytoplanktonic characteristics of the different regions in the bay (SOUISSI *et al.,* 2000), the same procedure of regionalization is computed, simultaneously using the partition in clusters by association analysis (computed referring to hydrological data) and the phytoplanktonic abundance data in the computation of the probabilities. In other words, the homogeneity in planktonic characteristics of clusters obtained by hydrological classification is estimated

If these preliminary groups of stations present homogeneity in plankton characteristics, a direct relationship between hydrology and phytoplanktonic characteristics in the Bay of Tunis can be highlighted.

Step 5.1. : Regionalized variables

Classical mapping takes into account only one variable. Multivariate mapping needs to define a regionalized variable computed as a multivariate stochastic model (HARFF & DAVIS, 1990), simultaneously allowing the estimation of confidence levels associated with classification, and making spatial predictions of class memberships at locations where no observations are available.

Assuming multinormality of the data, the used metric is based on the Mahalanobis' distance D^2 computed between characteristics of a station and the centroid of a group (HARFF *et al.,* 1993):

$$D^2_{(l)}\left(Y_{(l)}, m_G\right) = \left(Y_{(l)} - m_G\right) \bullet \Sigma^{-1} \bullet \left(Y_{(l)} - m_G\right)' \quad (10)$$

where $Y_{(l)}$ are the measures computed (estimated) at the station *(l)*
m_G the centroid of variables of the considered group of sites (G)
Σ^{-1} corresponds to the inverse of the variance - covariance matrix for the group *G*

The centroid is defined as a vector containing the average values of each variable among the stations belonging to the considered group.

Here, the probability values are regarded as regionalized variables and expressed by an i-model Bayes' relationship:

$$p\left(X_{(l)} \in G_i\right) = \frac{p_i \left|\Sigma_i\right| \exp\left(-D^2_{i(l)}\right)}{\Sigma_{i \in l\,K}\, p_j \left|\Sigma_j\right| \exp\left(-D^2_{j(l)}\right)},$$

$$i,j \in I^K,\ I^K = \{1,\ldots,K\} \quad (11)$$

with an *a priori* probability p_i and *K* is the number of station groups obtained after clustering.

The *a priori* probability p_i is defined as the expected value of a normal probability that a station *(l)* became a member of the cluster *(i)*. It corresponds to the ratio of the number of stations in the cluster *(i)* versus the total number of stations

Assuming that the dispersion matrices are equals (HARFF & DAVIS, 1990),

$$\Sigma_i = \Sigma_j = \Sigma_0 \qquad \forall i, j \in I^K$$

a pooled variance-covariance matrix Σp (COOLEY & LOHNES, 1971; LEGENDRE & LEGENDRE, 1998) was used as a substitute for the normal dispersion matrix Σ in the D^2 computation.

Step 5.2. : Mapping procedure

The data are assumed to be isotropic, i.e. for all geographic directions of stations considered to be influencing the estimation, the variogram function (see formula 12) is considered to be the same.

$$\gamma(d)_{est} = \frac{1}{2n(d)} \sum_i^{n(d)} \left(Y_i - Y_{i+d}\right)^2 \quad (12)$$

where $\gamma(d)$ is the value of variance as a function of distance *d* between observations;
$n(d)$ is the number of pairs of points separated by a distance *d*.
Y_i and Y_{i+d} are the observed values at location *i* and distance *d*.

The obtained values of probabilities were mapped applying a regular grid to the bay presenting 42 lines and 46 columns and values at these nodes were interpolated using the kriging method (MATHERON, 1962). Values on nodes located out of the bay frontiers are not displayed in the results.

The points contributing to the interpolation are located in a circle drawn around each considered node and its radius is determined using the 'distance of influence of the process' determined by the sill of the variogram function. A fitted mathematical spherical

function was adjusted to the curve of the experimental variogram to solve the kriging equation system.

Another constraint for mapping was to consider a minimum of three stations taking into account the interpolation procedure; this is to avoid an estimation based on a limited number of nearby stations and permit an estimation based on the whole of the surrounding area.

To compute all the pre-cited analysis, programs were developed under Matlab Software and run on a particular computer.

Results

As presented in the Fig. 3, and according to the equivalent vectors method, first, five hydrological parameters were selected: chlorophyll-a, temperature, nitrates, pheophytin and dissolved oxygen. The selected species, 29 taxons (16 diatoms and 13 dinoflagellates) (Table 1), are present throughout the year meaning that they are adapted to the environmental condition of the Bay of Tunis and are indeed the most dominant and frequent phytoplanktonic species in the bay.

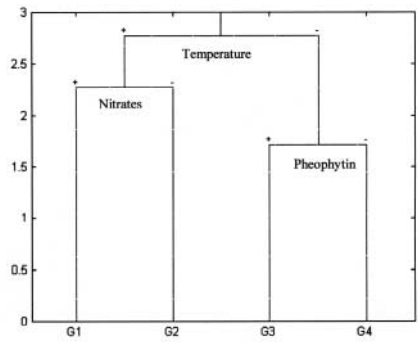


Fig. 5: Dendrogram of group of stations obtained for the hydrological characteristics by association analysis. Labels at nodal points mark the attribute that is responsible for the particular dichotomous division. The used metric is the khi square.

Association analysis

The binary transformed data of hydrology used in this classification produces the dendrogram of Figure 5. The labels at the dendrogram nodes identify the variable responsible for each division. All stations in the left-hand group (+) present high values of the attribute, whereas those on the right-hand side (-) present low values.

The ordinate value of the nodal points is the "associativity level" of the division, expressing the internal heterogeneity of the entire branch of the dendrogram before the partition (GILL, 1993).

Because of the limited number of stations the association analysis was limited to four final groups of stations and only three successive dichotomies were defined; the subdivision is based on three out the five considered variables, which in their order of significance are: temperature, nitrates and pheophytin, and using one variable for each partition.

Table 2 presents the partition of stations into the four groups (Fig. 6) and their characteristics regarding the attributes used during the association analysis.

The classification produced two coastal groups and two central groups of stations

Figure 6 presents a simple method of regionalization by gathering stations of the same

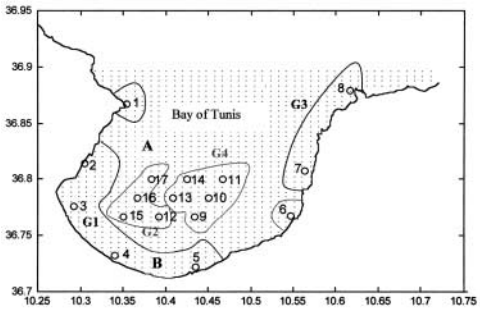


Fig. 6: A common regionalization applied to the bay of Tunis according to the classification obtained by association analysis.

Table 2
Partition of stations into the four groups and their relative characteristics with regard to the used parameters for each dichotomy.

Groups	Stations	First dichotomy	Next dichotomy
G1	2, 3, 4 and 5	Temperature	Nitrates +
G2	6, 12, 15, 16 and 17	+	Nitrates -
G3	1, 7 and 8	Temperature	Pheophytin +
G4	9, 10, 11, 13 and 14	-	Pheophytin -

Table 3
Probability values of each station belonging to each group of stations obtained by association analysis. In bold the maximum values for each station. The higher the probability value, the closer the station characteristics are to those of the group.

Station	Group 1	Group 2	Group 3	Group 4
1	0.0942	0.0042	**0.8773**	0.0243
2	**0.8272**	0.0061	0.1068	0.0599
3	**0.8440**	0.0096	0.0600	0.0865
4	**0.5026**	0.0348	0.0898	0.3729
5	**0.7710**	0.0030	0.2126	0.0133
6	0.0906	**0.8187**	0.0439	0.0468
7	0.0937	0.0034	**0.8995**	0.0034
8	0.0226	0.0807	**0.8397**	0.0570
9	0.3507	0.0111	0.0523	**0.5859**
10	0.0248	0.0728	0.0631	**0.8393**
11	0.0113	0.0066	0.0194	**0.9627**
12	0.0381	**0.9366**	0.0239	0.0014
13	0.0140	0.1389	0.0676	**0.7796**
14	0.0102	0.0124	0.0728	**0.9046**
15	0.1183	**0.8405**	0.0294	0.0118
16	0.0201	**0.9692**	0.0089	0.0018
17	0.0162	**0.9085**	0.0401	0.0351

group together. This representation does not cover the whole area of the bay. For instance, point A remains outside any estimation of its probability of belonging to a particular group; moreover point B of the bay is considered a part of region 1 only because of its geographic location being intermediate between two real stations of the same group.

Regionalization and Mapping

Probabilities of the 17 stations belonging to the 4 groups, based on hydrology, were calculated (Table 3). As might be expected, stations present their respective maximum probabilities towards the group in which they appear in the classification. However maximum values differ from one station to another.

Figure 7 shows, for each group, the map obtained by kriging of the probability values of a station belonging to that group, maximum values of interpolated probabilities are located around the stations contained by that group. Limits of estimation considered the coast land and the offshore area (Gulf of Tunis), where values are then obtained by extrapolation and not interpolation of true values of probabilities. Through these maps of "iso-probabilities", each point location of the bay had four probabilities belonging to the four groups of stations.

Points presenting maximums for the same group were gathered in the same zone, a final regionalization of the bay is then obtained and presents four homogeneous zones (Fig. 8).

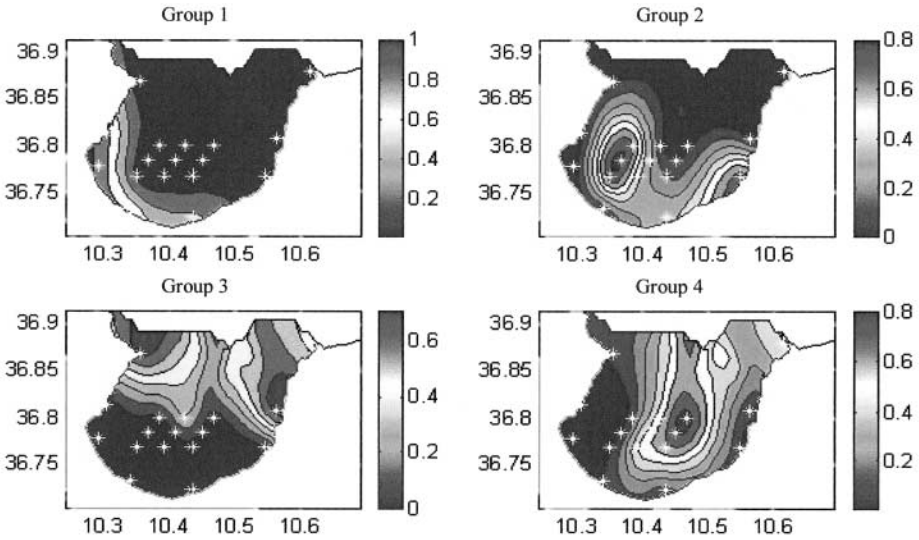


Fig. 7: Kriged probabilities of stations belonging to each group based on hydrological characteristics.

Table 4
Indicator Species for each group of stations. Species were retained according to their relative Indicator values computed using the IndVal method (Dufrêne and Legendre, 1997) Diat: Diatom; Dino: Dinoflagelate

Group 1	*Bellerochea malleus (Diat)*
	Lithodesmioides polymorphum (Diat)
	Bellerochea horologicallis (Diat)
Group 2	*Ceratium fusus seta (Dino)*
	Rhizosolenia delicatula (Diat)
	Gonyaulax spp. (Dino)
	Rhizosolenia fragilissima (Diat)
	Ceratium furca eugrammum (Dino)
Group 3	*Alexandrium spp. (Dino)*
	Rhizosolenia stolterfothii (Diat)
Group 4	*Nitzschia closterium (Diat)*

This final map presents homogenous zones relating to probabilities of each point of the bay gathered into the same group and so sharing its hydrological characteristics.

Using the same partition of stations into groups, obtained according to the hydrological variables by association analysis, the probabilities of stations are computed once again but using the values of abundance of the 29 phytoplankton species selected previously. The abundance values are consequently incorporated in the formulae of D^2.

Figures 9 and 10 respectively present maps of "iso-probabilities" obtained according to the probability values of stations and the final regionalization of the bay vis-à-vis their characteristics in phytoplankton abundance.

Indicator species

The indicator value method is computed for each group of stations obtained by association analysis. Table 4 presents the indicator species for each cluster.

Indicator values for the three species *Bellerochea malleus, Lithodesmioides polymorphum* and *Bellerochea horologicallis,* show a marked confinement to group 1 (south-western bay) by high values for that group and very low values for the others; these are stenotopic species (i.e. species with small niche breadth). Group 4 presents only one species that can be assimilated as an indicator and other species present quite similar but low indicator values. In contrast a high number of species present high indicator values for groups 2 and 3.

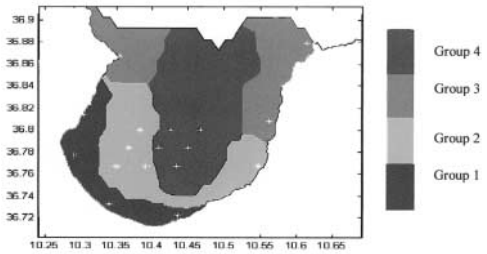


Fig. 8: Final regionalization of the bay based on hydrological characteristics.

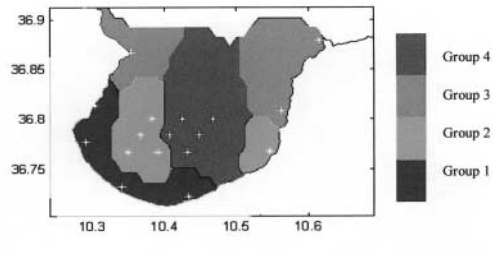


Fig. 10: Final regionalization of the bay based on phytoplanktonic characteristics.

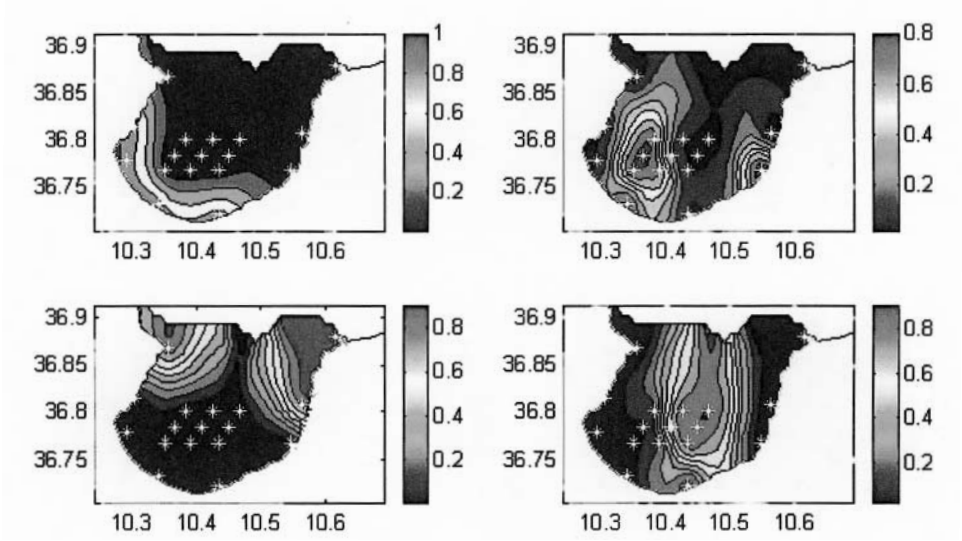


Fig. 9: Kriged probabilities of stations belonging to each group based on phytoplanktonic characteristics.

Discussion

Methodology

This study presents an original means of spatial analysis combining several methods (data normalization, clustering, mapping) and introducing new methods to marine ecological data analysis (Bayes' probability as a regionalized variable). This yields a strategy for marine regionalization involving the typification of ecological samples by multivariate classification, followed by the subdivision of the area of investigation. The result is a multivariate regionalization, or mapping of the classification onto the plane.

The main purpose of this study is to obtain a regionalization of the sampled area regarding multivariate data; generally, only one descriptor is mapped a time, using interpolation of its values in the area of investigation; here the whole data set, including time records, can be used to compute this regionalization, allowing a new spatial analysis approach for marine data.

The original data set, including hydrological and phytoplanktonic abundance data, was transformed and a selection of time-period and descriptors (variables and species) made.

The attribute selection takes into account the discriminatory power of the different variables with respect to the whole dataset. The selected descriptors are those which explain the major part of the variance.

Several transformations were identified from literature (SNEATH & SOKAL, 1973; HERNANDEZ ENCINAS, 1994; SOKAL &

ROHLF, 1995; LEGENDRE & LEGENDRE, 1998) intending to stabilize the variance between variables; in this way the multinormality of the hydrological and planktonic data is obtained, a necessary condition for the computation of the following method (Bayesian classification).

The choice of the suitable transformation goes through the test of multinormality of the obtained transformed data (DAGNELIE, 1975). In this way the best way in which data are normalized is chosen as the appropriate transformation procedure.

The multivariate classification used here is association analysis, a hierarchical, divisive and monothetic clustering. Association analysis being monothetic, using only one descriptor at a time to effect division, it is less than optimal (GILL, 1993; LEGENDRE & LEGENDRE, 1998) because of the use of only a part of the total descriptors. But it is nevertheless attractive because of the meaningful interpretation of the final partition, which is defined in a simple and clear way. This trait involves a direct characterization of obtained clusters regarding the attributes used during the successive partitions. Association analysis was successfully employed in environmental discrimination (EREZ & GILL, 1976; BUCHBINDER, 1979; AYALON *et al.,* 1981; GILL, 1993). The cited studies indicate that association analysis applied to binary data (presence/absence or transformed data) is adequate for decoding meaningful patterns in data.

For further work, the partition of observations into classes (in the process of association analysis) can be improved. For instance, the use of the codification of data into more than two classes or the use of percentiles is suggested. An optimal procedure could be the classification regarding a critical threshold presenting an ecological meaning.

The next step performed here is the computation of a Bayesian-type classification, a method originally used in geological applications and adapted in this work for marine ecological surveys. Values of probabilities of each station to be owned by each group are regarded as a regionalized variable and mapped as general characteristics of the groups towards the used variables. For this study this stochastic approach permits the coupling of multivariate classification with regionalization, allowing the detection of spatial homogeneity. Analysis of the probability values within the group of stations can provide an alternative way of identifying heterogeneity between stations and can be used in the detection of critical thresholds when computing classification.

The pre-transformation of the data involving their multinormality appears as a significant step for the whole treatment; in this way the numerical analyses presented in this work may well be appreciated with other data variables or with much longer multivariate time series.

The mapping procedure using the kriging method of interpolation presented here produces satisfactory results; applications to environmental data and ecology have been discussed by GILBERT & SIMPSON (1985), ARMSTRONG *et al.* (1989) and SOARES *et al.* (1992) and show rather good conclusions mainly regarding geostatistics. Establishing a region of influence, the kriging tool estimates value for any grid node, giving the neighbourhood influence a share in the computation.

This algorithm appears subsequently appropriate to other applications in ecological regionalization. Actually, to extend the possibilities of the methodology to other applications, other classification methods can be coupled with the regionalization and mapping procedures presented here.

Ecology

FROMENTIN *et al.* (1993) raised the problem of determining the association encountered in biological space. Authors explained the origin of the difficulty of the

diversity in marine organisms' behaviour (IBANEZ & BOUCHER, 1987; WILLIAMS & TITUS, 1988). In this study, coupling (1)the classification by association analysis obtained considers the best discriminatory power of each descriptor in stages, presenting a simple way of characterize the properties of resulting subgroups and (2)the Index Value characterization which is calculated combining the species' relative abundance with its relative frequency of occurrence in the corresponding group of stations, a study of biological characteristics with the abiotic specifications of each group is allowed.

The limitation in number of months processed in this work is enough to allow ecological interpretation; in fact, a regionalization based upon the full two years would produce maps which would be difficult to interpret because of the multiple processes that can take place and change during such a long period. However, the algorithm proposed is able to handle any size of dataset.

In this case study, the first partition is computed using chlorophyll-a as a dividing attribute separating the two groups of stations. The first, gathering stations 2, 3, 4, 5, 6, 12, 15, 16 and 17, presents values of temperature higher than the median corresponding to regions of high temperature because of the proximity to the coast and the influence of the power plant coolant outflow in the area; the second group, with stations 1, 7, 8, 9, 10, 11, 13 and 14, corresponds to regions of low temperature records. The same result was suggested by SOUISSI *et al.* (2000) using the box-plot or quartiles as a characterization method (TUKEY, J.W. 1977). Association analysis therefore appears a suitable method for coupling characterization with classification of marine multivariate ecological data.

The second partition association analysis divided these two groups into four subgroups using nitrate as the dividing attribute. Classification was imposed at this level because of the limited number of stations.

The first group, comprising the stations 2, 3, 4 and 5, forming the southwest region of the bay, presents an intense influx of perturbation and shows the presence of the three species *Bellerochea malleus, Lithodesmioides polymorphum* and *Bellerochea horologicallis.*

Indicating a high eutrophic level of the water (DALY-YAHIA KEFI *et al.,* 2001), the species *Lithodesmioides polymorphum* and *Bellerochea horologicallis* were first observed in the Mediterranean Sea in 1994 in this part of the bay. The location of three commercial ports on the coasts of this region may explain the presence of newly appearing species, in fact, HALLAGREAFF (1998) notes that ships' ballast appears as the main vector of aquatic introduction, principally of phytoplankton.

These species seem to be adapted to turbid water and presented average densities of 150 cells/l. (DALY-YAHIA KEFI *et al.,* in revision).

Group 2, confined specially to the western region of the bay, presents a low concentration of nitrates compared to group 1 explained by the multiple sources of inland water (lagoons and wadies) carrying the harvests of the intensive industrial activities and water purification existing in that region. The toxic dinoflagellate *Alexandrium spp,* present with 10 species and predominant in this region, comes forward as an important constituent of the algal population. Since it is dangerous with its paralysing toxin production observed in several cases (BALECH, 1995; CABRINI *et al.,* 1996) there appears to be a serious need for a monitoring programme to study the development and extension of these species along the nearby coasts. DALY-YAHIA KEFI (1998) notes that temperature appears as a precursor of their development; the presence of a thermal power station in a lagoon located side by side with this region may explain such a high occurrence of these species.

Group 3 representing the northern parts of the bay presents a low average of chlorophyll-a and nitrates. The absence of incoming water from the land seems to be a limitation on nutrient availability in this region

and may explains the poor richness of the phytoplankton population reported by DALY-YAHIA KEFI *et al.* (in revision). The diatom *Nitzschia closterium,* an opportunist and perpetuate species in the bay constitutes the only kind of phytoplankton that shows a real presence in this north-eastern part of the bay. The low numbers of diatoms and dinoflagellates makes the interpretation of that record difficult.

Group 4, comprising exclusively central stations, presents low concentrations of nitrates. The remark can be generalized to all nutrients (SOUISSI *et al.,* 2000), the relatively high presence of phytoplankton species suggested by the high values of chlorophyll-*a* can be explained by the current transports. This abiotic factor, establishing a continuous mixing of the water column, is a possible reason for the richness of the phytoplankton population; preventing a predominance of one or a limited number of species. The two predominant genres of phytoplankton are the dinoflagellate genus *Ceratium* and the diatom genus *Rhizosolenia. Ceratium* species in general and the species *Ceratium furca* in particular, which are common to all Mediterranean water (JACQUES, 1969) and present fairly high levels in this period of the year; spring efflorescence was observed by HALIM (1960) in Villefranche-Sur-Mer (Ligurian sea) and by MARGALEF (1969) in the waters of Barcelona and Castellon.

Diatom species *Rhizosolenia delicatula* and *Rhizosolenia fragilissima* constitute the dominant species within the population of diatoms in the occidental Mediterranean basin (Barcelona, Castellon and Banyuls) (DALY-YAHIA KEFI, 1998) but they occur differently in the diverse regions of the Mediterranean Sea.

The Bay of Tunis ecosystem shows a complex functioning, besides its temporal heterogeneity (SOUISSI *et al.* 2000), the spatial heterogeneity emphasized in this study can be explained firstly by industrial and urban influences, which play a role in the hydrology of the coastal waters, especially in the south-western region of the bay which is situated in a wind-protected area (BEN CHARRADA, 1997).

In this work, hydrological and phyto-planktonic data of the Bay of Tunis were studied. Using an original method of regionalization coupled with associative analysis, a mapping of homogeneous regions of the bay was performed. The application of this methodology can be extended to other kinds of dataset. SOUISSI *et al.* (2001) used this regionalization method to map demersal fish assemblages in the Bay of Biscay. Other applications can be envisaged concerning further kind of data (land or oceanographic, physical or chemical, biotic or abiotic components, etc.). In addition, the regionalization method can be used to study other time-scale variabilities, in consideration of which, detrending procedures can be used to simplify the ecological interpretation of the results.

This wide applicability of the methodology is due, also, to the possibilities of exchange of transformation and classification algorithms.

Acknowledgements

Data collected and borrowed from: Dr Ons Daly-Yahia Kéfi and Dr Néjib Daly-Yahia; during the research performed in the execution of their thesis work. We thank them for their availability and generosity.

This research was supported by both the "Programme Nationale de l'Environnement Côtier (PNEC art4)" and the "Comité Mixte de Coopération inter-Universitaire" (CMCU 2000 program (00/F0912)).

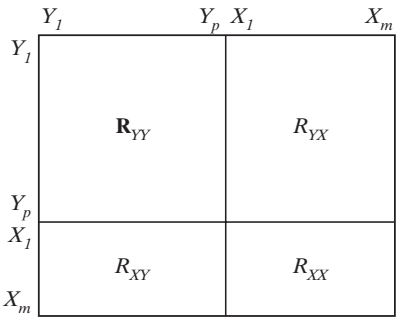
APPENDIX

EQUIVALENT VECTORS METHOD

The goal of the used Equivalent Vectors Method -EVM- (ESCOUFIER, 1970) is to extract from a set of *p* measured variables a subset of p variables ($q << p$) so that the

principal components of the subset are the closest to the ones of the original dataset. Then we obtain a classification of variables in decreasing order of the explained variance.

Let X and Y represent two multivariate data sets. We can calculate a correlation matrix between variables of Y and those of X so as:



Matrix of correlation between Y and X data sets

The correlation coefficient, RV, between the two series is (ESCOUFIER,1970)

$$RV(Y,X) = \frac{Tr\left(\mathbf{R}_{YX}\mathbf{R}_{XY}\right)}{\left[Tr\left(\mathbf{R}^2_{YY}\right)Tr\left(\mathbf{R}^2_{XX}\right)\right]^{1/2}}$$

The coefficient varies between 0 and 1. It is maximum (equal to 1) if the principal components of Y and X are proportional.

The selection procedure as used in our case study considers as Y the whole data set and as X only one descriptor. By permutation we detect the descriptor that presents the highest RV (the one most correlated to the first eigenvector of Y). We reiterate the same procedure changing just the X data set, which becomes the combination of the first selected descriptor and, step-by-step, one of the remaining variables. The RV increases at each addition of new descriptors to X if we consider the whole set of variables in X, we obtain RV equal to 1.

The RV curve as a function of the descriptors ranged in a decreasing order has a parabolic form. We retain the first descriptors corresponding to the increasing part of the curve or until the variance reaches a value of the total explained variance estimated to be optimal.

References

ARMSTRONG, M., 1989.- Geostatistics. *Vol 1and 2.* Kluwer Academic Publishers, Dordrecht. 491 + 546 pp.

AYALON, A., M. Bar-Mathews, & Y. Nathan., 1981.- Geochemistry of stream sediments along the western coast of the Gulf of Elat (Aqaba). *J. Geochem. Explor.* 15,393-404.

BALECH, E., 1995.- The genus Alexandrium Halim (Dinoflagellata). Eds., Sherkin Island Marine station.151 pp.

BEN CHARRADA, R. 1997. Etude hydro-dynamique et écologique du complexe petit Golfe-Lac de Tunis. Contribution à la modélisation de l'écosystème pélagique des eaux côtières - Petit Golfe de Tunis. Pages 384. *Ecole Nationale d'Ingénieurs de Tunis.* Université de Tunis I, Tunis.

BOURGAULT, G., D. MARCOTTE, & P. LEGENDRE., 1992.- The multivariate (Co) Variogram as a spatial weighting function in classification methods. *Mathematical Geology* 24,5,463-478.

BOYER, J. N., J. W. FOURQUREAN, & R. D. JONES., 1997.- Spatial characterization of water quality in Florida Bay and Whitewater Bay by multivariate analyses: Zones of similar influence. *Estuaries* 20,4,743-758.

BUCHBINDER, B., 1979.- Facies and environments of miocene reef limestones in Israel. *J. Sediment. Geol.* 49,1323-1344.

CABRINI, M., R. CHIURCO, S. COK, E. MARTECCHINI, I. PECCHIAR, & P. GANIS., 1996.- The occurence of Alexandrium spp, along the northern Adriatic coasts. IOC of UNESCO. 81-83.

COOLEY, W. W., & P. R. LOHNES., 1971.- Multivariate data analysis. John Wiley & Sons, New York.364 pp.

DAGNELIE, P., 1975.- L'analyse statistique à plusieurs variables. Les Presses Agronomiques de Gembloux, Gembloux (Belgique).362 pp.

DALY YAHIA, N. 1998. Dynamique saisonnière du zooplancton de la Baie de Tunis (Systématique, écologie numérique et biogéographie

Méditerranéenne). Pages 242. *Département des Sciences Biologiques.* Université de Tunis II, Tunis.

DALY YAHIA KEFI, O., E. NÉZAN, & M. N. DALY YAHIA., 2001.- Sur la présence du genre *Alexandrium Halim* (Dinoflagellés) dans la baie de Tunis (Tunisie). *Oceanologica Acta,* 24 (supplement), 17-25.

DALY YAHIA-KÉFI, O. 1998. Le phytoplancton de la Baie de Tunis (analyse systématique, biogéographique, quantitative et synécologiques des diatomées et des dinoglagellés). Pages 332. *Département des Sciences Biologiques.* Université de Tunis II, Tunis.

DALY YAHIA-KÉFI, O., S. SOUISSI, & M. N. DALY YAHIA., (submitted) Spatio -temporal distribution of dominant Diatom and Dinoflagellate species in the Bay of Tunis (south-western Mediterranean): An ecological framework.

DUFRENE, M., & P. LEGENDRE., 1997. Species assemblages and indicator species: The need for a flexible asymmetrical approach. *Ecological Monographs* 67,3,345-366.

EREZ, Y., & D. GILL., 1976. Multivariate analysis of biogenic constituents in recent sediments off Ras Burka, Gulf of Elat. *J. Math. Geol.* 8,6,77-98.

ESCOUFIER, Y., 1970.- Echantillonage dans une population de variables aléatoires réelles. *Pub. Inst. Stat. Univ.* Paris. 19,1-47.

FROMENTIN, J. M., F. IBANEZ, & P. LEGENDRE., 1993. A phytosociological method for interpreting plankton data. *Mar. Ecol. Prog. Ser.* 93,285-306.

GILBERT, R. O., & J. C. SIMPSON., 1985.- Kriging for estimating spatial pattern of contaminants: potential and problems. *Environ. Monit.* Assess. 5,113-135.

GILL, D., 1993. Discrimination of sedimentary facies by association analysis. *Mathematical Geology* 25,4,471-482.

HALIM, Y., 1960. Etude quantitative et qualitative du cycle écologique des dinoflagellés dans les eaux de Villefranche-sur-mer. *Annales de l'Institut Océanographique* 38,2,124-232.

HALLEGRAEFF, G. M., 1998. Transport of toxic dinoflagellates via ships' ballast water: bioeconomic risk assessment and efficacy of possible ballast water managment strategies. *Mar Ecol Prog Ser* 168,297-309.

HARFF, J., & J. C. DAVIS., 1990. Regionalization in geology by multivariate classification. *Mathematical Geology* 22,5,573-588.

HARFF, J. E., J. C. DAVIS, & W. EISERBECK., 1993.- Prediction of Hydrocarbons in Sedimentary Basins. *Mathematical Geology* 25,7,825-836.

HERNANDEZ ENCINAS, L., 1994. Partitioning large data sets: Use of statistical methods applied to a set of Russian igneous-rock chemical analyses. *Computers & Geosciences* 20,10,1405-1414.

IBANEZ, F., & J. BOUCHER., 1987. Anisotropie des populations zooplanctoniques dans la zone frontale de Mer Ligure. *Oceanologica Acta* 10,2,205-216.

JACQUES, G., 1969. Aspects quantitatifs du phytoplancton de Banyuls-sur-mer (Golfe du Lion).III. Diatomées et Dinoflagellés de juin 1965 à juin 1968. *Vie et Milieu* 20,1-B,91-16.

KOUKI, A. 1984. Contribution à l'étude de la dynamique sédimentaire dans le petit Golfe de Tunis. Pages 167. *Département des Sciences Biologiques.* Université de Tunis II, Tunis.

LANCE, G. N., & W. T. WILLIAMS., 1967. Mixed-data classificatory programs. II. divisive systems. *Aust. Comput. J.* 1,82-85,.

LEGENDRE, P., & L. LEGENDRE., 1998. Numerical ecology. Elsevier Science B.V., Amsterdam.853 pp.

MAHALANOBIS, P. C., 1936. On the generalized distance in statistics. *Proc. Natl. Inst. Sci. India* 2,49-55.

MARGALEF, R., 1969.- Composicion especifica del fitoplancton de la costa catalano-levantina (Mediterraneo occidental) en 1962-1967. *Inv. Pesq.* 33,1,345-380.

MATHERON, G., 1962. Traités de géostatistiques appliquée. Editions technip, Paris. 334 pp.

MATHERON, G., 1965.- Les variables régionalisées et leur estimation - Une application de la théorie des fonctions aléatoires aux sciences de la nature. Masson, Paris. 305 pp.

OLIVER, M. A., & R. WEBSTER., 1989. A geostatistical basis for spatial weighting in multivariate classification. *Mathematical Geology* 21,1,15-35.

RICHMAN, M. B., & P. LAMB., 1985. Climatic pattern analysis of three- and seven-day summer rainfall in the central United States: Some methodological considerations and a

regionalization. *American Meteorological Society* 24,1325-1343.

SNEATH, P. H. A., & R. R. SOKAL., 1973. Numerical taxonomy, the principles and practice of numerical classification. W. H. Freeman and Company, San Francisco, California.

SOARES, A., J. TAVORA, L. PINHEIRO, C. FREITAS, & J. ALMEIDA. 1992. Predicting probability maps of air pollution concentration - A case study on Barreiro/Seixal industrial area. *Fourth international geostatistics congress,* Troya, Portugal.

SOKAL, R. R., & F. J. ROHLF., 1995. Biometry. W. H. Freeman and Company, New York.

SOUISSI, S., F. IBANEZ, R. BEN HAMADOU, J. BOUCHER, A.C. CATHELINEAU, F. BLANCHARD & J-C. POULARD., 2001. A new multivariate mapping method for studying species assemblages and their habitats : example using bottom trawl surveys in the Bay of Biscay (France). *Sarsia* 86:, 527-542.

SOUISSI, S., O. DALY YAHIA-KÉFI, & M. N. DALY YAHIA., 2000. Spatial characterisation of nutrient dynamics in the Bay of Tunis (South-western Mediterranean) using multivariate analyses: consequences for phyto- and zoo-plankton distribution. *Journal of Plankton Research,* 22, 2039-2059.

THRONDSEN, J., 1995. Estimating cell numbers. *In* HALLAGRAEFF, G. M., ANDERSON, D. M. & CEMBELLA, A. D. (eds), manual on harmful marine microalgae. *IOC Manuals and Guides*-UNESCO 33,63-80.

TUKEY, J. W., 1977. Exploratory data analysis. xvi + 688 pp.

WHITE, D., M. RICHMAN, & B. YARNAL., 1991. Climate regionalization and rotation of principal components. *International Journal of Climatology* 11,1-25.

WILLIAMS, W. T., & J. M. LAMBERT., 1959. Multivariate methods in plant ecology. I. Association-analysis in plant communities. *J. Ecol* 47,83-101.

WILLIAMS, W. T., & J. M. LAMBERT., 1960. Multivariate methods in plant ecology. II. The use of an electronic digital computer for association-analysis. *J. Ecol* 48,689-710.

WILLIAMS, W. T., & J. M. LAMBERT., 1961. Multivariate methods in plant ecology. III. Inverse association-analysis. *J. Ecol* 49,717-729.

WILLIAMS, B. K., & K. TITUS., 1988. Assessment of sampling stability in ecological applications of discriminant analysis. *Ecology* 69,1275-1285.