Acta Biotheoretica
December 2003; 51(4) : 277 - 294
http://dx.doi.org/10.1023/B:ACBI.0000003984.95892.bb

Fairly Processing Rare and Common Species in Multivariate Analysis of Ecological Series. Application to Macrobenthic Communities from Algiers Harbour

C. Manté[1*], J. Claudet[2*] and C. Rebzani-Zahaf[3]

[1]Laboratoire d'Océanographie et de Biogéochimie, Campus de Luminy, Case 901, F, 13288 Marseille Cedex 09, France
[2]Laboratoire MAERHA, IFREMER, BP 21105, 44311 Nantes Cedex 03, France
[3]Institut des Sciences Biologiques, Université des Sciences et Techniques Houari Boumedienne, El Alia, Algérie

*: Corresponding author : C. Manté Email: mante@com.univ-mrs.fr J. Claudet Email: jclaudet@ifremer.fr

Abstract: Systematic sampling of communities gives rise to large contingency tables summing up possible changes in the assemblages' structure. Such tables are generally analysed by multivariate statistical methods, which are ill-suited for simultaneously analysing rare and common species (Field et al., 1982). In order to separately process species belonging to either of these categories, we propose a statistical method to select common species in a sequence of ecological surveys. It is based on a precise definition of rarity, and depends on a rarity parameter. In this work, this parameter will be optimised so that the sub-table of common species captures the essential features of the complete table as well as possible.
In this way we analysed the spatio-temporal evolution of macrobenthic communities from the Algiers harbour to study the pollution influence during a year. The examination of the communities' structuring was done through Principal Components Analysis (PCA) of the species proportions table. Environmental variables were simultaneously sampled. We show that the data structure can be explained by about 25% of the total number of present species. Two environmental gradients were brought to the fore inside the harbour, the first one representing pollution, and the second one representing hydrological instabilities.
Since rare species can also convey information, the complete table was also coded according to a generalised presence/absence index and submitted to Correspondence Analysis. The results were consistent with those of PCA, but they depended on more species, and highlighted the influence of sedimentology on the assemblages composition.

Keywords: Rare species - selection - macrobenthic communities - multivariate analyses - Algiers harbour - pollution - Hellinger distance

1. INTRODUCTION

Systematic sampling of marine benthic communities over time or space results in the collection and in the identification of a great variety of species, which can be divided into two different categories:

- common species: frequently sampled, and abundant
- rare species: their presence may be significant from an ecological viewpoint (hydrodynamism, substrate, depth, biotic and abiotic factors,...) but is highly subject to sampling errors: "...by the law of small samples, they have a good chance of being substantially over- or under- represented" (Preston, 1948). They also can be accidental species (Péres and Picard, 1964).

Putting a species in either of these categories is not easy: « extreme rarity and extreme commonness lie at opposite ends of the spectrum of abundances or range sizes, but where commonness ends and rarity begins remains entirely undefined » (Gaston, 1997). Ecologists appeal to various indices to select main species in their data. For instance, in marine ecology, Ibanez (1991) kept species sampled in half the surveys, Ibanez and Dauvin (1988) selected the 30 most abundant species, Fromentin et al. (1997) took both previous parameters into account, Field et al. (1982) studied species having above 4% dominance at any one station and Souprayen *et al.* (1991) studied only ten dominant species using the Sanders (1960) index. In his book dedicated to rarity, Gaston (1994) gave a long list of other criteria adopted by researchers to discard rare species. He suggested choosing as a cut-off the first quartile of the frequency distribution of species abundance. Nevertheless, he pointed out that, this criterion being relative, a species may move in and out a category, merely because of changes in the total number of sampled individuals.

Generally, rare species are eliminated at the beginning of the analysis. Is it appropriate? From an ecological viewpoint, there is no clear answer to this question (Gaston, 1994, pp. 158-159). From the statistical side, "quantitative" exploratory techniques such as Principal Components Analysis (PCA) or Multidimensional Scaling are well-suited to process common species while, because of their scarcity, rare ones only play a minor role in such analyses (Critchley, 1988). Reciprocally, analysing common species from a "qualitative" point of view through the usual presence/absence coding is inadequate, since much information (proportions, reproduction cycles,...) is lost. An appropriate species selection method will enable us to resolve this problem by splitting the set of present species into rare and common species. Species belonging to either of these categories can then be separately processed.

Manté *et al.* (2001) proposed several methods for selecting common species, based on computing a smoothed presence/absence index γ for each species in each survey. The "local" method separately operates on each survey. It depends on a pair (α_0, η_0) of parameters, which were arbitrarily chosen in previous works (Manté *et al.*, 1995; Manté and Durbec, 1995; Manté et al., 1997). Indeed, any value (α_0, η_0) of the pair of parameters is

associated with a sub-table of the contingency table crossing all the species with all the surveys. In the present paper, we propose a strategy for choosing (α_0, η_0), in order to build a "small" sub-table as close as possible to the whole table. For that purpose, we will make use of asymptotic properties of γ to establish a relation $\eta := \eta_0(\alpha)$ between both the parameters. This relation will transform the original bivariate optimisation problem into a one-dimensional problem.

The proposed methods were tested on macrobenthic communities from Algiers harbour. In the case of PCA, we obtained legible displays with a reduced (about 27% of the sampled species) number of descriptors. PCA brought to the fore the influence of pollution and hydrodynamism on the communities' structure. The data were afterwards coded through the smoothed presence/ absence index, and the obtained table was submitted to CA. This analysis highlighted a third factor impacting on the communities' composition: sedimentology.

2. LOCAL SELECTION AND SMOOTHED PRESENCE/ABSENCE CODING

The first subsection reminds statistical definitions of rare species in a single sample. The second one is mainly dedicated to the asymptotic properties of the smoothed presence/ absence index. We establish in this part a relation between the selection parameters, which will be used in subsequent parts.

2.1 Rare species in a sample

Consider a fixed sample composed of *N* (random) specimens, belonging to one of the *S* (random) sampled taxa. It is most probable that the true number of present species was *S'>S*, but several rare species were not collected while other ones were, due to sampling fluctuations. Let K_e be the number (random) of individuals belonging to the species "e", and k_e (resp. *n, s*) be some observation of K_e (resp. *N, S*). We will assume throughout this paper that, conditionally to *N=n*, K_e obeys the binomial law $B(n, \pi_e)$.

Fix the "rarity parameter" $\alpha_0 \in [0,1]$, and let $\pi_0(n)$ be the proportion (decreasing with *n* and α_0) such that: $P_{\pi_0(n)}(K_e = 0) = \alpha_0 = (1 - \pi_0(n))^n$.

Definition 1 The species "e" is "α_0-rare" (respectively α_0-threshold) if its true proportion π_e is such that: $\pi_e < \pi_0(n)$ (resp. $\pi_e = \pi_0(n)$).

In other words, the probability of not seeing such a species in a sample of size *n* is greater than α_0. This definition does not enable us to decide from the estimate $\hat{\pi}_e := k_e / n$ of π_e whether "e" is α_0-rare or not, but we can evaluate as follows the probability of this event.

Consider the upper confidence limit $\hat{\pi}_e^*(k_e,n)$ of the true proportion π_e at some fixed confidence level η_0. By definition, whatever π_e may be, $P_{\pi_e}\left(\hat{\pi}_e^*(k_e,n) \geq \pi_e\right) \geq \eta_0$. Thus, if η_0 is large, "e" will likely be α_0-rare when $\hat{\pi}_e^*(k_e,n) \leq \pi_0(n)$. We can now lay down an operational definition of rarity (Manté *et al.*, 1995).

Definition 2 The species "e" is (α_0,η_0)-rare if $\hat{\pi}_e^*(k_e,n) \leq \pi_0(n)$.

The number of (α_0,η_0)-rare species in a survey is a decreasing function of α_0 and η_0. Thus it is important to give appropriate values to these parameters. Indeed, η_0 must be "large" in order that $\hat{\pi}_e^*(k_e,n)$ is a valuable upper limit for each proportion.

Symmetrically, a species "e" is (α_0,η_0)-rare if and only if η_0 is smaller than the level $\eta(k_e,n)$ for which $[0,\pi_0(n)]$ is a confidence interval for π_e. This level is given by:

$$\eta(k_e,n) = \sum_{m=k_e+1}^{n} \binom{n}{m} \pi_0(n)^m \left(1-\pi_0(n)\right)^{n-m} \qquad (1)$$

We can now rewrite the previous definition:

Definition 2' The species "e" is (α_0,η_0)-rare if $\eta(k_e,n) > \eta_0$.

This version is more convenient than the first one, because we merely have to compute the distribution function of a binomial law.

2.2 The smoothed presence/absence coding and its asymptotic properties

According to (1), the probability for an α_0-threshold species to have a frequency less than or equal to k_e in an n-sample is $\gamma_n(k_e) := 1-\eta(k_e,n)$. Clearly, $\gamma_n(0) = (1-\pi_0(n))^n = \alpha_0$ and $\gamma_n(n) = 1$, but it is noteworthy that $\gamma_n(k) \approx 1$ for moderates values of *k* as soon as the sample size *n* is big enough (Manté and Durbec, 1995). Thus, if α_0 is "small", γ can be considered as a smoothed presence/absence coding, and could be used to generalise classical similarity or dissimilarity indices designed for presence/absence tables. In the case of Correspondence Analysis, this coding increased dramatically the stability of the eigenvalues and scores issued from the decomposition of presence/absence tables (Manté and Durbec, 1995; Manté *et al.*, 1997).

Introduce now additional notations:

$$\lambda_0 := -\ln(\alpha_0)$$
$$\theta_i := \sum_{k=0}^{i} e^{-\lambda_0} {\lambda_0}^k / k! = P\left(\wp(\lambda_0) \leq i\right)$$
$$\Theta_{\alpha_0} := \{\theta_0, \theta_1, ..., \theta_k, ...\} \in [0,1]$$

We proved (Manté *et al.*, 2001) the punctual convergence of the index:

Proposition 1 For any positive integer *k*, and any value of the rarity parameter $\alpha_0 \in]0,1[$

$$\gamma_n(k) \xrightarrow{n \to +\infty} \theta_k$$

Thus, except in the case of small samples, the values $\{\gamma_n(k_s) \mid 1 \le s \le q\}$ obtained from sampled data will gather round values of Θ_{α_0} (Manté *et al*, 1997; Manté and Durbec, 1995). As a consequence, if *n* is large, the species "e" is (α_0, η_0)-rare if and only if $\theta_{k_e} < 1 - \eta_0$. Therefore, it is natural to compare $1 - \eta_0$ with elements of Θ_{α_0}.

Consider now be the discrete distribution $\Re_{\alpha_0}$, supported by Θ_{α_0}, such that:

$$P(\Re_{\alpha_0} = \theta_k) := e^{-\lambda} \lambda_0^{\,k} / k!$$

Another convergence result was demonstrated, concerning the empirical distribution function of the index (Manté *et al.*, 2001):

Proposition 2 If the species "e" is α_0-threshold, $\gamma_n(K_e)$ converges in distribution towards $\Re_{\alpha_0}$.

That is why, since η_0 must be "large", we proposed to assign to $1 - \eta_0$ a value having great chance to be overrunned by α_0-threshold species:

$$1 - \eta_0(\alpha_0) := Quantile(\Re_{\alpha_0}, \tau),\ \tau \in]0,1[\text{ being « small »} \qquad (2)\ .$$

At last, it is noteworthy that no species can be $(\alpha_0, \eta_0(\alpha_0))$-rare if $\tau \le \alpha_0$ (Manté *et al.*, 2001). Thus, in order to actually select species, we must choose $\tau > \alpha_0$. Otherwise, the proposed algorithm could not discard any observation.

In conclusion, the asymptotic properties of the index enable us to associate to each value of the rarity parameter α_0 a realistic value $\eta_0(\alpha_0)$ of the second selection parameter, depending only on the level τ.

3. OPTIMISING THE SELECTION PARAMETERS

The local selection consists (Manté *et al.*, 1995) in discarding first from each survey all the (α_0, η_0)-rare species, which are pooled in a pseudo-taxon named "Background Noise". In a second step, species found (α_0, η_0)-rare in all the surveys are discarded from subsequent analyses, giving rise to a sub-table extracted from the complete $n \times s$ table *T* crossing the *n* surveys with the *s* species. This method is fairly conservative and well suited for processing by PCA or MDS the proportions associated to common species (Manté *et al.*, 1995; Manté *et al.*, 1997).

The problem now consists in finding a « good » value $(\alpha_{opt}, \eta_{opt})$ of the parameter, taking into account two conflicting objectives:

1. the number s_{opt} of rows (species) of T_{opt} should be as small as possible, in order to reduce computing expense – this could lead to discard all species

2. the main features of the optimal sub-table T_{opt} should be equivalent to those of *T* – this could lead to keep all the species.

We will make use of the results of §2 to build a sequence of candidate sub-tables, and propose criterions to choose a « good » sub-table, in accordance with the above objectives.

Let $0 < \alpha_1 < \cdots < \alpha_A \geq \tau$ (fixed) be an increasing sequence of rarity parameters, and T_a be the sub-table associated with the parameter $(\alpha_a, \eta_0(\alpha_a))$; the table T_1 may be zero, while $T_A = T$.

Since PCA of the $n \times s_a$ table T_a amounts to computing its Singular Value Decomposition (SVD), the first objective is considered equivalent to minimising the computation cost of the SVD of T_a, given by Golub and Van Loan (1996):

$$CC_a := 4{M_a}^2 m_a + 22{m_a}^3, \text{ where } M_a = Max(n, s_a) \text{ and } m_a = Min(n, s_a).$$

More precisely, we will minimise the relative complexity:

$$C_a := \frac{CC_a}{CC_A} \in [0,1].$$

Concerning the second objective, we adapted ideas of Krzanowski (1987), who proposed a general method for selecting the *q* (fixed) main variables in a $n \times s$ table like *T*. Suppose that the essential dimensionality of the problem is *k* (fixed). Let us denote *Y* the $n \times k$ matrix of the first principal component scores, yielding the best *k*-dimensional approximation of *T*. The method consists in eliminating in turn the variables, in such a way that a Procrustean squared distance between *Y* and the $n \times k$ matrix of principal component scores associated with the remaining variables is minimal. This process is iterated until a fixed number *q* of variables is left.

Krzanowski's method works for any type of variables, but the number *q* of variables to select is an input of his algorithm. It must be an output of ours. Nevertheless, the Procrustean distance used in that paper is well suited for our purpose.

Let Z_a denotes the $n \times k$ matrix of principal component scores of T_a, and consider the SVD of $Z_a Y'$:

$$Z_a Y' = U_a \Sigma_a V_a'$$

1. where U_a is a $n \times k$ matrix with orthonormal columns, V_a is a $k \times k$ matrix with orthonormal columns, and Σ_a is a nonnegative diagonal matrix. The Krzanowski's distance between the configurations of the *n* surveys associated with the tables *T* and T_a is:

$$D_{Kr}{}^2(T, T_a) := Trace\left(YY' + Z_a Z_a' - 2\Sigma_a\right).$$

It is invariant under translations, rotations and reflexions. Objective 2 will consist in minimising it, the "main features" of a table consisting in its *k* first principal components.

The proposed method can be summed up by the scheme below.

1. Inputs : the complete table T and the arbitrary level τ (e.g. 0.05)

2. Choose a sequence : $0 < \alpha_1 < \cdots \alpha_{A-1} < \tau \leq \alpha_A$

3. For a=1 to A

3.1. Eliminate all the $(\alpha_a, \eta_0(\alpha_a))$- rare species : $\mapsto T_a \subseteq T$

3.2. Compute $D_{Kr}{}^2(T, T_a)$ and C_a

4. Next a

5. Select a maximal curvature point : $\left(D_{Kr}(T, T_{opt}), C_{opt}\right) \mapsto \alpha_{opt}$

6. Perform PCA of the sub-table T_{opt}.

It has been programmed in R by the second author.

4. APPLICATION TO MACROBENTHIC COMMUNITIES FROM ALGIERS HARBOUR

4.1 Data description

4.1.1 Environment

Algiers harbour is located in the Northwest of Algiers bay and covers a stretch of water of Ha 184. It is composed of three basins called "Vieux Port", "Agha" and "Mustapha" (Rebzani-Zahaf *et al.*, 1997).

The Vieux Port basin is a Ha 74 area, and its depth ranges from 7 to 20 meters. The Northern Channel connects it with the open sea. Its purposes are basically goods and passengers transportation, yachting and fishing. The sediment is sandy mud. This basin collects the urban waste of the city centre, and the bay waters through the Northern Channel.

The Agha basin is a Ha 35 area, whose depth ranges from 6 to 15 meters. It is connected with the other basins by narrow channels. It is a transhipment basin for goods. The sediment is a reduced sandy mud. The waste collected is basically urban.

The Mustapha basin is a Ha 75 area, and its depth ranges from 7 to 11 meters. The Southern Channel connects it with the open sea. This basin has heavy industrial and business activities. Several types of sediments could be observed. Deep in the docks lies a black, foetid and sticky mud. In the basin's mouth and in the manoeuvres area the mud is reduced and less foetid, mixed with numerous fragments.

4.1.2 The sampled stations

A network of 34 stations was established, covering the whole harbour area. In the Mustapha basin, 13 sample stations were examined, in the Agha basin 7, and in the Vieux Port, 14. See the location of the station on the figure n° 1.

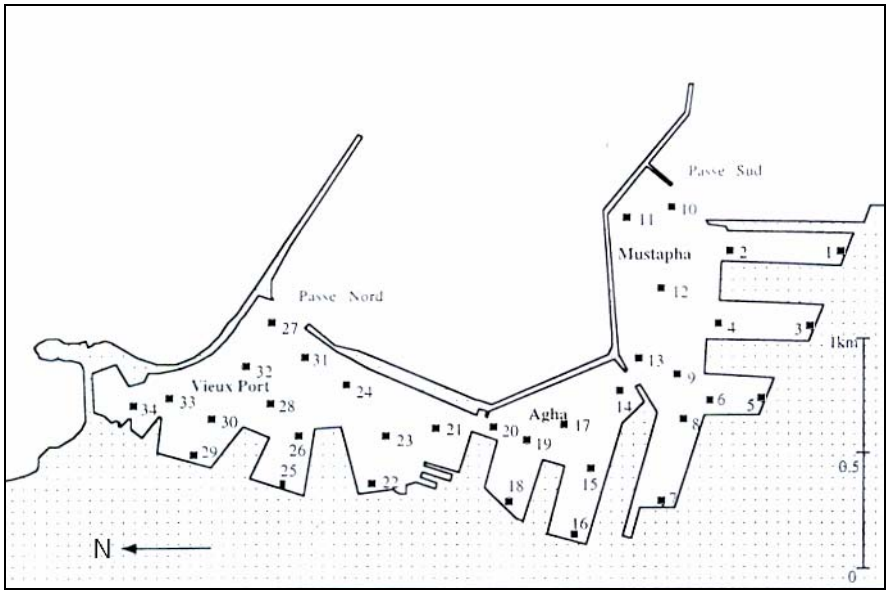

Fig 1

4.1.3 Sampling

The stations were seasonally sampled, in Fall (December 1983), in Winter (March 1984), in Spring (June 1984) and in Summer (September 1984).

The sampling device was an Orange Peel (1/12 m2) grab. In the 34 stations in Fall, Winter, Spring, and Summer, three replicates of the sediment were collected. The sediment was poured on a 1 mm meshed sieve. The benthic macrofauna was then fixed with formalin in order to be sorted out and classified in the laboratory.

The stations located at the bottom of the docks 1, 3, 5 and 7 of Mustapha basin were always azoïc, and were removed from the study. Two hundred and sixty one species were observed, giving rise to a *97x261* table. The sample size ranged from 16 to 30212 $ind.m^{-2}$; its average was 3280 $ind.m^{-2}$ and its median was 1964 organisms per square meter.

The physical and chemical parameters (temperature, salinity, pH, oxygen dissolved in sediment, organic matter in suspension) were measured at every season for the 34 stations. The nature of sediment in each sampled station was also recorded.

4.2 Selecting the optimal rarity parameter

Since the selection process has been designed for binomial counts, each survey can be straightforwardly linked to a multinomial law. Thus, before performing PCA, we first transformed the lines of each table T_a into a proportions vectors, and second by the square root application. The distance between two surveys is thus the Hellinger distance between the associated laws (Manté and Durbec, 1995).

The method exposed in §3 consists in progressively eliminating from the initial contingency table *T* all the observations of $(\alpha_a, \eta_0(\alpha_a))$-rare species, giving rise to a sequence $T_a, 1 \le a \le A$ of sub-tables. The sequence of rarity parameters used had the general form $\alpha_a = 10^{x_a}$, where x_a regularly ranged from –15 to $Log_{10}(0.052)$, and the number of computed sub-tables was *A*=100.

It is noteworthy that $D_{Kr}(T, T_a)$ is not necessarily a decreasing sequence, because $\eta_0(\alpha)$ is not a monotone function of α, while the number of selected species in each survey increases with both these parameters (see §2.1). The obtained sequence must exhibit only a decreasing trend.

4.2.1 The exhaustive case

In this case, the dimensionality of the problem (see §3) was very high: *k=77. N*early all the information from the initial table was kept: the sum of the first k eigenvalues of the original covariance matrix corresponded to 99.95% of the total variance. On figure 2 are plotted the successive distinct values of the criterions $(D_{Kr}(T, T_a), C_a)$. This curve displays a decreasing trend, from the point (0,1), corresponding to the whole table ($\alpha_A = 0.052$), to the point (3.331,0.0459) corresponding to $\alpha_1 = 10^{-15}$. In the latter case, only 34 species (plus the Background Noise) should be kept. Of course, intermediate points are more interesting, and we searched for points of maximal curvature, *i.e.* points where the steepness of the curve is much higher on the left than on the right. Two such points can be determined on the figure 2, corresponding to the indices 90 and 65. With $\alpha_{90} = 2.14\,10^{-3}$ we would select 175 species (about 67% of the total number), while with $\alpha_{65} = 7.36\,10^{-7}$, we would keep only 71 species (about 27% of the total number of species).

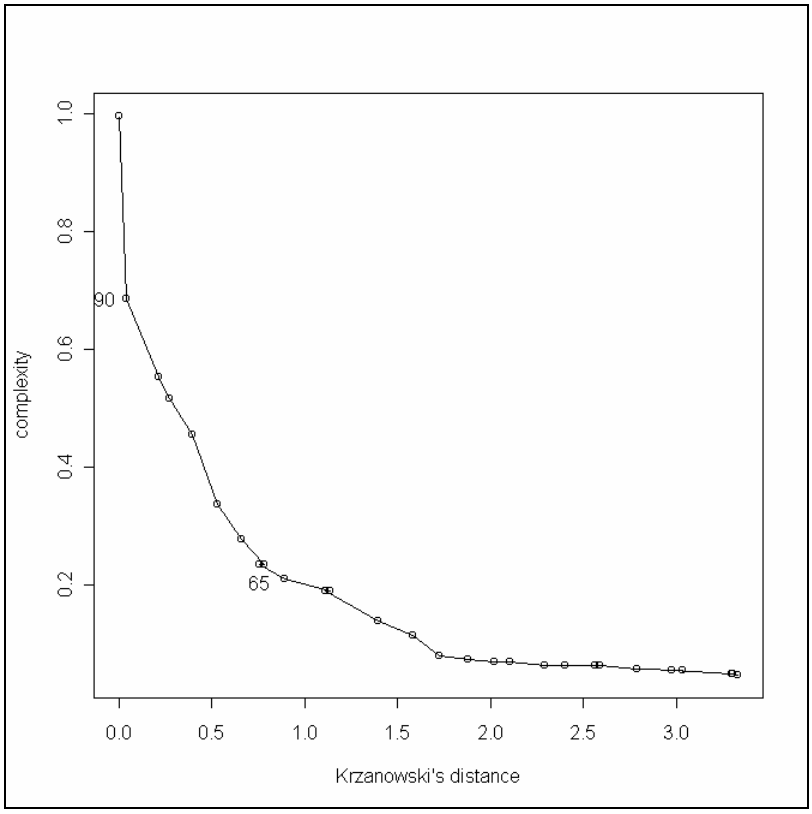

Fig 2

4.2.2 The five-dimensions case

Generally, exploratory analyses focus on the very first components extracted from the analysed tables. The suitable number *k* of components is determined from the structure of eigenvalues of the complete table. Here, the five first components corresponded to 72.5 % of the variance.

The associated curve, plotted on figure 3, also displays a decreasing trend from the point (0,1) to the point (0.2937,0.0459) corresponding to $\alpha_1 = 10^{-15}$. In this case, α_{opt} is easier to determine, and we find again $\alpha_{opt} = \alpha_{65} = 7.36\,10^{-7}$ (71 selected species); the second selection parameter is, according to (2), $\eta_{opt} := \eta_0(\alpha_{opt}) = 0.9416$.

What is, in this case, a α_{opt}-rare species? Depending on the sample size *n*, the proportion of such a species should be smaller than the critical proportion $\pi_0(n)$. This proportion ranged from 0.001868 (with *n*=7553) to 0.9707 (with *n*=4); its median value was 0.02835 (with *n*=491). It is interesting to consider the critical number obtained by taking the integer part of $n.\pi_0(n)$ (Manté & Durbec, 1995). This number ranges from 3 (*n*=4) to 14 (*n*=7553), and its median value is 13. Thus, we neglected species whose number of individuals was less than 13 in most surveys.

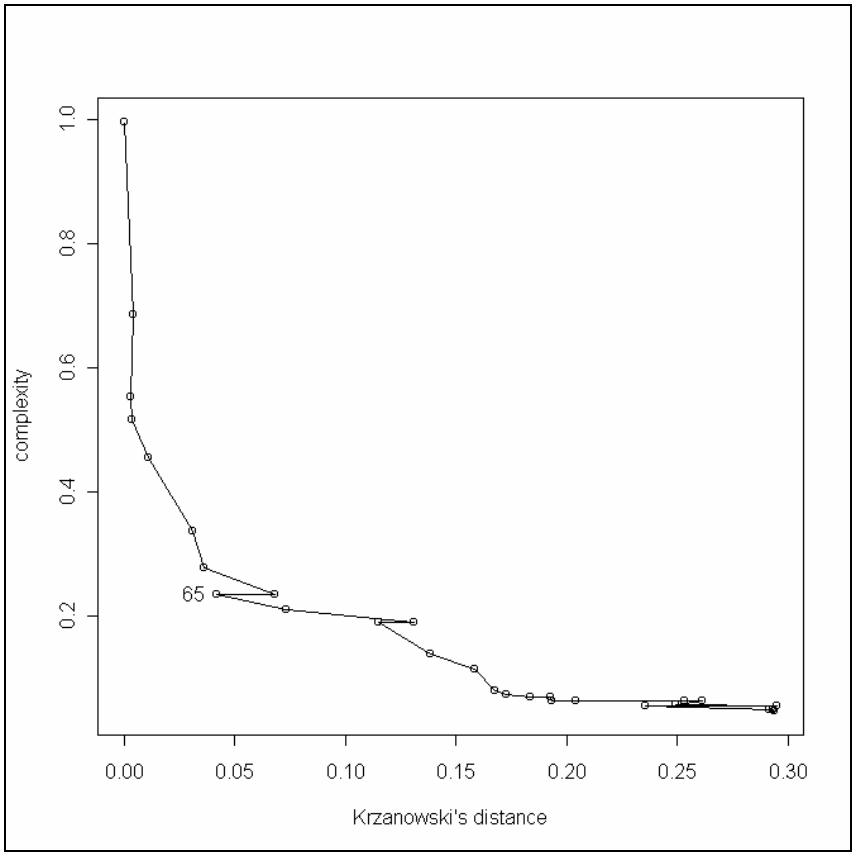

Fig 3

4.3 Exploratory analyses

4.3.1 Principal Components Analysis of the common species

The first two principal components extracted 54.76 % of the variance of the sub-table associated with α_{opt}. The next eigenvalues decreased very weakly and steadily. Consequently, we only analysed the first two principal components in order to give them an ecological signification. The seasonal effect was analysed through the representation of the factorial co-ordinates of the sampled stations. The external interpretation of the environmental variables will then be supervised. The study of the components is achieved when the analysis can be performed in terms of environmental gradients that impose a structure to the data (Kenkel and Orloci, 1986).

The first principal component (44.06 % of the variance)

The projection of the variables on the first factorial plane (components 1 – 2) is shown on the figure n° 4.

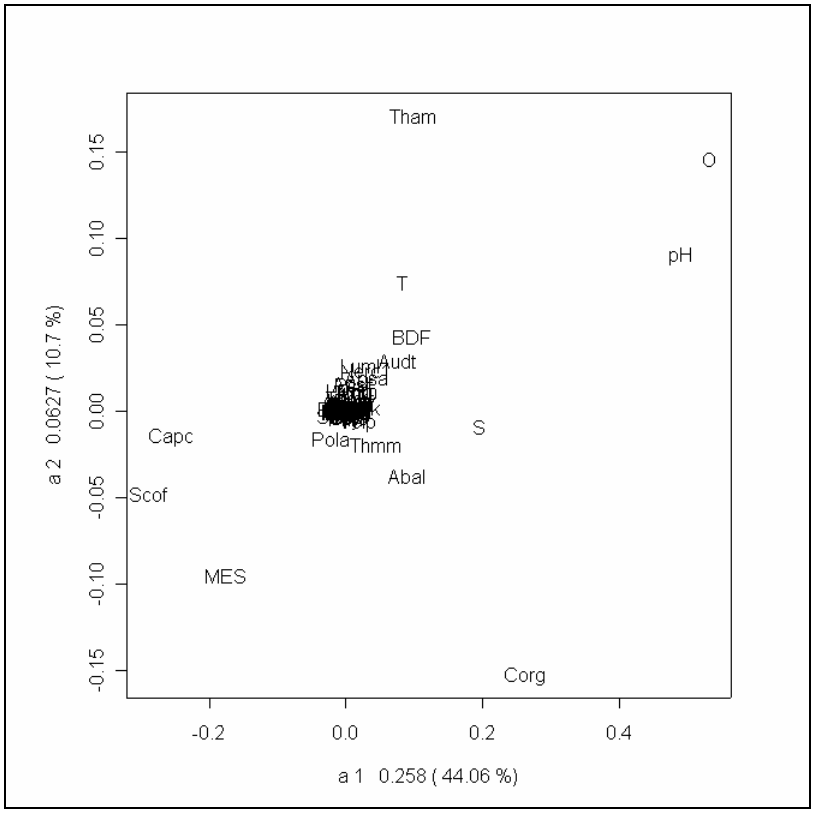

Fig 4

This axis put forward mostly *Capitella capitata*, *Scololepis fuliginosa* and *Corbula gibba*. The Mollusc *Corbula gibba* is a wide ecological range species of unbalanced polluted areas (Peres et Picard, 1964) whereas the Polychaetes *Capitella capitata* and *Scololepis fuliginosa* are first order pollution indicator species (Rebzani-Zahaf, 1990 and Rebzani-Zahaf *et al*., 1997). These two species were only located in the Mustapha basin.

Capitella capitata is a subsurface deposivore, which proliferates and resists to extreme environmental conditions. It is characteristic of a substrate enriched in organic matter, whose presence in great numbers indicates an area with an intense bacterial activity. *Scololepis fuliginosa*, a surface deposivore, is strictly correlated with the presence of organic matter in decomposition. A positive association with the “Background Noise”, BDF, means the presence of a large number of rare species. Consequently, the stations whose first co-ordinates were positive displayed a higher biodiversity than those with negative co-ordinates did.

Examine now the place and the significance of the sampled stations. Agha and Mustapha stations reached the most important absolute and relative contributions. The co-ordinates of the stations on the first axis are displayed on figure n° 5.

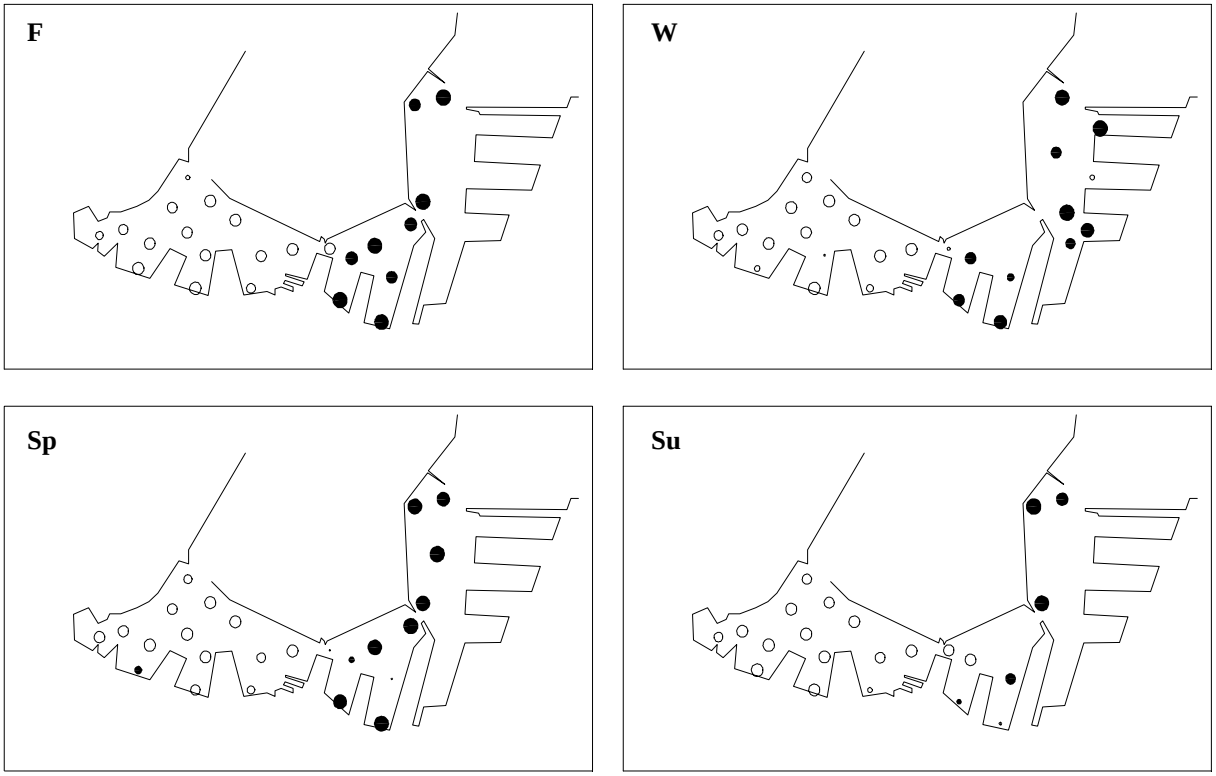

Fig 5

This component separates the Vieux Port stations (positive co-ordinates) from the Agha and Mustapha stations (negative co-ordinates). The most correlated stations were the ones of Mustapha basin, which is the most polluted. The Agha basin stations had an intermediate position between the previous stations and those of the Vieux Port, which is the less polluted basin. The spatial structure was nearly the same from Fall to Spring. In Summer, it seems the Agha basins became more homogeneous, probably due to the stagnancy of waters.

The external interpretation of the environmental variables will help us to conclude about the ecological signification of this axis.

The suspended matter had a negative correlation with this axis (- 0.175) whereas the oxygen content and the pH had a positive correlation (respectively 0.531 and 0.490). The suspended matter was more important in the Mustapha basin and the oxygen content was higher in the Vieux Port basin. The pH is linked with the oxygen content because the sedimentary organic material decomposition goes with oxygen consumption by the organisms, and the produced carbon dioxide lowers the pH.

To sum up, this component is associated with a nearly time-invariant pollution gradient.

The second principal component (10.70 % of the variance)

Examine first the internal interpretation of this component by analysing its correlation with the biotic variables (figure n° 4). The species which had an absolute contribution higher than the average absolute contribution are *Abra alba*, *Corbula gibba*, *Tharyx marioni* and *Scololepis fuliginosa*. This axis showed a high opposition between the Polychaete *Tharyx marioni*

in its positive part and *Corbula gibba* in its negative part. *Tharyx marioni* is a subsurface deposivore mud species, specific of sediment enriched in organic matter. It likes settling bottoms, where hydrodynamism is weak. *Corbulla gibba* is a wide ecological range species indicator of unstable soft bottoms. It is an indicator of sediment enriched in organic matter and necessitates a relatively important oxygenation. It can be found in areas where depth is less than 12 meters, and where the current coming from the harbour entrance or created by the ships propellers put back fine sediments in suspension. This species proliferates as long as the recycling of organic matter in the sediment is not disturbed by the hydrodynamism. The pseudo-taxon BDF was also positively associated with this component.

The co-ordinates of the sampled stations are shown in the figure n° 6. The Mustapha basin stations had absolute contributions lower than the average. The stations most negatively correlated with this axis were those presents in the ships evolution area, where hydrodynamism and depth are the most important.

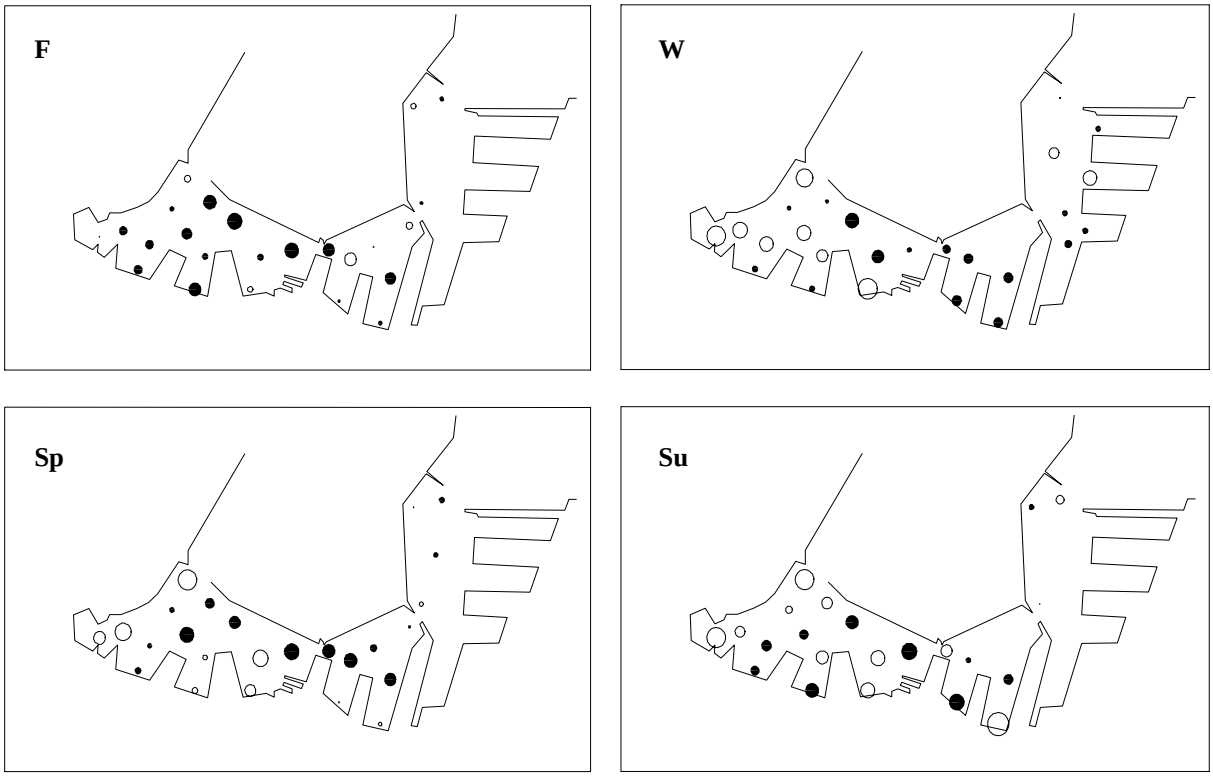

Fig 6

The main features highlighted by this component concerns the Vieux port basin, where the hydrodynamism issued from offshore waters had an important influence, except in Fall. This influence was very clear in winter, and persisted in Spring and Summer (stations 27, 32, 33 and 34, close to the North pass).

The environmental variables were weakly correlated with this component: the suspended matter was in the negative part of the axis (with a correlation of –0.094) whereas the oxygen and the pH were in its positive part (respectively 0.146 and 0.090). The suspended matter is in opposition with the oxygen content (see figure n° 4).

Thus, the second principal component defined two regions. A deep one with an important hydrodynamism (transition area) and another one where depth and hydrodynamism are weaker (more polluted area). It is a hydrodynamic gradient from

the docks to the ship evolution area. Species that had great contribution confirm again the role of the environmental factors (suspended matter, pH, dissolved oxygen).

4.3.2 Multivariate analysis of the smoothed presence/absence coding for all the species

The original counts table was encoded in smoothed presence/absence according to (1), with the same rarity parameter than in the previous section: $\alpha_{opt} = 7.36\,10^{-7}$. The corresponding asymptotic coding function (see §2.2) is displayed on figure 7.

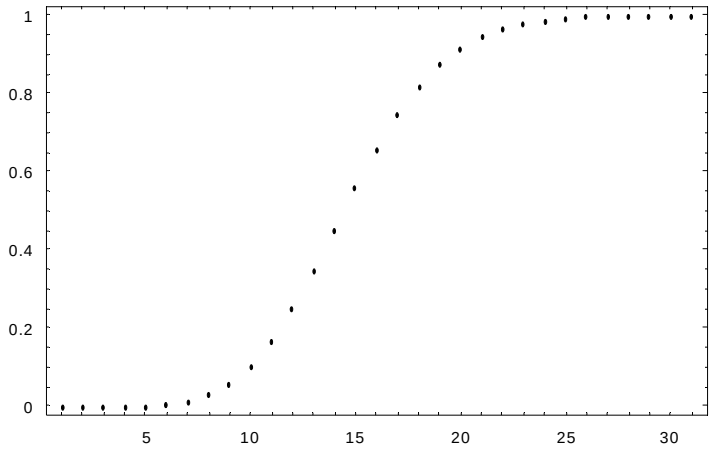

Fig 7

The smoothed presence/absence table was afterwards submitted to Correspondence Analysis. The first axis extracted 10.72 % of the variance, the second one 4.80 %, and the third one 4.36 %. From the second axis, those percentages decreased very weakly and steadily. We only give an ecological interpretation to the first axis.

A large number of species was scattered around the first axis, so only the species that had absolute contributions higher than the average absolute contribution (38) are displayed on the first factorial plane. Absolute and relative contributions are shown on Table 1.

Aponuphis bilineata, *Capitella capitata*, *Scololepis fuliginosa*, *Staurocephalus rudolphii* and *Polydora antennata* were positively correlated with this component (see figure n° 8). *Capitella capitata*, *Scololepis fuliginosa* and *Staurocephalus rudolphii* are species indicator of polluted areas that proliferate in mud sediment enriched in organic matter. *Polydora antennata* is a second order pollution indicator (Rebzani-Zahaf, 1991 and Rebzani-Zahaf *et al*., 1997). *Aponuphis bilineata* is a wide ecological range species but its relative contribution was very weak. The species negatively correlated with this axis were *Prionospio malmgreni* and *Tellina pulchella* which are exclusively specific of well sorted fine sands, *Abra alba* which is specific of the sandy-mud sediment, *Glycera convoluta*, specific of fine sands and *Heteromastus filiformis* which can be found in sandy-mud bottoms in sheltered areas. All of these species presented high relative contributions. The others species are either wide ecological range species, without specified signification, or species with very weak relative contributions.

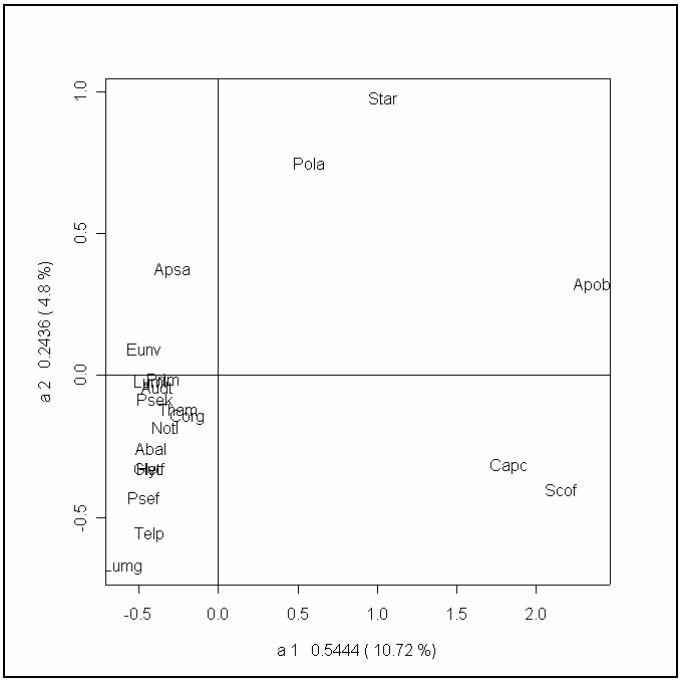

Fig 8

No seasonal effect was detected; consequently, the scores of the stations are not plotted. The highest relative contributions were those of Agha and Mustapha basins stations. All these stations were positively correlated with the first axis. The more they included plant and mollusc fragments the more they were correlated with the axis. The relative contributions of the Vieux Port stations negatively correlated with this axis were always lower than those observed for the two others basins. All of these stations present a sediment with reduced sandy-mud.

Thus, the first axis of this correspondence analysis of the smoothed presence/absence coding highlights a disruption gradient defined by a crescent sandy supply from its positive part to its negative part.

5. DISCUSSION

A large number of species are often represented in quantitative samples of marine benthic ecosystems. A great number of these species is only present in low abundance. Since their variance is not significant, their influence on PCA is very weak (Critchley, 1988), and they can only increase the noise and interpretation difficulties (Field *et al.*, 1982). Consequently, rare species are often eliminated before multivariate analyses; generally, the more relevant ones are selected according to some arbitrary frequency threshold.

To our knowledge, only Stephenson and Cook (1980) investigated the selection of representative species in the framework of data analysis. They proposed two methods. The first one was based on studying how removing in turn each species alters the structure of dissimilarity between samples. They rejected it, because of its excessive computational cost. The second one was based on the computation of the sum of dissimilarities between species. The eliminated species were in the first case those which poorly contribute to the dissimilarity structure, and in the second case those with low summed

dissimilarity. This work is similar to their first method, but it is not computing expensive. Moreover, the issues of the Stephenson and Cook's methods highly depend on the chosen dissimilarity index, and on subjective cut-off levels depending too on the dissimilarity index. In our case, the elimination of some species in a survey only depends on the number of collected individuals and on the selection parameters, which are indeed significance levels. Moreover, a species is eliminated only if it has been discarded from all the surveys. Thus, our selection process does not actually depend on the dissimilarity index; nevertheless, the optimal values of the selection parameters do.

The probabilistic method developed here allowed us to eliminate rare species in a data table considering their sampling probability. Moreover, the simultaneous representation of the relative complexity and the Krzanowski's distance gives an important information on the communities structuring.

The obtained results were different according to the data encoding (proportions, or smoothed presence/absence coding) and the method used (PCA or CA), but they were consistent with each other. The CA highlighted species whose absolute contributions to PCA were very weak, and enabled us to enhance the interpretation.

ACKNOWLEDGEMENTS

This work was carried out under the Ifremer/INSU program PNEC-ART4.

REFERENCES

CLARKE, K.R & AINSWORTH, M., 1993. A method of linking multivariate community structure to environmental variables. *Marine Ecology Progress Series*, 92: 205-219.

CRITCHLEY, F., 1988. Principal components analysis: some majorisation, perturbation and nonnegative matrix theory. *Statistique et Analyse des Données* 13, 1: 8-14.

FIELD, J.G., CLARKE, K.R. & WARWICK, R.M., 1982. A practical strategy for analysing multispecies distribution patterns. *Marine Ecology Progress Series*, 8: 37-52.

FROMENTIN, J.M., DAUVIN, J.C., IBANEZ, F., DEWARUMEZ, J.M. & ELKAIM, B., 1997. Long-term variations of four macrobenthic community structures. *Oceanologica Acta*, 20, 1: 43-53.

GASTON, K. J., 1994. Rarity. *Population and community biology series, 13*, Chapman & Hall, London.

GASTON, K. J., 1997. What is rarity ?, In : The biology of rarity. Causes and consequences of rare-common differences, W.E. Kunin and K.J. Gaston eds, Chapman & Hall, London, 30-47.

GOLUB, G.H. & VAN LOAN, F.V.L., 1996. *Matrix computation (third edition).* The John Hopkins University Press, Baltimore and London.

IBANEZ, F. & DAUVIN, J.C., 1988. Long-term changes (1977 to 1987) in a muddy fine sand *Abra alba* - *Melinna palmata* community from the Western English Channel: multivariate time-series analysis. *Marine Ecology Progress Series*, 49: 65-81.

IBANEZ, F., 1991. *Treatment of the data deriving from the COST\ 647 project on coastal benthic ecology: the within-site analysis.* In: Space and Time Series Data Analysis in Coastal Benthic Ecology. An analytical exercise organized within the framework of the COST 647, Project of Coastal Benthic Ecology, KEEGAN, B.F. (ed), Commission of the European Communities: 5-41.

KENKEL, N.C. & ORLOCI, L., 1986. Applying metric and nonmetric Multidimensional Scaling to ecological studies: some new results. *Ecoloy*, 67, 4, 919-928.

KRZANOWSKI, W.J., 1987. Selection of variables to preserve Multivariate Data Structure, using Principal Components. *Applied Statistics*, 36, 1, 22-33.

MANTE, C. & DURBEC, J.P., 1995. Sélection d'espèces dominantes et présences- absences lissées. Une application à l'analyse exploratoire de données écologiques. In *Actes des* XXVII^{èmes} *Journées de Statistique:* 436-442.

MANTE, C., DAUVIN, J.C. & DURBEC, J.P., 1995. Statistical method for selecting representative species in multivariate analysis of long-term changes of marine communities. Applications to a macrobenthic community from the Bay of Morlaix. *Marine Ecology Progress Series*, 120: 243-250.

MANTE, C., DAUVIN, J.C. & ELKAIM, B., 2001. Methods for selecting dominant species in ecological series. Application to marine macrobenthic communities from the English Channel. *J. Rec. Océanographique*, 26, 1-2, 29-36.

MANTE, C., DURBEC, J.P. & DAUVIN, J.C., 1997. Analyse de l'évolution temporelle de communautés macrobenthiques à partir des probabilités de présence des espèces. *Oceanologica Acta*, 20, 1, 71-79.

PERES J.M. & PICARD, J., 1964. Nouveau manuel de bionomie benthique de la Mer Méditerranée. *Rec. Trav. St. Mar. End.*, 47, 31, 3-137.

PRESTON, F.W., 1948. The commonnness, and rarity, of species. *Ecology*, 29, 254-283.

REBZANI-ZAHAF, C., 1990. Les peuplements macrobenthiques du port d'Alger. Evolution spatio-temporelle. Impact de la pollution. *Thèse de Magistère. ISN/USTHB, Alger*, 199p. et annexes A146p.

REBZANI-ZAHAF, C., BELLAN, G., BAKALEM, A. & ROMANO J.C., 1997. Cycle annuel du peuplement macrobenthique du port d'Alger. *Oceanologica Acta* 20, 2, 461-477.

SANDERS, H.L., 1960. Benthic studies in Buzzard Bay III. The structure of the soft-bottom community. *Limnol. Oceanogr.*, 5: 139-153.

SOUPRAYEN, J., DAUVIN, J.C., IBANEZ, F., LOPEZ-JAMAR, E. O'CONNOR, B. & PEARSON, T.H., 1991. *Long-term trends of macrobenthic communities: numerical analysis of four north-western european sites.* In: Space and Time Series Data Analysis in Coastal Benthic Ecology. An analytical exercise organized within the framework of the COST 647, Project of Coastal Benthic Ecology, KEEGAN, B.F. (ed), Commission of the European Communities: 265-438.

STEPHENSON, W. & COOK, S.D., 1980. Elimination of species before cluster analysis. *Australian Journal of Ecology*, 5: 263-273.

Figure 1: Algiers harbour map and sampled stations positioning

Figure 2: The successive distinct values of $\left(D_{Kr}(T,T_a),C_a\right)$ in the exhaustive case.

Figure 3:The successive distinct values of the criterions $\left(D_{Kr}(T,T_a),C_a\right)$ in the five-dimension case.

Figure 4: Projection of the common species and supplementary variables on the first plane of PCA. Abal : *Abra alba* ; Audt : *Audouinia (Cirriformia) tentaculata* ; *Capc* : *Capitella capitata* ; Corg ; *Corbula gibba* ; Scof : *Scololepis (Malacoceros) fuliginosa* ; Tham : *Tharyx marioni* ; MES : suspended matter; O : disolved oxygen content ; pH : pH ; S : salinity ; T : temperature.

Figure 5:Co-ordinates of the sampled stations on the first axis of the PCA on the common species. F : Fall; Sp : Spring ; Su : Summer ; W : Winter. Black dots: negative scores; white dots: positive scores; the symbol size depends on the absolute score.

Figure 6: Co-ordinates of the sampled stations on the second axis of the PCA on the common species. Same legend as figure 5.

Figure 7: The first values of $\Theta_{\alpha_{opt}}$

Figure 8: Projection of the species on the first plane of CA. For the species, same legend as in Table 1.

Table 1 Absolute and relative contributions of the species. Only the species that had absolute contributions higher than the average absolute contribution are presented. Abal : *Abra alba*, Apob : *Aponuphis bilineata*, Apsa : *Apseudes africanus orientalis*, Audt : *Audouinia (Cirriformia) tentaculata*, Capc : *Capitella capitata*, Corg : *Corbula gibba*, Eunv : *Eunice vittata*, Glyc : *Glycera convoluta*, Hetf : *Heteromastus filiformis*, Lumg : *Lumbrineris gracilis*, Luml : *Lumbrineris latrelli*, Notl : *Notomastus latericeus*, Pola : *Polydora antennata*, Prim : *Prionospio malmgreni*, Psef : *Pseudoleiocapitella fauveli*, Psek : *Pseudolirius kroyeri*, Scof : *Scololepis (Malacoceros) fuliginosa*, Star : *Staurocephalus rudolphii*, Telp : *Tellina pulchella*, Tham : *Tharyx marioni*.

Species	Abal	Apob	Apsa	Audt	Capc	Corg	Eunv	Glyc	Hetf	Lumg
Absolute contributions	197	139	44	187	3259	61	74	64	76	49
Relative contributions	1937	433	368	3422	8162	645	1018	761	1036	449
Species	Luml	Notl	Pola	Prim	Psef	Psek	Scof	Star	Telp	Tham
Absolute contributions	99	41	156	51	72	45	4137	429	84	72
Relative contributions	1216	407	1288	541	688	422	8448	1812	594	986