Reprinted from *Estuarine and Coastal Marine Science* (1976) 4, 179–205

Art. N° 326
Contribution COB N° 452

Application of Inertia Methods to Benthic Marine Ecology: Practical Implications of the Basic Options

P. Chardy[a], M. Glemarec[b], A. Laurec[a]*

[a] *Centre Oceanologique de Bretagne, 29273 Brest, Cedex, and*
[b] *Biological Oceanography Laboratory, University of West Brittany, 29283 Brest, Cedex, France.*

Received 11 June 1974 and in revised form 3 March 1975

The various so-called inertia methods, the aim of which is to summarize the relationships between points by a configuration of reduced dimension, may all be considered variants of a more general method. In this regard, the links existing between Principal Component Analysis, Principal Co-ordinates Analysis and the Analysis of Correspondences appear clearly. Three fundamental options (inherent in all inertia methods) determine the divergences among these techniques: choice of distance, choice of weights attributed to the points, choice of position of the origin. This viewpoint also shows up the relationships between the analysis of the **R** and **Q** matrices of one and the same set of data. The application of these techniques to data of the benthic bionomy of the North Gascony continental shelf illustrates the ecological implications of the three possible theoretical choices. Analysis of the differences observed at the level of the structures obtained produces an overall ecological interpretation that one method taken in isolation could not reveal.

Introduction

The capacity of electronic computers and increasingly easy access to computing facilities give biologists a greater choice of analytical methods to solve their data processing problems. With this advantage goes the need for profound consideration of the methodology, because the use of an inappropriate method is in fact a bar to progress. Being aware of this need we have attempted to discuss within the field of inertia techniques the various possible theoretical choices and to elucidate their ecological significance.

Any inertia method consists of fitting a set of points with given weights and distances into a sub-space of reduced dimension. These techniques are widely used in structural ecology and include Principal Components Analysis, the Analysis of Correspondences and Principal Co-ordinates Analysis. These methods have an essentially descriptive purpose and are based on techniques of multivariate analysis that ecologists, following Bray & Curtis (1957), refer to by the now standard name of Ordination. Specific and Common Factor Analysis is based on a completely different point of view (explanatory model founded on restrictive hypotheses) and will not be considered here.

There have been notable recent contributions expounding the theory to ecologists and illustrating the possibility of applying methods of ordination. Orloci (1966) and Austin & Orloci (1966) have analysed the consequences of the choice of methods, coefficients and

*This paper is cosignated in alphabetical order.

transformations in the field of vegetative ecology, Ibanez & Seguin (1972) have shown that the application of the three methods mentioned above to one and the same set of data can give rise to very different structures. These differences are a consequence of three fundamental options:

choice of distance
choice of weights attributed to the points
choice of position of the origin.

We propose now, so as to better understand the ecological implications of these fundamental options, to consider the classical inertia methods not as a juxtaposition of specific techniques, but rather as so many variants of the General Method defined by Lebart & Fenelon (1971). The approach will be on both the theoretical and the practical level, the practical application being the processing of benthic data from the North Gascony Continental Shelf. Our concern will, however, remain essentially methodological; the practical applications have been carried out to give a clearer understanding of a necessarily abstract theoretical exposition and not within the framework of a particular ecological investigation. The objective is to provide ecologists with the elements of analytical strategy in a field where there is an increasing number of people wishing to use such methods.

Mathematical principle of the processing methods

Notation

The treatment will be based on Imax observations characterized by Jmax variables. Whenever we represent the Imax observations by a number of points in a space with Jmax axes, we shall speak of observation points. When speaking on a purely geometrical level we shall talk simply of points and not of observations or of variables. The co-ordinates will be contained within the matrix$X$ (Imax, Jmax) having Imax rows and Jmax columns. $X\ (i,j)$ will be the jth co-ordinate of the ith point. X' (Imax, Jmax) will denote the transposed matrix ($X'\ (j,i) = X\ (i,j)$). $TI\ (i)$ will be the sum of the terms in the ith row and $XI\ (i) = TI\ (i)/I$max will be the mean of this row, $SI\ (i)$ being the corresponding standard deviation. Similarly for $TJ\ (j)$,$XJ\ (j)$ and $SJ\ (j)$. Finally, $P\ (i)$ will be the weight attributed to the point i.

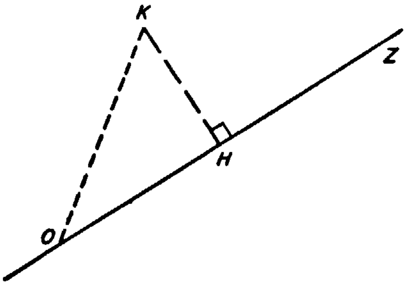
Inertia analyses: Euclidean representation of a set of points having weights and distances (Benzecri et al., 1973)

The general analysis**: **fitting into a sub-space of reduced dimension. The representation of a set of Imax observation points characterized by Jmax variables should be based on the pairwise examination of the corresponding Jmax axes. Such a task rapidly exhausts our capabilities for a high value of Jmax, hence the need to find a space of reduced size distorting the initial configuration as little as possible, which means taking account of the relative position of the Imax observation points. For instance, if O is the origin, OZ an axis passing through the origin and a point K, and H is the projection of K on the OZ axis, we may write:

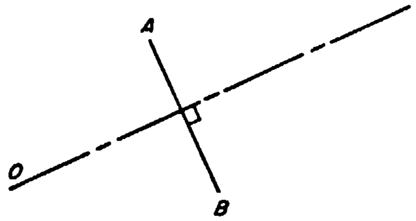
$$^{a}OK^2 = OH^2 + HK^2 \quad \text{(Scheme A)}$$

The smaller HK^2 is in relation to OH^2, the more information the position of H on the OZ axis provides on the true position of K. The simplest plan therefore, if we have Imax points

[a] OK^2 = square of the length of the segment OK.



Scheme A



Scheme B

is to choose the axis maximizing $\sum_{i=1}^{I\max} OH\,(i)^2$. The expression $\sum_{i=1}^{I\max} OH\,(i)^2$ will be a measure of the relevance of this axis, compared with $\sum_{i=1}^{I\max} OK\,(i)^2$. In mechanics, the axis obtained would be the first axis of inertia of the *I*max points having unit weights. If we resolve:

$$\sum_{i=1}^{I\max} OK\,(i)^2 = \sum_{i=1}^{I\max} OH\,(i)^2 + \sum_{i=1}^{I\max} H(i)K(i)^2.$$

Then $H\,(i)K(i)^2$ = square of the length of segment $H(i)K(i)$.

By projecting parallel to the *OZ* axis the points $K(i)$ in the perpendicular hyperspace we will have a set of points in a space of dimension of *J*max—1. The procedure may be repeated on these projections. We obtain a new system of axes in respect of which we will place *I*max points. This method constitutes the General Analysis as described by Lebart & Fenelon (1971). If, following the given notation, **X** (*I*max, *J*max) is the matrix of the data and **X**′ (*J*max, *I*max) the transposed matrix to obtain the axes of inertia we must form the matrix $\mathbf{C} = \mathbf{X}'\mathbf{X}$ (by which we recognize the form of inertia) and diagonalize it.

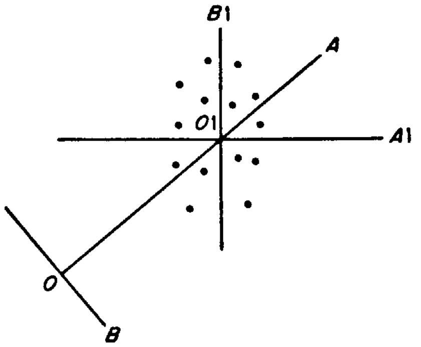
The eigenvector corresponding to the greatest eigenvalue will give the direction of the first axis and the eigenvalue will be the corresponding inertia (and similarly for the other eigenvectors and other axes). It will be noted that if $d^2\,(j1, j2)$ is the square of the distance between two variable points $j1$ and $j2$, $d^2\,(j1, j2) = \sum_i [X\,(i, j1) — X\,(i, j2)]^2$, $d^2(j1, j2) = c(j1, j1) + c\,(j2, j2) — 2\,c\,(j1, j2)$.

The basic scheme proposed by the General Analysis comprises various extensions, corresponding, as mentioned in the introduction, to displacement of the origin, choice of different weights for the points, or the introduction of non-Euclidean distances.

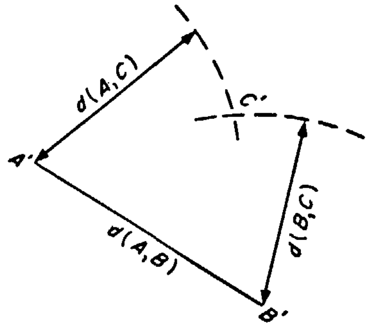
Choice of weights attributed to the various points. It may be desired to give greater importance to certain points than to others. A weight $P\,(i)$ is then given to each point and an attempt is then made to maximize $\sum_i P\,(i).\, OH\,(i)^2$. Calculation is performed in a similar manner. The form of inertia is given by:

$$c\,(j1, j2) = \sum_i P\,(i)\, X\,(i, j1).X\,(i, j2).$$

An interesting case arises when certain weights are zero. The corresponding points do not take any part in defining the axes. For instance: if the two points *A* and *B* are equidistant from the origin, if *PA* is the weight of *A* and *PB* the weight of *B*, and if *PA* = *PB* = 1,



Scheme C



Scheme D

the first axis is the mid-perpendicular of AB, If $PB = 0$ and $PA \neq 0$ this axis passes through A. If $PA = 0$ and $PB \neq 0$, this axis passes through B (Scheme B).

Choice of origin. The results obtained by the General Analysis obviously depends on the position of the origin. In the scheme C we can see the advantage to be had in placing the origin at the centre $O1$ of our cluster of points. The axes $O1\ A1$ and $O1\ B1$ are of much greater relevance than the axes OA and OB. It is the intrinsic form of the cluster that we are investigating. Note that displacing the origin to the centre of gravity of the Imax points is equivalent to centring the variables.

Choice of distances. The distance is a measure of the dissimilarity between two points[a]. The most classical distance is obviously the Euclidean distance $d^2\,(i1, i2) = \sum_j\, [X(i1, j) - X(i2, j)]^2$.

This is the distance implicit in our initial geometric consideration. There are other measures of the difference between two points:

$$\text{the } \chi^2 \text{ distance: } d^2\,(i1,\, i2) = \sum_j \frac{1}{TJ(j)} \cdot \left[\frac{X(i1, j)}{TI\,(i1)} - \frac{X(i2, j)}{TI\,(i2)}\right]^2;$$

the angular distances: $1 - c(i1, i2)$, where $c(i1, i2)$ can be a correlation coefficient or, more generally, an index of similarity.

This brings us to the case of Euclidean distance using Euclidean representations. If we start with a set (E) in which are defined distances $d(i1,\, i2)$ as being the distance between the points $i1$ and $i2$, a Euclidean representation will be a set (E') of points such that the Euclidean distances of these points are the given distances. For instance, assume three points A, B, C and the distances defined in any manner, $d(A, B)$, $d(A, C)$, $d(B, C)$. Mark a point A' then a point B' at a distance $d\,(A, B)$ from A'. We will mark C' at one of the intersections of the circles of centre A' and B' and the repective radii $d\,(A, C)$ and $d\,(B, C)$. We will then have a Euclidean representation in two dimensions of the set (A, B, C) having distances $d\,(A, B)$, $d\,(A, C)$, $d\,(B, C)$ (Scheme D).

By displacing the origin to the centre of gravity of the points A', B', C' we will have a centred Euclidean representation. We can carry out an inertia analysis of this representation and mark A', B', C' in respect of the two axes of inertia obtained.

There are several other ways of constructing a centred Euclidean representation, but it has been shown (Benzecri *et al.*, 1973) that the result after the inertia analysis does not depend on the method of construction of the Euclidean representation.

[a] In strict mathematical acceptance the term distance should be used only in case triangle inequality is satisfied.

A Euclidean representation may come about naturally: in the case of the χ^2 distance, using the notation defined already, the Euclidean representation having as its matrix the co-ordinates XE (Imax, Jmax) with $XE\,(i,j) = \dfrac{X\,(i,j)}{\sqrt{TJ(j)\;TI(i)}}$ will correspond to the Imax points. In certain cases it may be much more difficult to construct a Euclidean representation. (cf. Method of Principal Co-ordinates).

Representation of points and of variables

Suppose that we have chosen Euclidean distance and equal weights (General Analysis). We have presented our investigation as the search for new axes inside which are placed the Imax observation points.

For this purpose we have sought the axes of inertia of the cluster. We may attach to each of these axes an a priori abstract variable. If the projection of the ith point is $H(i)$ on the axis of inertia already having a unit vector, the variable in question will have the value $c\,\overline{OH}\,(i)$. The value of the multiplicative constant c will be fixed so as to 'normalize'[a] the variable corresponding to the various axes of inertia: c is chosen so that $\sum_i c^2\, OH\,(i)^2 = 1$ (c will then equal $1/\sqrt{\lambda_k}$ if λ_k is the eigenvalue corresponding to the axis k). This variable will be called a normalizing factor. To each axis of inertia (geometric concept, also called factorial axis) there thus corresponds a factor (variable-source of variation).

It is possible to write each of the initial Jmax variables as a linear combination of standardized factors. If VD_1 is the first initial variable ($VD_1\,(i) = X\,(i, 1)$, $i = 1$, Imax) we may write $VD_1 = \sum_k \mu_k\,(1)\; V_k$, V_k being the kth factor. We will have: $X(i, 1) = \sum_k \mu_k(1)\;\; V_k\,(i)$ for $i = 1$, Imax. The μ_k (1) will be the co-ordinates of the variable VD_1, in the factor space. We can represent the co-ordinates $V_k\,(i)$ of the Imax observation points in respect of the factorial axes I and II, II and III . . . . (thus using the 'factor scores' of English and American writers). Similarly, we can represent the co-ordinates $\mu_k(j)$ of the Jmax initial variables in respect of the factors I and II and III, etc . . . . We would then be using the 'factor loadings' of the English-speaking authors.

Thus far we have assumed that we had Imax observations in a system of Jmax variables, for example Imax samples with Imax species, giving us a matrix $\mathbf{X}$ (Imax, Jmax). We could have studied the dual problem by using the transposed (Jmax, Imax) matrix of the preceding matrix. Continuing with our example, we would have represented the Jmax species in the space of Imax samples. Within the framework of the General Analysis, it can be shown that (except for the conventions of representation) we would have reached the same configurations, the role of observations and of variables being changed. Note from a practical point of view that to find the axes of inertia we were obliged to diagonalize a square matrix of dimension Jmax. The dual problem makes it necessary to diagonalize the matrix of dimension Imax. Given this consequent equivalence it is possible to choose the method leading to diagonalization of the smaller matrix.

In the case of a non-Euclidean distance defined by a quadratic form, General Analysis of the Euclidean representation is performed. If we consider the case when it is proposed to reduce the variables space, it is possible to represent the variables before and after centring.

[a] 'Normalize' = to divide each series of values by the square root of the sum of the squares of all the values, so that the sum of the squares of the new series is equal to 1.

The choice of conventions becomes more complex but on the practical level we can always resort to the instance which involves diagonalizing the smaller matrix.

The allocation of weights does not complicate the problem. On the other hand, when the distance is not defined by a quadratic form, the duality of variables and observations disappears: we then have only observations and distances.

The symmetry envisaged may seem to contradict the experience of the ecologist with problems of choosing between what have been called the **R** and **Q** techniques. However, displacing the origin of the centre of gravity of the points, i.e. centering the rows, is in no way equivalent to centering the columns. Similarly, reducing the columns in no way implies that the rows are reduced by the same token. It is at this level that the problems arise.

The classic variants of inertia methods

Three standard types of inertia methods are currently used in ecology:

Principal Component Analysis
Principal Co-ordinates Analysis
Analysis of Correspondences.

Principal Component Analysis (Hotelling, 1933). Our representation of the search for a space with configuration of reduced dimensions has placed the accent on points, starting from geometrical considerations. Historically, the accent was initially on the variables. *J*max variables were considered to find the influences, first, of one virtual variable (unifactorial model of Spearman, 1904), then, still historically, of several unknown variables which were linear combinations of concrete initial variables. The effort to find these abstract variables assumed the investigation of the variance-covariance matrix or the variable correlation matrix. Whenever the ecologist considered samples as points and species as variables, he was working on a variance-covariance matrix or a correlation matrix of the species. We have seen that in General Analysis it was necessary to diagonalize the matrix $\mathbf{C} = \mathbf{X'X}$, where $\mathbf{c}\,(i1, i2) = \sum_{j} X(i1, j)\, X(i2, j)$.

When the species have been centred first, **C** is a variance-covariance matrix; when the species have been centred and reduced, **c** is a correlation matrix.

Similarly, when the ecologist deals with the dual problem, he is operating on centered, or centered and reduced samples. At this stage of the treatment, it is relevant to specify the nature of the relationships existing between the analysis of the **R** and **Q** matrices of one and the same set of data. This terminology, introduced by psychologists using Factor Analysis, then extended to ecology, means that a matrix of intersample distances is of type **Q** and a matrix of interspecific distances is of type **R** (Williams & Dale, 1965; Orloci, 1967). The fact that there is a certain confusion between the terms **R** and **Q** in the literature (Ivimey-Cook, Procter & Wigston, 1969) is of little relevance; the main thing is that the choice of the resolution space (species space or sample space) should be capable, if necessary, of producing various factorial structures. Principal Component Analysis, like certain other methods, supposes this kind of alternative. As we have just seen above, the differences between the **R** technique and **Q** technique are due essentially to standardization (centering and reduction) of the data. The problem has already been dealt with by Orloci (1967) in respect of centering. The connection between the **R** and **Q** analyses is thus clarified from the theoretical point of view, but in practice, the ecological significance of standardization of the species or plots remains an important problem. This subject is developed in detail later, under the heading referring to the applications of inertia methods to bionomic data from the North Gascony Continental Shelf.

Note also that the classical presentation of Principal Component Analysis often goes with the hypothesis of multinormality of the distribution of the variables, particularly in testing the significance of the results obtained. This hypothesis seems to us to be very restrictive; at most one could expect to normalize the marginal laws. Finally, as Benzecri *et al.* (1973) point out, the hypothesis of statistical independence of the samples is rarely satisfied in real problems. In ecology it is often unreasonable. The problem of the inference of the results obtained is extremely interesting, but classical statistics is not of such great help in this instance.

Analysis of Correspondences (Cordier, 1965). This method is strictly limited to the case when the Table$\mathbf{X}$ (Imax, Jmax) is positive or null, each row and each column comprising at least one strictly positive term.

The starting point of the analysis of correspondence consists in using the χ^2 distance. Retaining our notation, the distance of two points is given by:

$$d^2 (i1, i2) = \sum_j \frac{1}{TJ(j)} \cdot \left[\frac{X(i1, j)}{TI(i1)} - \frac{X(i2, j)}{TI(i2)}\right]^2.$$

An obvious Euclidean representation is given by:

$$XE(i, j) = \frac{X(i, j)}{\sqrt{TJ(j)\, TI(i)}}.$$

The χ^2 distance is well known by statisticians and its choice has been amply commented upon by the initiator of the method of Correspondences (Benzecri *et al.*, 1973), in respect of essentially probabilistic considerations.

The Analysis of Correspondences assumes moreover the allocation of a weight $TI(i)$ to the ith point. The justification for this weighting is based on the principle of distributional equivalence: if two points $i1$ and $i2$ are such that $\frac{X(i1, j)}{TI(i1)} = \frac{X(i2, j)}{TI(i2)}$ these two points may be combined into a single point of weight $TI(i1)+TI(i2)$, without affecting the search for axes of inertia. Moreover, this allocation of weights makes is possible to retain the symmetry of observations and variables (weights can be given to variables as to observations). By giving the weight $m(j)$ to the variable j, the Euclidean distance becomes:

$$d^2 (i1, i2) = \sum_j m(j) (X(i1, j) - X(i2, j)]^2.$$

The Analysis of Correspondences can be regarded as an inertia analysis of points of which the co-ordinates are $\frac{X(i, j)}{TI(i)\, TJ(j)}$, the *ith* point being allocated the weight $TI(i)$, the jth variable the weight $TJ(j)$ (for further details see Appendix). Finally, the Analysis of Correspondences is an inertia analysis carried out from the centre of gravity of the points. It is extremely interesting to note that the symmetry between observations and variables is retained by this displacement of the origin.

Principal Co-ordinates Analysis (Gower, 1966). We have seen that the relationship between the inertia form and the matrix of the distances is given by:

$$d^2 (i1, i2) = c(i1, i1) + c(i2, i2) - 2c(i1, i2).$$

In general when there is a matrix **C**, we may attempt to construct a Euclidean representation such as:

$$d^2(i1, i2) = c(i1, i1) + c(i2, i2) - 2c(i1, i2).$$

A particular case of interest is when: $c(i1, i2) = c(i2, i2) = 0$ and when $c(i1, i2) = -d^2(i1, i2)/2$. Gower (1966) has perfected a method of constructing a centred Euclidean representation having equal weight. It should be noted that we no longer have the duality of observations and variables. From the beginning we have only one set of points. This method opens up the way to using a field as wide as possible of metric densities. Furthermore, Benzecri *et al.* (1973) have generalized the investigation to cover the case of inequal masses.

This paper will deal with only certain of the possible variants, namely those which are most widely used in ecology. A later paper will be devoted to a more extended investigation, using in particular the possibilities offered by the method of Principal Co-ordinates.

Ecological implications of the fundamental options

Choice of distances. We must define a distance which is a measure of the [difference] between two samples. We have chosen to work on the $X(i, j) = \log[N(i, j) + 1]$ where $\mathbf{N}(i, j)$ is the number of the species j in the sample i. (We are thus avoiding a part of the discussion, relating to the choice of the transformation, which we will deal with in a later publication.) The Euclidean distance will be given by $d^2(i1, i2) = \sum_j [X(i1, j) - X(i2, j)]^2$. Certain considerations may lead us to not using this distance, such as follows.

Effect of abundance of the species. Certain species have numbers which may fluctuate very widely from one sample to another. The part taken by the species in the evaluation of the distances between the samples may mask the influence of other species. The ecologist may then consider it necessary to restore to each species the same influence, or at least to moderate the heterogeneity of the influences. For this purpose it is possible, amongst other things, to operate with reduced species. The distance is then given by the expression:

$$d^2(i1, i2) = \sum_j \frac{1}{SJ(j)^2} [X(i1, j) - X(i2, j)]^2.$$

This distance is that of an analysis in the **R** mode of correlations (centred and reduced species: centring of the species does not affect the intersample distances since it corresponds to displacement of the origin). Reduction of the species thus corresponds to weighting of the type

$$\frac{1}{SJ(j)^2}.$$

Other weightings are conceivable, in particular that of the χ^2 distance using the expression:

$$\frac{1}{TJ(j)}.$$

Effect of density of samples on heterogeneity of distances. If we have four samples $i1$, $i2$, $i3$, $i4$ with:

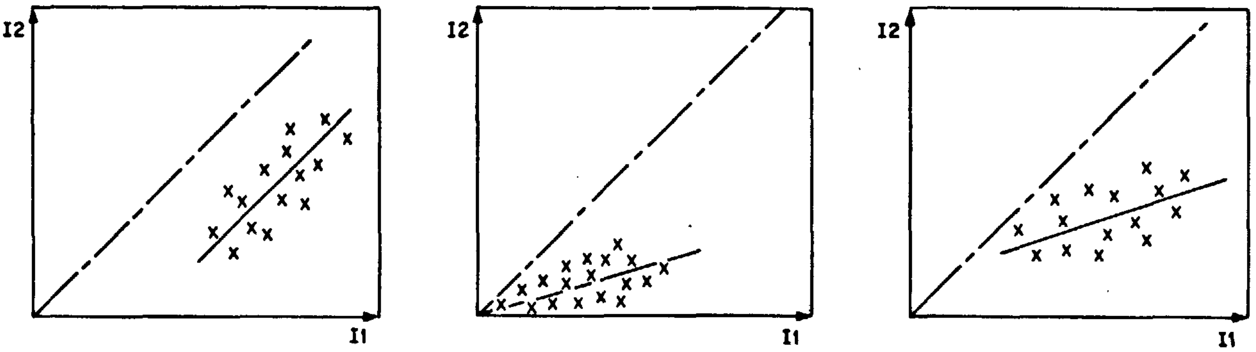
$$X(i3, j) = \lambda X(i1, j) \text{ and } X(i4, j) = \lambda X(i2, j) \text{ for } j = 1,$$

*J*max we will have:

$$d\,(i3,\, i4) = \lambda\; d\,(i1,\, i2).$$

In general, the distance between two samples, at least one of which is dense, will tend to be greater than the distance between two sparse samples. The rich samples will thus have a great deal of influence on the first axes. We will mention below the solutions provided by the classical methods for this problem.

Concept of sample profile. Given two samples, $i1$, and $i2$, and within this system of axes the species j at the abscissa point $X\,(i1, j)$, ordinate point $X\,(i2, j)$, we can conceive configurations 1, 2 and 3.



Configuration 1
$X(i1,j)=A+X(i2,j)+\epsilon(j)$

Configuration 2
$X(i1,j)=BX(i2,j)\epsilon(j)$

Configuration 3
$X(i1,j)=A+BX(i2,j)+\epsilon(j)$

The euclidean distance between the two samples $i1$ and $i2$ is: $\underset{j}{\Sigma}[X(i1,j) - X(i2,j)]^2$. This distance is null if, and only if, $X(i1,j) = X(i2,j)$ for every j. But if, for instance, $X(i1,j) = A + X(i2,\, j) + \varepsilon(j)$ for any j (configuration 1) $\varepsilon(j)$ being a residual the Euclidean distance between $i1$ and $i2$ may be extremely large and the obvious faunistic affinity somewhere cancelled. Such a kind of affinity should be better expressed by the distance $\underset{j}{\Sigma}\{[X(i1,\, j) - XI(i1)] - [X(i2, j) - XI(i2)]\}^2$ which is equivalent to the Euclidean distance on centred data. It is this definition of the profile used in fact in investigations in the **Q** mode of covariances (centred samples). Note that the problems broached above (heterogeneity of the distances between samples) does exist in practice. At any rate, configuration 1 is not very likely, particularly owing to the double absences (species represented neither in sample $i1$, nor in sample $i2$).

In the same way, when the affinity relationships between $i1$ and $i2$ are of the type $X(i1, j) = B.X(i2, j)\,\varepsilon(j)$ for any j (configuration 2) it is more appropriate to work with $\dfrac{X(i, j)}{TI(i)}$ than with $X(i, j)$. This is produced by the distance of χ^2. Note in this instance that the problem of the influence of the density of the samples on the heterogeneity of the distances is resolved. Note that configuration 2 is the most probable, it alone being compatible with the numerous simultaneous absences.

For configuration 3 affinity between the samples would be better revealed by using $\dfrac{X(i, j) - XI(j)}{SI\,(i)}$ instead of $X(i, j)$. Such a profile is in fact used by an analysis in the **Q** mode of correlations (reduced centred samples). Note once again that the problem of the influence of the density of the samples on the heterogeneity of the distances is solved.

It should be noted these three preliminary transformations are in fact the mathematical illustration of the profile concept in ecology.

Many definitions of the profile have been proposed. All of them are an attempt to show the affinities of two samples beyond the fact that one of them has more abundant occurrences of species.

Effect of double absences on the stability of distances. If species absent in samples *i*1 and *i*2 are added to the list of species studied, the Euclidean distance of these two samples does not change. This is a property which is retained in an analysis using the **R** technique and in an analysis using the distance of χ^2. On the other hand, in the case of a **Q** analysis, because of centring the samples, the distances are greatly affected by the double absences.

Choice of weights. The most natural choice is obviously that giving equal weights to each point. The only exception is in Analysis of Correspondences. The principle of distributional equivalence has no ecological significance in our case. It is thus simply the mean of preserving symmetry between samples and species.

Choice of origin. It is always useful to place the origin at the centre of gravity of the points. The first axis extracted is more discriminating and the structure obtained gains in neatness. Analysis of Correspondences and Principal Co-ordinate Analysis place the origin at the centre of gravity in all cases of problems. General Analysis on the contrary does not move the origin in any case at all, which by inference must lead to the obtainment of a unipolar axis (with limited discriminating power). The case of Principal Component Analysis is more complicated. An analysis in the **R** mode (in accordance with the conventions of the section entitled Principal Component Analysis) gives a structure in which the origin is at the centre of gravity of the species points and not at the centre of gravity of the sample points. Conversely, an analysis in the **Q** mode displaces the origin to the centre of gravity of the sample points. An interesting variation consists in recentring the data (see the section on applications below).

Description of data

The raw data have been borrowed from the work of one of the authors (Glemarec, 1969) and concern 30 benthic samples taken in identical sediments. These are silted sands on the North Gascony Continental Shelf. With regard to the climatic factors, the samples are distributed schematically as follows.

> Ten dredge hauls in the infra-littoral zone (between 0 and 20 m) of which which 5 were off the Vendean coast (samples 1 to 5) and 5 in the Baie de la Forêt, near Concarneau (samples 6 to 10).
>
> Ten dredge hauls in the shore circa-littoral zone (between 20 and 30 m), 5 in the Baie de Concarneau (samples 11 to 15) and 4 in the Baie d'Etel (samples 16 to 20).
>
> Ten dredge hauls in the seaward circa-littoral zone (between 80 and 140 m) (samples 21 to 30) of the Grande Vasière.

Previously (Chardy & Glemarec, 1974) Principal Component Analysis has brought to light a certain amount of information on the combined role of the climatic and edaphic factors.

In these new examples the edaphic factors connected with the granulometry of the sediment were homogeneous and therefore the aim of the analysis will be to obtain a structure representing as best as possible the faunistic affinities between samples, so as to clarify the relationships existing between the three climatic areas defined above (zones) and to estimate overall, in respect of the climatic factors, the importance of the geographical factors connected with the locality. In fact within one and the same zone the five combined samples are extremely wide apart (250 km between the Vendean coast and the Baie de la Forêt, 100 km between the Baie de Concarneau and the Baie d'Etel). Similarly, the samples on the Grande Vasière may be far apart from each other. The investigation centres on a faunistic assembly of 76 species.

The data have been subjected to the classical transformation $y = \log(x+1)$. The aim is not to normalize the distributions (a vain hope considering the large number of zeros in the matrix of raw data), but to heed a classical piece of intuition in ecology, by which more importance is given to the differences in numbers for small values than for large values. The problem of the choice of transformations (connected with the problem of the choice of matrix) will be discussed in a later work.

Results of applications

In order to provide a concrete illustration of the formal exposition outlined in the previous section we will comment upon the factorial structures obtained by application on the same data of Principal Component Analysis, Analysis of Correspondences and Principal Coordinates Analysis, in respect of those deduced from General Analysis (Table 1).

The differences observed will be discussed step by step with a view to attributing an ecological significance to the various options chosen. In accordance with the theoretical exposition, the distance, weight allocated to the points and the position of the origin will be defined for each of the methods. The problem considered is that of the representation of the sample observations in the reduced space of the species variables.

General Analysis and Principal Component Analysis

General Analysis on non-standardized data. Euclidean distance:

$$d^2(i1, i2) = \sum_{J=1}^{J\max} [X(i1, j) - X(i2, j)]^2.$$

Equal weight for all points.

Origin not displaced.

The Euclidean representation is the same as the initial set, $XE(i, j) = X(i, j)$.

The distribution of the 30 samples in the planes defined by the axes I and II (Figure 1) and the axes II and III (Figure 2) suggests the following comments:

The first axis is unipolar (general factor) and combines all the samples into positive values. Although not very discriminating, axis I isolated towards its positive pole the infralittoral samples from Vendée (1 to 5). The samples 1 to 5 are by far the richest in total number of individuals. The infra-littoral samples from Vendée are for this reason clearly separate from those of the Baie de la Forêt. On the other hand, the shoreward circa-littoral samples form a set not distinguished by the axis I. This axis I thus contains a large proportion of triviality, hence its high percentage of inertia extracted (41%).

Axis II separates the infra-littoral (1 to 10) into negative values from the shore circa-littoral and the seaward circa-littoral (11 to 30) in the positive values (Figures 1 and 2).

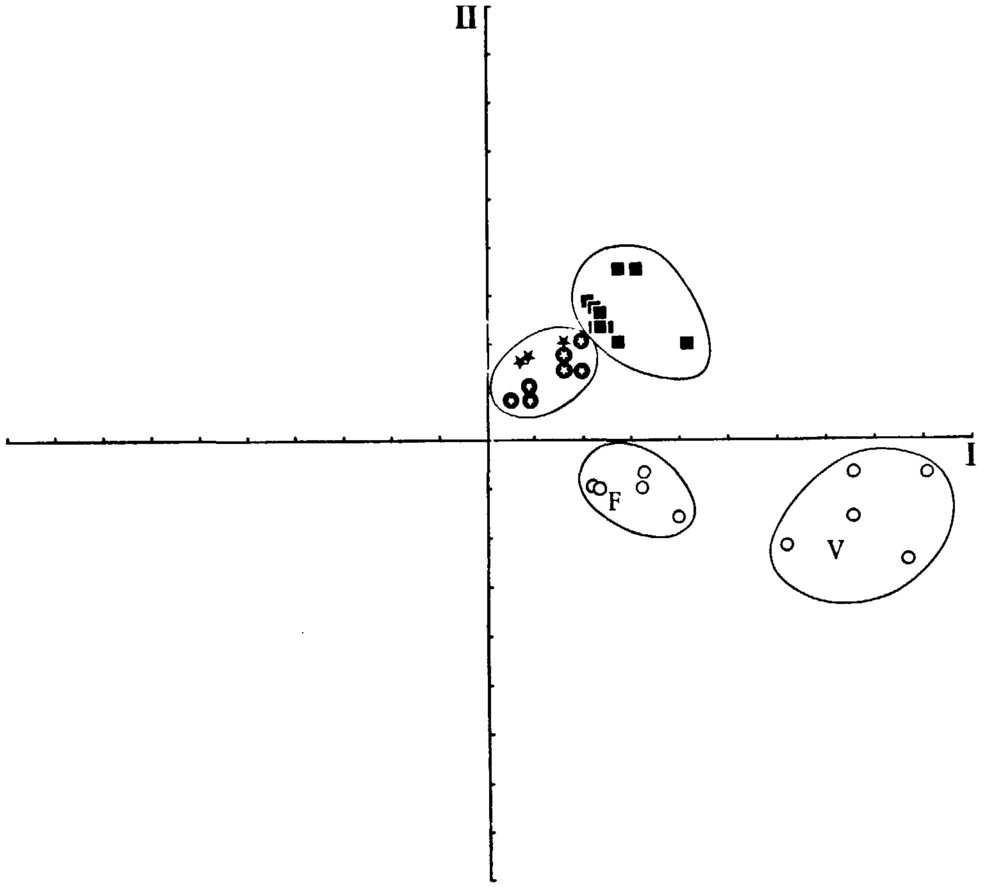


Figure 1. General Analysis—data not standardized—Plane I–II. Key to figures: C, Baie to Concarneau; E, Off Etel; F, Baie de la Forêt; V, Off Vendée.
Round symbols, infra-littoral zone; square symbols, shore circa-littoral zone; starred symbols, seaward circa-littoral zone; black starred, Chaetopterid facies.

The axis III (Figure 2) makes it possible to distinguish the shore circa-littoral (11 to 20) from the seaward circa-littoral (21 to 30). Note that the samples 23, 24, 26, extremely isolated towards the positive pole of the axis III, belong to the same facies; the Chaetopterid tube-dwellers.[a].

We can see, therefore, that the structure of the General Analysis is considerably affected by the richness of the stations and the large abundances of certain species at the heart of the samples from the Vendean intra-littoral zone and of the Chaetopterid facies for instance.

Centred species, samples not standardized

Euclidean distance.

Equal weight for all points.

Origin displaced to the centre of gravity.

Euclidean representation: $XE(i, j) = X(i, j) - XJ(j)$.

The structure obtained is identical to that of a Principal Component Analysis of the interspecific variance-covariance matrix.

[a] The term 'facies' according to Peres (1961) means: superabundance of one or a small number if species without the qualitative composition of the biocoenosis being affected. The reference is to a variety of sub-species of *Chaetopterus variopedatus* to which the provisional variety name, *nana*, which is extremely abundant in this particular habitat, is given.

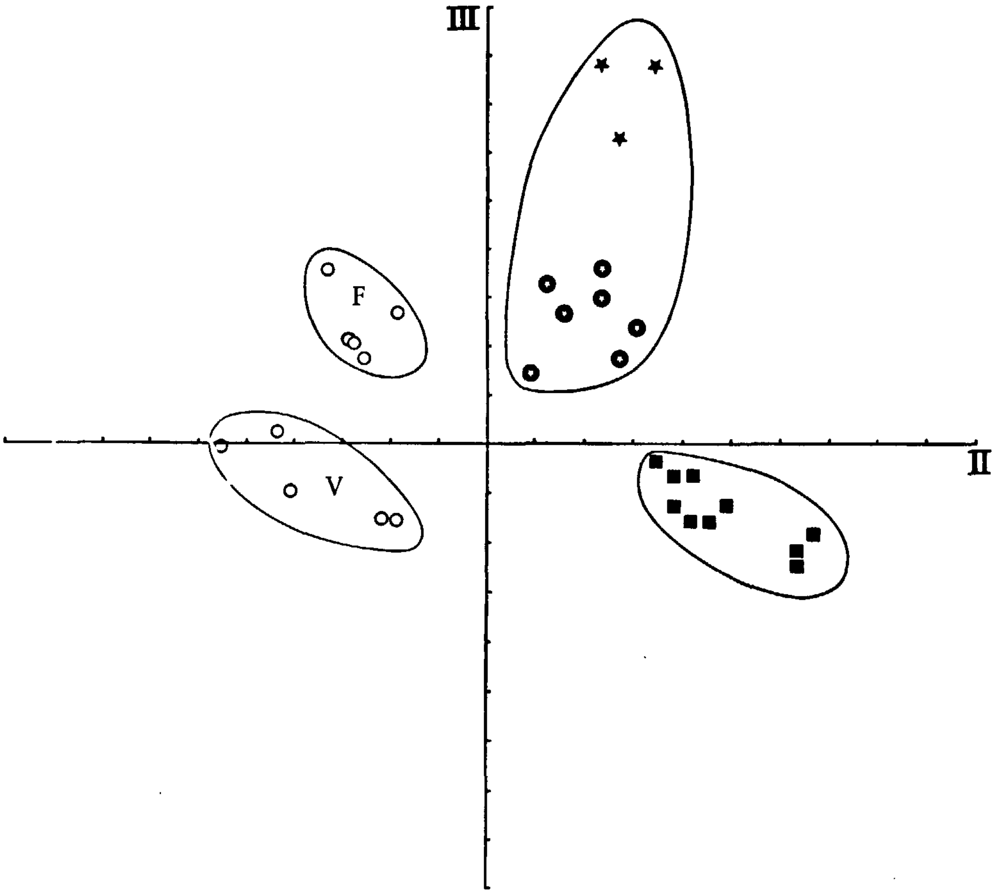


Figure 2. General Analysis—data not standardized—Plane II–III. See legend to Figure 1 for key.

Table 1.

	General Method (After Lebart & Fenelon, 1971) Euclidean distance. Equal weight for all points. Origin not displayed. Nature of symmetry: Perfect symmetry between variables and observations.	
Principal Co-ordinates (Gower, 1966)	Principal components (Hotelling, 1933)	Correspondences Cordier, (1965)
Any angular coefficient (from indices of similarity) Equal weight for all points Origin displaced to centre of gravity Nature of symmetry: None (impossible to represent variables and observations in the same sub-space)	Special angular coefficient: Covariance or correlation coefficient Equal weight for all points Origin displaced: =to centre of gravity of observations **Q** mode =to centre of gravity of variables **R** mode Imperfect symmetry: changing over variables and observations is equivalent to changing the mode from**Q** to **R** and *vice versa*	χ^2 distance Unequal weight for all points Origin displaced to centre of gravity Perfect symmetry between variables and observations

In our case we arbitrarily assume observations = samples
variables = species

The structure defined by the first two axes (Figure 3) provides a summary of the structure with three axes described in the previous example. The net effect, therefore, is the economy of an axis. The first axis, bipolar this time, is still affected by the richness of the Vendean infra-littoral samples (1 to 5), but it discriminates the infra-littoral quite broadly from the other two zones. At the same time axis II separates, as in the previous analysis, the shore circa-littoral from the seaward circa-littoral and shows the Chaetopterid facies within the latter. The richness of the stations (Vendean infra-littoral and Chaetopterid facies) still profoundly affects this structure.

The essential difference of this structure as compared to the above is the disappearance of a first trivial axis. The proportion of variance attributed to axis I is, moreover, much smaller (32·2%).

Centred and reduced species; samples not standardized

$$\text{Distance: } d^2(i1, i2) = \sum_{j=1}^{Jmax} \frac{1}{SJ(j)^2} \left[X(i1, j) - X(i2, j) \right]^2$$

Equal weights.

Origin displaced to the centre of gravity

$$\text{Euclidean representation: } XE(i, j) = \frac{X(i, j) - XJ(j)}{SJ(j)}.$$

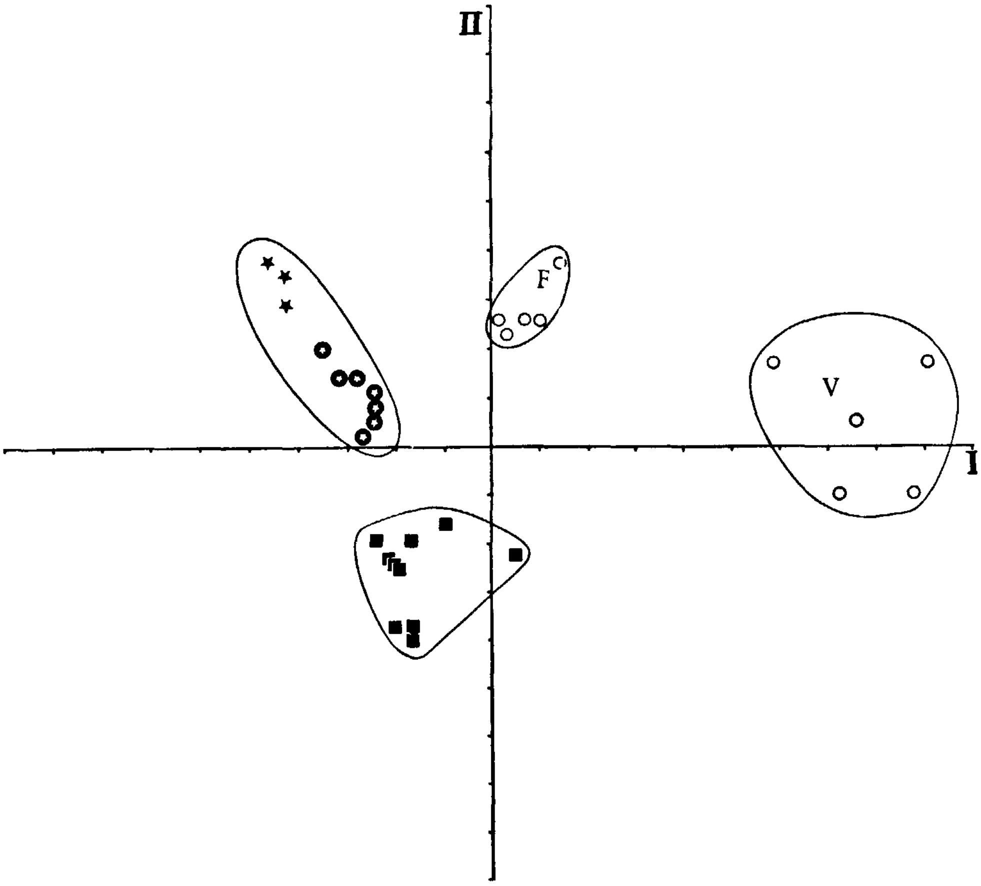


Figure 3. Principal Component Analysis—Centred species—Plane I–II. See legend to Figure 1 for key.

The structure obtained is identical to that of a Principal Component Analysis of the interspecific correlation matrix (Bravais-Pearson).

The bipolar axis I [Figure 4(a)] is still a factor connected with the richness of individuals in the samples and it discriminates, like the preceding structures, the stations of the Vendean infra-littoral (1 to 5). However, the infra-littoral of the Baie de la Forêt is not separated from the shore circa-littoral by axis II, in contrast with the previous structure. These two sets, moreover, have a fairly similar faunistic position and are distinguished only by their two or three dominant species such as *Amphiura filiformis* for the shore circa-littoral, *Magelona alleni*, *Clymene oerstedi* for the infra-littoral. Extraction of the third factorial axis [Figure 4(a)] is now necessary (the proportion of variance of which is only 8·8%) to show this separation into two stages. However, the Chaetopterid facies is no longer singled out in the seaward circa-littoral. In fact, the structure of Figure 4(b) is extremely close to that of Figure 3, the latter being defined only by the two initial axes and yet giving us a little more information. The influence of abundant species (*Amphiura filiformis*, *Clymene oerstedi*, *Magelona alleni*, *Chaeopterus*) is therefore minimized in the case of reductions of species [Figure 4(a) and (b)].

Centred samples, species not standardized

$$\text{Distance: } d^2(i1, i2) = \sum_{J=1}^{Jmax} \{[X(i1, j) - XI(i1) - (X(i2, j) - XI(i2))]\}^2$$

or:

$$= \sum_{J=1}^{Jmax} [X(i1, j) - X(i2, j)]^2 - [XI(i1) - XI(i2)]^2$$

Equal weight.

Origin displaced, not to the centre of gravity.

Euclidean representation: $XE(i, j) = X(i, j) - XI(i)$.

This new case corresponds to a Principal Component Analysis of the variance-covariance matrix of the samples.

The distribution of the samples in the space of the first three factors (Figures 5 and 6), very similar to the structure obtained by the General Analysis (Figures 1 and 2), can be summarized as follows:

a first unipolar axis affected by the richness of the stations;

a second axis separating the infra-littoral from the other two zones;

a third axis singling out the Chaetopterid tube-dwellers;

Centring of the stations does not therefore produce appreciable modifications of the structure in relation to analysis of the raw data (cf. General Analysis).

Centred and reduced samples; *species not standardized*

Distance:

$$d^2(i1, i2) = \sum_{J-1}^{Jmax} \left[\frac{X(i1, j) - X(i1)}{SI(i1)} - \frac{X(i2, j) - XI(i2)}{SI(i2)} \right]^2$$

or $= [1 - r(i1, i2)]/2$

where r is the Bravais-Pearson correlation coefficient.

Equal weights

Origin not displaced.

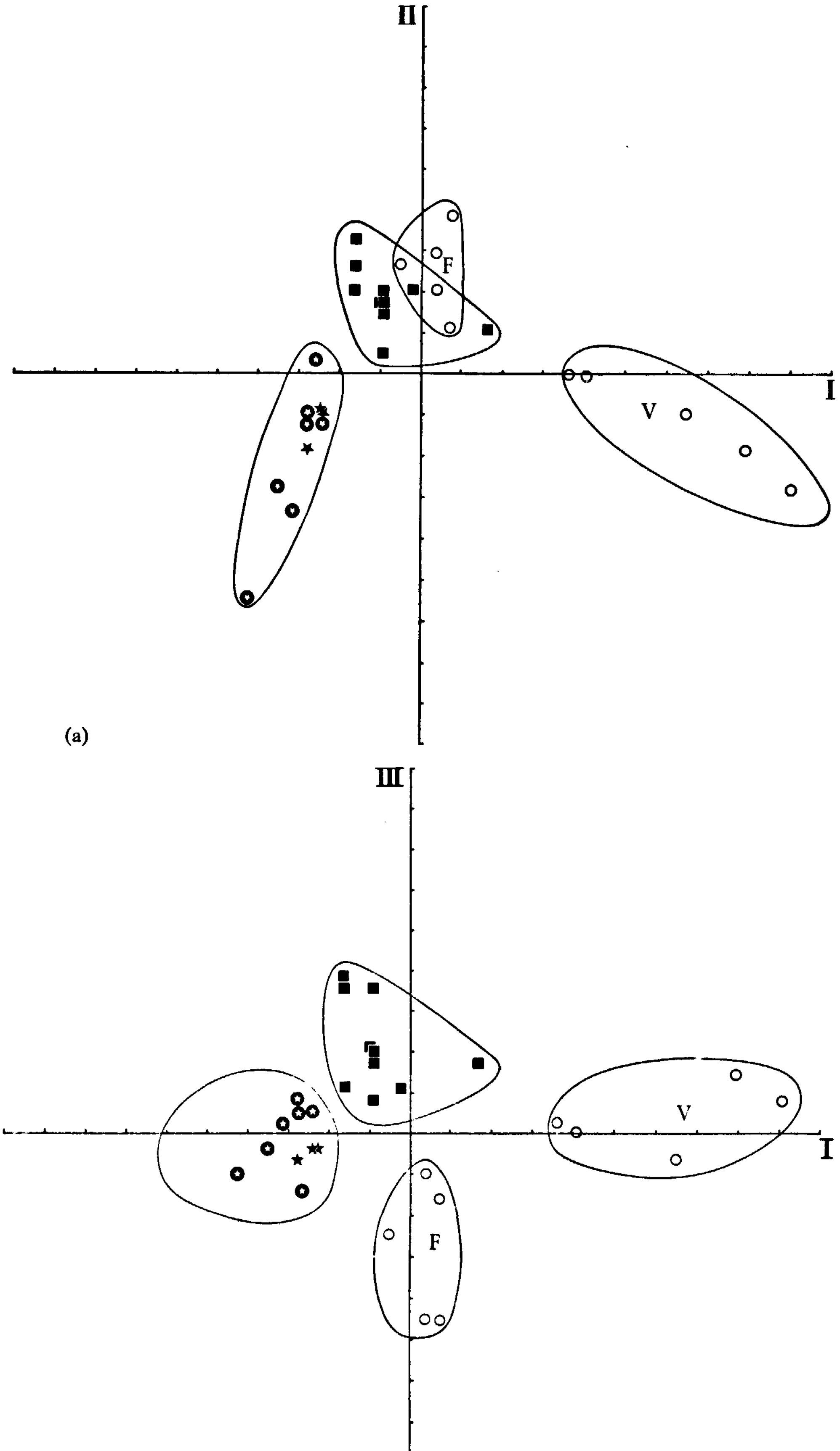


Figure 4. (a) Principal Component Analysis—centred and reduced species—Plane I–II. (b) Principal Component Analysis—centred and reduced species—Plane I–III. See legend to Figure 1 for key.

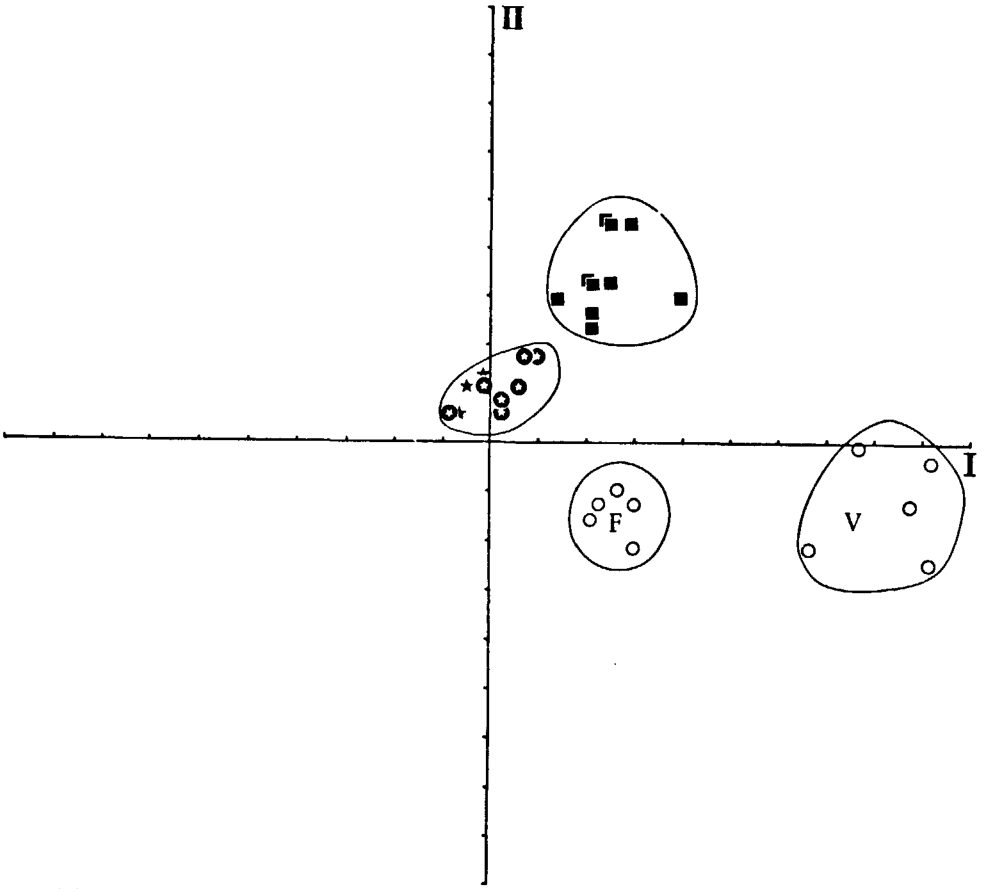


Figure 5. Principal Component Analysis—centred samples—species not standardized Plane I–II. See legend to Figure 1 for key.

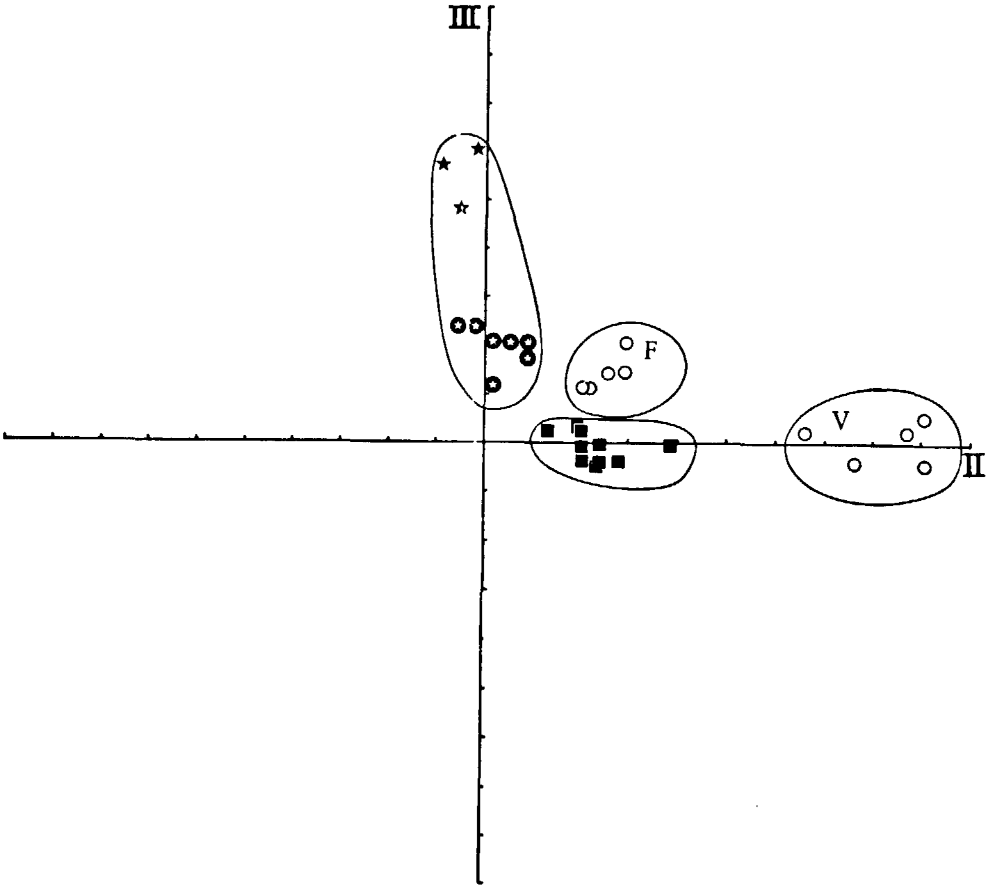


Figure 6. Principal Component Analysis—centred samples—Plane II–III. See legend to Figure 1 for key.

$$\text{Euclidean representation: } XE\,(i,j) = \frac{X\,(i,j) - XI\,(i)}{SI\,(i)}$$

This case corresponds to the Principal Component Analysis of the inter-sample correlation matrix (**Q** mode) (Bravais-Pearson).

The structure defined by the first two axes (Figure 7) shows for the first time the combination of the infra-littoral samples on one and the same set brought to light by axis II. The richness in individuals of the Vendean populations thus extremely attenuated. Axis I is virtually unipolar just as may be feared in cases of analysis where the species are not centred. Note that for the first time we can single out station 25, which appears extremely isolated from the other samples: this station has a coarse silted sand, its fauna is impoverished and is not characteristic of the fine silted sands out to sea. This pecularity is even more evident in the projection of the samples in the plane II–III (Figure 8) which shows the basic structure with 4 groups already illustrated above. Although the species are not reduced this structure does not discriminate the Chaetopterid facies or the Vendean population. The fifth axis has to be extracted, the proportion of variance of which (6%) is very small, in order to see for instance the Chaetopterid samples. By attenuating the effect of richness of individuals we prevent one of the initial axes from being devoted to the singling out of some samples where the individuals are extremely abundant.

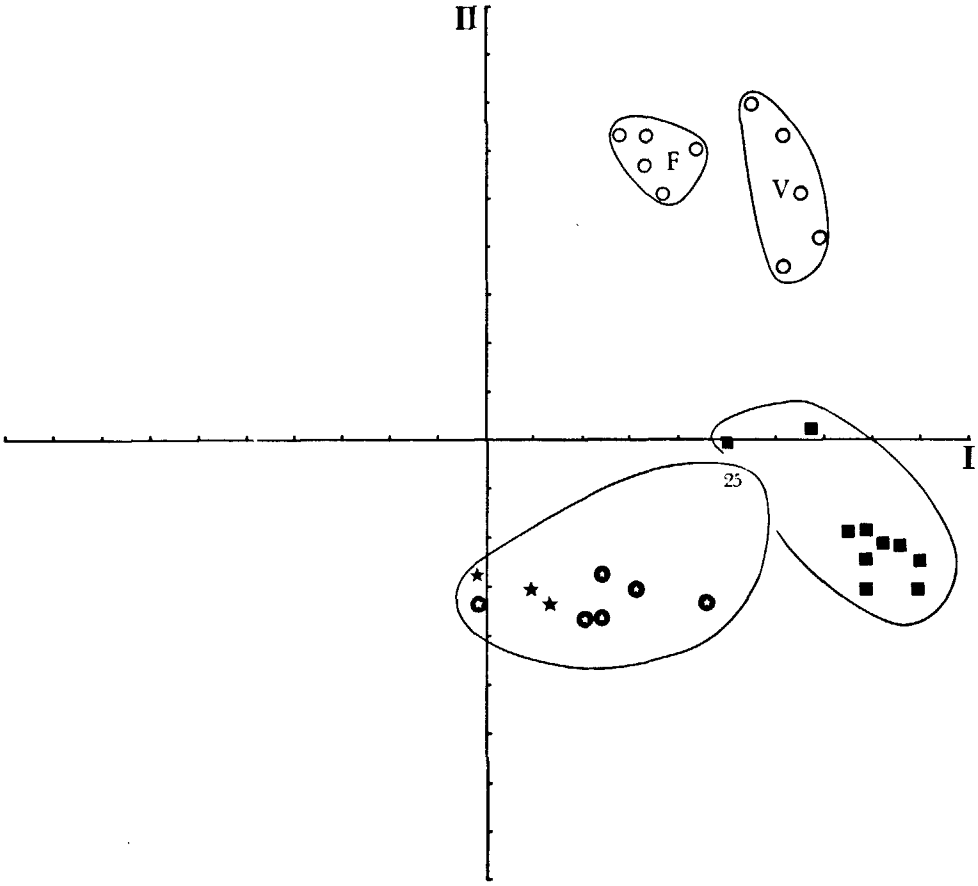


Figure 7. Principal Component Analysis—centred and reduced samples—Plane I–II. See legend to Figure 1 for key.

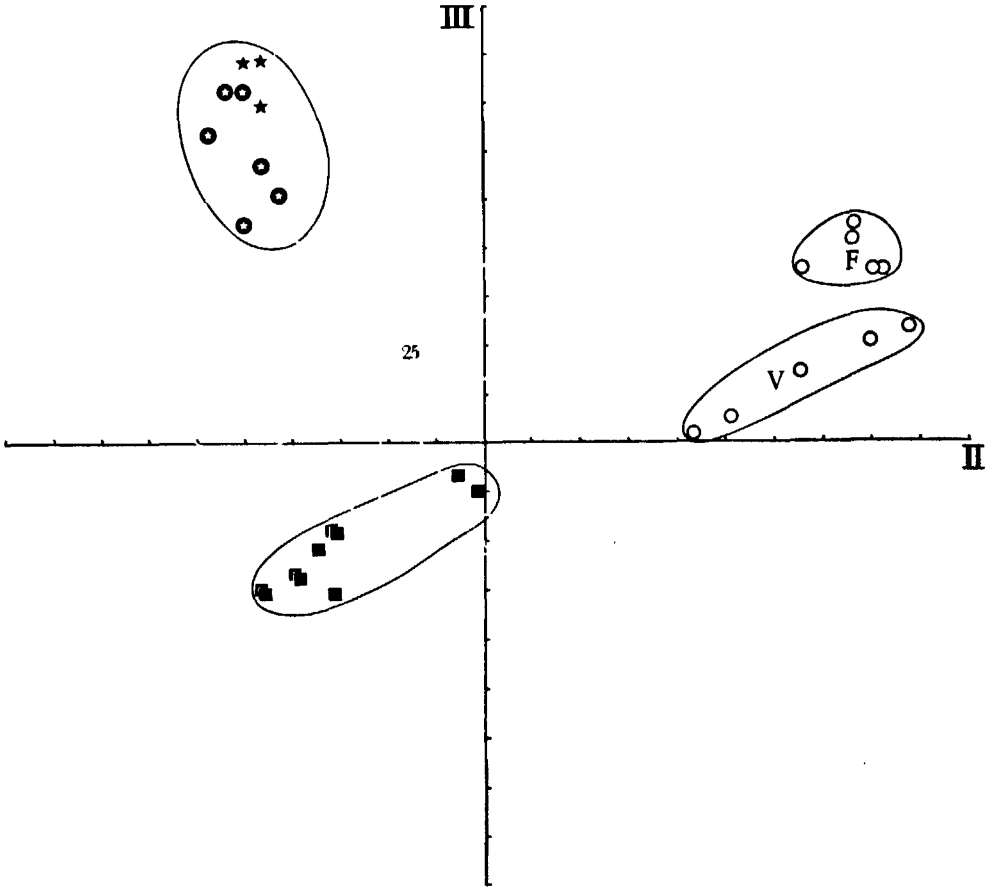


Figure 8. Principal Component Analysis—centred and reduced samples—Plane II–III. See legend to Figure 1 for key.

Principal co-ordinates analysis

It is out of the question to exhaust the possibilities offered by this method. It is nevertheless of interest to investigate the connections with the methods of Principal Component Analysis. We have said that if we had a matrix **C** (Imax, Jmax) the method of Principal Co-ordinates allows us to construct a Euclidean representation where two samples have a distance $d\,(i1,\, i2)$ with:

$$d^2\,(i1,\, i2) = c\,(i1,\, i1) + c\,(i2,\, i2) - 2c\,(i1,\, i2)$$

and to carry out an inertia analysis of this Euclidean representation around its centre of gravity.

If $c\,(i1,\, i2) = \sum_j X\,(i1, j)\; X\,(i2, j)$, $d\,(i1,\, i2)$ is simply the Euclidean distance of $i1$ and $i2$, the Euclidean representation is then the same as the initial set.

If we displace the origin to the centre of gravity of the samples, the Euclidean representation is centred. A General Analysis after such centring (or **Q**-mode Principal Component Analysis of the covariances) thus gives the same result as a Principal Co-ordinates Analysis applied to the matrix **C** (Imax, Imax) with $c\,(i1,\, 2) = \sum_j X\,(i1,\, j) X\,(i2,\, j)$. Readers familiar with Principal Co-ordinates will easily understand this convergence.

If $\mathbf{X}$ (Imax, Jmax) is the matrix of the data, $\mathbf{C} = \mathbf{XX}'^{a}$. To centre the species is equivalent to pre-multiplying the matrix $\mathbf{X}$ on the left by $\mathbf{A}$ with $\mathbf{A} = (\mathbf{I}-\mathbf{B})$ where $\mathbf{I}$ is the identity matrix and $\mathbf{B}$ given by $B(i, j) = 1/I\text{max}$ for any i and j.

If $\mathbf{X}_1$ is the matrix after centring the species, $\mathbf{X}_1 = (\mathbf{I}-\mathbf{B})\mathbf{X}$. Instead of $\mathbf{C}$ we will then have:

$$\mathbf{C}_1 = \mathbf{X}_1\,\mathbf{X}'_1 = (\mathbf{I}-\mathbf{B})\;\mathbf{XX}'\,(\mathbf{I}-\mathbf{B})' = (\mathbf{I}-\mathbf{B})\;\mathbf{C}\,(\mathbf{I}-\mathbf{B}).$$

Multiplication on the right and left hand side of $\mathbf{C}$ by $\mathbf{I}-\mathbf{B}$ is the double centring required by the calculations of the method of Principal Co-ordinates.

In the same way, to apply the method of Principal Co-ordinates to $c\,(i1, i2)$ with $c\,(i1, i2) = \sum_{j}[X\,(i1, j) - XI\,(i1)]\;[X\,(i2, j - XI\,(i2)]$ will give the same results as a General Analysis after double centring. The samples are explicitly centred since the origin is placed at the centre of gravity of the samples by a study of the Principal Co-ordinates.

Finally, application of the method of Principal Co-ordinates in the case where $c\,(i1, i2)$ is the coefficient of correlation will give the same result as a General Analysis around the centre of gravity of the reduced and centred samples. Obviously, the General Analysis has the advantage in the cases presented here of respecting the duality between observations and variables which does not exist in the method of Principal Co-ordinates.

If it is desired to compare, as has already been done in ecology, a classical Principal Component Analysis (interspecific correlation) with a Principal Co-ordinates Analysis using any index of similarity, it is important to recognize that the differences in the results obtained have a double origin:

different metric;
displacement of the origin (an often ignored aspect).

Centring of samples and species (double centring of non-reduced data)

Distance: Euclidean distance.
Equal weight.
Origin displaced to the centre of gravity.
Euclidean representation:

$$XE\,(i, j) = X\,(i, j) - \{XI\,(i) - [XJ\,(j) - XM]\}$$

WhereXM is the mean of$X\,(i, j)$.

The symmetry of the species and the samples in this new representation is perfect. This analysis corresponds to a Principal Component Analysis applied to the variance-covariance matrix of the samples. This is thus a very specific case of Principal Co-ordinates Analysis. The choice of distance explains the property of symmetry of the variables and of the observations. In the general case of Principal Co-ordinates, this properly no longer exists because the main point of this method is to utilize distances which are not defined by quadratic forms (any indices of similarity.)

As may have been expected the structure obtained (Figure 9) is very close to that deduced from the analysis carried out on the centred species (Figure 3).

The richness in individuals of the Vendean samples and the Chaetopterid facies appears.

Axis II separates the shore circa-littoral and seaward circa-littoral.

[a] Note that we have usually operated on the form$\mathbf{X'X}$ but the result of equivalence outlined in the section on 'Representation of points and variables' allows us to argue about $\mathbf{C} = \mathbf{XX}'$.

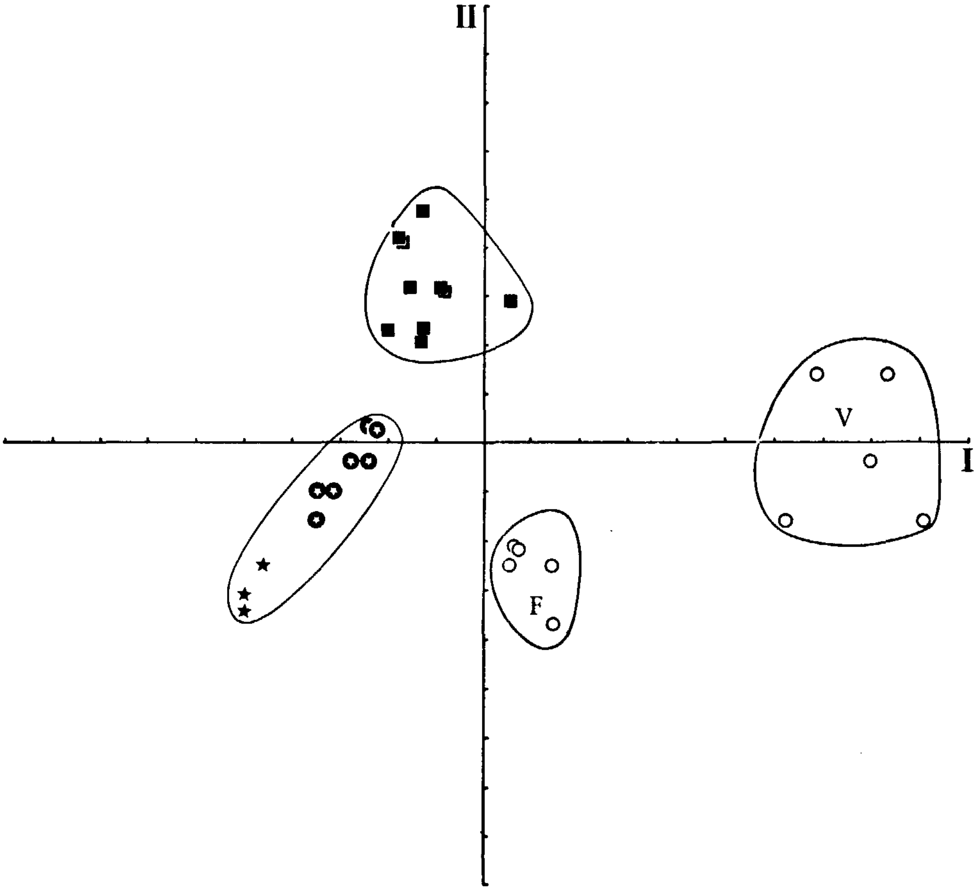


Figure 9. Principal Co-ordinates Analysis—double centring of data—Plane I–II. See legend to Figure 1 for key.

The preceding examples has shown that centring of the plots introduced virtually no modifications.

Standardized samples; *recentred species*. Distance $[1-r(i1, i2)]/2$ where r is the coefficient of correlation. This analysis is identical to that of a Principal Component Analysis of the intersample correlation matrix (**Q** mode) with displacement of the origin.

Equal weights
Origin at the centre of gravity of the points.
Euclidean representation:

$$XE(i,j)=\frac{X(i,j)-XI(i)}{SI(i)}-X\,2\,\mathcal{J}(j)$$

where $X\,2\,\mathcal{J}(j)$ is the mean of $\dfrac{X(i,j)-XI(i)}{SI(i)}$.

The structure obtained in the plane of the first two axes (Figure 10) is identical to that obtained in the plane II–III of the Principal Component Analysis of the intersample correlation matrix (Figure 8). The stations rich in individuals are not singled out. Station 25 is extremely isolated. By avoiding the extraction of a first trivial axis, this analysis has the merit

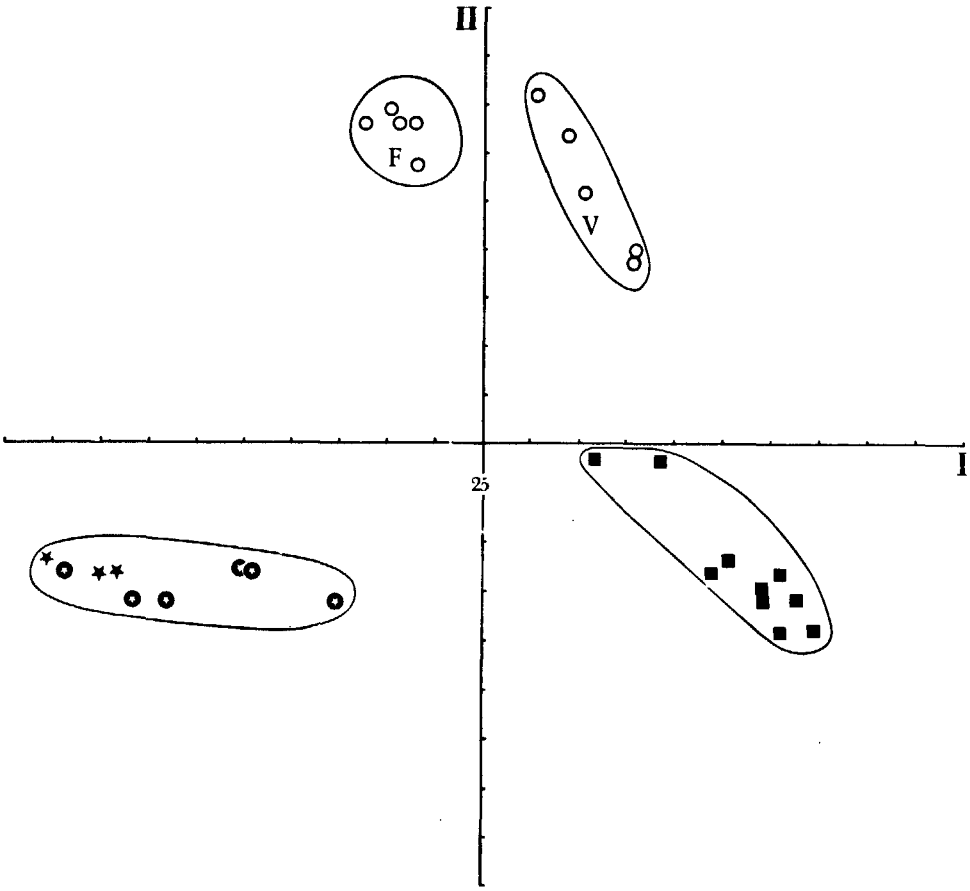


Figure 10. Principal Co-ordinates Analysis—standardized samples—Plane I–II. See legend to Figure 1 for key.

of showing the essential ecological profiles in the plane I–II and so in a way are we economizing on one axis.

Analysis of Correspondences
Distance: χ^2
Weight allocated to the points: $TI\,(i)$
Origin displaced to the centre of gravity
Euclidean representation of the observations:

$$XE\,(i, j) = \frac{1}{\sqrt{TJ\,(j)}} \cdot \frac{X\,(i, j)}{TI\,(i)}$$

The structure obtained in the plane I–II suggests the following ecological remarks (Figure 11).

Axis I isolates very clearly the seaward circa-littoral from the shore circa-littoral and from the infra-littoral.

The axis II separates the infra-littoral (negative values) from the shore circa-littoral (positive values). Moreover, the infra-littoral, despite its regional division, constitutes a well individualized entity. In this structure, the samples rich in individuals are not normally separated. Thus the region of Vendée, is very similar to that of the Baie de la Forêt. The chaetopterid facies remains isolated among the silted sands of the seaward circa-littoral;

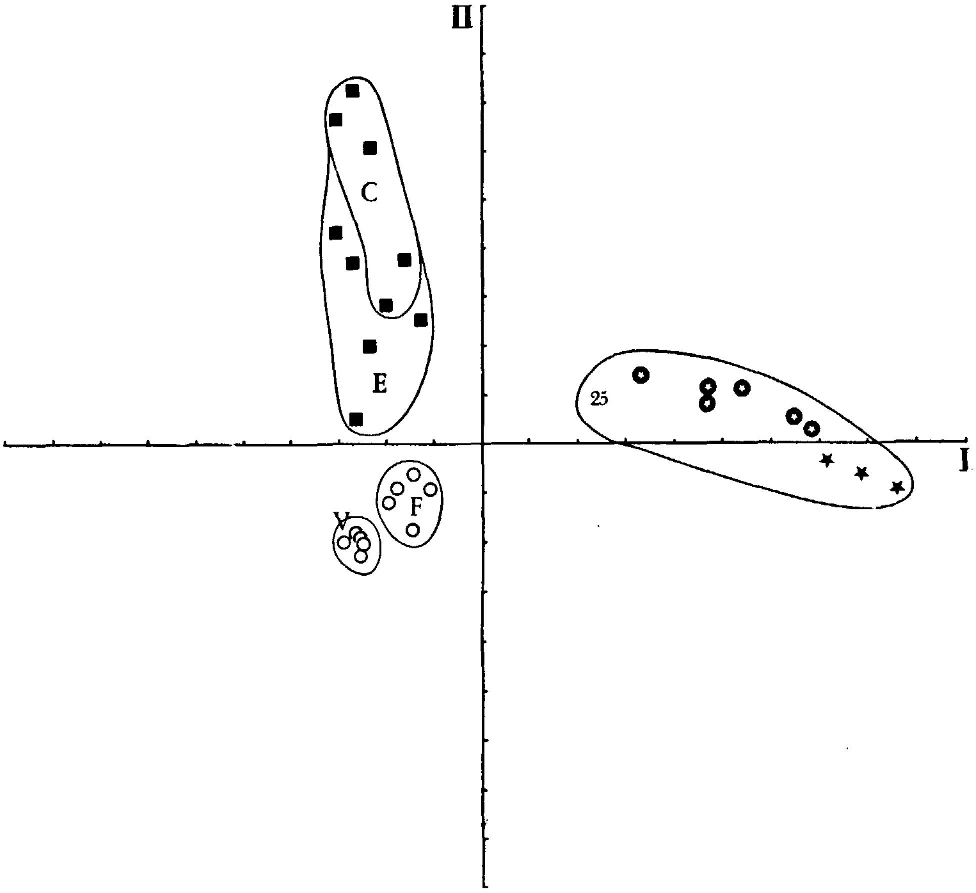


Figure 11. Analysis of Correspondences—Plane I–II. See legend to Figure 1 for key.

without doubt it is more than one facies and its qualitative faunistic composition is a little different from that of the standard population.

The aberrant nature of station 25 is less marked than in other cases; its position displays a faunistic composition closer to the shore circa-littoral zone than that of the other stations of the seaward circa-littoral group.

In this analysis, and here for the first time, it is possible to isolate within the lower circa-littoral zone the samples of the Baie de Concarneau and those of Etal. These two regions are less clearly separated than those of the infra-littoral and display a gradient according to the positive values of axis II. In this connection it is important to emphasize that the infra-littoral populations of the Baie de la Forêt, which are followed topographically by the shore circa-littoral populations of the Baie de Concarneau, are further apart from each other than those of the Bair de la Forêt and the Baie d'Etel which are extremely far apart geographically.

The grouping of the infra-littoral samples (Vendée and Forêt) appears already in the plane I–III of the Principal Component Analysis (**Q** mode) (Figure 8) and also in the plane I–II of the recentred Principal Component Analysis (Figure 10), given here as a special case of the Principal Co-ordinates Analysis. The structures defined by these two methods, like the structure resulting from the Analysis of Correspondences, are not affected by the density of the samples (an essential difference between the Vendée bottoms and those of the Baie de la Forêt). It is interesting to note that the preliminary multiplication by $\frac{1}{TI\,(i)}$ of χ^2 is more effective in our example to moderate the heterogeneity of the distances between samples than the multiplication by $\frac{1}{SI\,(i)}$ of the Principal Component Analysis; examination of the compactness of the infra-littoral samples in Figure 11 shows this clearly.

However, the structure resulting from the Analysis of Correspondences introduces a new element in respect of the other analyses: the first axis isolates the seaward circa-littoral from the other two less deep zones. This axis represents graphically the role of the climatic factors, as they have been shown by Glemarec (1969), by classifying the samples according to two areas: a stenothermous area (seaward circa-littoral) towards the negative pole and a eurythermous area (shoreward circa-littoral and infra-littoral) towards the positive pole. As opposed to Principal Component Analysis, (recentred or otherwise) using the **Q** technique, the Analysis of Correspondences is not affected by the double absences of the species. Now this criterion, depending on whether account is taken of it or not, determines to a large extent the faunistic relationship of the shoreward circa-littoral in respect of the other two zones. Reference to the raw data is sufficient to show that numerous species are absent both from the shoreward circa-littoral and from the seaward circa-littoral: only the ecologist can determine the importance of such a criterion for the concept of classification in tiers. Disregarding the double absences leads, in the example with which we are concerned, to a clearer illustration of the role of the climatic factors.

Discussion of results

On the basis of the example chosen there is no doubt that the Analysis of Correspondences seems to be the most attractive method of processing for the ecologist. The structure obtained is closest to the results inferred from a more traditional analysis carried out on the same data (Glemarec, 1969), which for us constitutes a test.

In fact, the χ^2 distance conforms absolutely with the pre-occupations that the author set himself in this investigation: weighting of the effect of abundance of the species, attenuation of the effect of richness of samples, structure not altered by double absences, conditions which allow definition of the relationships between samples by relying on the concept of profile (cf. Section on Ecological implications of the fundamental options). On the other had it is clear that within the framework of an investigation of an economic nature (bearing in mind the quantitative biological richness of the bottom of the sea), the Analysis of Correspondences does not seem advisable.

Table 2 summarizes the ecological implications of the various methods, as they appear to us as the result of the processing in this work. The convergence of theoretical and practical considerations advanced in this investigation allow us to envisage the use of such a Table for the purpose of analytical strategy.

Reference to Table 2 must not, howcvei lead to the belief that the influences affecting the structures are necessarily independent of each other. For instance, to reduce the stations (to moderate the influence of density on the heterogeneity of distances) may lead in certain cases to giving more weight to the abundance of species of facies samples (where the heterogeneity of the numbers is strongest). Sin.ilarly, to weight the density of the samples leads implicitly to displaying the concept of faunistic profile.

Conclusions

The ecological implications inherent in the classical variations of inertia methods have led us to set up several types of concrete problems that may affect structures more or less profoundly.

TABLE 2.

Methods and metrics	General Analysis	Principal Components Analysis				Correspondences
		Centred species **R** mode	Centred and reduced species **R** mode	Centred samples **Q** mode	Centred and reduced samples **Q** mode	
'Influences' capable of affecting structures	Euclidean distance	Covariance between species	Correlation between species	Covariance between samples	Correlation between samples	χ^2
Abundance of species	+	+	—	+	+	—
Density of samples	+	+	+	+	—	—
Profiles of samples	+	+	+	+	—	—
Double absences of species	—	—	—	+	+	—
Proportion of triviality of first axis	+	—	—	+	+	—

Note that we are always concerned with the representation of the samples in the factor space

+means a method is affected by the corresponding 'influences'.

The case of Principal Co-ordinates, for which we have clearly shown the convergence with Principal Components Analysis in the case of the Euclidean distance, is not mentioned here.

Influence of the abundance of species.

Influence of the density of samples on the heterogeneity of distances.

Influences of double absences on the stability of distances.

Definition of faunistic profiles.

Profiles of triviality of the first axis.

These problems are connected with the choice of the distance, the position of the origin and the weighting of the points.

However, the choice of the distance must require the greatest attention from the user since the first four points listed above depend on it directly.

Theoretical considerations followed by practical applications on concrete data made it possible to show schematically the impact of the 'influences' considered with regard to the various inertia methods: Principal Analysis, Principal Co-ordinates Analysis and Analysis of Correspondences.

From the point of view of analytical strategy, reference to Table 2 clearly shows that to neutralize or accentuate each of these 'influences' constitutes a means of illustrating a quality of information, in accordance with the objectives of the investigation. Seeking the optimum method in absolute terms is obviously a mistake and each of the analytical methods illustrated in this paper represents a possibility for different typesof investigations according to the aims of the ecologist.

To place Principal Component Analysis within the general framework of inertia methods shows that the choice of the representation space (**Q** or **R** mode of the ecologists) has no significance. The real choice is that of centring and reducing the species or samples.

This is in fact a choice of the distance and of displacing the origin. Choice of the distance is largely outside the realm of inertia methods since it is the starting point of any multivariate analysis.

Acknowledgement

The authors wish to thank Dr M. J. R. Fasham for reading and commenting upon a first draft of this manuscript.

Appendix

Distances and weights attribution in an inertia analysis and especially in the analysis of Correspondences

The problems of dealing with the choice of weights and distances are somewhat overlapping and a confusion may arise with readers more familiar with classical statistical multivariate analysis than with inertia methods. We say for instance that the Analysis of Correspondences assumes a preliminary transformation of the $\dfrac{X(i,j)}{TI(i)\,.\,TJ(j)}$ and a weighting of the observations by $TI(i)$ and of variables by $TJ(j)$. It should be noted that preliminary transformation and weighting correspond to two different operations having distinct influences.

(1) The weighting of observations does affect the definitions of axes because it modifies the inertia form. It also affects the distance between variables as follows:

$$d^2(j1, j2) = \sum_i TI(i)\left[\frac{X(i, j1)}{TI(i)\,TJ(j1)} - \frac{X(i, j2)}{TI(i)\,TJ(j2)}\right]^2$$

$$= \sum_i \frac{1}{TI(i)}\left[\frac{X(i, j1)}{TJ(j1)} - \frac{X(i, j2)}{TJ(j2)}\right]^2$$

instead of (without weighting of observations)

$$d^2(j1, j2) = \sum_i \left[\frac{X(i, j1)}{TI(i)\,TJ(j1)} - \frac{X(i, j2)}{TI(i)\,TJ(j1)}\right]^2$$

$$= \sum_i \frac{1}{TI(i)^2}\left[\frac{X(i, j1)}{TJ(j1)} - \frac{X(i, j2)}{TJ(j2)}\right]^2$$

In contrast the weighting of observations does not affect the distance between the observations.

(2) Weights attributed to variables affect axes definition and distance between observations but does not affect distance between variables.

Briefly, giving weights $P(i)$ to observations or premultiplying$X(i, j)$ by $\sqrt{P(i)}$ leads to the same axes system, to the same variables co-ordinates (or factor loading) but to different observations co-ordinates (or factor scores). This is due to the fact that weighting of the observations does not modify the distances between observations while premultiplying does modify the distances.

References

Austin, M. P. & Orloci, L. 1966 Geometric models in ecology. II. An evaluation of some ordination techniques. *Journal of Ecology* **54**, 217–227.

Benzecri, F. *et al.* 1973 L'analyse des données. I. La Taxinomie. II. L'Analyse des Correspondances. *Dunod,* Paris 615 pp. and 619 pp.

Bray, J. R. & Curtis, J. T. 1957 An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs* **27,** 325–349.

Chardy, P. & Glemarec, M. 1974 Contribution au problème de l'étagement des communautés benthiques du plateau continental Nord-Gascogne. *Comptes rendus hebdomadaire des seances de L'Académie des Sciences t.* **278,** *Série D,* pp. 313–316

Cordier, B. 1965 L'Analyse des Correspondances. *Thèse, Fac. Sc. Rennes* 100 pp.

Glemarec, M. 1969 Les peuplements benthiques du plateau continental Nord-Gascogne. *Thesè d'Etat, Univ. Paris,* 170 pp.

Gower, J. C. 1966 The distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53,** 325–338.

Hotelling, H. 1933 Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **24,** 417–441, 498–520.

Ibanez, J. J. & Seguin, X. 1972 Etude du cycle annuel du zooplancton d'Adibjan Comparaison de plusieurs méthodes d'analyse multivariable. Composantes principales, correspondances, coordonnées principales. *Investigacion Pesquera* **36** (1), 81–108.

Ivimey-Cook, R. B., Proctor, M. C. F. & Wigston, D. L. 1969 On the problem of the **R/Q** terminology in multivariate analysis of biological data. *Journal of Ecology* **57,** 673–675.

Lebart, X. & Fenelon, X. 1971 Statistique et informatique appliquées. *Dunod, Paris* 426 pp.

Orloci, L. 1966 Geometric models in ecology. I. The theory and application of some ordination methods. *Journal of Ecology* **54,** 193–215.

Orloci, L. 1967 Data centring: a review and evaluation with reference to component analysis. *Systemic Zoology* **16,** 208–212.

Peres, J. M. Océanographie biologique et biologie marine. I. La vie benthique. *P.U.F., Paris* 540 pp.

Spearman, C. 1904 General intelligence objective by determined and measured. *American Journal of Psychology* **15,** 201–293.

Williams, W. T. & Dale, M. B. 1965 Fundamental problems in numerical taxonomy. *Advances in Botanical Research* **2,** 35–38.

Printed in Great Britain by Henry Ling Ltd., at the Dorset Press, Dorchester, Dorset