ADE-4

Between- and within-groups principal components analyses

Abstract

In this volume, the statistical analysis of a multivariate environmental array is described. Quantitative variables collected at s locations for t sampling dates are analysed. To have a distinct view of the respective influence of the seasonal succession and the sample location on the variability of the measures, principal components analyses were used on tables from the linear model of variance analysis in a two-way layout with one observation per cell. Moreover, this volume introduces to the use of multivariate techniques using projection onto a subspace.

Contents

S. Dolédec & D. Chessel

1 - Introduction

An important step in ecological data analyses consists in taking into account experimental objectives, i.e. experimental conditions (such as time and space), within linear multivariate analyses in order to solve problems such as: (i) What in a multivariate set of data depends only on time, space and what can be explained by an interaction between space and time? (ii) What in a faunistic table does not depend on the sampling conditions (see for example Usseglio-Polatera & Auda, 1987)[1]? We will focus here on the question of spatial-temporal design and its influence in hydrobiological studies.

This type of design is obviously not specific to hydrobiology. Most of the ecological studies search for the temporal evolution of systems. For this purpose, the same locations are sampled repeatedly at appropriate time intervals. For example, the study of the distribution and dynamics of animal and/or plant communities as well as the study of the environment (physical and chemical variables) involves the description of three-dimensional data table (site-time-species or site-time-environmental variables).

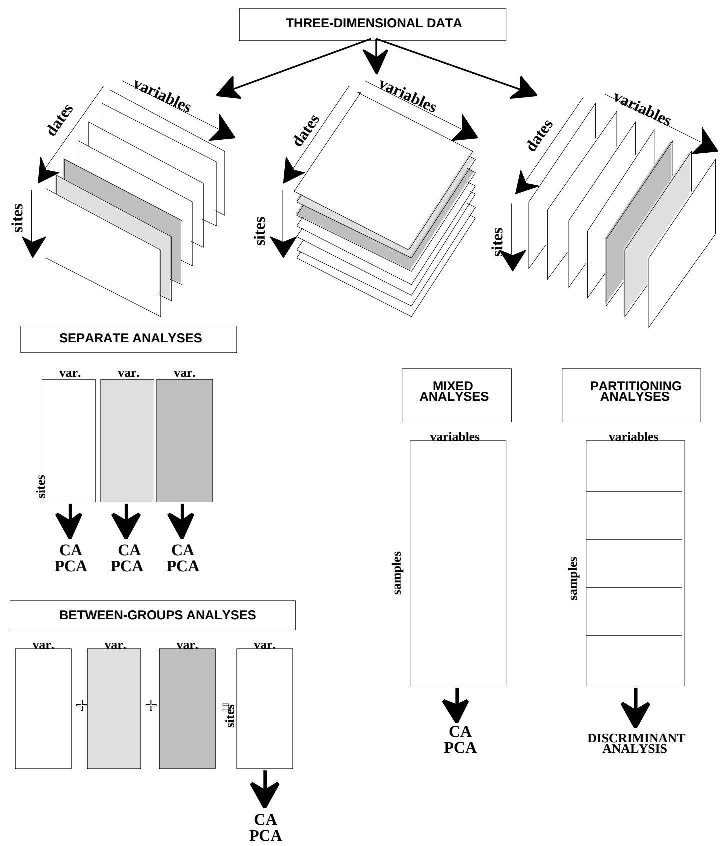


Figure 1 Spatial and temporal structure of ecological data. Data can be either analysed from the spatial point of view or from the temporal point of view or from the variable (taxa in our example) point of view.

Four ordination options may be found in the literature[2] (Fig. 1). They are respectively called: (i) separate analyses, (ii) between-groups analyses, (iii) mixed analyses, (iv) partitioning analyses.

2 - Classical approach

2.1 - Data

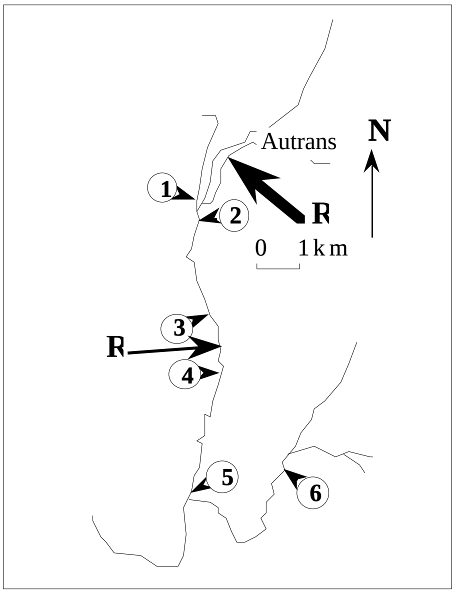


Figure 2 Study sites. The arrows indicate effluents of organic pollution.

The Méaudret is a small river from the Vercors receiving effluents from two villages (Autrans, Méaudre). It is a tributary of the Bourne River. Five sites were selected from upstream to downstream the Méaudret (Fig. 2). A sixth site was situated on the Bourne River as a non-polluted site. Physical and chemical data were sampled at these six sites for four occasions (Pegaz-Maucet, 1980)[3]. Ten physical and chemical variables were taken into account:

01	`Temp`	Water temperature (°C)
02	`Debit`	Discharge (l/s)
03	`pH`	pH
04	`Condu`	Conductivity (µS/cm)
05	`Oxyg`	Oxygen (% saturation)
06	`Dbo5`	B.D.O.5 (mg/l oxygen)
07	`Oxyd`	Oxydability (mg/l oxygen)
08	`Ammo`	Ammonium (mg/l NH_4^+)
09	`Nitra`	Nitrates (mg/l NO_3^-)
10	`Phos`	Orthophosphates (mg/l PO_4^{---})

2.1 - File creation

Create a data folder. Go to the **ADE•Data** selection card and select « Méaudret » (Fig. 3). With the left-hand data field, create an ASCII file named `Mil.txt` that contains the physical and chemical data. With the right-hand data field, create an ASCII file named `Code_Var` that contains the labels of variables.

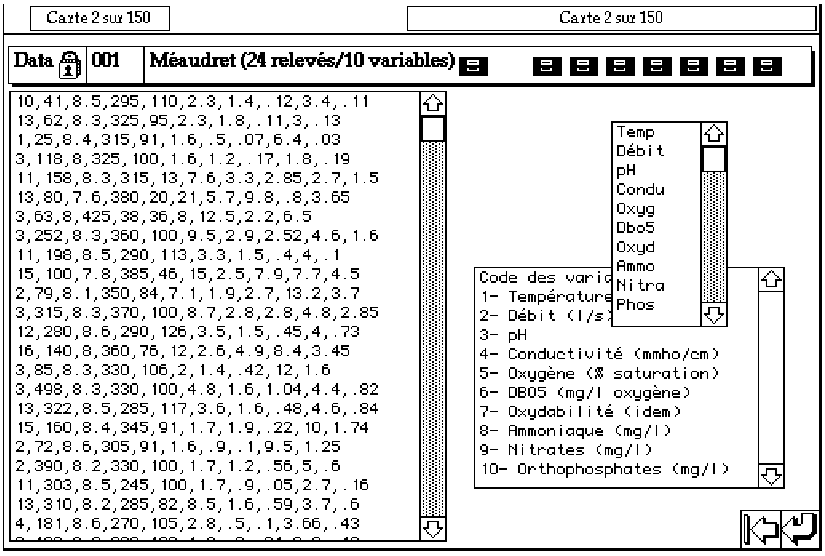


Figure 3 The « Méaudret » data card from the ***ADE•Data*** *stack.*

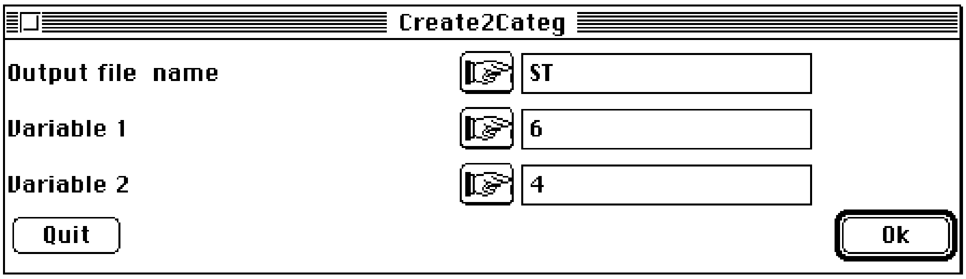
Transform `Mil.txt` into a binary file `Mil` (24-10). List the data using the **Edit with** option (Table 1).

Table 1 Data consist of 24 samples (rows) distributed over 6 stations and 4 occasions. The ten physical and chemical variables are represented in columns.

	1	2	3	4	5	6	7	8	9	10
1	10	41	8.5	295	110	2.3	1.4	0.12	3.4	0.11
2	13	62	8.3	325	95	2.3	1.8	0.11	3	0.13
3	1	25	8.4	315	91	1.6	0.5	0.07	6.4	0.03
4	3	118	8	325	100	1.6	1.2	0.17	1.8	0.19
5	11	158	8.3	315	13	7.6	3.3	2.85	2.7	1.5
6	13	80	7.6	380	20	21	5.7	9.8	0.8	3.65
7	3	63	8	425	38	36	8	12.5	2.2	6.5
8	3	252	8.3	360	100	9.5	2.9	2.52	4.6	1.6
9	11	198	8.5	290	113	3.3	1.5	0.4	4	0.1
10	15	100	7.8	385	46	15	2.5	7.9	7.7	4.5
11	2	79	8.1	350	84	7.1	1.9	2.7	13.2	3.7
12	3	315	8.3	370	100	8.7	2.8	2.8	4.8	2.85
13	12	280	8.6	290	126	3.5	1.5	0.45	4	0.73
14	16	140	8	360	76	12	2.6	4.9	8.4	3.45
15	3	85	8.3	330	106	2	1.4	0.42	12	1.6
16	3	498	8.3	330	100	4.8	1.6	1.04	4.4	0.82
17	13	322	8.5	285	117	3.6	1.6	0.48	4.6	0.84
18	15	160	8.4	345	91	1.7	1.9	0.22	10	1.74
19	2	72	8.6	305	91	1.6	0.9	0.1	9.5	1.25
20	2	390	8.2	330	100	1.7	1.2	0.56	5	0.6
21	11	303	8.5	245	100	1.7	0.9	0.05	2.7	0.16
22	13	310	8.2	285	82	8.5	1.6	0.59	3.7	0.6
23	4	181	8.6	270	105	2.8	0.5	0.1	3.66	0.43
24	3	480	8.2	290	100	1.3	0.8	0.04	2.2	0.13

It is then necessary to create a file that indicate which sampling site and which sampling date a given sample (row) belongs to, i.e., to create a file that describe the experimental design.

Select the option **Create2Categ** of the **TextToBin** module as follows:



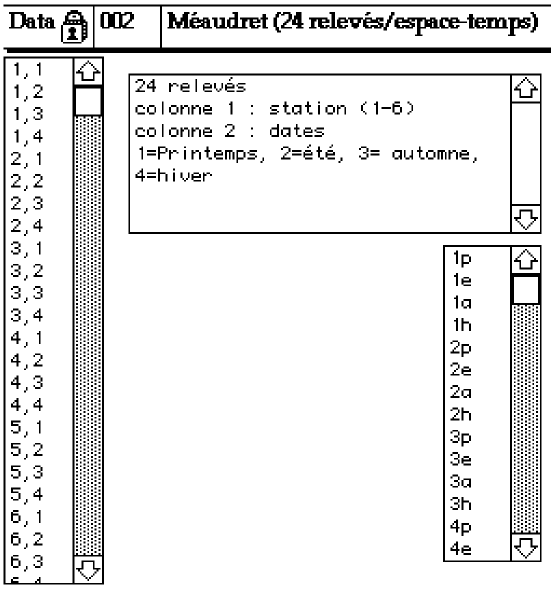
The new file `ST` (24-2) indicates that sampling units are distributed into six groups (6 sites) and four replicates (1-Spring; 2-Summer; 3-Autumn; 4; Winter).

This results in a listing as follows:

```
File ST contains two categorical variable
for a complete two-way layout without repetition
Row number: 24  Column number: 2
------------------------------------------
|   Description of a coding matrix  |
------------------------------------------
Qualitative variables file: ST
Number of rows: 24, variables: 2, categories: 10

Description of categories:
----------------------------------------------
Variable number 1 has 6 categories
----------------------------------------------
[  1]Category:    1 Num:  4 Freq.:  0.1667
[  2]Category:    2 Num:  4 Freq.:  0.1667
[  3]Category:    3 Num:  4 Freq.:  0.1667
[  4]Category:    4 Num:  4 Freq.:  0.1667
[  5]Category:    5 Num:  4 Freq.:  0.1667
[  6]Category:    6 Num:  4 Freq.:  0.1667

Variable number 2 has 4 categories
----------------------------------------------
[  7]Category:    1 Num:  6 Freq.:   0.25
[  8]Category:    2 Num:  6 Freq.:   0.25
[  9]Category:    3 Num:  6 Freq.:   0.25
[ 10]Category:    4 Num:  6 Freq.:   0.25

----------------------------------------------
Auxiliary binary output file STModa: Indicator vector of modalities
It contains variable number for each modality
It has 10 rows (modalities) and one column

Auxiliary ASCII output file ST.123: labels (two characters) for 10
modalities
It contains one label for each modality
It has 10 rows (modalities) and labels 1a,1b, ..., 2a, 2b, ...
Variable number 1,2, ..., A, ..., Z,+, Modality number a,b, ..., z,+
```

Create a label file `Code_Rel` that contains character strings (`1p,1e,1a,1h...,` `2p,2e,...,6e,6a,6h`) to identify samples. Figures identify sites and letters identify seasons (`p` for Spring, `e` for Summer, `a` for Autumn, `h` for winter).

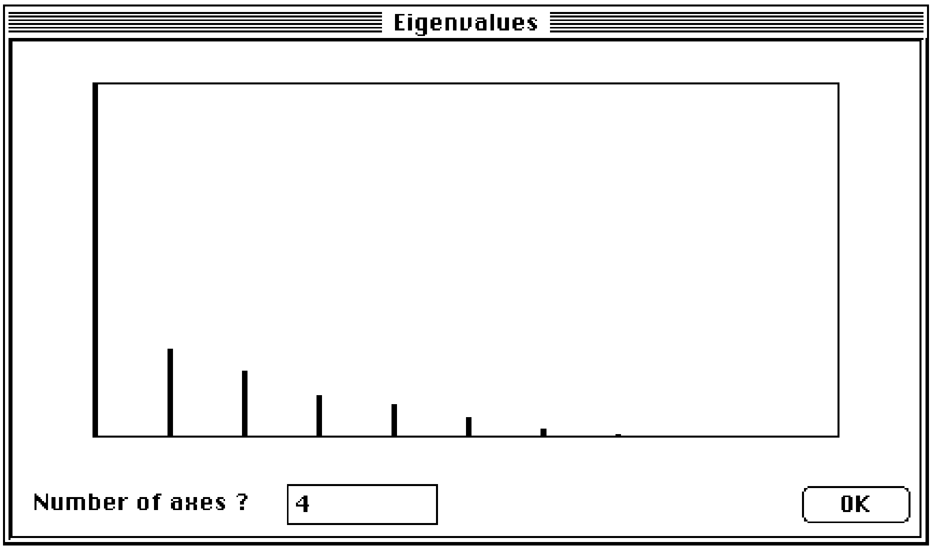
These labels must be picked up from the right hand field of the « Méaudret+1 » example of the **ADE•Data** stack:



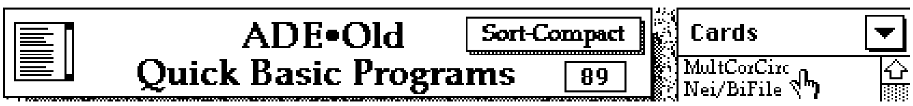
You should note that instead of processing the above option **Create2Categ** to create file `ST`, you can copy the left-hand field of the above card (`Plan.txt` by default) and transform the data into binary as usual. In that case, you have to read the resulting file using the **Read Categ File** option of the **CategVar** module (this option was incorporated in the **Create2Categ** option of the **TextToBin** module).

2.2 - Normalised principal components analysis

Run the classical normalised PCA using the **Correlation matrix PCA** option of the **PCA** module and select four axes as follows:

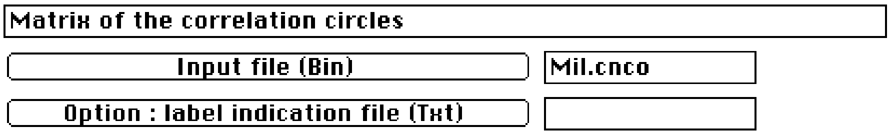


Select the Quick Basic program **MultCorCirc** via the **ADE•Old** selection card:



[See now ADEScatters : Draftman's display]

Fill in the dialog boxes as follows:



Select the default limits of the graphics and a window size of 400 pixels. This results in Fig. 4. Such graphics depicts the geometry of the 10 points (variables) in the multidimensional space R^{24}. They demonstrate a clear redundancy among variables no. 4 (Conductivity), no. 6 (B.O.D), no. 7 (Oxydability), no. 8 (ammonia concentration), and no. 10 (Phosphates), which all describe organic pollution.

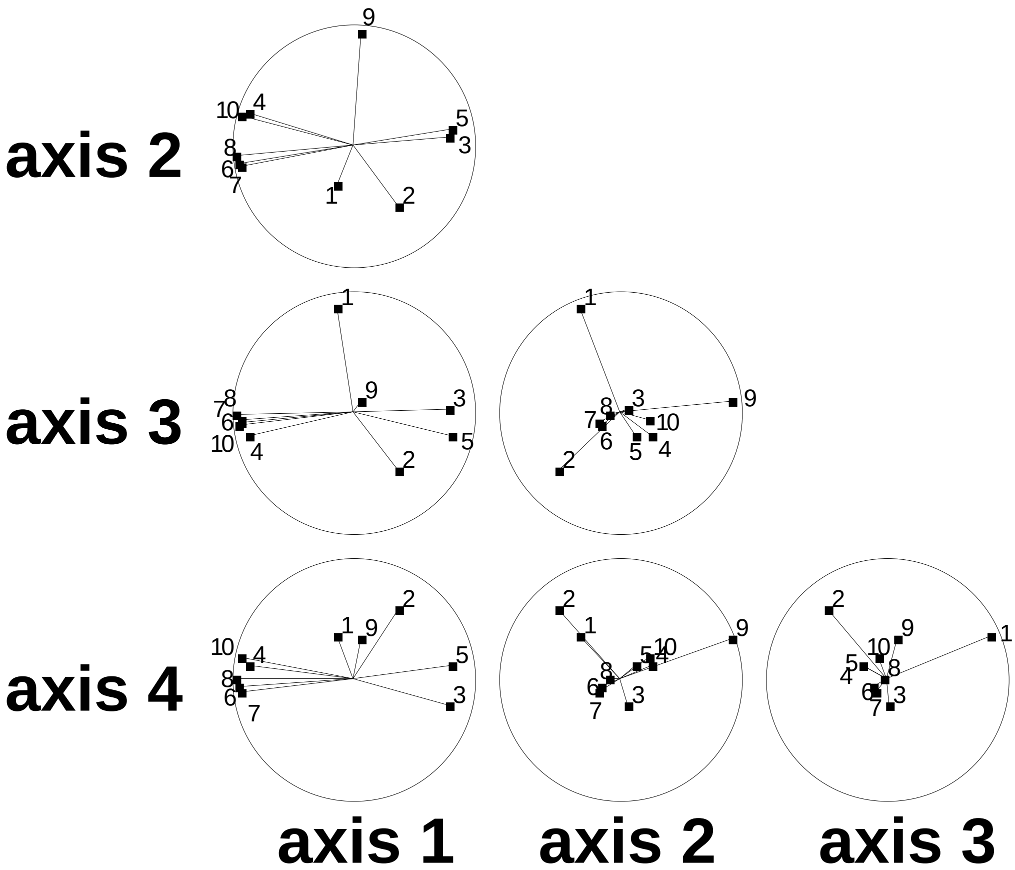
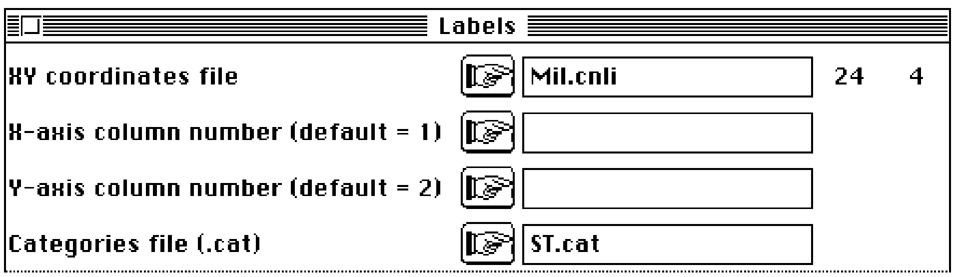


Figure 4 Correlation circles (factorial maps of variables for all combination of two factors among the four selected). The plane F1-F2 is at the top left of the figure.

The factorial maps of samples summarise the normalised PCA and may be elaborated using a simple two-axis diagram. The **ScatterClass** module of ADE allows to plot the centres of gravity for each group and the link between a sample and a group.



Select the **Stars** option of the **ScatterClass** module and fill in the boxes as above. One may choose to represent the variability among sites (Fig. 5A) or among seasons (Fig. 5B). Note that the **ScatterClass** module enables the simultaneous drawing of the two. The same options can be applied to the factorial plane 3-4 (Fig. 6).

You should note that the **Stars** option uses automatically the `ST.123` file for the identification of groups. Consequently in Fig. 5 and Fig. 6, the labels were replaced by more easily understandable ones.

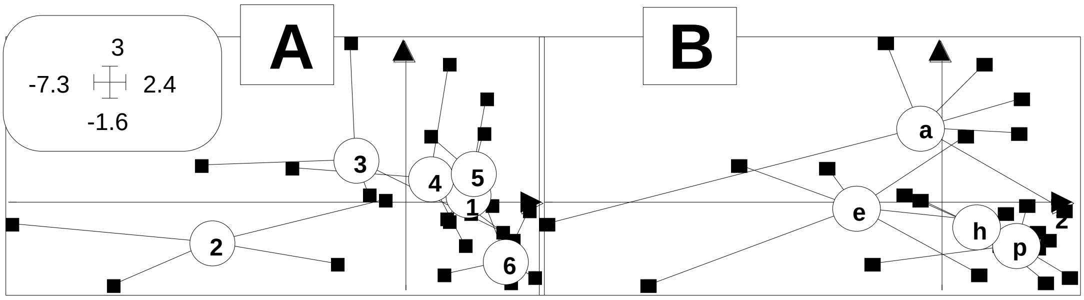


Figure 5 Factorial plane 1-2 of the sampling units. A - Variability of scores among sites. B - Variability of scores among season. Samples are identified by squares. Lines link samples to the corresponding site or season.

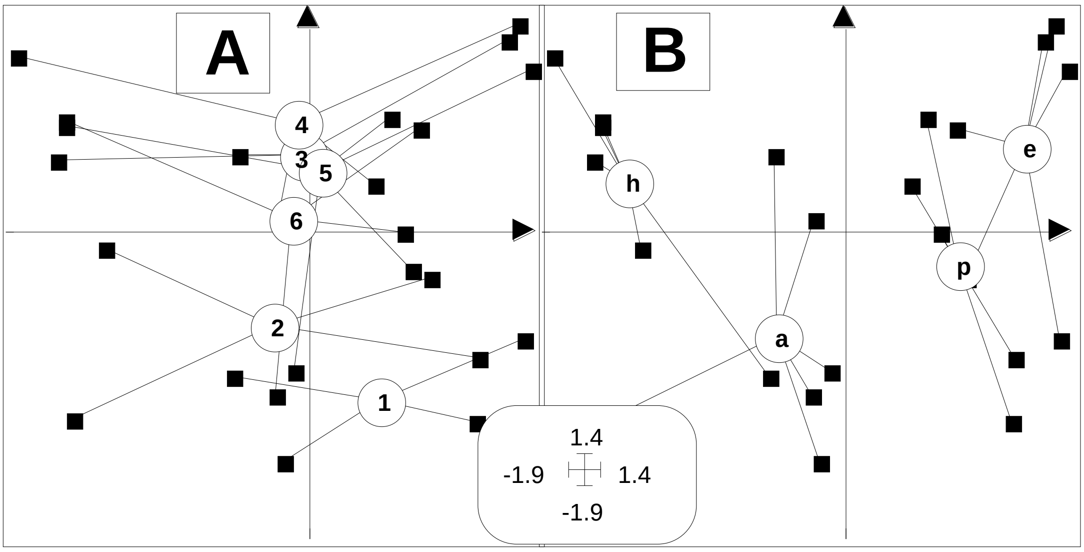


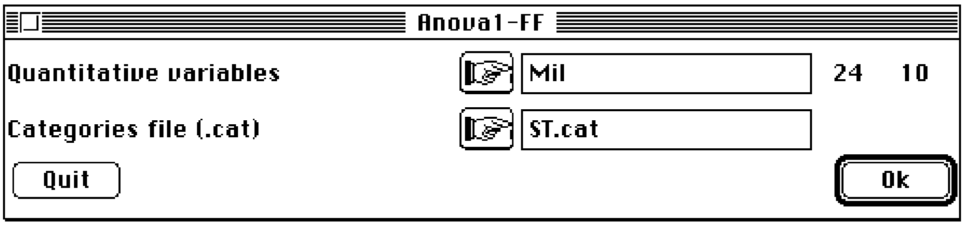
Figure 6 Factorial plane 3-4 of the sampling units. A - Variability of scores among sites. B - Variability of scores among season. Samples are identified by squares. Lines link samples to the corresponding site or season. A typology of seasons is demonstrate by axis 3 (B) whereas axis 4 depicts a typology of sites (A).

As a result, the four first axes of the normalised PCA of `Mil` are needed to depict the correlation among variables that are linked to the spatial-temporal structure. The first axis (57% of inertia) takes into account pH, conductivity, oxygen concentration, B.O.D, oxydability, ammonia and phosphates concentrations. It may be interpreted as a mineralization gradient and also indicate the high level of pollution in site 2 during summer. Such a pollution induces an acidification (lower pH), a lower oxygen concentration, a higher B.O.D and oxydability. High concentrations in ammonia and phosphates also are characteristics of a high organic pollution. A restoration of the river can be observed from site 3 towards site 5. Sites 1 and 6 represent the unpolluted sites. The temporal evolution of pollution is different from the seasonal cycle defined by water temperature (third axis). Furthermore, the restoration of water quality along the river course is not exactly related to the gradient of discharge from up- to downstream (fourth axis).

Consequently this analysis mixed together the seasonal typology and the spatial typology, which control the spatial-temporal process produced by water flowing and the evolution of air temperature. This process may be decomposed (in a geometric sense), i.e., one can choose to focus on a given component (space, time) of the sampling design or to eliminate this component.

2.3 - Coming back to raw data

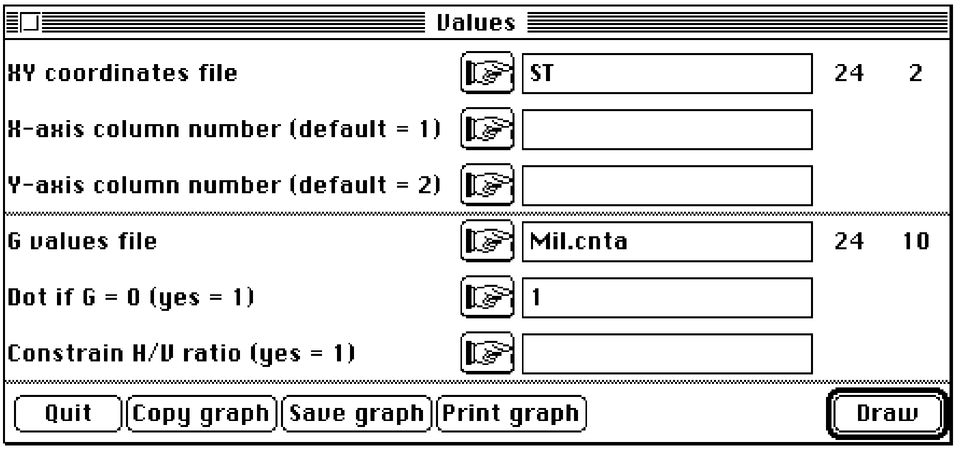
To test the spatial and the temporal effect, a One-way Analysis of Variance may be processed. Select the option **Anova1-FF** of the module **Discrimin** and fill in the dialog boxes as follows:



This module enables the computation of a one-way layout with a fixed effect of each quantitative variable of a file upon each categorical variables of another file. This results in a listing as follows:

```
variable 1 from Mil versus variable 1 from ST
--------------------------------------------------------------
Source *      SS* d.f.*     MS*     F*   Proba*
Between * .6708E+01*  5* .1342E+01* .3725E-01*  0.99911*
Within * .6482E+03*  18* .3601E+02*
Total  * .6550E+03*  23*
--------------------------------------------------------------
variable 1 from Mil versus variable 2 from ST
--------------------------------------------------------------
Source *      SS* d.f.*     MS*     F*   Proba*
Between * .6345E+03*  3* .2115E+03* .2063E+03*  0.00000*
Within * .2050E+02*  20* .1025E+01*
Total  * .6550E+03*  23*
--------------------------------------------------------------
....................................................................
vari 2 * vari 3 * vari 4 * vari 5 * vari 6 * vari 7 * vari 8 *
0.10056* 0.34913* 0.00816* 0.01385* 0.01066* 0.00056* 0.00937*
0.00227* 0.01495* 0.05375* 0.24410* 0.47147* 0.74409* 0.36242*

vari 9 * vari 10*
0.04892* 0.01882*
0.07868* 0.18287*
```

The experimental design can be used to plot the normalised values. Use the option **Value** of the module **Scatters**:



This results in Fig. 7.

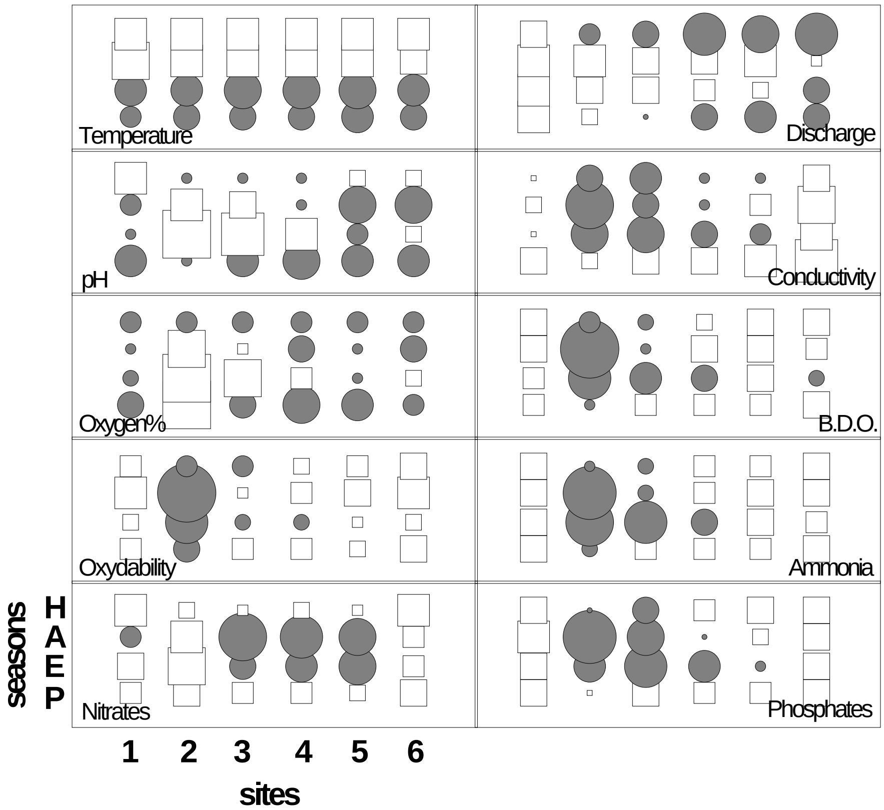


Figure 7 Spatial-temporal representation of the normalised values of 10 physical and chemical variables. The surfaces of circles (values>mean) and squares (values<mean) are proportional to the normalised data.

3 - Removing an effect: within-groups PCA

In this analyses, all centres of classes are plotted at the origin of the factorial maps and the sampling units are scattered with the maximal variance around the origin. Hence, the aim of such analysis is to enable the simultaneous study of the spatial typologies or to make a collection of spatial typologies (Fig. 8).

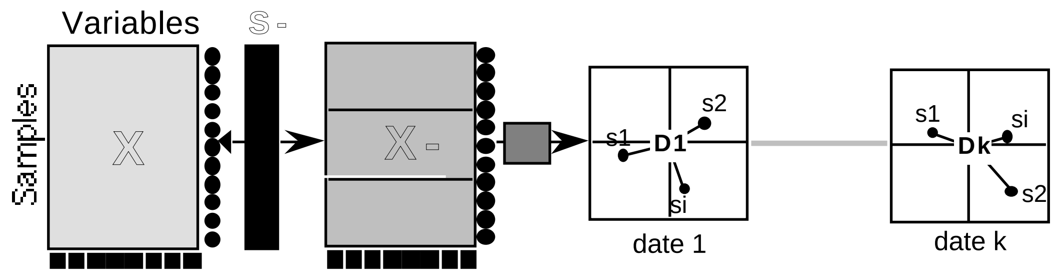
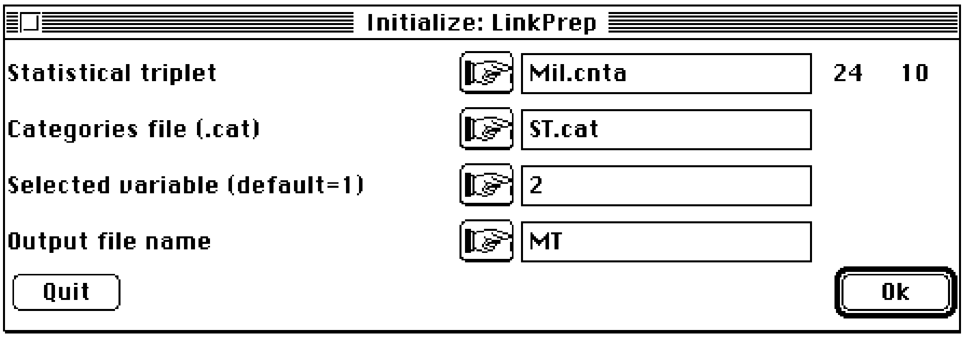


Figure 8 Eliminating a temporal effect connected to the margin (noted S-). The row and column weights are indicated by circles and squares. In the array noted X- lay the residuals (by dates). The multivariate analysis of this table results in a collection of spatial typologies (after Dolédec & Chessel, [4]*).*

There are at least two ways for doing a within-groups analysis in ADE. The simplest one consists in using the module **Discrimin** that enables the study of the link between a table and a categorical variable that identifies the groups.

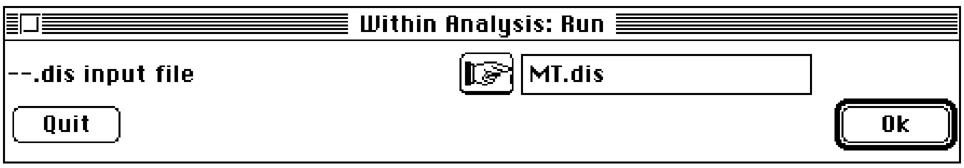
This analysis has to be prepared using the **Initialize: LinkPrep** option as follows:

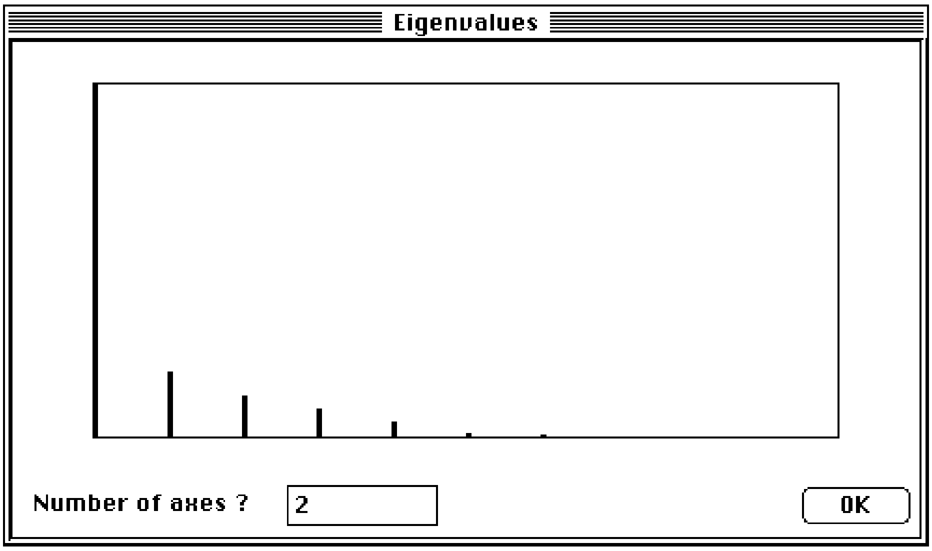


The above selection means that the normalised PCA of `Mil` is linked to the partition of sampling units by season. The resulting listing is as follows:

```
-------------------------------------------------------
New TEXT file MT.dis contains the parameters:
   input file: Mil.cnta
   categorical variable file: ST.cat
   n° of categorical variable used: 2
-------------------------------------------------------
-------------------------------------------------------
Between and Within-class inertia
Categories defined by column 2 of file ST
Input statistical triplet: table Mil.cnta
total inertia: 10.000000
between class inertia 3.185858 (ratio: 0.318586)
within class inertia 6.814142 (ratio: 0.681414)
-------------------------------------------------------
```

To remove the temporal effect, the within-seasons PCA is computed via the option **Within Analysis** of the **Discrimin** module as follows:





This results in a listing as follows:

```
-------------------------------------------------------
Within-class analysis
Categories defined by column 2 of file ST
Input statistical triplet: table Mil.cnta
Number of rows: 24, columns: 10
total inertia: 10.000000

-------------------------------------------------------
File MT.whta contains the block-centered array
It has 24 rows and 10 columns
File MT.whpc contains the column weights
It has 10 rows and 1 column
File MT.whpl contains the row weights
It has 24 rows and 1 column
within-class inertia 6.814142 (ratio: 0.681414)

-------------------------------------------------------
Num. Eigenval.  R.Iner. R.Sum    |Num. Eigenval.  R.Iner.   R.Sum |
01  +4.6505E+00 +0.6825 +0.6825  |02  +8.7006E-01 +0.1277 +0.8102 |
03  +5.5652E-01 +0.0817 +0.8918  |04  +3.9004E-01 +0.0572 +0.9491 |
05  +2.0546E-01 +0.0302 +0.9792  |06  +6.5492E-02 +0.0096 +0.9888 |
07  +3.1483E-02 +0.0046 +0.9935  |08  +2.2419E-02 +0.0033 +0.9968 |
09  +1.2484E-02 +0.0018 +0.9986  |10  +9.6367E-03 +0.0014 +1.0000 |

File MT.whvp contains the eigenvalues and relative inertia for each
axis. It has 10 rows and 2 columns

File MT.whls contains scores of the rows of the initial table (lambda
norm).  It has 24 rows and 2 columns
................................................etc.
File MT.whli contains standard scores of the rows of the centered
table (lambda norm). It has 24 rows and 2 columns
................................................etc.
File MT.whc1 contains column scores with unit norm
It has 10 rows and 2 columns
................................................etc.
File MT.whco contains standard column scores with lambda norm
It has 10 rows and 2 columns
................................................etc.
```

In terms of inertia, the total PCA of `Mil` results in an inertia equal to 10 (number of variables in a normalised PCA). The within-groups inertia is equal to 6.81, i.e., 68.1% of the total inertia is attributed to the within-groups PCA. Moreover, 68.3% of the within-groups inertia is given by the first axis. Doing a within-groups PCA is quite similar to doing the simultaneous PCA of the four variables-by-sites tables defined by the four sampling seasons. As a result, it is possible to search for a graphical representation according to four different factorial maps related by the within-groups PCA. Use the option **Labels** of the module **Scatters** and select the groups of rows via the **Row & Col. selection** option of the **Windows** menu (Fig. 9):

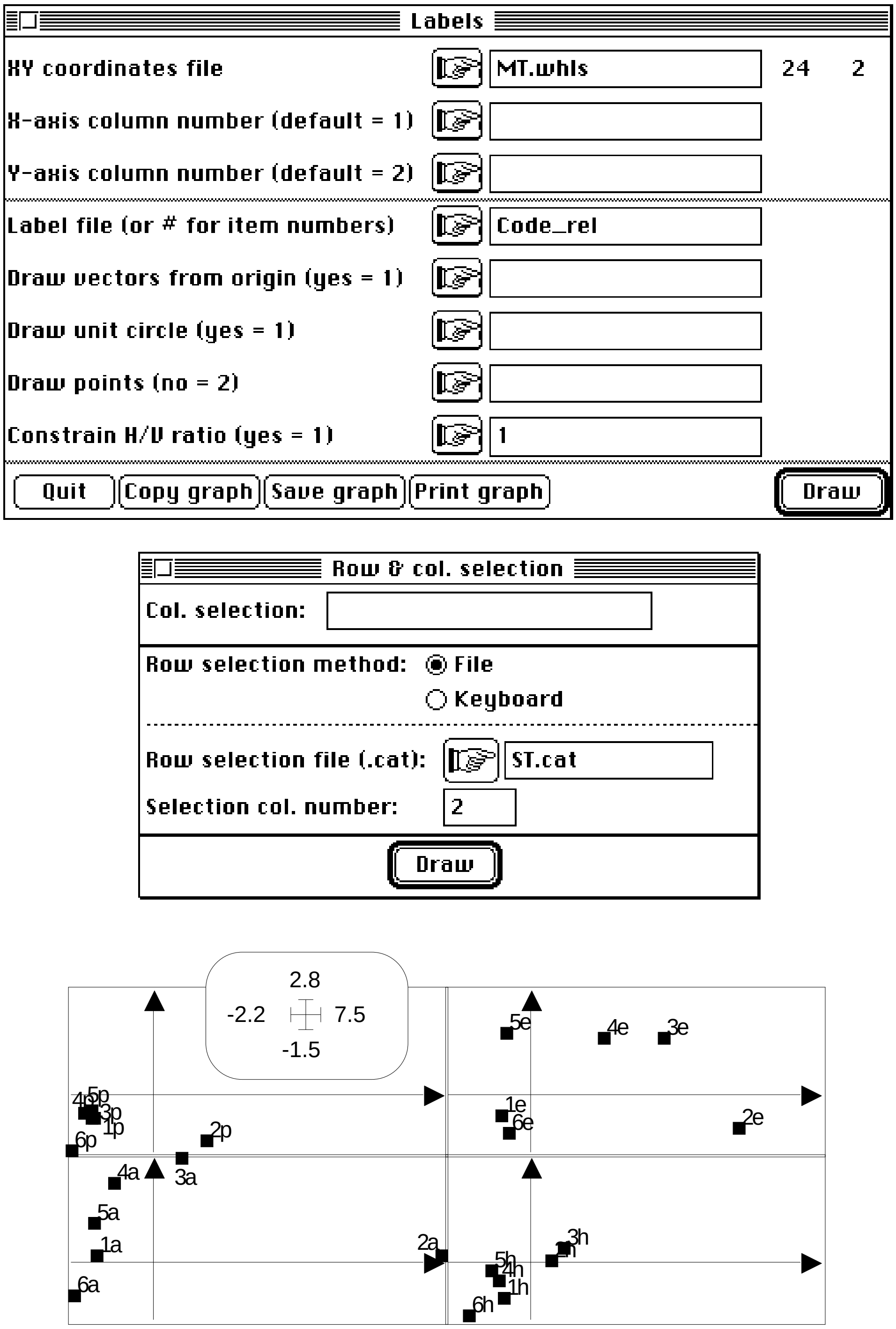


Figure 9 Projection of the normalised multidimensional space onto the factorial plane (1-2) of the within-groups PCA. A character string identifies the sampling site (figures 1-6) and the seasons (letters).

The spatial typology is not similar from one sampling season to another (Fig. 9). In this figure, we have used the file `MT.whls` whereas in the Fig. 10, we have used the file `MT.whli`. The within-groups inertia axes represent axes produced by the superposition of the four groups (dates) of six sites centred by dates. The corresponding multidimensional space may be projected using a two-axis representation. The resulting coordinates are in `MT.whli`. Using this latter file each of the four graphics is centred by date (Fig. 10).

By contrast, `MT.whls` represents the 24 raw sampling units projected onto the within-groups inertia axes. In that case, the sub-spaces are no more centred by date. However, in Fig. 8 and Fig. 10, one can see how the longitudinal gradient behave from site 1 towards site 5 using the same reference point (origin of graphics).

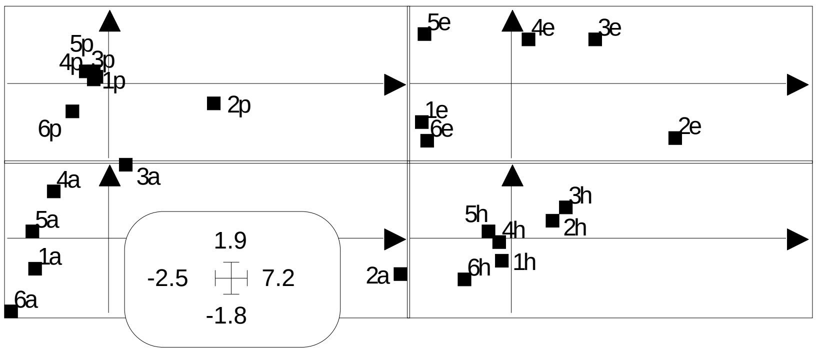


Figure 10 First factorial plane (1-2) of the within-groups analysis. A character string identifies the sampling site (figures 1-6) and the seasons (letters).

Use the **Labels** option of the **Scatters** module and `MT.whc1` or `MT.whco` to draw the correlation circles in Fig. 11. In `MT.whc1` the column scores are normalised. In `MT.whco`, the variance of scores is equal to the eigenvalue. The factorial scores of variables in `MT.whco` do not represent correlation between axes and variable; they are covariances. Consequently, the representation of a correlation circle with these values is not absolutely correct.

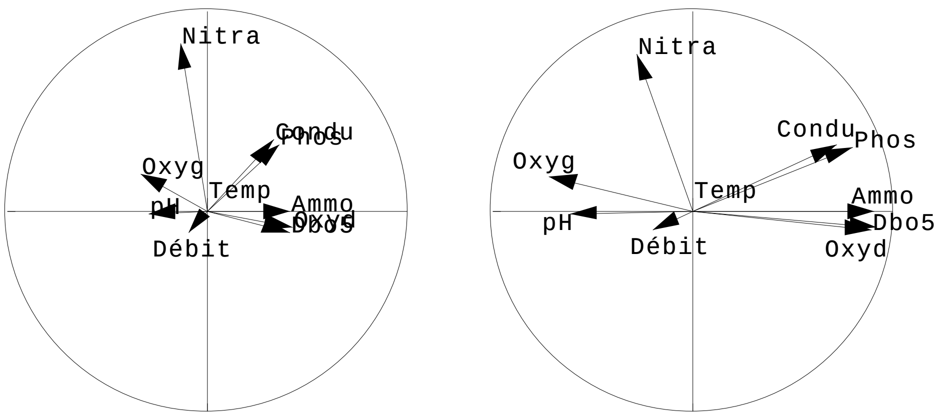


Figure 11 Correlation circles using file `MT.whc1` (on the left) and file `MT.whco` (on the right).

These two files (`.c1` and `.co`) represent two viewpoints concerning the analyses using projections. Let us consider a n-p matrix **X**, and two diagonal matrices **D** and **Q** associated to the rows and columns of **X** respectively. Furthermore, let us consider a subspace **A** defined by a categorical variable (site or season). The classical inertia analysis of the triplet ($\mathbf{P_A(X)}$, **Q**, **D**) with $\mathbf{P_A(X)}$ being the projection table of **X** on to subspace **A** results in scores for columns (noted `.co`) and scores for rows (noted `.li`).

Another point of view is to consider that the within-groups PCA aims to compute a linear combination of variables (noted `.li`) using coefficient for variables ((noted `.c1`) so that the projected inertia is maximum. This means that the variance should be as high as possible as well as the % of projection according to the following equation:

projected inertia = inertia x % of projection.

The latter point of view introduces to PCA with respect to instrumental variables (see volume 5).

Finally, the interpretation of the within-dates PCA can be as follows: during the Spring period (low pollution) sites 1, 3 and 5 are grouped together (Fig. 9 and Fig. 10). Sites 6 (Bourne river unpolluted site) and 2 (polluted) are separated. In August, as pollution increases, site 2 goes farther along axis 1; site 1 separates from sites 3 and 5 and gets closer to site 6. The succession of sites 3 to 5 suggests a restoration from up to downstream. In Autumn, pollution increases. In Winter, sites 4 and 5 get closer to unpolluted sites whereas sites 2 et 3 are always influenced by the effluents of Autrans village.

Symmetrically, on can compute a within-sites PCA, which seeks for elements common to the six dates-by-variables tables (Fig. 12).

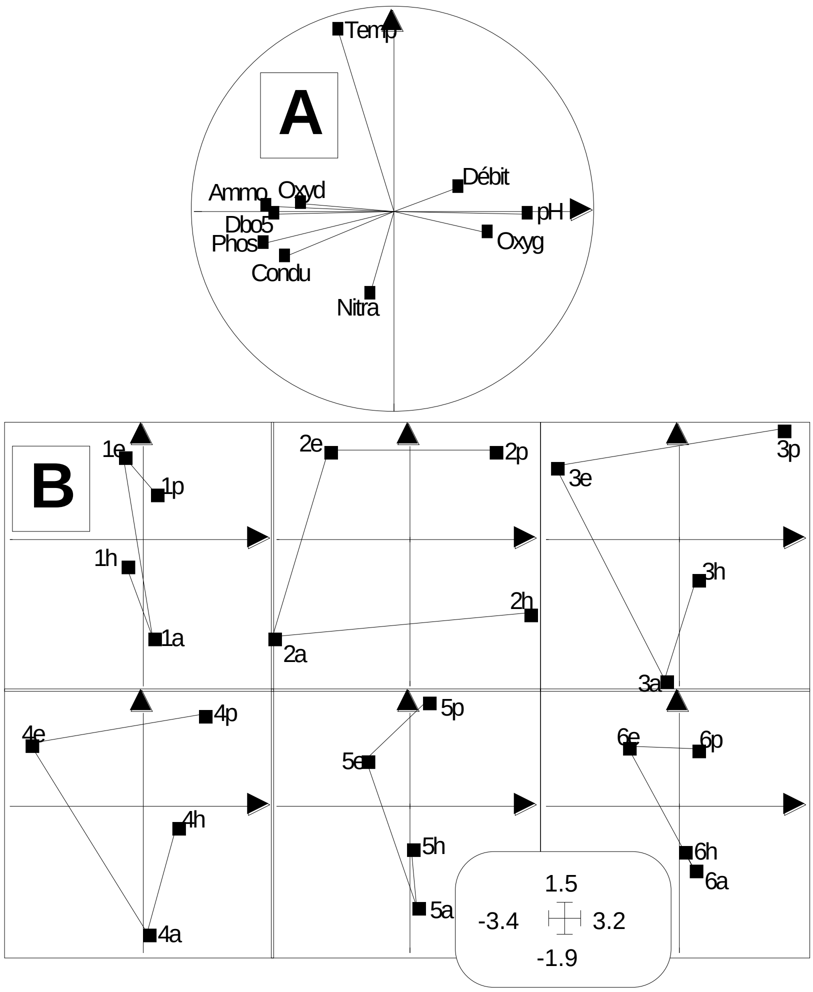


Figure 12 Results of the within-sites PCA. Projection of variables and samples on the first factorial plane (1-2). A - Correlation circle. B - Factorial plane (1-2) of samples separated by sites.

In that case, the sampling chronology is represented at each station. The pollution is clearly identified by axis 1 and the seasonal cycle occurs on axis 2 (Fig. 12A). Trajectories indicate that the variation among sampling dates is maximum close to the pollution effluent (site 2) and less important for sites 3 to 5. The non-polluted sites 1 and 6 are relatively stable (Fig. 12B).

4 - Focusing on the effect: between-groups PCA

A between-groups PCA may be associated to the within-groups analysis. The second one seeks for axes shared by the subspaces. The first one seeks for axes of the centre of gravity space and focuses on the between-groups difference, in that case the temporal variations (Fig. 13).

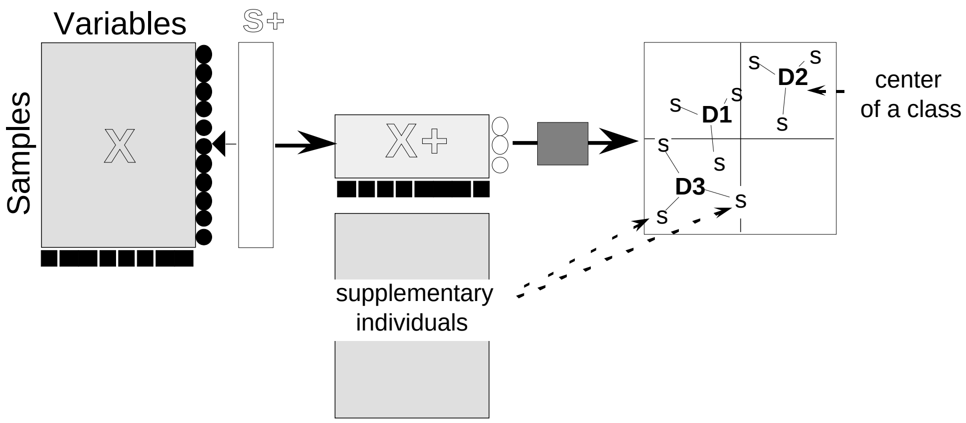
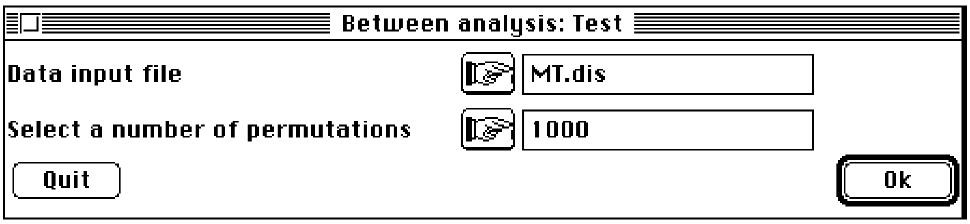


Figure 13 Taking into account a temporal effect (noted S+). Row and column weights are indicated by circles and squares. In the array noted X+, data are cumulated by dates. The supplementary individuals are generated by the initial data table (X). After the central procedure (hatched rectangle), a typology of sampling dates is displayed on factorial maps (after Dolédec & Chessel, op. cit.).

The statistical significance of the dispersion of the centres of gravity may be tested using the **Between analysis: test** option of the module **Discrimin** as follows:



This results in a listing as follows:

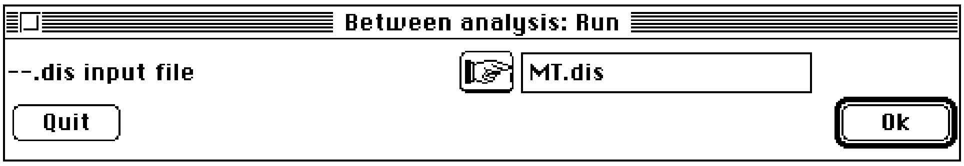
```
number of random matching: 1000  Observed: 3.185858
Histogram: minimum = 0.336260, maximum = 3.406161
number of simulation X<Obs: 998 (frequency: 0.998000)
number of simulation X>=Obs: 2 (frequency: 0.002000)
```

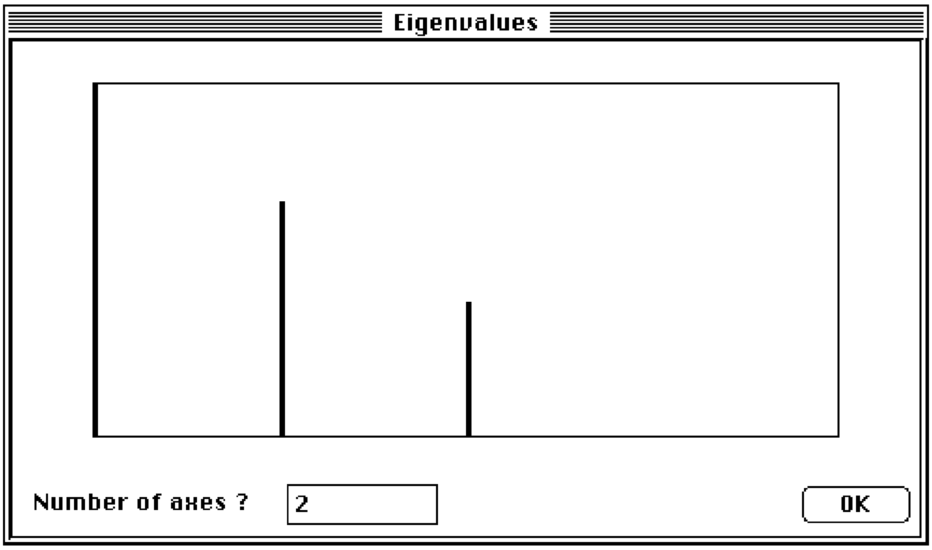
This means that only two random values out of 1000 random permutations are higher than the observed value. an histogram is also associated with these values (see below)

```
|*****
|****************
|*******************************
```

```
    |******************************************
    |**************************************************
    |*************************************************
    |***************************************
    |****************************
    |************************
    |********************
    |*******************
    |********
    |*******
    |*****
    |***
    |*
    |
    |*
 •->|*
```

Consequently, the between-groups PCA can be further investigated. After using the **Initialize: LinkPrep** option of **Discrim** (see the within-groups paragraph), use the **Between analysis: Run** option as follows:






```
---------------------------------------------------------
Between-class analysis
Categories defined by column 2 of file ST
Input statistical triplet: table Mil.cnta
Number of rows: 24, columns: 10
total inertia: 10.000000
---------------------------------------------------------
File MT.beta contains multivariate means by classes
It has 4 rows and 10 columns
File MT.bepl contains class weights
It has 4 rows and 1 column
File MT.bepc contains column weights
It has 10 rows and 1 column
This statistical triplet is the gravity centres one
between-class inertia 3.185858 (ratio: 0.318586)
---------------------------------------------------------
```

```
Num. Eigenval.  R.Iner. R.Sum  |Num. Eigenval.  R.Iner. R.Sum |
01  +1.5551E+00 +0.4881 +0.4881  |02  +1.0390E+00 +0.3261 +0.8143 |
03  +5.9176E-01 +0.1857 +1.0000  |04  +0.0000E+00 +0.0000 +1.0000 |

File MT.bevp contains the eigenvalues and relative inertia for each
axis
It has 4 rows and 2 columns

File MT.bec1 contains column scores with unit norm
It has 10 rows and 2 columns
File :MT.bec1
-----------------------Minimum/Maximum:
Col.:  1 Mini = -0.33749 Maxi = 0.41569
Col.:  2 Mini = -0.87323 Maxi = 0.33048

File MT.beli contains standard gravity centre scores with lambda norm
It has 4 rows and 2 columns

File :MT.beli
-----------------------Minimum/Maximum:
Col.:  1 Mini = -1.8209  Maxi = 1.1864
Col.:  2 Mini = -1.2083  Maxi = 1.3519

File MT.bels contains standard row scores with lambda norm
It has 24 rows and 2 columns

File :MT.bels
-----------------------Minimum/Maximum:
Col.:  1 Mini = -5.1795  Maxi = 2.5444
Col.:  2 Mini = -1.643 Maxi = 2.2857

File MT.beco contains standard column scores with lambda norm
It has 10 rows and 2 columns
File :MT.beco
-----------------------Minimum/Maximum:
Col.:  1 Mini = -0.42086 Maxi = 0.51839
Col.:  2 Mini = -0.89008 Maxi = 0.33686
---------------------------------------------------------
```

In that analysis the between-groups inertia is equal to 3.19, i.e., 31.9% of the total inertia (simple PCA of `Mil`) is attributed to between-groups PCA. As a result of the complementary nature of within-groups and between-groups analyses, the total inertia of the initial table can be decomposed into two parts. Each part is then decomposed into axes.

Consequently, considering the total inertia of a table **X** noted I_t, the inertia of table $\mathbf{X}^-$ (within-groups model) noted I_t^-, and the inertia of table $\mathbf{X}^+$ (between-groups model) noted I_t^+, we have the following relation:

$$I_t = I_t^+ + I_t^-$$

We can apply the equation to our example using the "season" effect:

$$10 = 3.185 + 6.815$$

or the site effect:

$$10 = 4.413 + 5.587$$

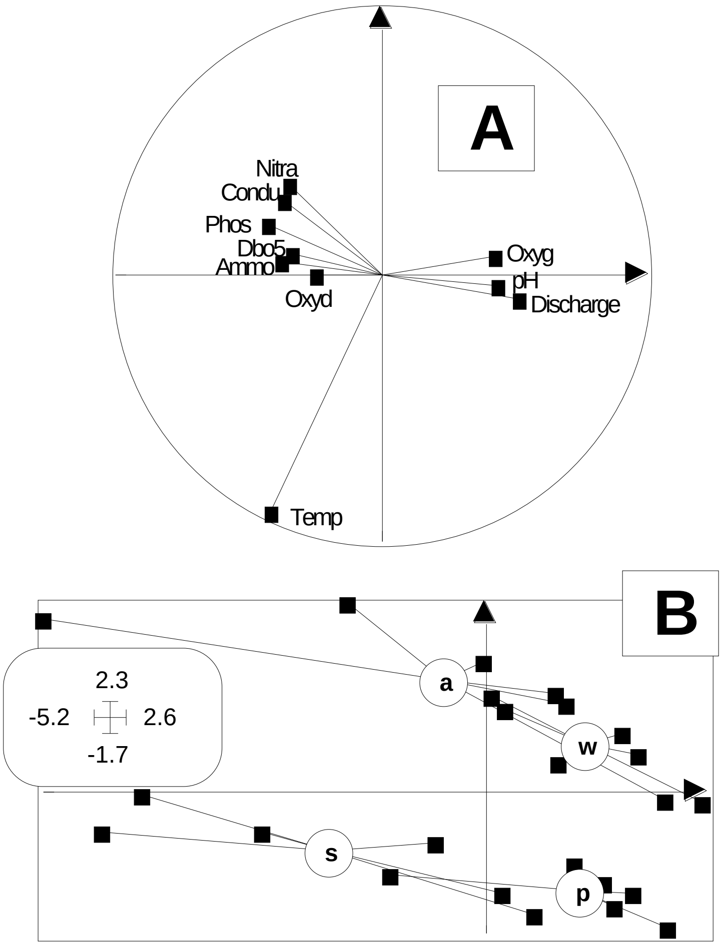
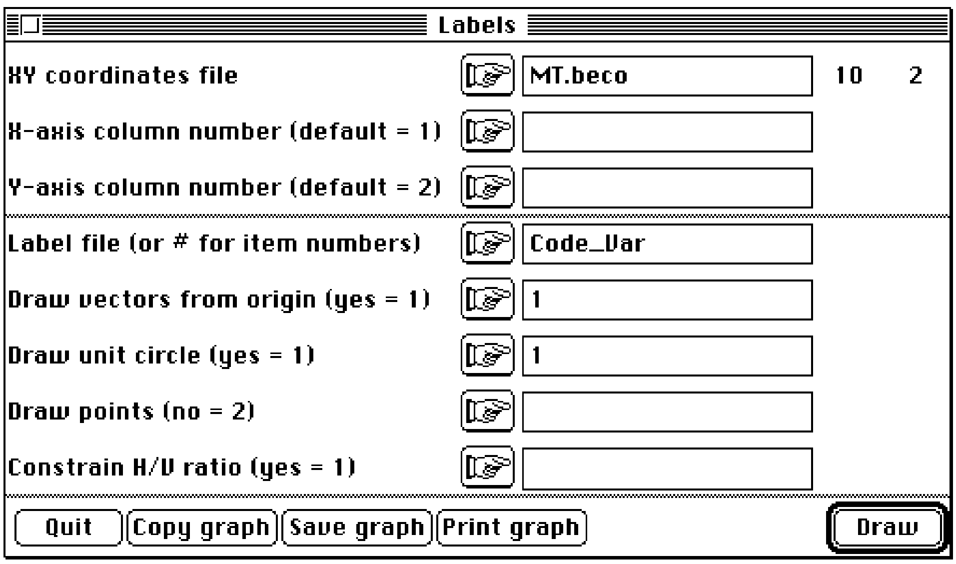
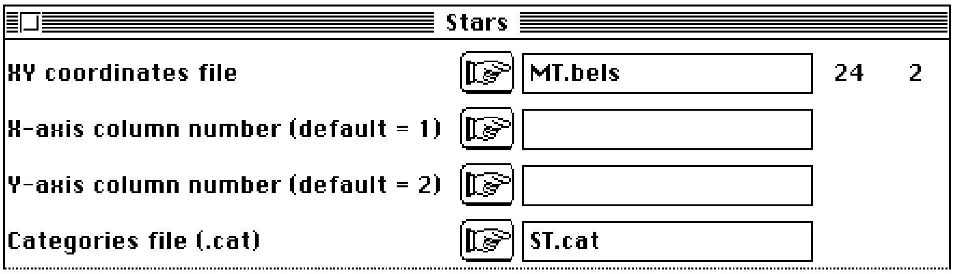


Figure 14 Results of the between-date PCA. A - Correlation circle. B - First factorial plane (1-2). The centres of gravity (letters in a circle) are distributed in the best way compared to Fig. 5B and 6B.

Use the **Labels** option of the **Scatters** module as follows:



After some modification with ClarisDraw™, this results in Fig. 14A. Use the **Stars** option of the **ScatterClass** module as follows:



This results in two graphics. Select the second one. After some modification with ClarisDraw™, this results in Fig. 14B. The temporal evolution is thus summarised (Fig. 14).

In average, according to axis 1, pollution is higher in Autumn (a) and Summer (s). During Winter (w) and Spring (s), the higher discharge induce a dilution of the organic pollution in the river. The axis 2 describes the seasonal rhythm influence by water temperature. Consequently, Autumn and Winter period are opposed to Spring and Summer.

Similarly, a between-sites PCA can be processed using the same procedure (Fig. 15). Variables describing pollution are again prominent on axis 1, whereas axis 2 takes into account the restoration process (evolution of nitrate concentration). Three groups of sites can be identified (Fig. 15): (1) site 2 (polluted site), (2) sites 1 and 6 (unpolluted sites), and sites 3 to 5 (sites under restoration).

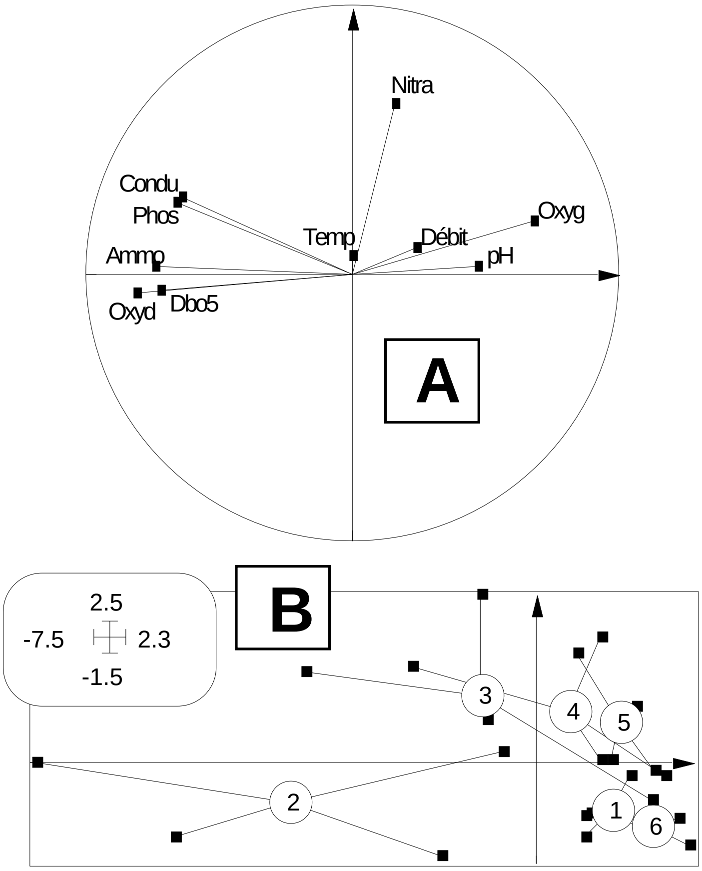


Figure 15 Results of the between-sites PCA. A - Correlation circle. B - First factorial plane (1-2). The centres of gravity (figure in a circle) are distributed in the best way compared to Fig. 5A and 6A.

5 - Decomposition of the variance

5.1 - Use of eigenvalues

Consequently, each analysis decomposes the total variability into spatial and temporal variability. The major part of this variability is taken into account by the first eigenvalue of each analysis (Fig. 16). A lower part of the total variability is lost by removing the temporal effect (within-dates PCA in Fig. 16). By contrast, a higher part of the total variability is lost by removing the spatial effect (within-sites PCA). Such a prominence of the spatial effect is also represented by the first eigenvalue of the between-sites PCA that is higher than the first eigenvalue of the between-dates PCA.

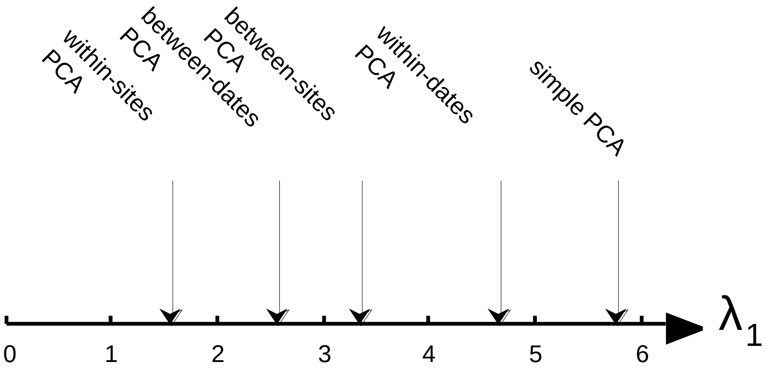
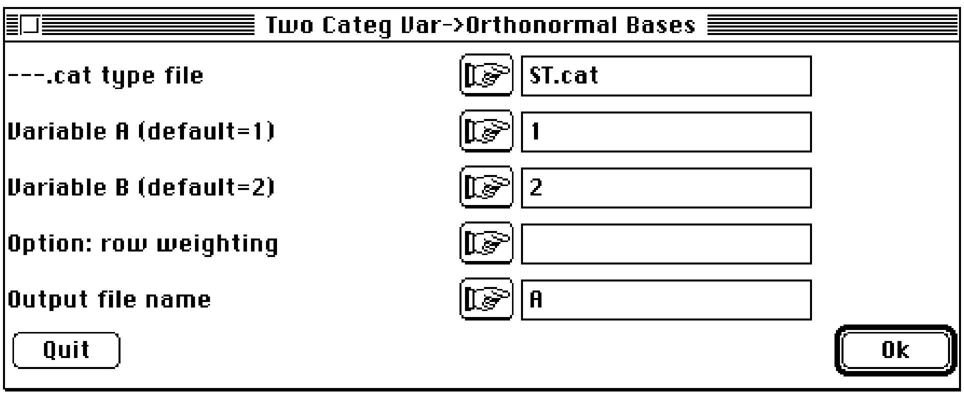


Figure 16 First eigenvalue (λ_1) of each analysis.

5.2 - Projection on to subspaces

The decomposition of the variance is associated to the Pythagorean theorem. Let **Z** a normalised variable be a unitary vector of R^n. In a geometric sense, the length (or squared length) of **Z** equals 1. In a statistical sense, the variance of **Z** equals 1. The norm (squared length of the vector) of a vector projected on to a subspace is equal to the variance of its components (if we are in the orthogonal subspace generated by $\mathbf{1}_n$ (centring)). The ratio of the second variance and the first variance is the percentage of variance explained by the projection. Consequently, the procedure incorporates two projection steps: (1) projection on to a subspace, (2) projection on to factorial axes of the PCA (reduction of the dimensions of the initial table).

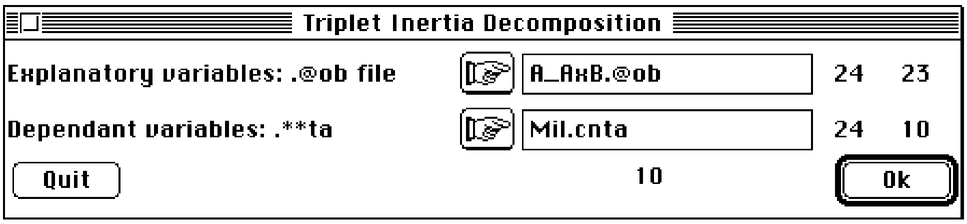


Several decomposition of the variance among various analyses are available in ADE. The module for using projections is called **Projectors**. First of all, use the **Two Categ Var->Orthonormal bases** option of that module as above. This option allows the construction of the orthonormal basis of seven subspaces as follows:

```
Subspaces from two categorical variables
```

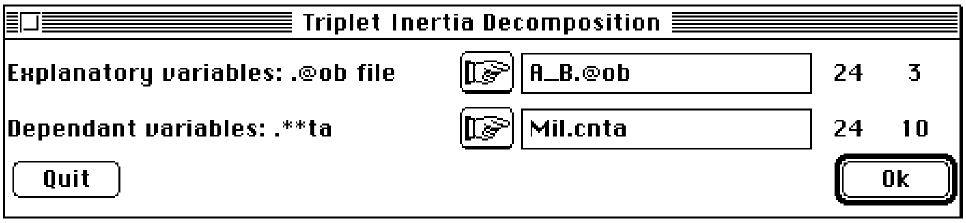
```
------------------------------------------
Input file: ST
It has 24 rows and 2 columns
Generic output file name: A
Crossing variable A (n° 1) and B (n° 2)
------------------------------------------
file A_AxB.@ob contains an orthonormal basis of subspace AxB
It has 24 rows and 23 columns
file A_A+B.@ob contains an orthonormal basis of subspace A+B
It has 24 rows and 8 columns
file A_A•B.@ob contains an orthonormal basis of subspace A•B
It has 24 rows and 15 columns
file A_A.@ob contains an orthonormal basis of subspace A
It has 24 rows and 5 columns
file A_B.@ob contains an orthonormal basis of subspace B
It has 24 rows and 3 columns
file A_A/B.@ob contains an orthonormal basis of subspace A/B
It has 24 rows and 5 columns
file A_B/A.@ob contains an orthonormal basis of subspace B/A
It has 24 rows and 3 columns
```

A projection of the normalised data (`Mil.cnta`) on to these subspaces takes into account a part of the variability. The total variability of `Mil.cnta` is taken into account by the subspace `AxB`. This can be verified by using the **Triplet Inertia Decomposition** option of the module **Projectors** as follows:




```
Orthonormal basis: A_AxB.@ob
It has 24 rows and 23 columns
Dependant variable file: Mil.cnta
It has 24 rows and 10 columns
------------------------------------------
|---|----------|----------|----------|   |------|-----|
|   |Subspace A| A Orthogo| Total    |   |  A+  |  A- |
|---|----------|----------|----------|   |------|-----|
| 1 |1.0000e+00|0.0000e+00|1.0000e+00|   | 10000|  0  |
| 2 |1.0000e+00|0.0000e+00|1.0000e+00|   | 10000|  0  |
| 3 |1.0000e+00|0.0000e+00|1.0000e+00|   | 10000|  0  |
| 4 |1.0000e+00|0.0000e+00|1.0000e+00|   | 10000|  0  |
| 5 |1.0000e+00|0.0000e+00|1.0000e+00|   | 10000|  0  |
| 6 |1.0000e+00|0.0000e+00|1.0000e+00|   | 10000|  0  |
| 7 |1.0000e+00|0.0000e+00|1.0000e+00|   | 10000|  0  |
| 8 |1.0000e+00|0.0000e+00|1.0000e+00|   | 10000|  0  |
| 9 |1.0000e+00|0.0000e+00|1.0000e+00|   | 10000|  0  |
| 10|1.0000e+00|0.0000e+00|1.0000e+00|   | 10000|  0  |
|---|----------|----------|----------|   |------|-----|
|Tot|1.0000e+01|0.0000e+00|1.0000e+01|   | 10000|  0  |
|---|----------|----------|----------|   |------|-----|
```

The variability associated to the effect (noted `Subspace A`) is uniformly equal to 1. The total inertia of `Mil.cnta` associated to such a projection is equal to 10. We can consider the seasonal effect, which corresponds to the decomposition of variance according to the subspace `B` and associated orthonormal basis. Use the **Triplet Inertia Decomposition** option of the module **Projectors** as follows:




```
Orthonormal basis: A_B.@ob
It has 24 rows and 3 columns
Dependant variable file: Mil.cnta
It has 24 rows and 10 columns
-----------------------------------------
```

	Subspace A	A Orthogo	Total	A+	A-
1	9.6870e-01	3.1300e-02	1.0000e+00	9687	312
2	5.0843e-01	4.9157e-01	1.0000e+00	5084	4915
3	4.0050e-01	5.9950e-01	1.0000e+00	4004	5995
4	3.1188e-01	6.8813e-01	1.0000e+00	3118	6881
5	1.8405e-01	8.1595e-01	1.0000e+00	1840	8159
6	1.1580e-01	8.8420e-01	1.0000e+00	1158	8841
7	5.8602e-02	9.4140e-01	1.0000e+00	586	9413
8	1.4446e-01	8.5554e-01	1.0000e+00	1444	8555
9	2.8246e-01	7.1754e-01	1.0000e+00	2824	7175
10	2.1099e-01	7.8901e-01	1.0000e+00	2109	7890
Tot	3.1859e+00	6.8141e+00	1.0000e+01	3185	6814

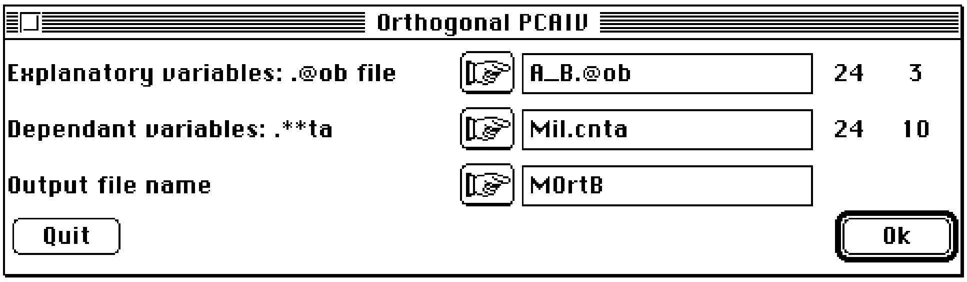
You can verify that the total inertia associated to this projection (`Subspace A, Tot=3.186`) is equal to the between-dates inertia (see above) whereas the inertia of the orthogonal projection (`A Orthogo, Tot=6.814`) corresponds to the within-dates inertia. Moreover, the importance of variable no. 1 (water temperature = 96.9%) in the definition of the seasonal effect is underscored.

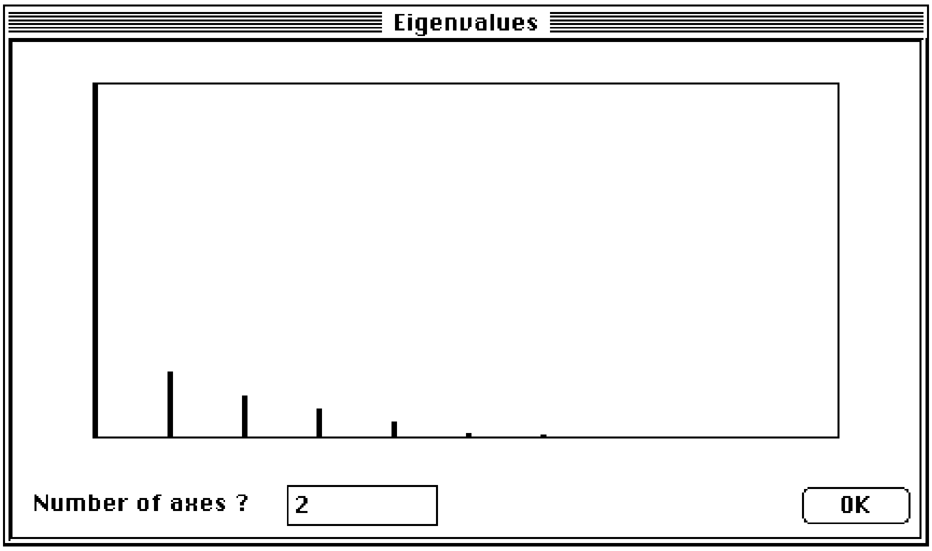
Table 2 Example of decomposition of each normalised variable (A: site effect, B: seasonal effect; A•B: interaction.). Values are multiplied by 1000.

A	B	A•B	Tot		
102	9687	210	1000	01	Temperature
3783	5084	1131	1000	02	Discharge
2497	4004	3497	1000	03	pH
5528	3118	1353	1000	04	Conductivity
5220	1840	2939	1000	05	Oxygen
5375	1158	3466	1000	06	B.D.O.
6774	586	2639	1000	07	Oxydability
5450	1444	3105	1000	08	Ammonia
4367	2824	2807	1000	09	Nitrates
5029	2109	2860	1000	10	Phosphates
4412	3185	2401	10000		

Using these projections, it is possible to make a decomposition of the original data similar to a one-way layout (Table 2). As a second step, the projected variance may be further decomposed according to each analysis (between, within, interaction, etc.). The projected variance on each axis for a given variable is equal to the square of the corresponding factorial coordinate. For example, let us consider the within-dates PCA.

You may either use the coordinates incorporated into `MT.whco` computed before or use the option **Orthogonal PCAIV** of the **Projectors** module as follows:





Note that the values incorporated into `MOrtB.ivco` are similar to those of `MT.whco`. You may import these values into Excel™ (**Edit with**), and put them to the square (Table 3). You also can import the `A-` column resulting from the projection of `Mil.cnta` on to subspace `B` (B in Table 2 appears as projected in Table 3).

Table 3 Example of variance decomposition of each normalised variable projected on to subspace generated by the within-date PCA. Values are multiplied by 1000.

		VARIANCE				
		initial	projected	axis 1	axis 2	residuals
1	Temperature	1000	31	0	3	28
2	Discharge	1000	492	37	11	443
3	pH	1000	600	374	0	225
4	Conductivity	1000	688	524	105	59
5	Oxygen	1000	816	505	29	282
6	B.O.D.	1000	884	810	10	64
7	Oxydability	1000	941	855	7	79
8	Ammonia	1000	856	826	0	29
9	Nitrates	1000	718	77	608	33
10	Phosphates	1000	789	641	96	51

As a result, about 80% of the variance of B.O.D, oxydability, and ammonia concentration participate to axis 1 (pollution) in the within-dates PCA (Fig. 11).

5.3 - Alternative centring

More generally, an experiment incorporates factors that interferes. For example, temporal replicates were recorded but the sampling chronology is of no interest for the experimenter; or a lot of sites were investigated but the role of the spatial distribution is of no interest. The corresponding interference factor (site, time) can be removed in average; this is the case in the classical within-groups PCA of this volume. Moreover, the given interference may be removed on the average effect and on its variance. This is the case when a within-groups PCA is processed on a table which variables are normalised by groups of individuals (Dolédec & Chessel, 1987). Such an option is available via the **Within group normalised PCA** option of the **PCA** module (Fig. 17).

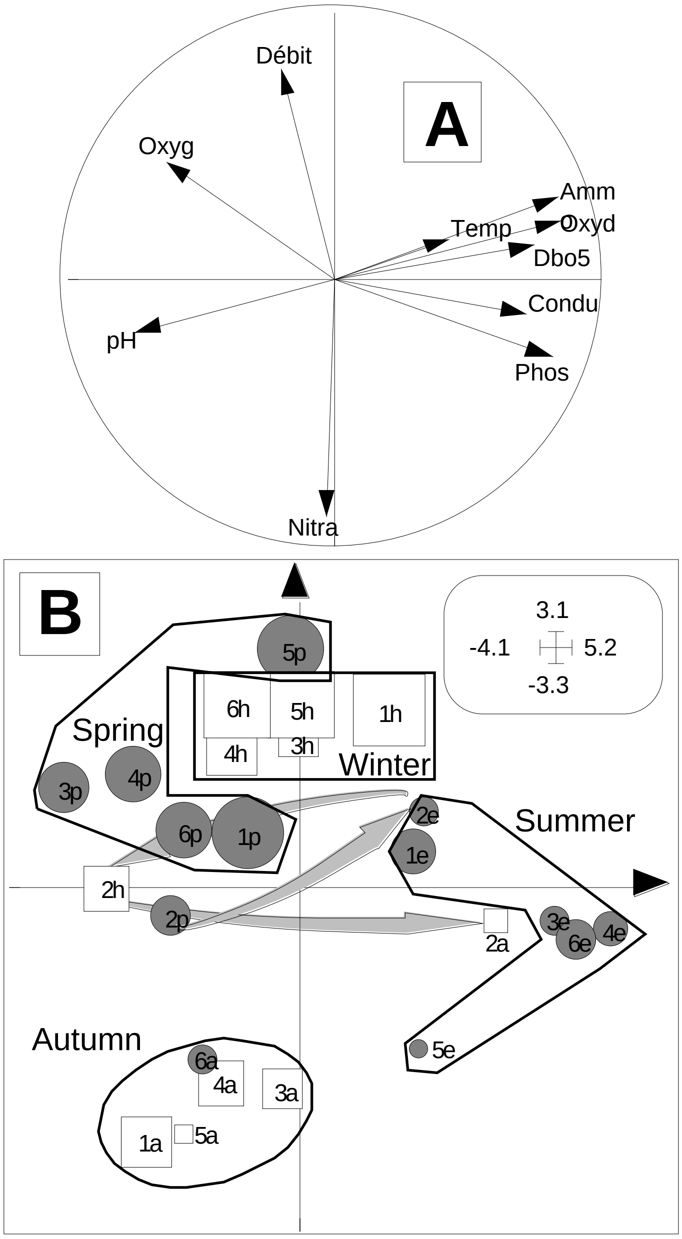


Figure 17 Within groups PCA of the table normalised by groups of sites (normalised deviations from the average by sites). A - Correlation circle of variables. B - F1-F2 factorial map of sampling units. Squares and circles are proportional to the SUs scores on axis 3. The centre of gravity of each group of sampling dates is at the origin.

In that case variables (columns) are normalised by groups of sampling units (according to sites or dates). As a consequence, the average values for a variable and for each group equal 0. As seen before, *total inertia = within-groups inertia + between-groups inertia*. In this option, the between-groups inertia equals 0. Consequently, this type of analysis is also normalised by variables as *total inertia = within-groups inertia*

= *average variance by groups = 1*. As a result, the values contained in the analysed table are equal to

$$x_{ijk} = \left(z_{ijk} - z_{i.k}\right) / s_{i.k}$$

with z_{ijk} being the raw value, $z_{i.k}$ being the average value, and $s_{i.k}$ being the standard deviation of the k*th* variable at the i*th* site.

The **Partial normed PCA** option of the **PCA** module is quite different (Bouroche, 1975)[5]. In that option, variables (columns) are centred by groups of sampling units and normalised using the average variance by groups (within-groups inertia). Consequently, this type of analysis also is normalised by variables as in a standard PCA (Fig. 18).

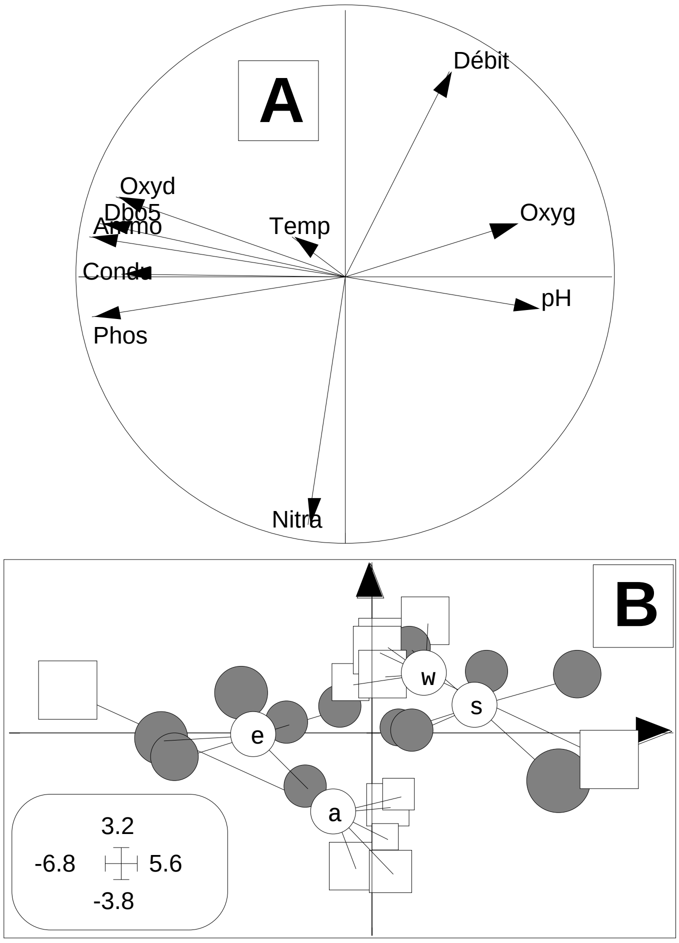


Figure 18 Partial normalised PCA. A - F1-F2 Correlation circle. B - F1-F2 factorial map of sampling units. Squares and circles are proportional to the SUs scores on axis 3. The centre of gravity of each group of sampling dates is at the origin. The centre of gravity of each group of sites is indicated by a letter in a circle (w = Winter; s = Spring; e = Summer; a = Autumn).

These two options both result in a normalisation by variables (columns). The first option removes the within-groups variability which equals 1 whereas the second option preserves such a variability. The interest may be apparent to study the matching between two tables (e.g., a faunistic table and an environmental table). Therefore, one can either reduce the diversity of the environmental menu and the heterogeneity in species composition to a value of 1. By contrast these features can be taken into account if the experimenter consider the heterogeneity of the environmental characteristics as compared to the heterogeneity of the species composition.

This example underlines the plasticity of ADE program library and the need for defining clearly the objectives in using projection on to subspaces because a lot of options are available. Furthermore, these between- and within analyses can be used in the context of correspondence analysis[6,7].

Finally, multiway layout analyses can be found in ADE program library incorporating partial triadic analysis[8] and STATIS[9,10] (see volume 6).

Références

[1] Usseglio-Polatera, P. & Auda, Y. (1987) Influence des facteurs météorologiques sur les résultats de piégeage lumineux. Annales de Limnologie : 23, 1, 65-79.

[2] Dolédec, S. & Chessel, D. (1991) Recent developments in linear ordination methods for environmental sciences. Advances in Ecology, India : 1, 133-155.

[3] Pegaz-Maucet, D. (1980) Impact d'une perturbation d'origine organique sur la dérive des macroinvertébrés d'un cours d'eau. Comparaison avec le benthos. Thèse de 3e cycle, Université Lyon 1, 130 pp.

[4] Dolédec, S. & Chessel, D. (1987) Rythmes saisonniers et composantes stationnelles en milieu aquatique I- Description d'un plan d'observations complet par projection de variables. Acta Œcologica, Œcologia Generalis : 8, 3, 403-426.

[5] Bouroche, J.M. (1975) Analyse des données ternaires: la double analyse en composantes principales. Thèse de 3° cycle, Université de Paris VI. 1-57 + annexes.

[6] Dolédec, S. & Chessel, D. (1989) Rythmes saisonniers et composantes stationnelles en milieu aquatique II- Prise en compte et élimination d'effets dans un tableau faunistique. Acta Œcologica, Œcologia Generalis : 10, 3, 207-232.

[7] Beffy, J.L. & Dolédec, S. (1991) Mise en évidence d'une typologie spatiale dans le cas d'un fort effet temporel : un exemple en hydrobiologie. Bulletin d'Ecologie : 22, 3-11.

[8] Thioulouse, J. & Chessel, D. (1987) Les analyses multi-tableaux en écologie factorielle. I De la typologie d'état à la typologie de fonctionnement par l'analyse triadique. *Acta Œcologica, Œcologia Generalis* : 8, 4, 463-480.

[9] L'Hermier des Plantes, H. (1976) Structuration des tableaux à trois indices de la statistique. Théorie et applications d'une méthode d'analyse conjointe. Thèse de 3° cycle, USTL, Montpellier.

[10] Escoufier, Y. (1982) L'analyse des tableaux de contingence simples et multiples. Metron : 40, 53-77.