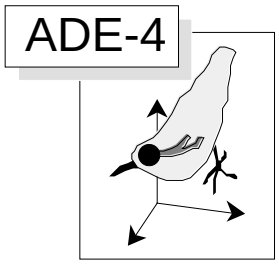


Principal components analysis

Abstract

This volume describes the computation and usual graphical display of a normalised principal components analysis (PCA) processed on physical and chemical data (Carrel et al., 1986, Approche graphique de l'analyse en composantes principales normée : utilisation en hydrobiologie. *Acta Œcologica, Œcologia Generalis* : 7, 189-203) Alternative computation and graphical representations (canonical graphs, cartography of scores, data reconstitution) also are presented.

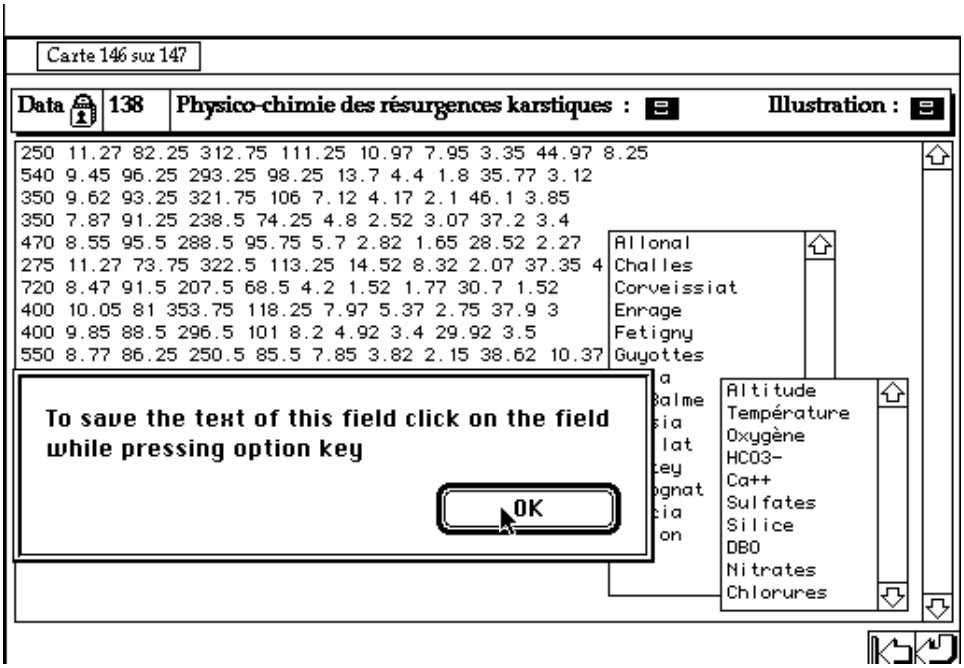
Contents

Jean-Michel Olivier

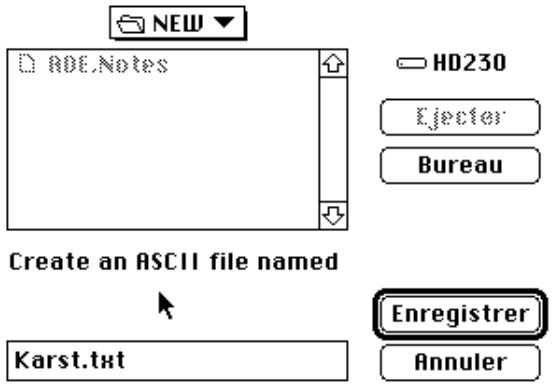
1 - Data input

The analyses and graphics made from the data of Carrel [1] will be done in the same order as in the paper. First of all we consider the data that describe the physical and chemical characteristics of 14 karstic springs (Barthélémy, 1984)[2].

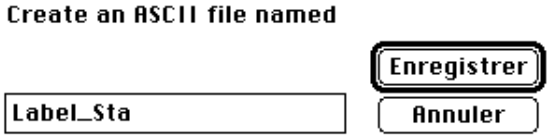
Create a data folder and go to the **ADE-4•Data** selection card. Select «Karst» in the right-hand data menu. A card shows up as follows:

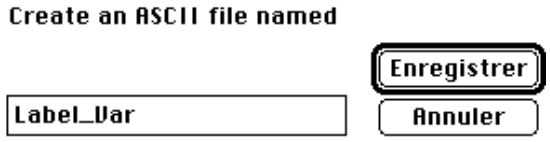


Copy the data as indicated. A new ASCII file (`Karst.txt`) appears in the data folder.



Similarly, copy the two other files describing sites and variables as follows:



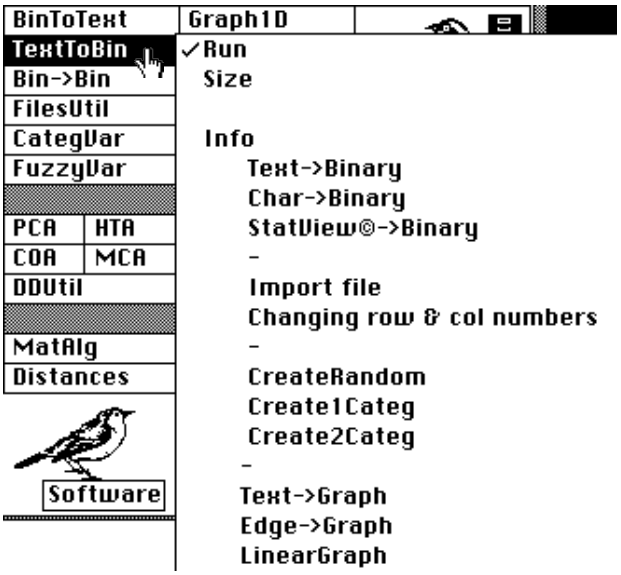


Two new ASCII files are created in the data folder: `Label_Sta` (14 rows, 1 character string) and `Label_Var` (10 rows, 1 character string).

You can control the content of a created file by clicking its name and selecting the **Open** option as follows:

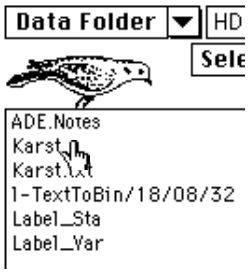
The BBEdit lite software is automatically opened and you can read the file `Karst.txt`.

After this verification, return back to the **ADE•Base** selection menu. Use the **TextToBin** module to transform the ASCII file `Karst.txt` into binary.



A dialog window shows up. Click the **Text input file** hand icon and select the file `Karst.txt` in your data folder. Give a name to the **Binary output file** as follows:

Click **OK** and quit the application with or without saving the listing. The binary file `Karst` will be further treated by PCA. You can verify its content directly by selecting the file in the **File list** of the **ADE•Base** selection card and opening it by double-click:



A listing shows up as follows (see **QuickStart** volume for more information).

```
*---------------------------------------
| ADE THINK C™ library * CNRS-Lyon *
| ADEBin: List BIN file
*---------------------------------------
---------------------------------------
Input file: Karst
Row: 14 Col: 10

   1 | 250.0000 | 11.2700 | 82.2500 |3
7.9500 |  3.3500 | 44.9700 |  8.2500 |
   2 | 540.0000 |  9.4500 | 96.2500 |2
4.4000 |  1.8000 | 35.7700 |  3.1200 |
```

To transform a binary file into an ASCII file you can use the **BinToText** module and the **Binary->Text** option. As an alternative, you can also use the **Edit with** command of the Data Folder menu. A `Karst-t` file is created (see **QuickStart** volume). This file can be open with a spreadsheet as Excel™. You can thereby improve the presentation of the data set as follows:

Karst.Excel

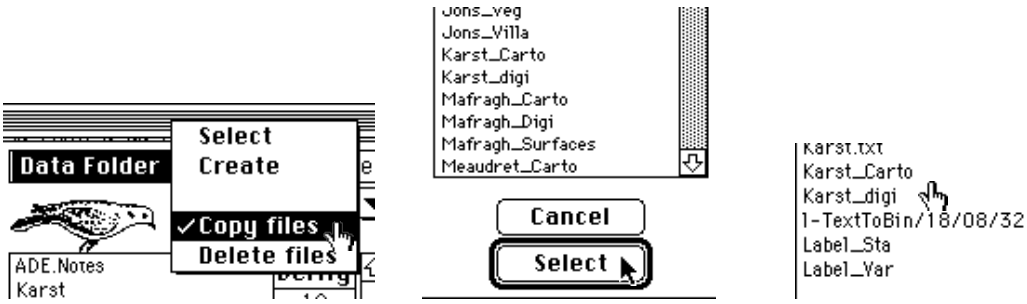
	A	B	C	D	E	F	G	H	I	J	K
1	Stations\Variables	Altitude	Température	Oxygène	HCO3-	Ca++	Sulfates	Silice	DBO	Nitrates	Chlorures
2	Allonal	250	11.27	82.25	312.75	111.25	10.97	7.95	3.35	44.97	8.25
3	Challes	540	9.45	96.25	293.25	98.25	13.7	4.4	1.8	35.77	3.12
4	Corveissiat	350	9.62	93.25	321.75	106	7.12	4.17	2.1	46.1	3.85
5	Enrage	350	7.87	91.25	238.5	74.25	4.8	2.52	3.07	37.2	3.4
6	Fetigny	470	8.55	95.5	288.5	95.75	5.7	2.82	1.65	28.52	2.27
7	Guyottes	275	11.27	73.75	322.5	113.25	14.52	8.32	2.07	37.35	4.47
8	Heria	720	8.47	91.5	207.5	68.5	4.2	1.52	1.77	30.7	1.52
9	La_Balme	400	10.05	81	353.75	118.25	7.97	5.37	2.75	37.9	3
10	Loisia	400	9.85	88.5	296.5	101	8.2	4.92	3.4	29.92	3.5
11	Maillat	550	8.77	86.25	250.5	85.5	7.85	3.82	2.15	38.62	10.37
12	Nantey	400	10	80	350	116.75	8.57	5.27	3.9	22.25	2.65
13	Samognat	480	8.92	96.25	290	92.75	6.25	2.67	1.82	31.1	2.07
14	Tarcia	420	10.4	85.5	313.25	108.5	7.35	4.6	1.8	23.2	2.37
15	Verjon	230	11.07	52.5	320.75	108.5	12.95	6	1.77	33.62	3.25
16											
17											
18											

Reminder elementary operations to begin with

- Create an ASCII file from a data card field of **ADE•Data**. Click the field while pressing the option key,
- Transform this ASCII file into binary using **Text->Binary** option of **TextToBin** module,
- Open any file from the Data Folder menu of the **ADE•Base** selection card.

2 - Definition of the geographical space

Select **Copy files** from the Data Folder menu. The entire list of the files available in the ADE/Files folder shows up. Copy the files `Karst_Carto` and `Karst_Digi`. These two files are duplicated into your data folder and show up in the file list of the **ADE•Base** selection card.



You can open these two files to control their contents. These two files have a PICT format. Do not change their format. These graphics must remain in PICT format to be used by ADE modules. The final graphics derived from ADE modules can be transformed into another format (MacDraw™, MacDrawPro™, ClarisDraw™, SuperPaint™, etc.) to improve the presentation (add labels, change fonts, etc.).

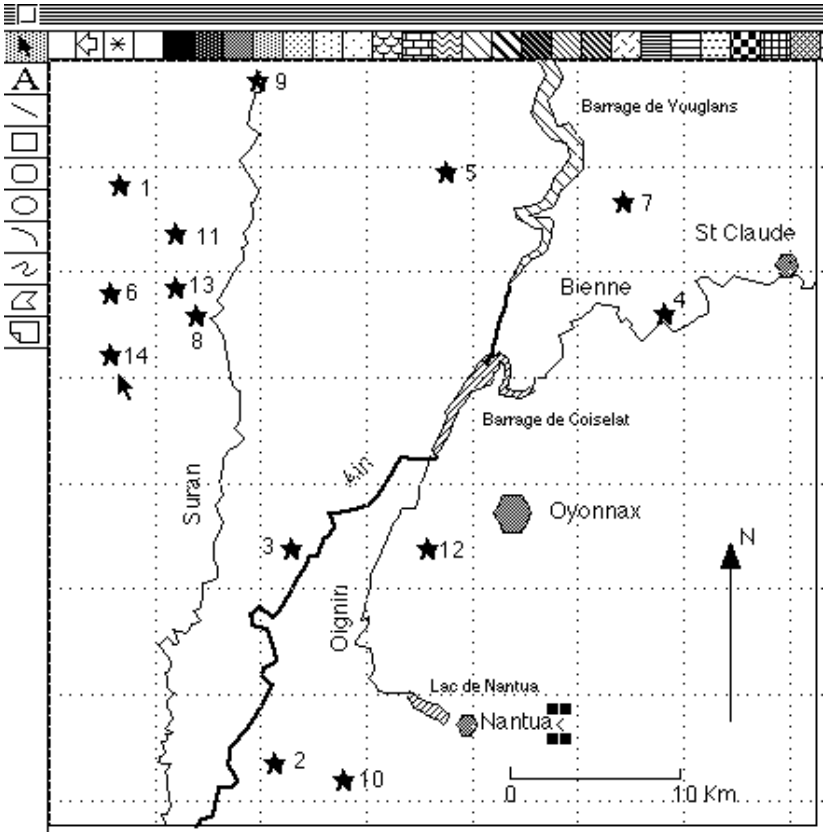
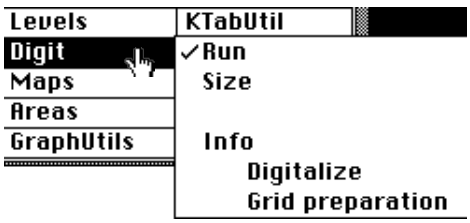
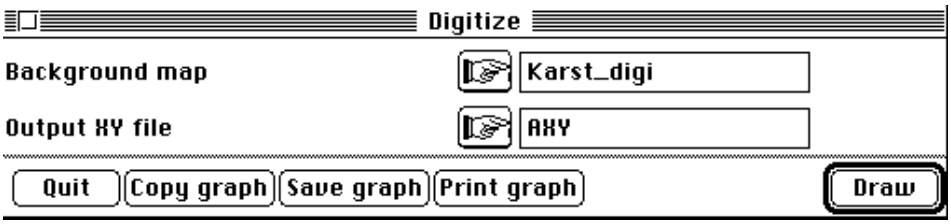


Figure 1 Content of file `Karst_digi`*. Stars drawn on the geographical map indicate the location of the sampling sites.*

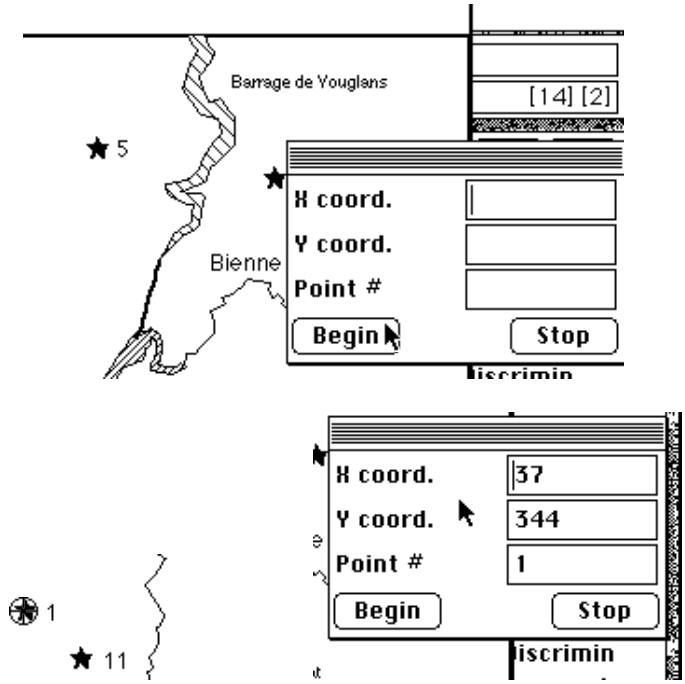
Use the **Digit** module to create the file that will contain the spatial coordinates of each site as follows:





Choose the file `Karst_Digi` as background map and type in a name into the **Output XY file** box (`AXY` in our example).

Click **Draw** to get the following screen:



Click **Begin**. Using the new cursor click each star successively from site 1 towards site 14. After the last site, click **Stop** and quit the application.

You can read the file `AXY` from the **Data Folder** menu (**ADE•Base** selection card).

```
Input file: AXY
Row: 14 Col: 2

 1 |  37.0000 |344.0000 |
 2 | 120.0000 | 34.0000 |
 3 | 129.0000 |149.0000 |
 4 | 330.0000 |275.0000 |
.............................etc.
 9 | 111.0000 |401.0000 |
10 | 157.0000 | 25.0000 |
11 |  68.0000 |318.0000 |
12 | 202.0000 |149.0000 |
13 |  67.0000 |289.0000 |
14 |  32.0000 |252.0000 |
```

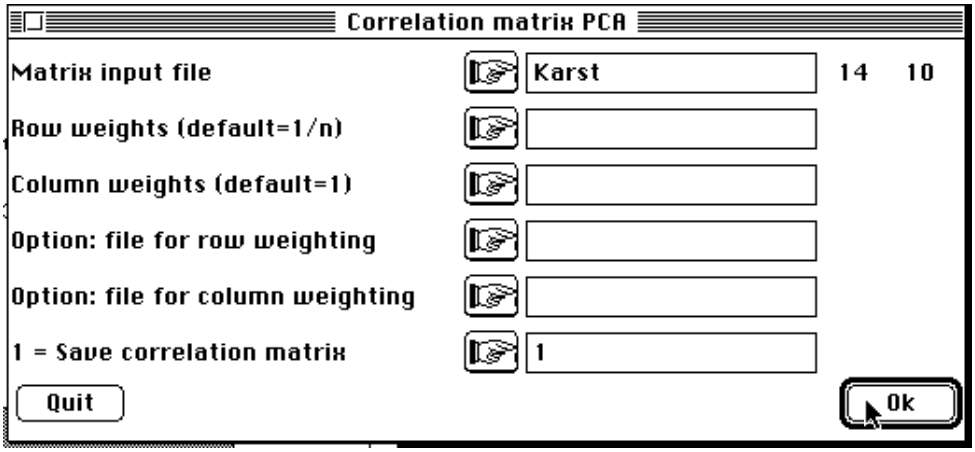
The output file `AXY` has 14 rows (sites) and 2 columns (coordinates X and Y of each site).

3 - Computation

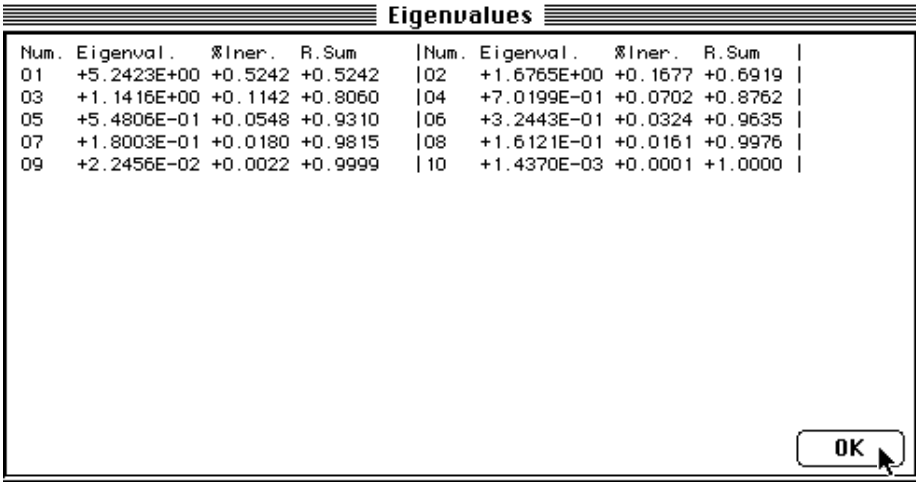
Because data are quantitative and variables are not expressed in the same units, normalised principal components analysis will be used to analyze this data set. The input table is `Karst` (binary). Choose the **PCA** menu of the **ADE•Base** selection card and run the module. Select `Karst` to fill in the **Matrix input file** box as follows:

By default, row weights are uniform (1/n with n = number of rows, here n = 14) and the column weights are unitary (1). Note that you can import other values as row and column weights.

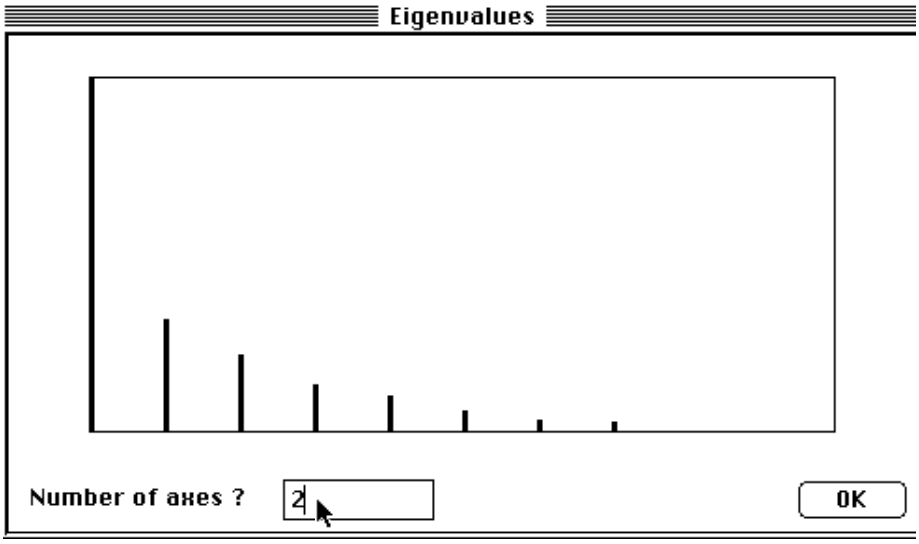
One can save the correlation matrix, which is the diagonalized matrix in PCA, by typing 1 in the **Save correlation matrix** box:



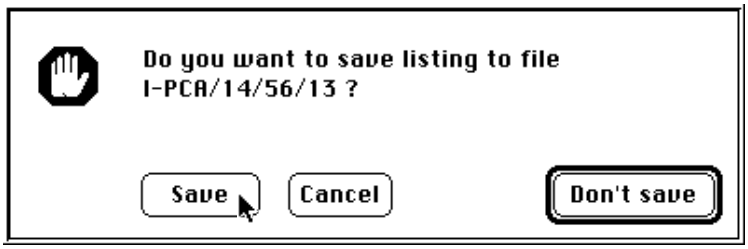
The eigenvalues, the inertia ratio of each axis, and the cumulated inertia shows up as follows:



Click **OK** to get the eigenvalues graph as follows:



This graph helps to choose the number of axes to be stored into the output files. Type in 2 and click **OK**. When you quit the **PCA** program, a dialog window shows up as follows:



Save the listing to get information about the results of this computation as follows:

```
*-----------------------------------------------------------*
| ADE THINK C™ library * CNRS-Lyon *            JT/DC/MH  |
| PCA: Correlation matrix PCA                               |
*-----------------------------------------------------------*
Classical Principal Component Analysis (Hotteling 1933)
Input file: Karst
---- Row weights:
File Karst.cnpl contains the row weights
It has 14 rows and 1 column
Each row has 0.0714286 weight (Sum = 1)
---- Column weights:
File Karst.cnpc contains the column weights
It has 14 rows and 1 column
Each column has unit weight (Sum = 10)
---- Table:
File Karst.cnta contains the centred and normed table
Zero mean and unit variance for each column
It has 14 rows and 10 columns
File :Karst.cnta ----------- Minimum/Maximum -----------
Col.:  1 Mini = -1.47575 Maxi =  2.39562
Col.:  2 Mini = -1.73674 Maxi =   1.5205
Col.:  3 Mini = -2.92485 Maxi = 0.980263
Col.:  4 Mini =  -2.2554 Maxi =  1.42569
Col.:  5 Mini = -2.13167 Maxi =   1.2465
Col.:  6 Mini = -1.39477 Maxi =  1.88992
Col.:  7 Mini = -1.63273 Maxi =  1.97618
Col.:  8 Mini = -1.01228 Maxi =  2.08352
Col.:  9 Mini = -1.73907 Maxi =  1.76488
```

```
Col.:  10   Mini = -0.987092   Maxi =  2.74045
---- Info: means and variances
File Karst.cnma contains the descriptive of the analysis
It contains successively:
     Number of rows: 14
     Number of columns: 10
     means and variances:
     Col.:  1 Mean:   416.786    Variance:     16020
     Col.:  2 Mean:   9.68286    Variance:   1.08958
     Col.:  3 Mean:   85.2679    Variance:   125.513
     Col.:  4 Mean:   297.107    Variance:   1578.48
     Col.:  5 Mean:   99.8929    Variance:   216.881
     Col.:  6 Mean:   8.58214    Variance:   9.87122
     Col.:  7 Mean:   4.59643    Variance:   3.55031
     Col.:  8 Mean:   2.38571    Variance:  0.528225
     Col.:  9 Mean:   34.0871    Variance:   46.3299
     Col.:  10   Mean:   3.86357    Variance:   5.63691
-----------------------------------------------------
File Karst.cn+r contains the Correlation matrix
from statistical triplet Karst.cnta
It has 10 rows and 10 columns
---------------------- Correlation matrix ------------------
[  1] 1000
[  2] -682 1000
[  3]  607 -693 1000
[  4] -638  722 -417 1000
[  5] -651  825 -470  979 1000
[  6] -501  760 -566  498  583 1000
[  7] -745  926 -616  680  787  806 1000
[  8] -344  158  -69  284  275  -32  319 1000
[  9] -309  128    1  -19    8  234  291   -2 1000
[ 10] -183  182 -126  -79   32  251  384  206  564 1000
-----------------------------------------------------------------------
-----------------------------------------
DiagoRC: General program for two diagonal inner product analysis
Input file: Karst.cnta
--- Number of rows: 14, columns: 10
-----------------------
Total inertia:       10
-----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +5.2423E+00 +0.5242 +0.5242   |02   +1.6765E+00 +0.1677 +0.6919 |
03   +1.1416E+00 +0.1142 +0.8060   |04   +7.0199E-01 +0.0702 +0.8762 |
05   +5.4806E-01 +0.0548 +0.9310   |06   +3.2443E-01 +0.0324 +0.9635 |
07   +1.8003E-01 +0.0180 +0.9815   |08   +1.6121E-01 +0.0161 +0.9976 |
09   +2.2456E-02 +0.0022 +0.9999   |10   +1.4370E-03 +0.0001 +1.0000 |
File Karst.cnvp contains the eigenvalues and relative inertia for each
axis
--- It has 10 rows and 2 columns
File Karst.cnco contains the column scores
--- It has 10 rows and 2 columns
File :Karst.cnco ----------- Minimum/Maximum -----------
Col.:  1 Mini = -0.823093   Maxi = 0.957283
Col.:  2 Mini = -0.383589   Maxi = 0.838496

File Karst.cnli contains the row scores
--- It has 14 rows and 2 columns
File :Karst.cnli ----------- Minimum/Maximum -----------
Col.:  1 Mini = -4.57004  Maxi =  3.41192
Col.:  2 Mini = -2.04113  Maxi =  2.70311
-----------------------------------------------------
```

The file `Karst.cncl` contains the row contributions to the trace; it is a 14 rows and 1 column file. You can then control the contents of the three files `Karst.cnpl`, `Karst.cnpc` and `Karst.cnta` with **ADEBin** module as follows:

```
Input file: Karst.cnpl
Row: 14 Col: 1

 1 |   0.0714 |
.............................................etc.
14 |   0.0714 |

Input file: Karst.cnpc
Row: 10 Col: 1

 1 |   1.0000 |
.............................................etc.
10 |   1.0000 |

Input file: Karst.cnta
Row: 14 Col: 10

   1 |  -1.3177 |  1.5205 | -0.2694 |  0.3937 |  0.7712 |  0.7600 |
1.7798 |  1.3268 |  1.5989 |  1.8475 |
   2 |   0.9735 | -0.2231 |  0.9803 | -0.0971 | -0.1116 |  1.6289 |
-0.1042 | -0.8059 |  0.2472 | -0.3132 |
   3 |  -0.5277 | -0.0602 |  0.7125 |  0.6203 |  0.4147 | -0.4654 |
-0.2263 | -0.3931 |  1.7649 | -0.0057 |
.............................................etc.
  13 |   0.0254 |  0.6870 |  0.0207 |  0.4063 |  0.5845 | -0.3922 |
0.0019 | -0.8059 | -1.5995 | -0.6291 |
  14 |  -1.4757 |  1.3289 | -2.9248 |  0.5951 |  0.5845 |  1.3902 |
0.7449 | -0.8472 | -0.0686 | -0.2584 |
```

The file `Karst.cnta` has the same the number of rows and columns than the input file `Karst`. Each value in the normalised table are equal to

$$x_{ij} = \frac{z_{ij} - \bar{z}_j}{s_j}$$

where z_{ij} is the original value, $\bar{z}_j$ is the mean for the jth variable and s_j its standard deviation. Consequently, the columns of this file have a mean equal to 0 and a variance equal to 1.

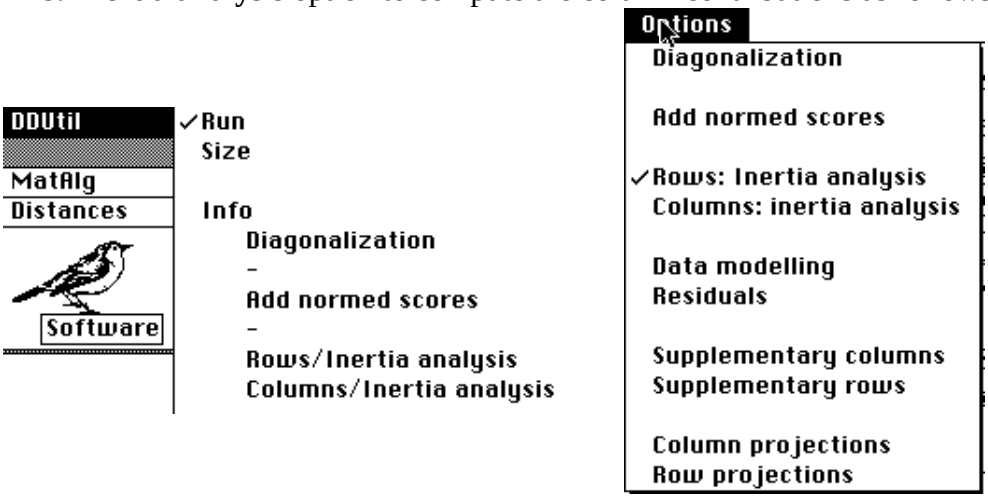
The correlation matrix allows to verify the consistency of the data. For instance, water temperature (variable n°2) is inversely related to the altitude (variable n°1) (correlation coefficient = -0.682) and concentration of dissolved oxygen (variable n°3) increases as temperature decreases (correlation coefficient = -0.693).

Select `Karst.cnvp` in the file list of the **ADE•Base** selection menu and open it to get the following listing:

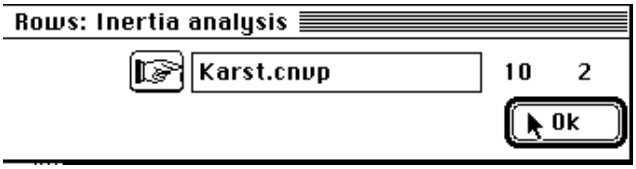
```
Input file: Karst.cnvp
Row: 10 Col: 2

   1 |   5.2423 |  0.5242 |
   2 |   1.6765 |  0.1677 |
   3 |   1.1416 |  0.1142 |
.............................................etc.
   9 |   0.0225 |  0.0022 |
  10 |   0.0014 |  0.0001 |
```

The first column of this file contains the eigenvalues of the analysis that represent the variance of scores associated to each axis (rows). The second column contains the inertia ratio for each axis. One can verify that ¬1/total inertia = 5.2423/10 = 0.542. Note that because of the unitary variance of variables, the total inertia is equal to 10.

Inertia analysis allows to add statistical aids in the interpretation of data. Use **DDUtil** and choose the **Rows: Inertia analysis** option to compute the row contributions and the **Columns: Inertia analysis** option to compute the column contributions as follows:



Select the file `Karst.cnvp` by clicking the hand icon as follows:



The listing of row inertia analysis give the following information:

```
Input file: Karst.cnta
Number of rows: 14, columns: 10
Inertia: Two diagonal norm inertia analysis
Total inertia:       10 - Number of axes: 2

File Karst.cncl contains the contribution of rows to the trace
It has 14 rows and 1 column
Row inertia
All contributions are in 1/10000

----------Absolute contributions----------
|Num  |Fac 1|Fac 2|
|    1| 1371| 2034|
|    2|   24|    8|
|    3|    5|  296|
|    4|  925|  401|
|    5|  580|  380|
|    6| 1586|   52|
|    7| 2845|    8|
|    8|  299|  147|
|    9|    4|   81|
|   10|  293| 3113|
|   11|  275| 1775|
|   12|  516|  226|
|   13|    3| 1356|
```

```
|   14| 1267|  117|

----------Relative contributions----------
|Num  |Fac 1|Fac 2||Remains| Weight | Cont.|
|    1| 6155| 2921||  922  |  714   | 1167 |
|    2|  331|   38|| 9630  |  714   |  389 |
|    3|   79| 1425|| 8495  |  714   |  348 |
|    4| 5397|  748|| 3853  |  714   |  898 |
|    5| 6887| 1444|| 1667  |  714   |  442 |
|    6| 8426|   89|| 1484  |  714   |  986 |
|    7| 8857|    8|| 1134  |  714   | 1684 |
|    8| 4600|  726|| 4672  |  714   |  341 |
|    9|  134|  757|| 9108  |  714   |  180 |
|   10| 1722| 5848|| 2429  |  714   |  892 |
|   11| 1811| 3730|| 4457  |  714   |  797 |
|   12| 7622| 1066|| 1311  |  714   |  355 |
|   13|   54| 6721|| 3223  |  714   |  338 |
|   14| 5646|  166|| 4186  |  714   | 1176 |
-------------------------------------------------
```

The listing of row inertia analysis give the following information:

```
Input file: Karst.cnta
Number of rows: 14, columns: 10
Inertia: Two diagonal norm inertia analysis
Total inertia: 10 - Number of axes: 2

File Karst.cncc contains the contribution of columns to the trace
It has 10 rows and 1 column

Column inertia
All contributions are in 1/10000

----------Absolute contributions----------
|Num  |Fac 1|Fac 2|
|    1| 1292|   26|
|    2| 1673|   34|
|    3|  947|   37|
|    4| 1269|  877|
|    5| 1484|  556|
|    6| 1172|   81|
|    7| 1748|  124|
|    8|  176|    9|
|    9|  101| 4059|
|   10|  135| 4193|

----------Relative contributions----------
|Num  |Fac 1|Fac 2||Remains| Weight | Cont.|
|    1| 6774|   43|| 3181  |10000   |  999 |
|    2| 8773|   57|| 1168  |10000   | 1000 |
|    3| 4968|   63|| 4968  |10000   | 1000 |
|    4| 6653| 1471|| 1875  |10000   | 1000 |
|    5| 7781|  932|| 1286  |10000   | 1000 |
|    6| 6145|  137|| 3717  |10000   | 1000 |
|    7| 9163|  208||  627  |10000   | 1000 |
|    8|  923|   15|| 9061  |10000   | 1000 |
|    9|  529| 6805|| 2665  |10000   | 1000 |
|   10|  709| 7030|| 2259  |10000   | 1000 |
-------------------------------------------------
```

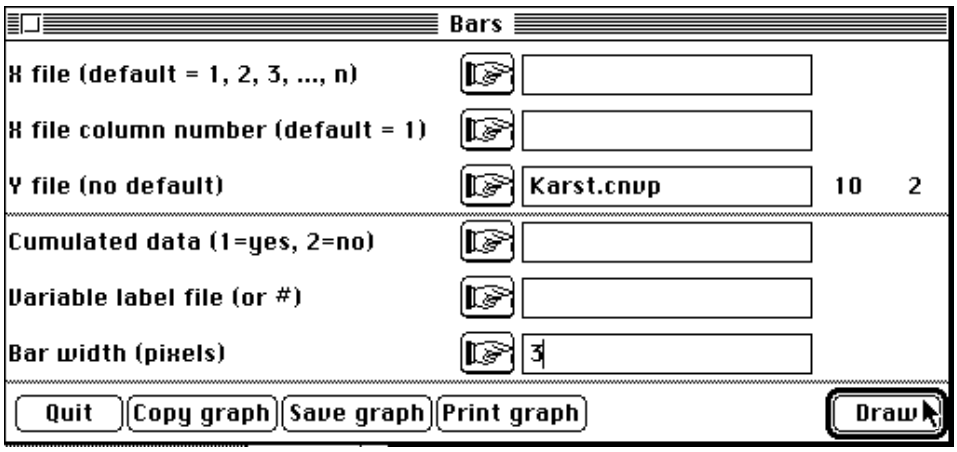
In these tables, the column "Num" identifies the rows (= 14 sites) or the columns (= 10 variables). The absolute contributions are the contributions of sites or variables to the definition of axes; for example, site 7 (Heria, underlined in the above listing)

contributes for 28.45% to the definition of the first axis and only for 0.08% to the definition of the second axis. The relative contributions indicate the contribution of each axis that explains the position of a row (site) or a column (variable). For instance, the first axis contributes for 88.57% to the position of site 7 (column "Fac 1"). Note that the eigenvalues ($\neg_1$, $\neg_2$) equal the sum of each column ("Fac 1" and "Fac 2"). The column "Remains" indicates the percentage of variability not taken into account by the two first axes (11.34% for site 7). The values presented in the columns "Weight" and "Cont." are those presented in the files `Karst.cnpl` and `Karst.cncl` respectively.

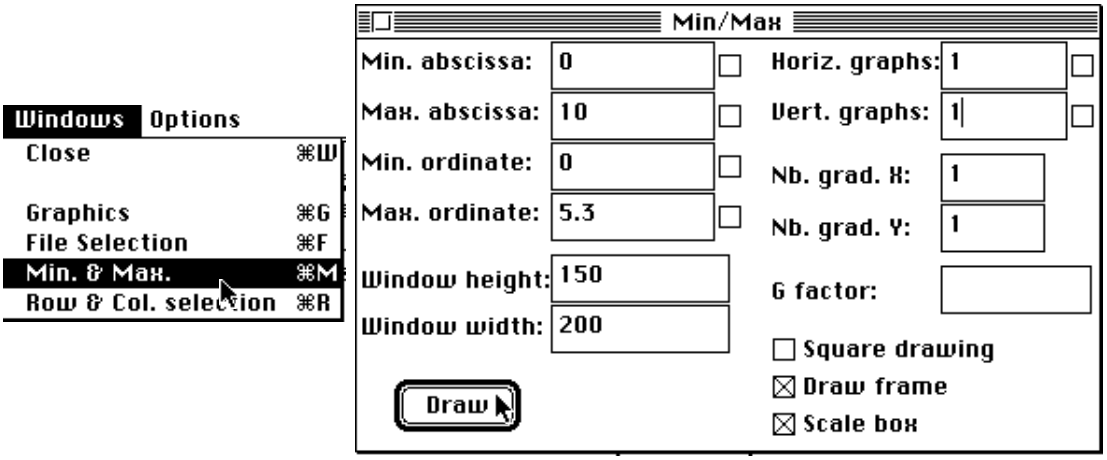
4 - Interpretation

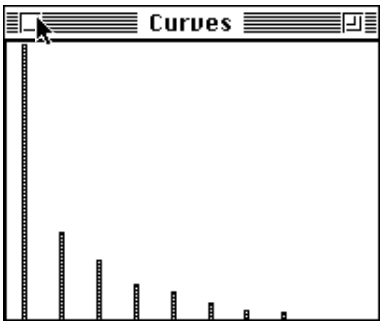
4.1 - Eigenvalues

Select **Curves** module to draw the eigenvalues diagram (first column of the file `Karst.cnvp`). Select the **Bars** option. Select `Karst.cnvp` as **Y file** and indicate the **Bar width** (optional) as follows:

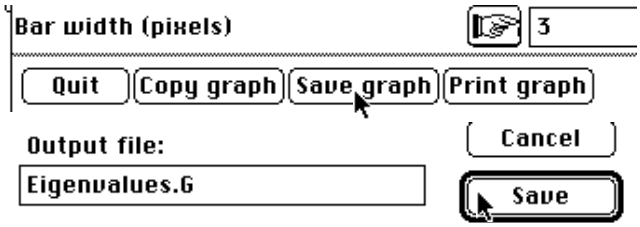


File `Karst.cnvp` has two columns and ten rows. To get the graphic concerning only the first column, select **Min & Max** in the **Windows** menu and change the dimensions of the window. Choose one graph per window:





To save this graphic click **Save graph** and give a name (e.g., **Eigenvalues.G**) as follows:

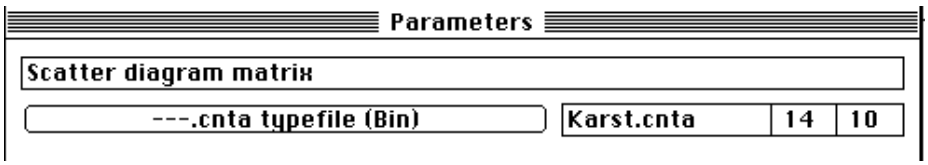


The illustration is stored as a PICT file, which can be opened with a drawing software (MacDraw™, ClarisDraw™, etc.).

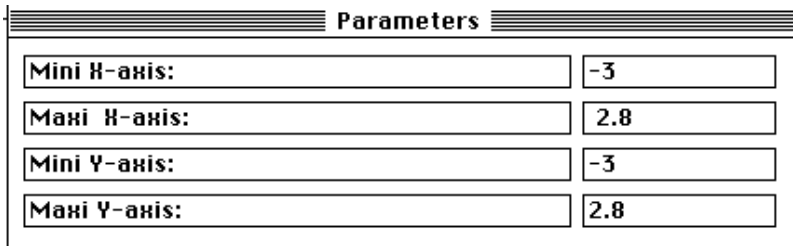
4.2 - Draftsman's display (See ADEScatters : Draftman's display)

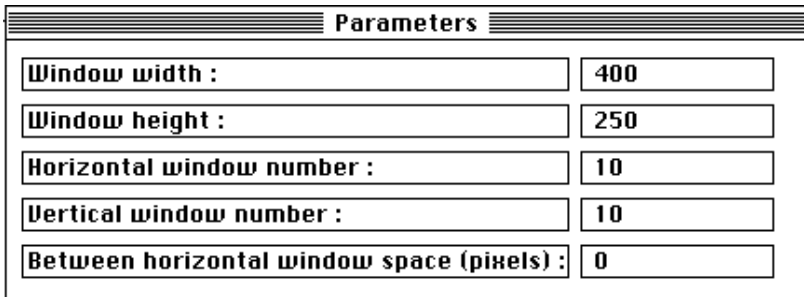
Such representation is used to show the correlation among variables. Go to the **ADE•Old** selection card and select **PCA/COA Graphical Display** as follows:

You may also use the **QuickBasic** button of the **ADE•Base** selection card for a quick selection of a Basic program. Select the **CorreMat** QuickBasic application and click button **B,** select the file `Karst.cnta` and click **OK** or return as follows:



For each dialog window you can change the input values:





This results in Fig. 2 which both indicates the values of correlation coefficients and give the linear regressions of variables.

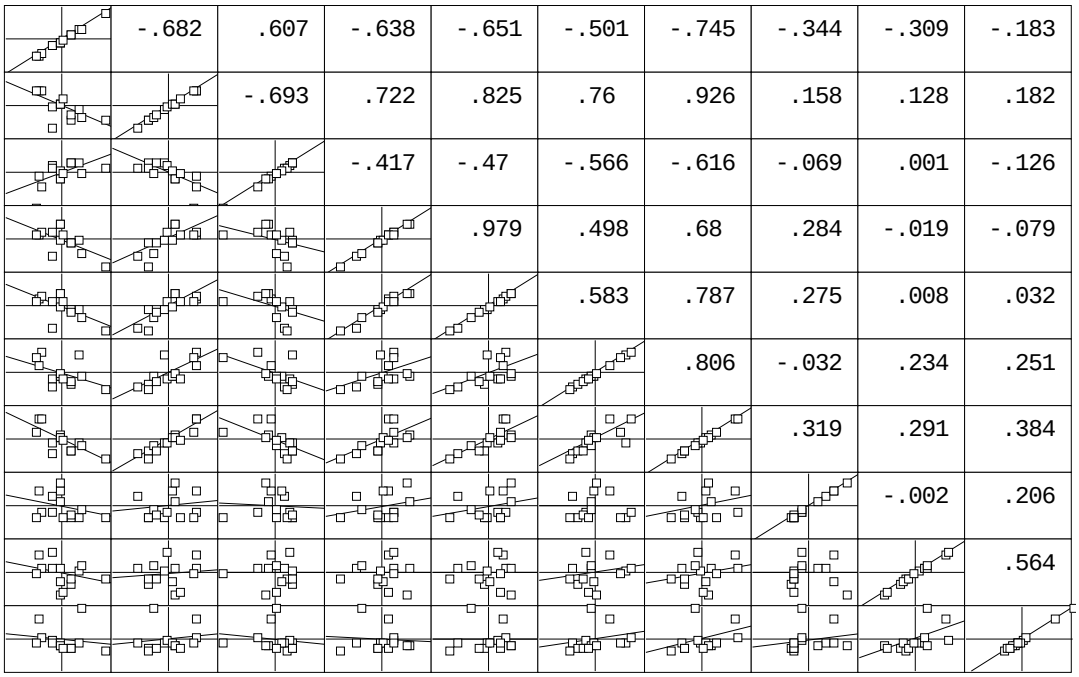


Figure 2 Draftsman's diagram of 10 variables measured in 14 karstic springs. Lines shown in the left graphs is from the linear regression of the vertical variable on the horizontal variable.

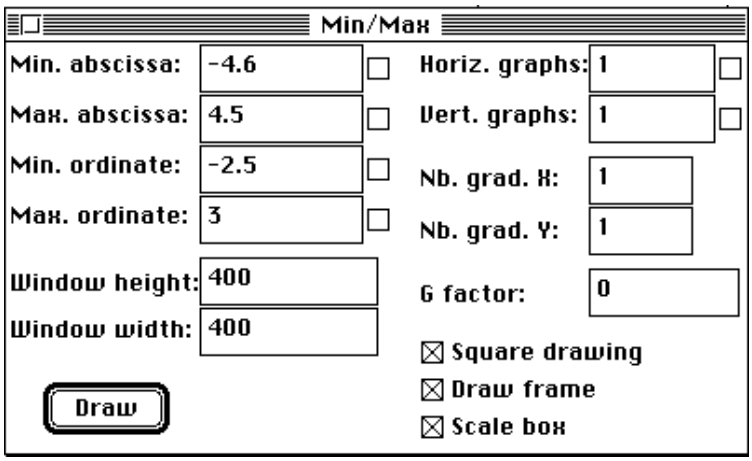
The correlation matrix (`Karst.cndi` file) is stored as indicated on the listing of the **CorreMat** application:

```
File  Karst.cndi contains correlation matrix
It is a  10  row and  10  column file
```

4.3 - Factorial map

Use the **Scatters** module with the **Labels** option to draw the factorial map of sites (Fig. 3). Choose the file `Karst.cnli` (site scores) and the file `Label-Sta` as Labels and Click **Draw**.

Use the **Min. & Max.** option of the **Windows** menu to modify the limits of the graphics. This makes all labels be included into the drawing frame:



This results in a pict file that may be stored into your data folder (Fig. 3).

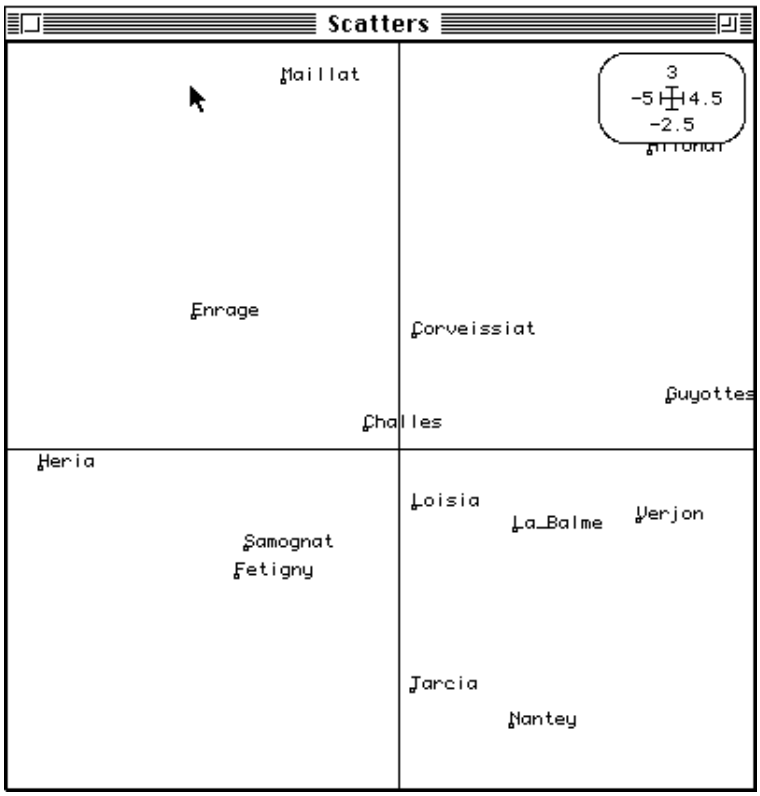
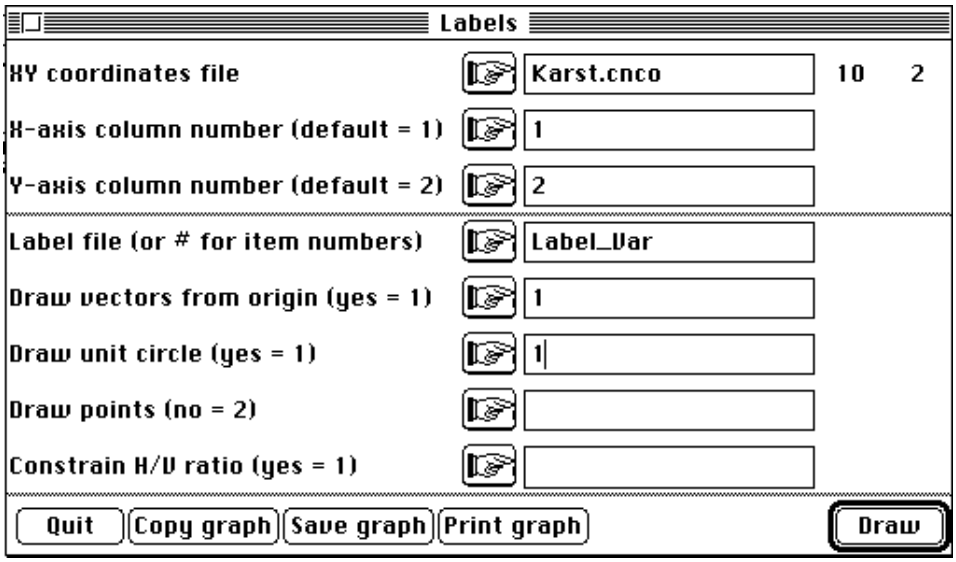


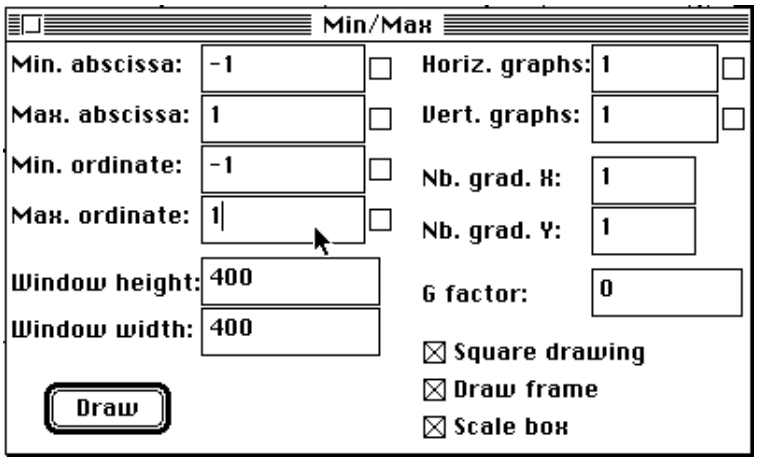
Figure 3 F1-F2 factorial plane of sites.

4.4 - Correlation circle

Use **Scatters** (**Labels** option) to draw the correlation circle (Fig. 4). Choose the file `Karst.cnco` as the **XY coordinates file** and `Label_Var` as **Label file** (Txt). Here we draw the correlation circle according to the first and second axes. Type in 1 (yes) into the **Draw unit circle** option.

Change the minimal and maximal values to give the minimal and maximal limits of the correlation coefficients for each axis (between 1 and -1).





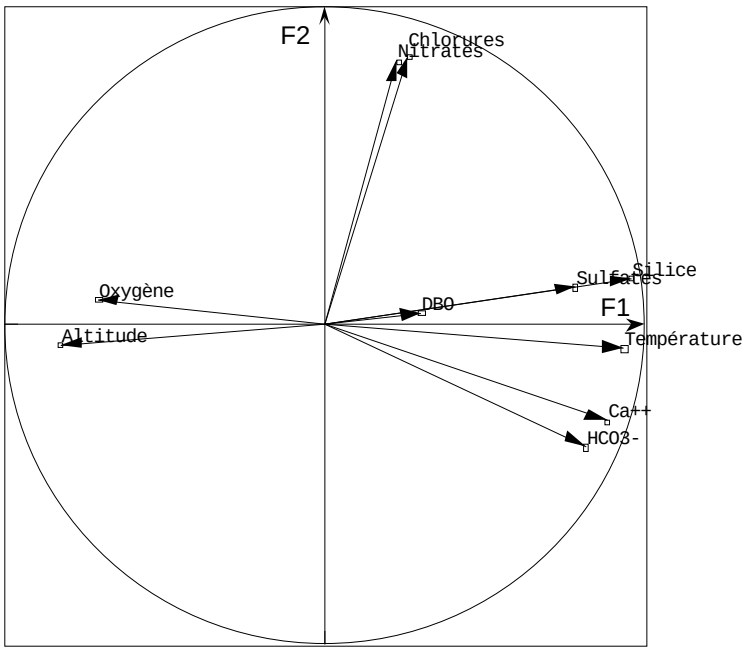


Figure 4 Correlation circle variables-axes (F1-F2 factorial plane).

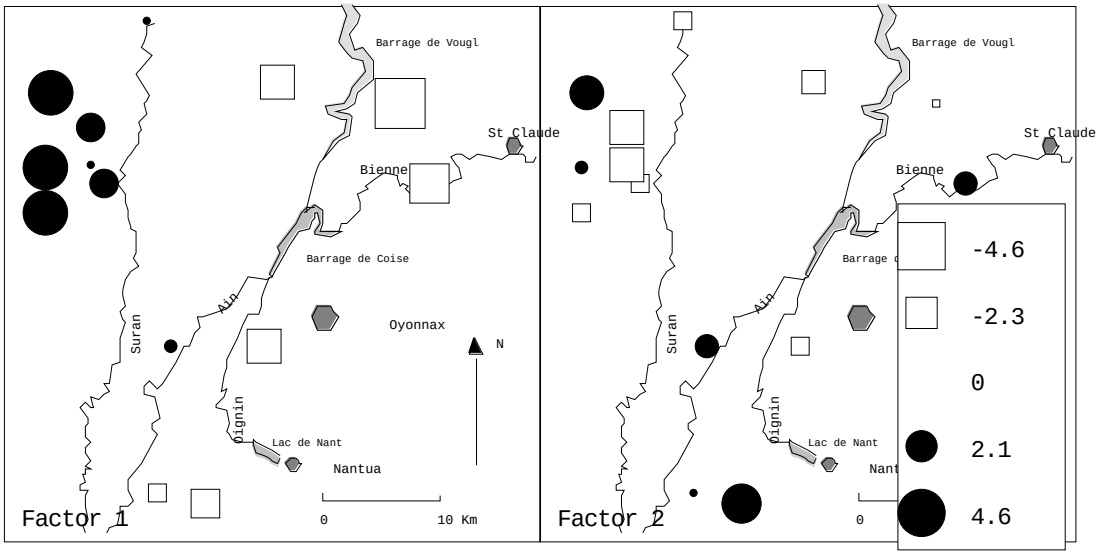


Figure 5 Site scores (first and second axis) plotted on the geographical map. Circles represent positive values and squares represent negative values.

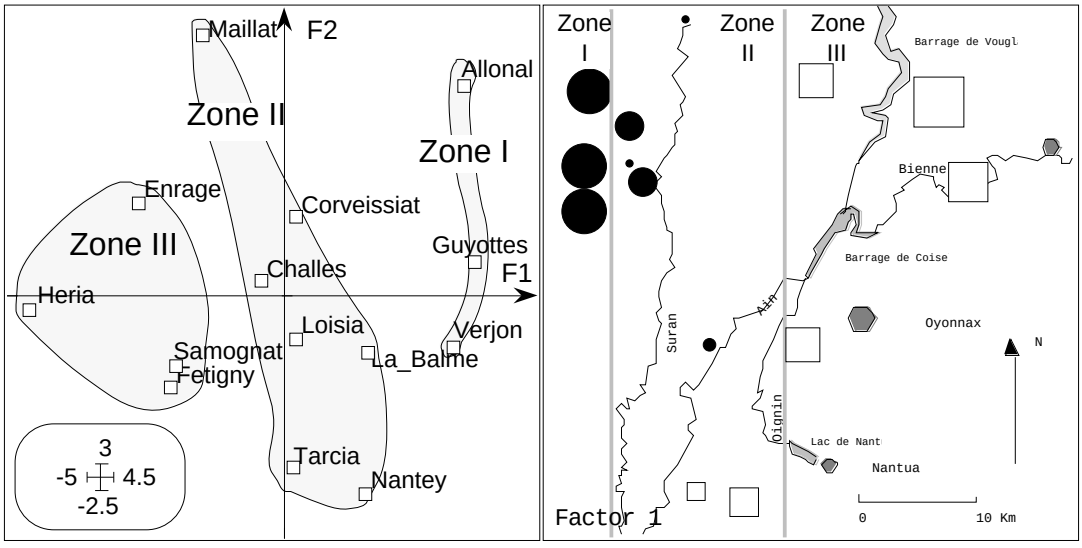


Figure 6 Interpretation of the F1-F2 factorial map. Three zones are identified according to the first factor. These zones are easily distinguishable on the geographical map where F1 scores are plotted (graphic on the right).

Two gradients can be identified from Fig. 3 and Fig. 4:

- the first axis describes the global mineralization, which takes into account the simultaneous variations of bicarbonates, calcium, sulphates and silica. Mineralization is related to the altitude as it is higher in the lowland sites than in the upland sites having steep slope and fast flow. Temperature and dissolved oxygen are also related to the altitude and consequently strongly associated to the first axis;

- the second axis underscores the effect of human activities. On that axis, the ordination of springs is strongly associated to nitrates and chlorides concentrations.

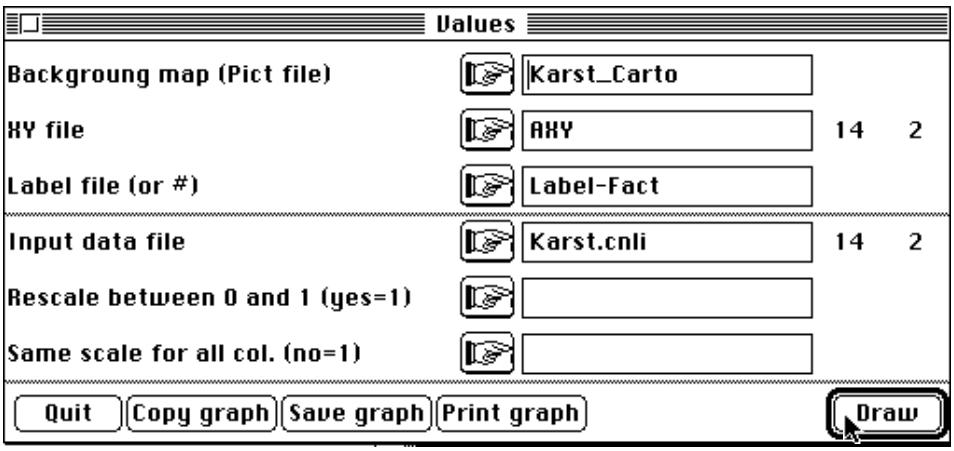
4.5 - Cartography of scores (See Maps : Values)

Scores result from the linear combination of the normalised values of variables under the constraint that the sum of the squared coefficients is equal to 1. They also represent numerical codes (to more or less one constant) maximising the sum of their squared correlation with variables. Because of their maximum variance, the factorial score can be used as any quantitative variable.

Use the `Maps` program with the **Values** option to plot the scores on the geographic map already prepared:

Use the file `Karst-carto` as **Background map (Pict file)** and `AXY` as **XY file**. You can create a label file (`Label_Fact`, ASCII file) that contains the labels of each graph (the axis number, i.e., "factor 1", "factor 2").

Use the file that contains the site scores (`Karst.cnli`) as **Input data file**. Do not use the **Rescale between 0 and 1** option if you want to plot the true scores:

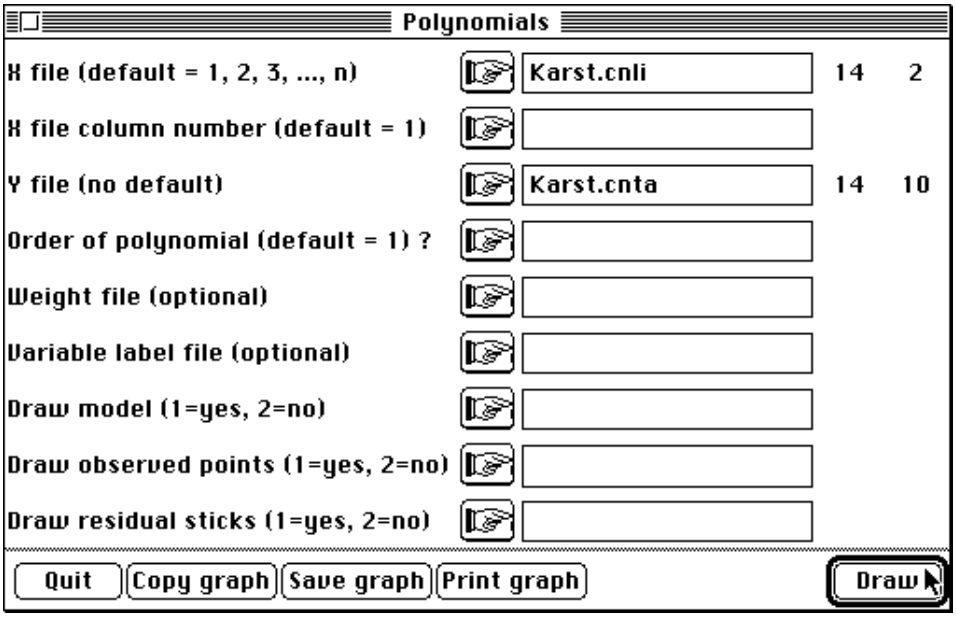


This results in Fig. 5. Save this graph as a PICT file. Using Fig. 3 and Fig. 5 results in a synthetic illustration depicting the East-West gradient of mineralization among springs (Fig. 6).

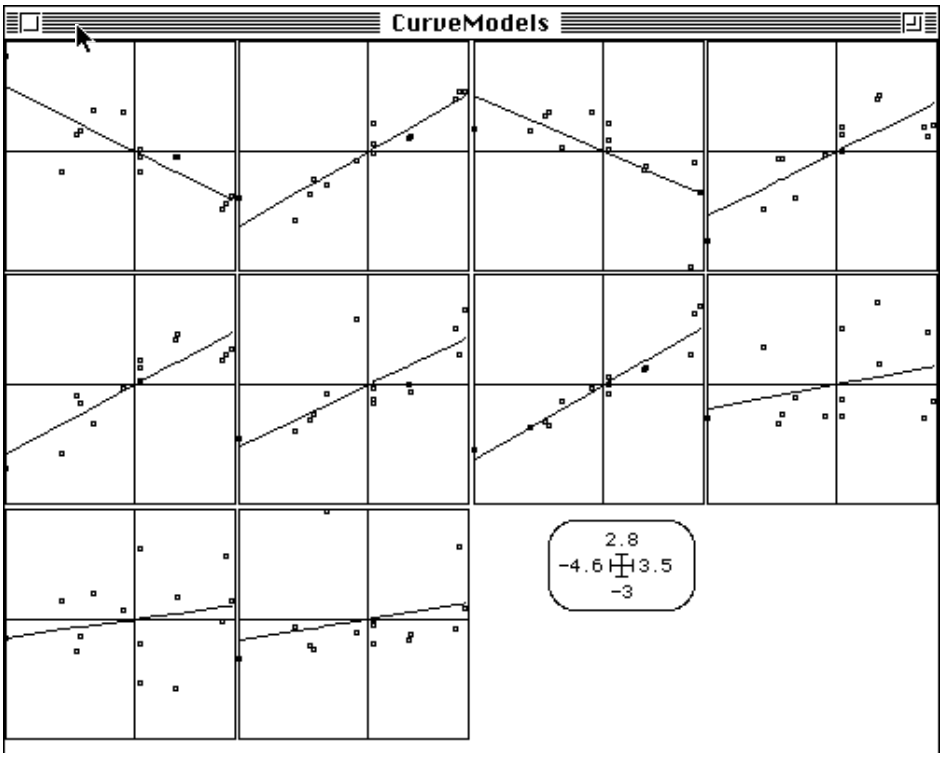
4.6 - Canonical graph

For each variable one can draw the linear regression of the normalised values on the site scores along the first axis. This ordination common to all graphs maximises the squared correlation and allows the simultaneous display of the relationships between variables and a synthetic variable (the first axis). Such a graph is called a canonical graph.

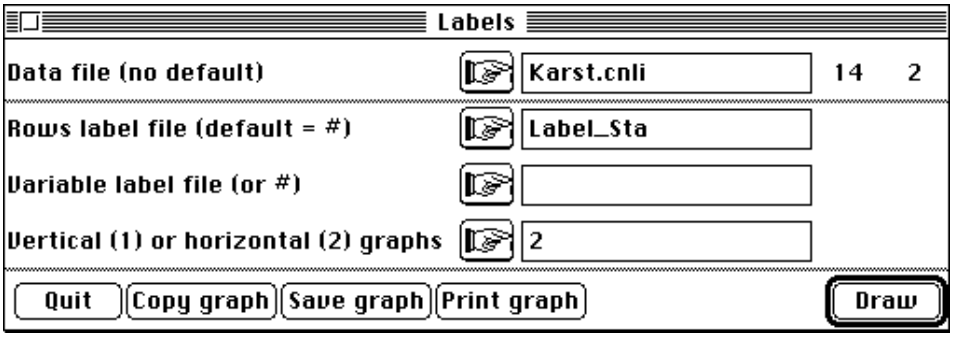
Select the **CurveModel** module and the **Polynomials** option to plot this graph. Type in the dialog window `Karst.cnli` as **X file** and `Karst.cnta` as **Y file**:

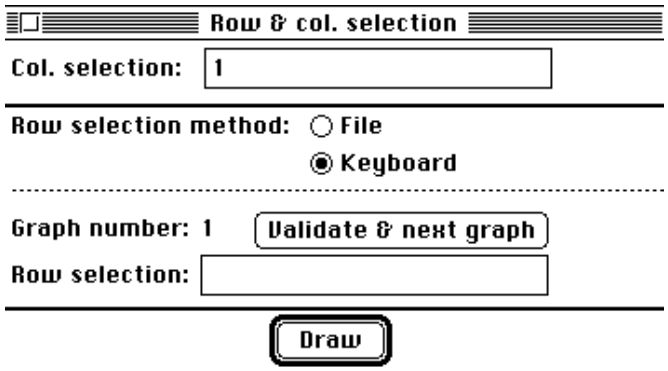


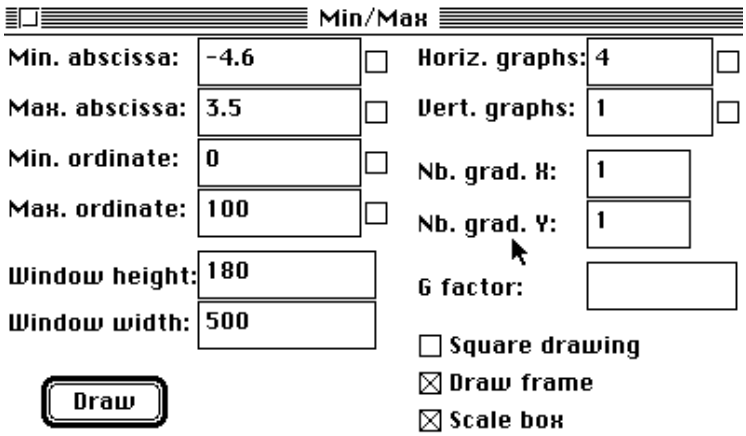
This results in the following graphics:



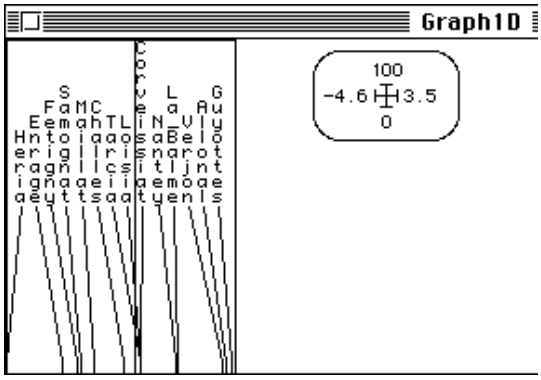
Use the **Graph1D** module with the **Labels** option to plot the ordination of sites along the first axis. Modify the dialog boxes as follows:







This results in the following graph:



The final graphic (Fig. 7) results from the rearrangement using a drawing software.

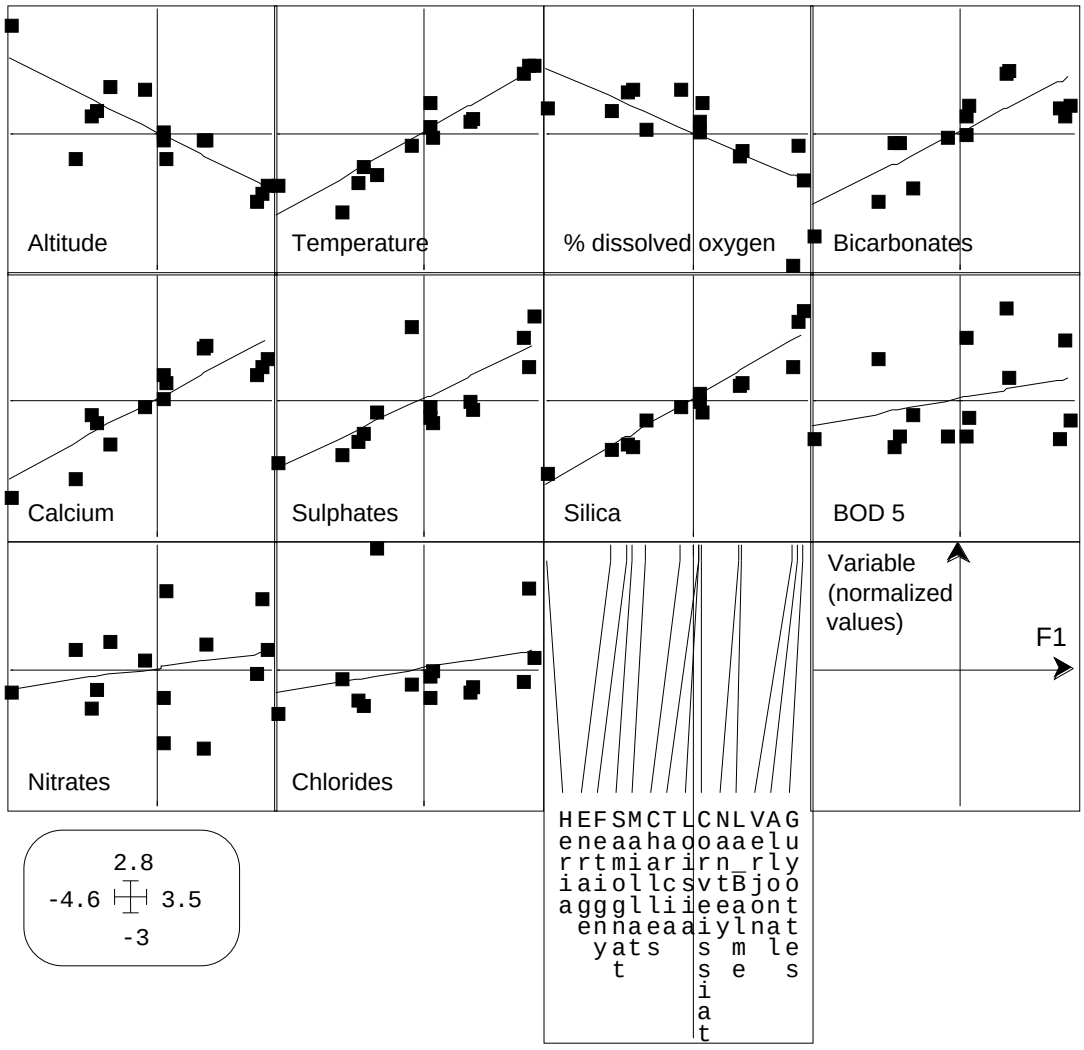


Figure 7 Canonical graphs.

These graphs allow the study of the relationships of each variable according to the mineralization described by the first axis. For example, water temperature and silica are closely related to the mineralization (squared correlation equal +0.94 and +0.96 respectively) and consequently appears as major variables along axis F1.

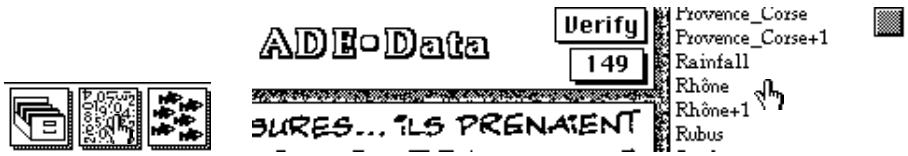
5 - Data reconstitution

5.1 - Objective

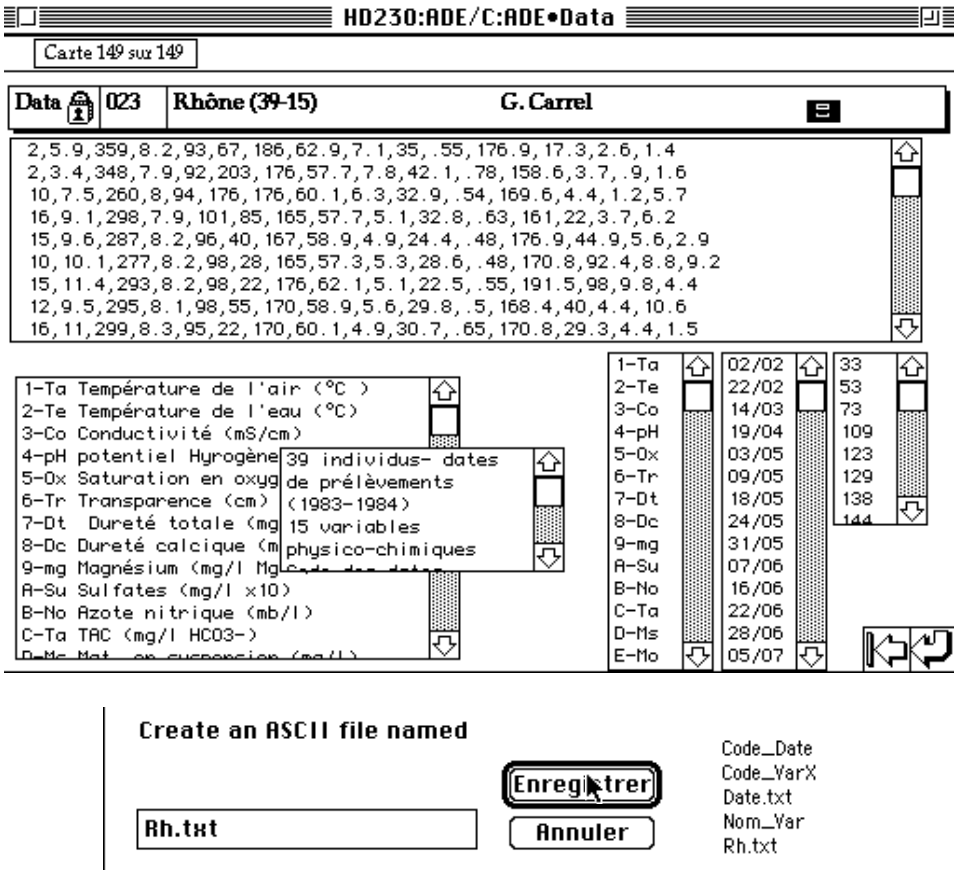
Data reconstitution is very useful when data are recorded chronologically (e.g. growth curves). The aim of this procedure is to superpose the data (here the normalised values) and a model using eigenvalues and row and column scores. It allows to visualise simultaneously the variations of axes and variables. In the following we describe an example where data were recorded at 39 dates in one site located on the Rhône River (France) over the period 1983-1984 (Carrel, 1986)[3]. Fifteen physical and chemical variables were taken into account.

5.2 - Data set

Go to the **ADE•Data** selection menu. Select «Rhône» in the example list:



Copy the data fields to your data folder:



Five files can be incorporated to your data folder:

(1) `Rh.txt` contains the data (39 rows = sampling dates and 15 columns = physico-chemical variables). You must transform `Rh.txt` into `Rh` (binary).

```
    1     2    3    4   5   6    7     8    9    10  11    12    13   14   15
1   2   5.9  359  8.2  93  67  186  62.9  7.1    35 .55 176.9  17.3  2.6  1.4
2   2   3.4  348  7.9  92 203  176  57.7  7.8  42.1 .78 158.6   3.7   .9  1.6
3  10   7.5  260    8  94 176  176  60.1  6.3  32.9 .54 169.6   4.4  1.2  5.7
4  16   9.1  298  7.9 101  85  165  57.7  5.1  32.8 .63   161    22  3.7  6.2
5  15   9.6  287  8.2  96  40  167  58.9  4.9  24.4 .48 176.9  44.9  5.6  2.9
6  10  10.1  277  8.2  98  28  165  57.3  5.3  28.6 .48 170.8  92.4  8.8  9.2
.....................................................................etc.
35 11   9.1  292  7.9 109  19  191  66.1  6.3  26.6 .54 151.3 258.3 16.5  5.6
36 -5   4.2  342  8.1  89 115  227  75.8  9.2    30 .78 174.5  10.3  1.7   .7
37  6   5.9  343  7.9  78 150  209  70.1  8.3  34.5 .84 152.5     9  1.1   .3
38  7   6.7  330  7.7  86  43  211  73.4  6.8  25.3  .6 173.2  29.5  2.2   .8
39  2   5.8  349  7.8  77 165  176  59.7  6.6  34.5 .45 170.8   5.9   .6   .7
```

(2) `Nom_Var` identifies the variables:

```
1-Ta Température de l'air (°C )     air temperature (°C )
2-Te Température de l'eau (°C)      water temperature (°C)
3-Co Conductivité (µS/cm)           conductivity (µS/cm)
```

```
4-pH potentiel Hydrogène (pH)       pH
5-Ox Saturation en oxygène (%)      percentage of oxygen saturation
6-Tr Transparence (cm)              water transparency (cm)
7-Dt Dureté totale (mg/l CaCO3)     total hardness (mg/l CaCO3)
8-Dc Dureté calcique (mg/l Ca++)    calcium hardness (mg/l Ca++)
9-mg Magnésium (mg/l Mg++)          magnesium (mg/l Mg++)
A-Su Sulfates (mg/l x10)            sulphates (mg/l x10)
B-No Azote nitrique (mg/l)          nitric nitrogen (mg/l)
C-Ta TAC (mg/l HCO3-)               total alkalinity (mg/l HCO3-)
D-Ms Mat. en suspension (mg/l)      suspended matter (mg/l)
E-Mo Mat. organique (mg/l)          organic matter (mg/l)
F-Ch Chlorophylle a (mg/l)          chlorophyll a
```

(3) `Code_VarX` (Txt, label file) contains the code of variables:

```
1-Ta
2-Te
3-Co
4-pH
5-Ox
6-Tr
7-Dt
8-Dc
9-mg
A-Su
B-No
C-Ta
D-Ms
E-Mo
F-Ch
```

(4) `Date.txt` (Txt) contains the sampling dates as the number of days since the 1st of January:

```
01-33  02-53  03-73  04-109 05-123
06-129 07-138 08-144 09-151 10-158
11-167 12-173 13-179 14-186 15-192
16-200 17-207 18-213 19-221 20-228
21-235 22-242 23-249 24-255 25-262
26-269 27-276 28-283 29-291 30-297
31-304 32-311 33-318 34-325 35-332
36-339 37-353 38-360 39-367
```

(5) `Code-Date` contains the sampling dates as day of sampling.

5.3 - Data analysis and interpretation

Compute a PCA (option **Correlation Matrix PCA**) on the table `Rh`:

```
Classical Principal Component Analysis (Hotteling 1933)
Input file: Rh
---- Row weights:
File Rh.cnpl contains the row weights
It has 39 rows and 1 column
Each row has 0.025641 weight (Sum = 1)
---- Column weights:
File Rh.cnpc contains the column weights
It has 39 rows and 1 column
Each column has unit weight (Sum = 15)
---- Table:
File Rh.cnta contains the centred and normed table
```

```
Zero mean and unit variance for each column
It has 39 rows and 15 columns
File Rh.cn+r contains the Correlation matrix
from statistical triplet Rh.cnta
It has 15 rows and 15 columns
---------------------- Correlation matrix -----------------
[  1] 1000
[  2]  884 1000
[  3] -856 -892 1000
[  4]  204  187 -201 1000
[  5]  346  192 -311  254 1000
[  6] -309 -108  169 -206 -505 1000
[  7] -831 -905  870 -159 -177   26 1000
[  8] -781 -895  844 -146  -87  -97  986 1000
[  9] -679 -524  587 -145 -531  609  590  447 1000
[ 10]  222  443 -255 -108 -219  707 -483 -571  170 1000
[ 11] -509 -585  553  -40 -178  229  554  511  497    6 1000
[ 12] -653 -860  758   -1   35 -250  814  867  165 -694   352 1000
[ 13]    0 -149   10   69  496 -591  207  274 -223 -588   -78  295
1000
[ 14]   79 -109  -42  196  386 -678  119  202 -346 -634 -110   326
883 1000
[ 15]  229   96 -283  173  448 -292 -127  -74 -328 -243 -252    92
316  286 1000
-------------------------------------------------------------------
----------------------------------------
DiagoRC: General program for two diagonal inner product analysis
Input file: Rh.cnta
--- Number of rows: 39, columns: 15
-----------------------
Total inertia:       15
-----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +6.2743E+00 +0.4183 +0.4183   |02   +4.1409E+00 +0.2761 +0.6943 |
03   +1.0082E+00 +0.0672 +0.7616   |04   +8.5920E-01 +0.0573 +0.8188 |
05   +7.6219E-01 +0.0508 +0.8697   |06   +6.6565E-01 +0.0444 +0.9140 |
.....................................................................etc.
13   +2.9567E-02 +0.0020 +0.9984   |14   +2.3569E-02 +0.0016 +1.0000 |
15   +0.0000E+00 +0.0000 +1.0000

File Rh.cnvp contains the eigenvalues and relative inertia for each
axis
--- It has 15 rows and 2 columns

File Rh.cnco contains the column scores
--- It has 15 rows and 6 columns

File Rh.cnli contains the row scores
--- It has 39 rows and 6 columns
```

The two first axes take into account 69% of the total inertia. An important question is: "How many axes should be stored?". From a statistical point of view and according to the aim of PCA (description of the correlation among variables) the answer is 2. From a biological point of view, a variable not correlated with the others may be of interest and should be taken into account. This is why we select 6 axes in this analysis (Fig. 8):

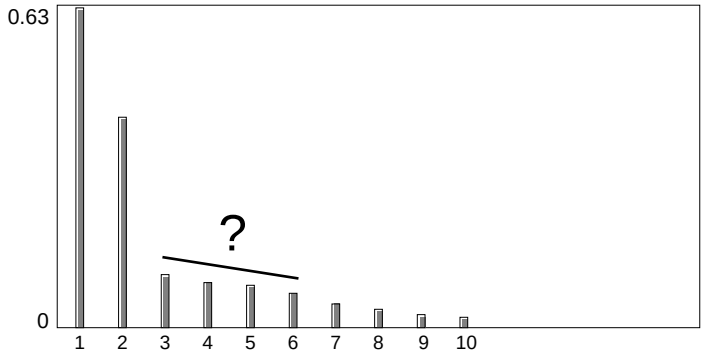
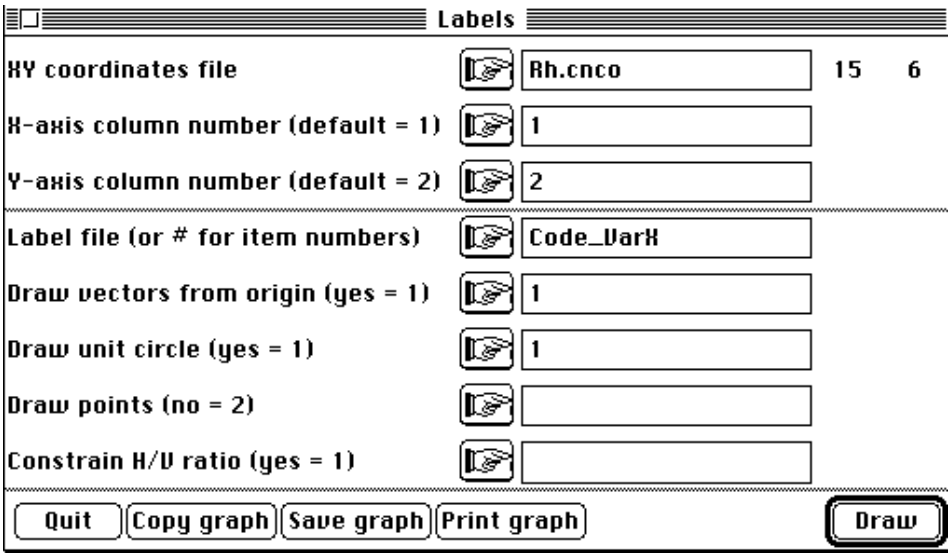


Figure 8 Eigenvalues diagram. This graph shows that a large part of the variability is taken into account by the two first axes.

Use the **Scatters** module to draw the F1 x F2 correlation circle (Fig. 9):



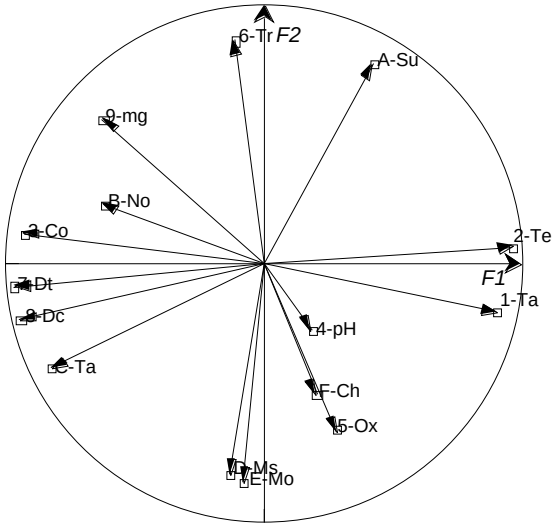
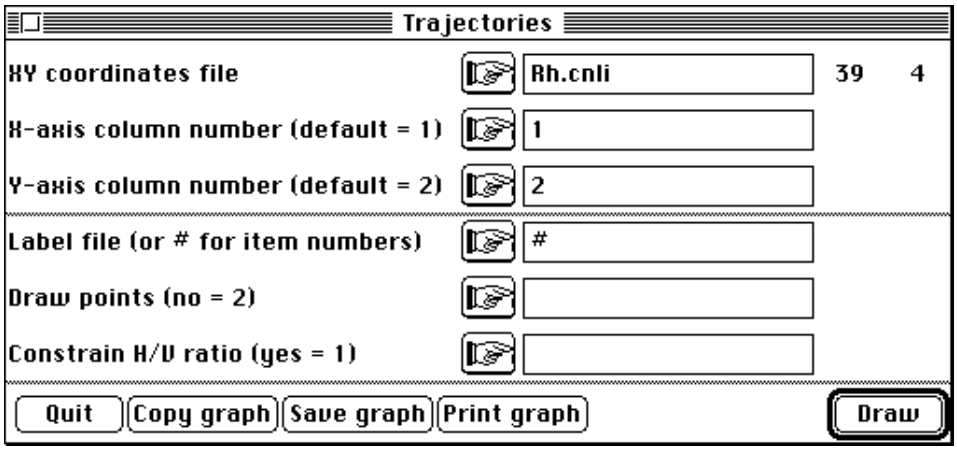


Figure 9 Correlation circle along the F1 x F2 plane.

Five variables are totally encountered for by the first axis (1, 2, 3, 7 and 8).

Two redundant groups of variables can be identified: (i) the first group is composed of air and water temperatures which best describe the seasonal cycle; and (ii) the second group deals with the mineralization (mainly related to the calco-carbonic equilibrium). Seasonal variations of the temperature involve seasonal variations of the $Ca^{++}/HCO_3^-/CO_3^{--}$ equilibrium by modifying the CO_2 solubility and the dissociation constant of the carbonic acid. The second axis deals with the variations of the variables 6, D and E, which follow discharge variations. Other variables are more or less associated with these two axes (C, A...) but the relationships of 1 or more axes contributing to the explanation of the variations of a variable cannot be clearly seen on such a graph. Other parameters not taken into account by the two first axes, because they are not related either to the season or the discharge in our example, are almost entirely missed by this type of graph. Use the **Scatters** module with the **Trajectories** option to draw the factorial map of sampling dates (Fig. 10):



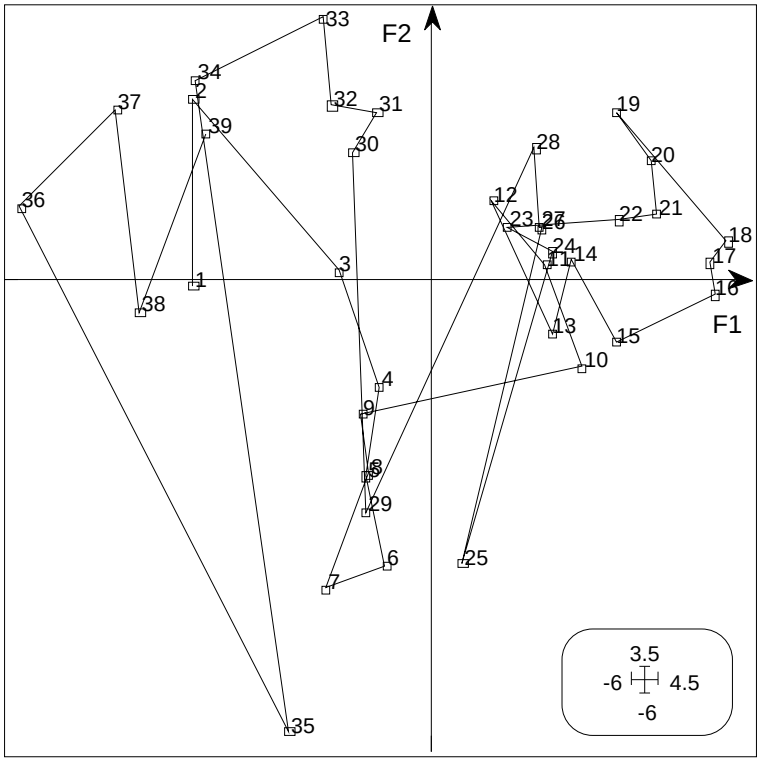
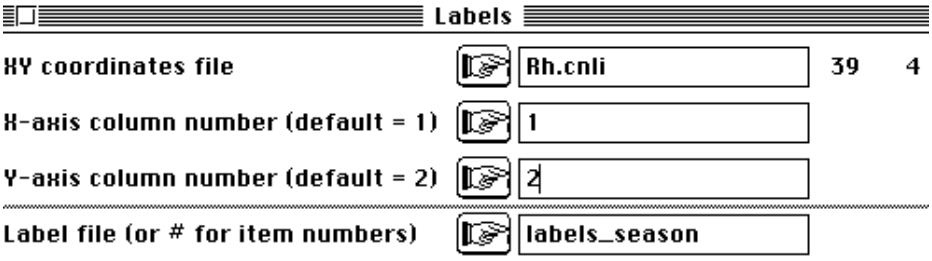


Figure 10 Sampling dates along the F1 x F2 plane. Dates are numbered and linked by a line by chronological order.

Create a label file (`Labels_season`, ASCII format, 39 rows, 1 character string) where each date is replaced by a label indicating the season as follows:

```
02/02  Wi   14/03  Sp   07/06  Su   06/09  Au   05/12  Wi
22/02  Wi   19/04  Sp   16/06  Su   12/09  Au   19/12  Wi
03/05  Sp   22/06  Su   19/09  Au   26/12  Wi   09/05  Sp
28/06  Su   26/09  Au   02/01  Wi   18/05  Sp   05/07  Su
03/10  Au   24/05  Sp   11/07  Su   10/10  Au   31/05  Sp
19/07  Su   18/10  Au   26/07  Su   24/10  Au   01/08  Su
31/10  Au   09/08  Su   07/11  Au   16/08  Su   14/11  Au
23/08  Su   21/11  Au   30/08  Su   28/11  Au
```

Use **Scatters** to get an illustration of the seasonal process (Fig. 11):



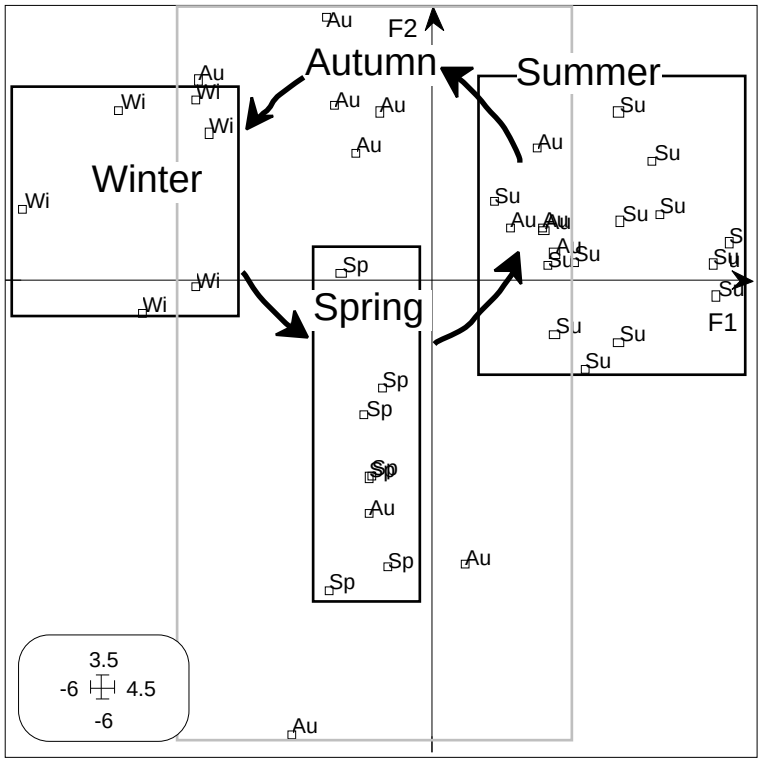
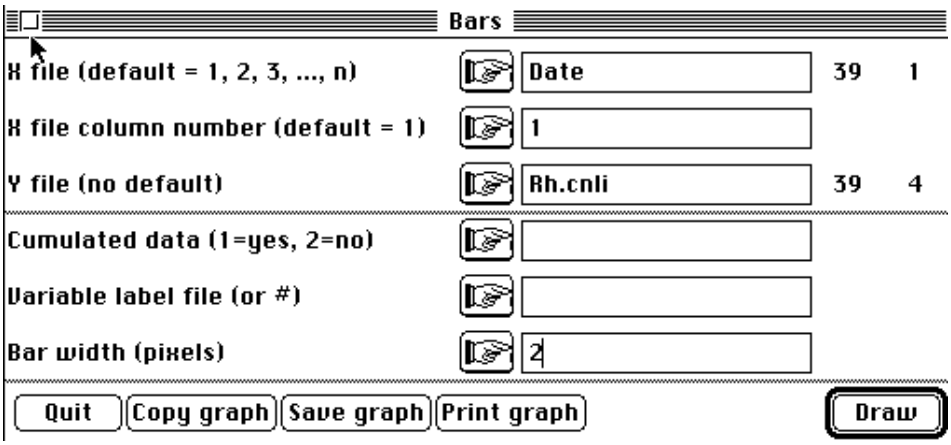


Figure 11 F1-F2 factorial map of the sampling dates. This graph shows the seasonal cycle and highlights the dispersion of data recorded in Autumn.

The temperature regime is thus highlighted in Fig. 11 (scores on the first axis) and the relationships of sampling dates and discharge is depicted by the second axis. The discharge of the Rhône river is characterised by a relatively regular summer (snow melting) and winter (low water level) discharge. During the 1983's spring, exceptional floods occurred. During autumn, periods of high and low discharge alternated.

This temporal dynamic can be easily demonstrated on a "functional graph" where the factorial scores of the sampling dates are plotted against time. Use the **Curves** module and the **Bars** option to draw such "functional graph". Select `Date` as **X file** (after transformation of `Date .txt` into `Date` (binary)):



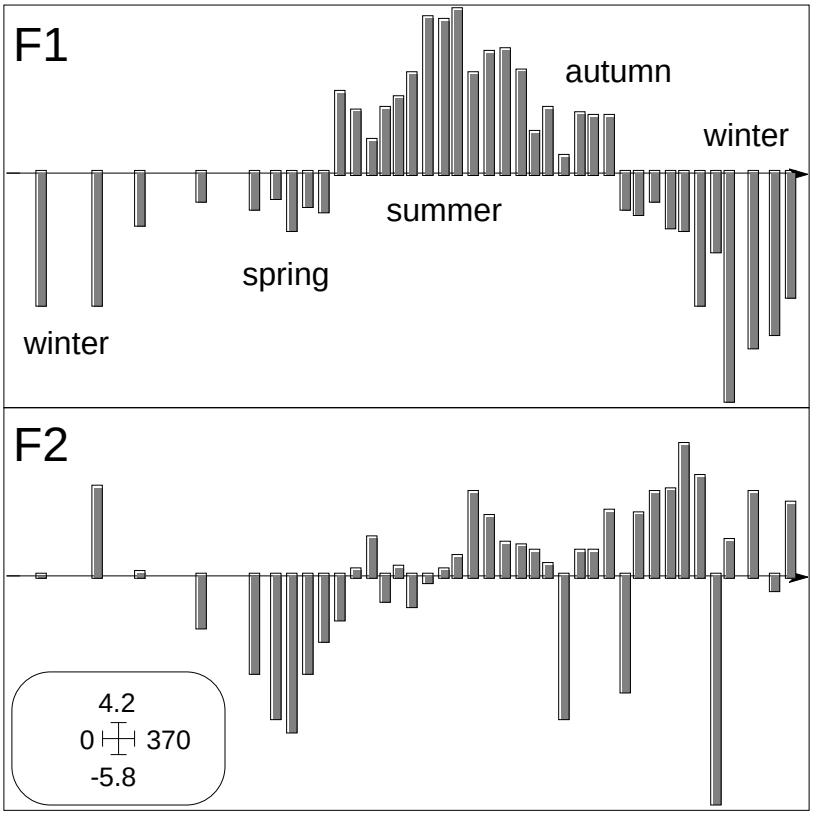


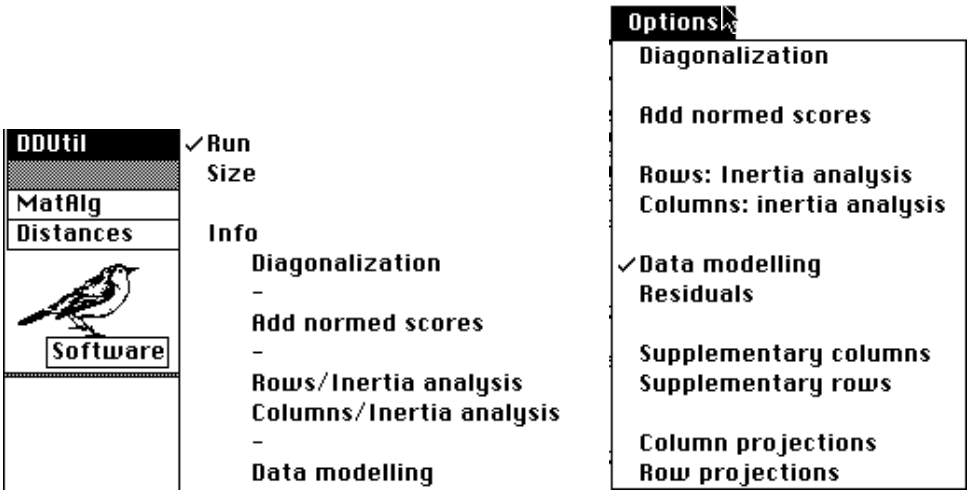
Figure 12 Temporal evolution of the sampling date scores. The seasonal cycle is shown along the first axis and the discharge variations are demonstrated along the second axis.

Factorial maps provide a basis for the interpretation of multivariate data sets along the first axes. Nevertheless, they are not very efficient for the analysis of further axes that describes particular variable or sample. Actually, on a factorial map the experimental point M_i (true value) at the end of vector $OM_i = (y_{i1},..., y_{ip})$ is replaced by its orthogonal projection m_i onto the F1-F2 factorial plane, for instance. This graph does not provide any measure of the distance between the points M_i and its projection m_i. As a result, the computation of inertia statistics permits to measure such a distance. Moreover, this distance can be viewed by plotting together the reconstituted data (by

linear auto-modelling) and the original data. This is the aim of data reconstitution (see **Statistical basis** volume).

5.4 - Data reconstitution (See DDUtil : Data modelling)

Use the **DDUtil** program and the **Data modelling** option to reconstitute (or recreate) the data as follows:



Select `Rh.cnvp` as **Input file**:


```
Modelling: Data reconstitution after PCA/CA
Title of the analysis: Rh.cnta
Number of rows: 39, columns: 15

File Rh.cnrA contains the model computed with 1 axe
It has 39 rows and 15 columns

File Rh.cnrB contains the model computed with 2 axes
It has 39 rows and 15 columns

File Rh.cnrC contains the model computed with 3 axes
It has 39 rows and 15 columns

File Rh.cnrD contains the model computed with 4 axes
It has 39 rows and 15 columns

File Rh.cnrE contains the model computed with 5 axes
It has 39 rows and 15 columns

File Rh.cnrF contains the model computed with 6 axes
It has 39 rows and 15 columns
```

Six files were created `Rh.cnrA`, `Rh.cnrB`, `Rh.cnrC`, `Rh.cnrD`, `Rh.cnrE`, and `Rh.cnrF`. Each file corresponds to the reconstructed data with the first axis for `Rh.cnrA`, the first and the second axes for `Rh.cnrB`,..., and the six first axes together for `Rh.cnrF`.

From a geometrical point of view the reconstitution corresponds:

- to an approximation of the original array by a rank 1 matrix. The form of this operation is equal to

$$\mathbf{Rh}_1 = \lambda_1^{-1/2}\mathbf{R}_1\mathbf{C}_1^t \quad \text{(decomposition into singular values)}$$

- to the modelling of the table where y_{ij} (measurement of the jth variable at the ith sampling date) is replaced by a model y_{ij}^1 as follows:

$$y_{ij}^1 = ka_ib_j$$

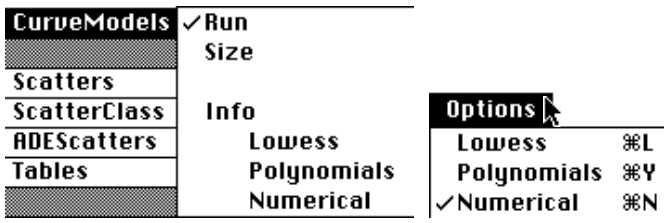
which minimises the reconstitution error E_1 equal to

$$E_1 = \sum_{i,j}(1/n)\left(y_{ij} - ka_ib_{ij}\right)^2$$

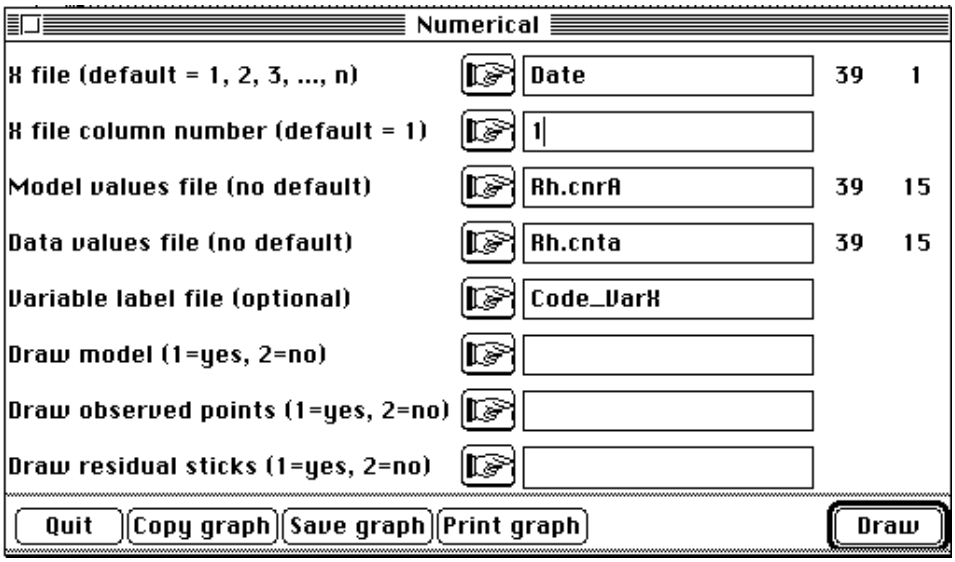
under the constraint that $\sum a_i^2 = 1$ and $\sum b_j^2 = 1$ b_j.

When data are multivariate chronicles, the simultaneous plot for each variable ($1 \leq j \leq p$) at the date t_i of the value y_{ij} and its model $\lambda_1^{-1/2}$ $R_{i1}C_{j1}$ corresponds to the simultaneous reading of the point and its projection. The temporal succession of each variable compared to a reference curve (at more or less one optimal dilation for each variable) is thereby shown.

Use the **Curvemodels** module with the **Numerical** option to superpose the normalised values (table `Rh.cnta`) and the reconstructed values (file `Rh.cnrA` for the model computed with the first eigenvalue and the first scores).



Fill in the boxes of the dialog window as follows:



This results in Fig. 13. All curves (reconstructed data) are identical (R_{11}, R_{21}, ..., R_{n1}) at more or less one scale coefficient ($l_1^{-1/2}$ C_{11}, $l_1^{-1/2}$ C_{21}, ..., $l_1^{-1/2}$ C_{P1}), which almost equals 0 when the temporal variations of the variable is not related to the model (e.g., the variable E-Mo).

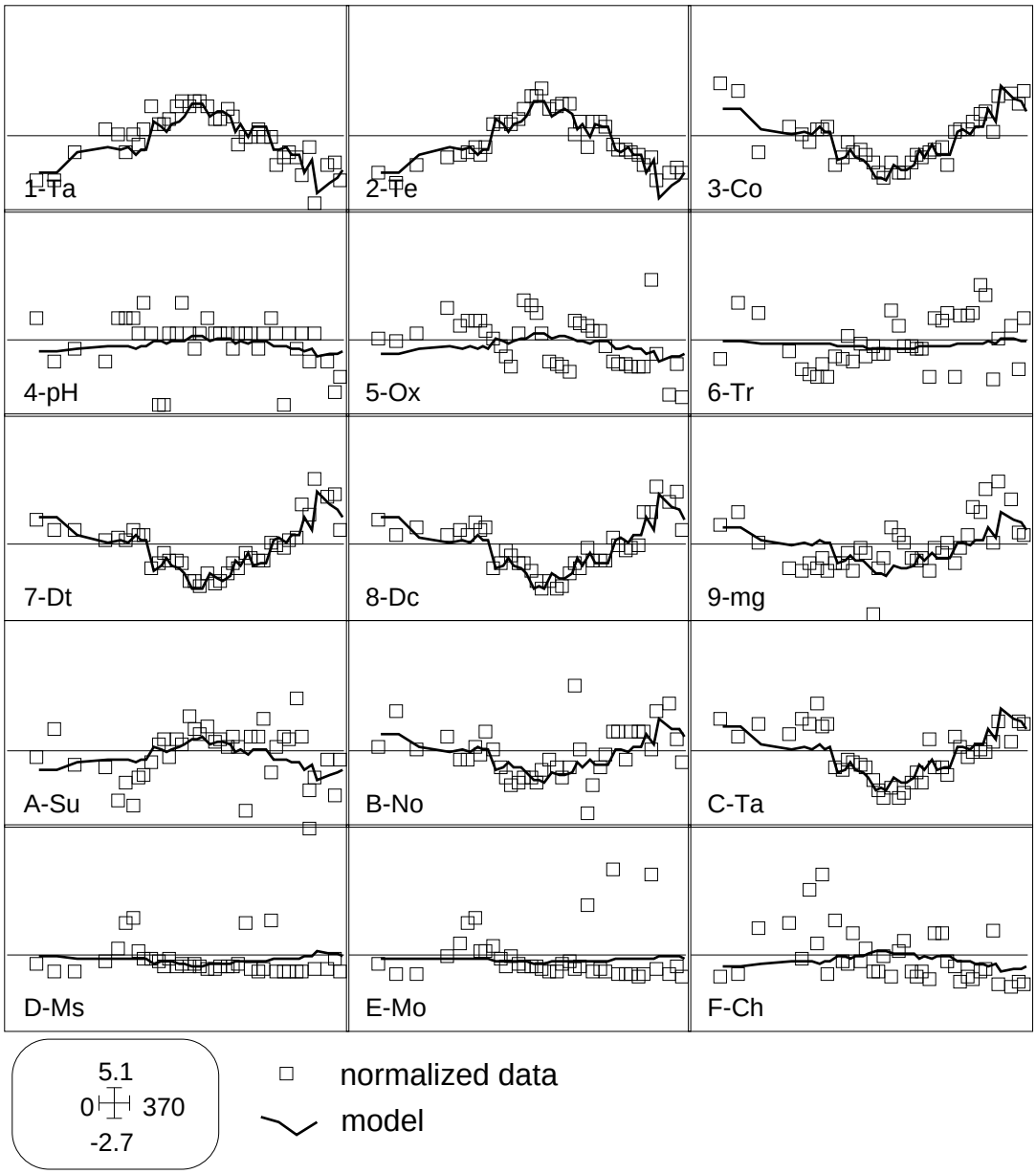


Figure 13 Superposition of the normalised values and the data reconstructed by the first axis of the PCA.

The two previously highlighted groups associated with the first axis (temperature and mineralization) are recognised.

Other variables are partially described by the model: total alkalinity (C-Ta), magnesium (9-mg), nitric nitrogen (B-No), sulphates (A-Su). The illustration clearly shows the unsuitability of the model for the variables described by other axes (suspended materials (D-Ms), organic matter (E-Mo), water transparency (6-Tr), pH (4-pH), or chlorophyll a (F-Ch)).

Consequently, the data reconstitution must be carried on for these poorly described variables. This gradual reconstitution of the normalised data is carried out by successive models as follows:

$$y_{ij}^{s} = \sum_{k=1}^{s} \lambda_k^{-1/2} R_{ik} C_{jk}$$

Using the same logic the original data can be reconstituted by:

$$x_{ij}^{s} = m_j + s_j \left(\sum_{k=1}^{s} \lambda_k^{-1/2} R_{ik} C_{jk} \right)$$

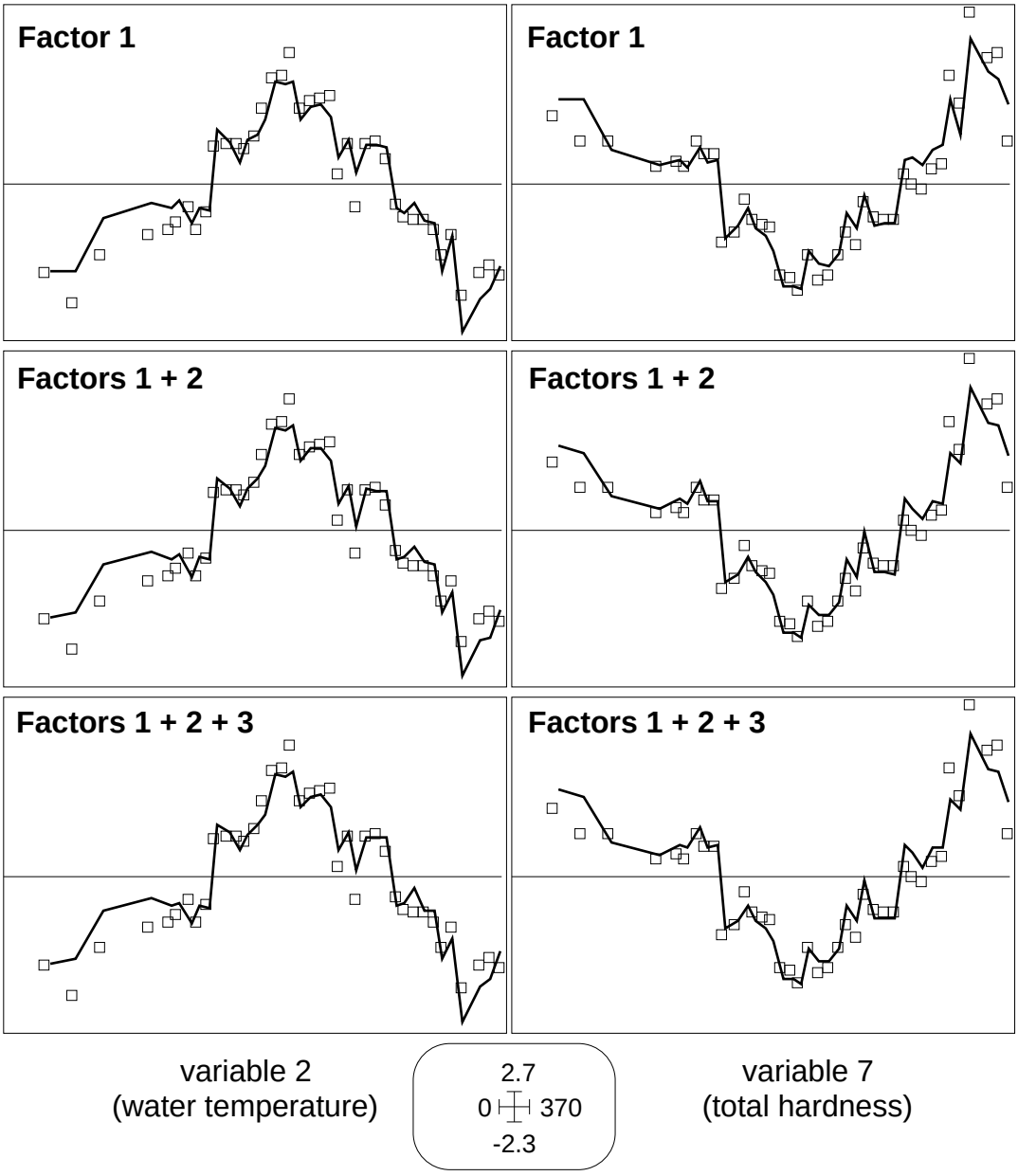


Figure 14 Data reconstitution (variables no. 2 and no. 7) with first axis, first and second axes, and first, second and third axes resulting from the normalised PCA.

As a result, the data reconstitution is achieved by the rank 1 model for the variables no. 2 (water temperature) and no. 7 (total hardness) (Fig. 14).

Generally, several axes are necessary for the reconstruction of a variable. In such a case, you can use the different files created by the data modelling module to complete the data reconstitution.

Because of the independence among axes, you must select a curve into the set of curves that are not correlated. As a result, successive curves (models) may have the

same general shape or by contrast, they can describe a new process. This occurs in Fig. 15 for variables no. 6 (water transparency) and no. 14 (organic matter).

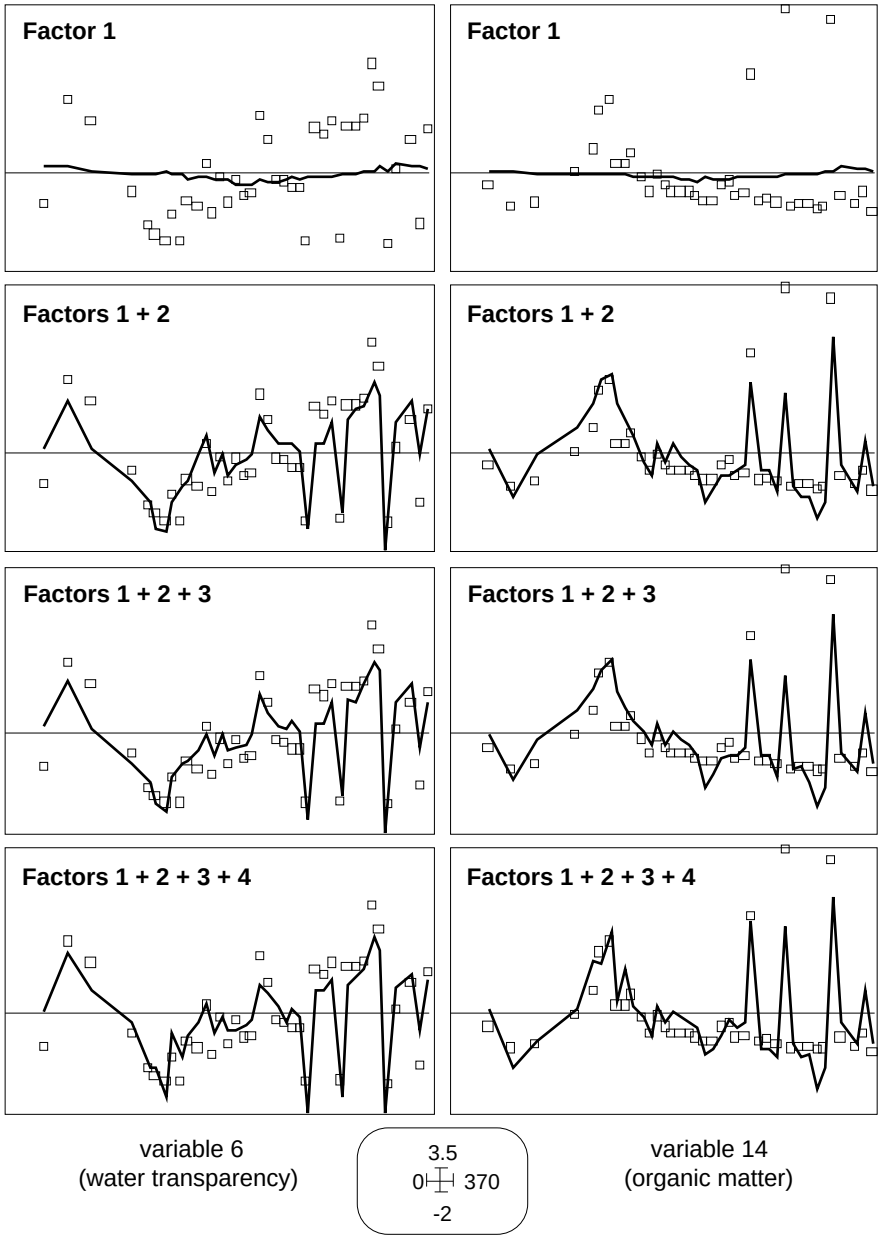


Figure 15 Gradual data reconstitution for variables no. 6 and no. 14 using the four first axes of the normalised PCA.

The variables no. 10 (sulphates) and no. 13 (suspended matter) behave similarly as particular sampling dates are taken into account by the fifth factor (Fig. 16).

One variable can be statistically independent of the others and one further axis can be associated with such a variable. As a result, the experimenter has to choose such axis if it is relevant despite that a statistical choice leads to eliminate such axis (Legendre & Legendre, 1984)[4]. In all the cases, the experimenter has to decide if the processes described by the different axes are (or not) relevant.

For example, the third axis describes the variable no. 4 (pH) and the fourth axis describes the variable no. 15 (chlorophyll a) (Fig. 17).

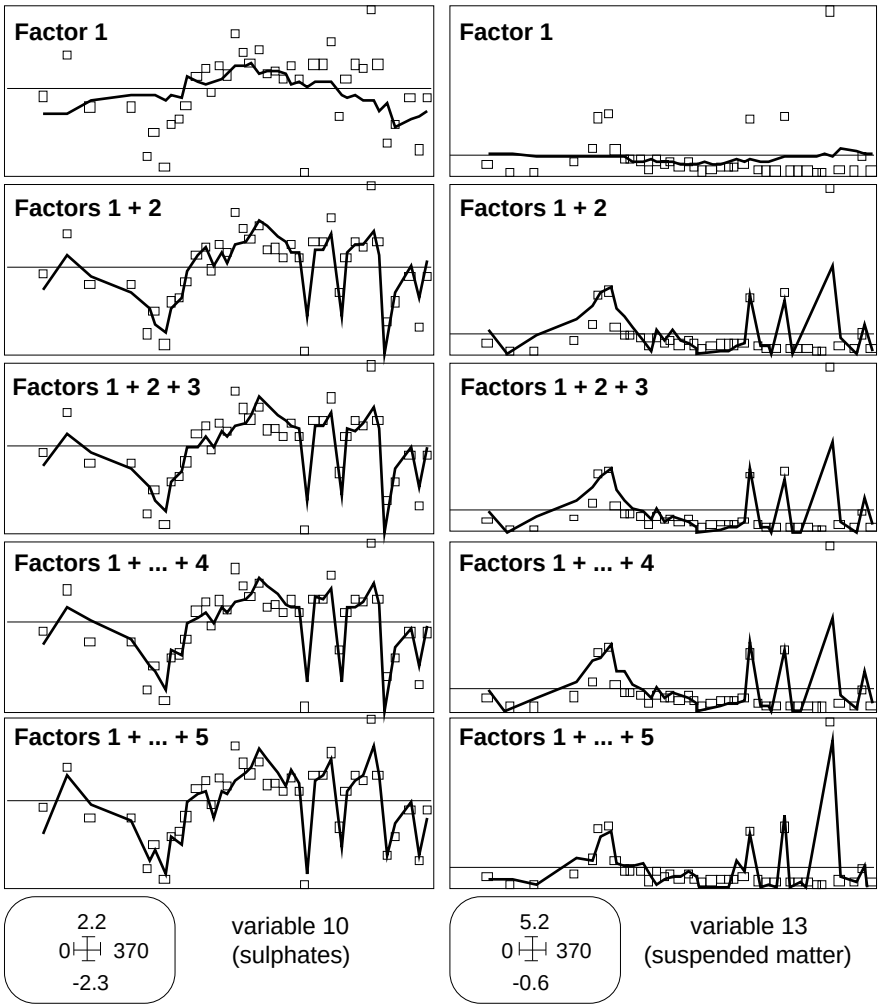


Figure 16 Gradual data reconstitution for variables no. 10 and no. 13 using the five first axes of the normalised PCA.

Finally, one can verify the coherence of the inertia statistics and the gradual reconstruction of the data.

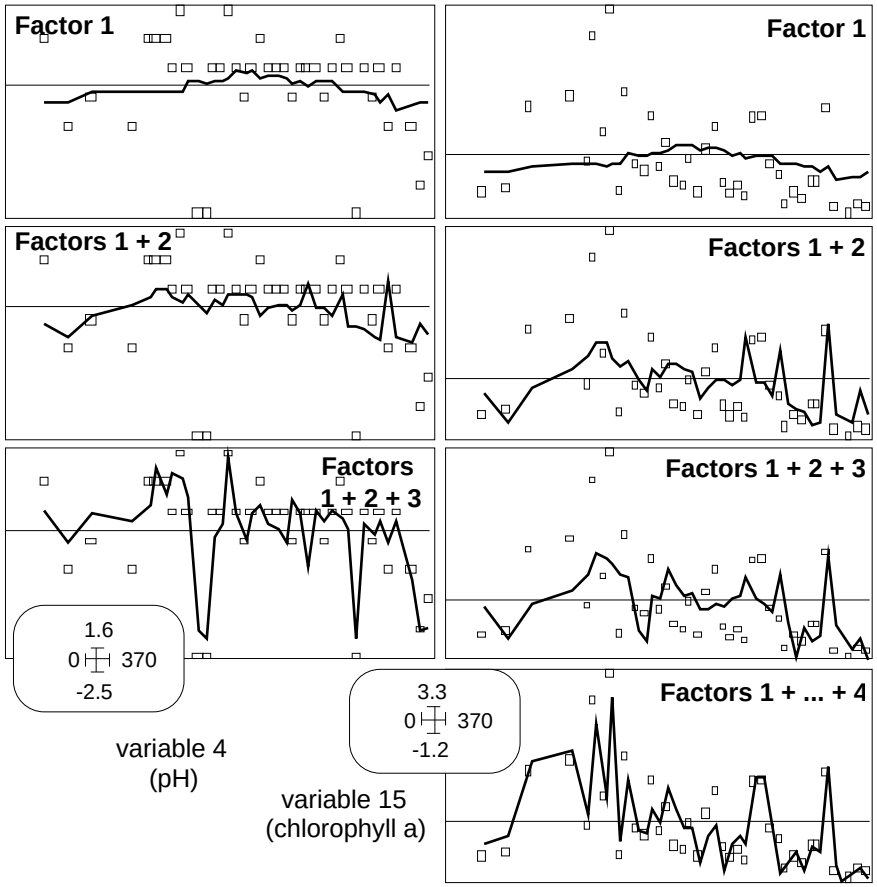


Figure 17 Gradual data reconstitution for variables no. 4 and no. 15 using the three and the four first axes of the normalised PCA.

Références

[1] Carrel, G., Barthelemy, D., Auda, Y. & Chessel, D. (1986) Approche graphique de l'analyse en composantes principales normée : utilisation en hydrobiologie. *Acta Œcologica, Œcologia Generalis* : 7, 2, 189-203.

[2] Barthélémy, D. (1984) Impact des pollutions sur la faune stygobie karstique : approche typologique sur seize émergences des départements de l'Ain et du Jura. Thèse de 3° cycle, Université Lyon 1. 1-182.

[3] Carrel, G. (1986) Caractérisation physico-chimique du Haut-Rhône français et de ses annexes : incidences sur la croissance des populations d'alevins. Thèse de doctorat. Université Lyon 1. 1-186.

[4] Legendre, L. & Legendre, P. (1984) La structure des données écologiques. Tome 2 - Masson, Paris. 2ème édition: 1-344.