# GRAPHICAL TECHNIQUES FOR MULTIDIMENSIONAL DATA ANALY

THIOULOUSE J. (1), DEVILLERS J. (2, 3), CHESSEL D. (4), and AUDA Y. (5)

(1) Laboratoire de biométrie, génétique et biologie des populations, U.R.A. CNRS
Université Lyon 1, 69622 Villeurbanne CEDEX, France.
(2) CTIS, 21 rue de la Bannière, 69003 Lyon, France.
(3) European Group for the QSAR Studies.
(4) Ecologie des Eaux Douces, U.R.A. CNRS 367, Université Lyon 1, 6
Villeurbanne CEDEX, France.
(5) Maison de l'Orient Méditerranéen, CNRS, Université Lyon 2, 7 rue Raulin, 6
Lyon, France.

## 1. Graphics as an interdisciplinary communication language.

The need for graphical data analysis tools is not to be demonstrated, and has been rec recently by Morgenthaler and Tukey (1989) in a conference about the future of analysis.

Scientific graphics, as defined by Bertin (1967), can play two basic and compleme roles. First, it is a particularly effective communication tool, and is used as such i whole scientific literature. But in the precise scope of biometry, the part of communic tool is made specially more crucial by the fact that exchanges are basically n disciplinary. Auda (1983, p.122) has already underlined this point, by noting that "I context, not only can Graphics settle a dialogue between researchers working in fiel different as Statistics, Informatics and Biology, but it also plays a prominent part i understanding of the scientific message". We are going to give some examples o point.

### 1.1. BASIC PRINCIPLES.

The credit must be given to Bertin (1967) for having defined the laws of scie graphics. Auda (1983) has derived from them the basic principles which can be appli multidimensional analysis of ecological data sets. It is interesting to first recall principles: universality, hypotheses economy, and existence of several levels of perce of a graph.

According to the universality principle, a graph should not use conventions (e.g.; ling or cultural conventions), and should only be based upon the following three fundam relationships:

- similarity/dissimilarity relationship
- order relationship
- proportionality relationship

The relations between the elements of a graph should reflect the relations betwee elements which are symbolized.

*J. Devillers and W. Karcher (eds.), Applied Multivariate Analysis in SAR and Environmental Studies,* 153–205.

According to the principle of hypotheses economy, the introduction of new unnecessary elements should be avoided (for example the representation of a quantity should be done with a circle rather than with a non-geometrical picture).

Lastly, a graph should be readable for different levels of perception (global, mean and local levels for example).

The three main types of variables which can be used for the construction of a graph are variables of separation, value, and size. Separation variables, like the different hatching patterns of the zones on a geographical map, should only be used to display similarity/dissimilarity relationships. Value variables, like gray levels on a map, allow the introduction of an order relationship. Size variables can be used to display a proportionality relationship, for example on a map with circles having an area proportional to the value of a variable.

To these basic concepts can be added the possibility to fit together elementary graphs. This is often necessary to render an account of the structuration of the data set (i.e.; groups of rows, columns, etc.). Collection and superimposition are the principal ways to assemble elementary graphs. The multiwindowing technique, introduced by operating system of computers with advanced graphical user interface is particularly suitable to these kinds of elementary graphs assembly.

Far from being only a set of theoretical considerations without practical consequences, these remarks have been at the origin of a computer program which has evolved from the "Graphique" program from Auda (1983), written on a Data General Eclipse S/140 mini-computer with a Tektronix plotter, to the "GraphMu" program (Thioulouse, 1989, 1990) for the Apple Computer Macintosh™ microcomputer.

Figure 1 shows examples of graphs using the above principles. These graphs have been drawn on a Macintosh with GraphMu and completed (numbering and positioning) with an object-oriented graphics software (e.g.; Claris MacDraw™)

### 1.2. EXPLORATORY DATA ANALYSIS (EDA).

EDA finds its origins in the works from Tukey (Tukey, 1977; Tukey and Tukey, 1981a, b, c; Chambers *et al.*, 1983). It arises from an exploratory approach, and is complementary to confirmatory data analysis, represented by inferential statistics and particularly generalized linear modeling. Starting from the fact that the hypotheses on which are based classical inferential statistics are often difficult (or impossible) to verify, the EDA proposes the use of simple graphical representation of data (scatter diagrams), or more elaborated (whisker box, cellulation technics), together with basic descriptive statistical parameters (distribution histograms, quantiles, smoothing, linear regression, residuals study). According to Tukey and Tukey (1981a) a difference can be made between "archival graphics", which are useful to store informations in a compact way, and "impact graphics", intended for putting in evidence a precise characteristic of data. Moreover, they distinguish three additional aims: merging data points, for example by using circles instead of points on a scatter diagram, separating, for example by using different symbols to represent points belonging to different groups, and alternating, for example by using symbols (for data points) positioned on a smoothly varying background (contour curves) so that both separation and merging can be achieved on the same graph.

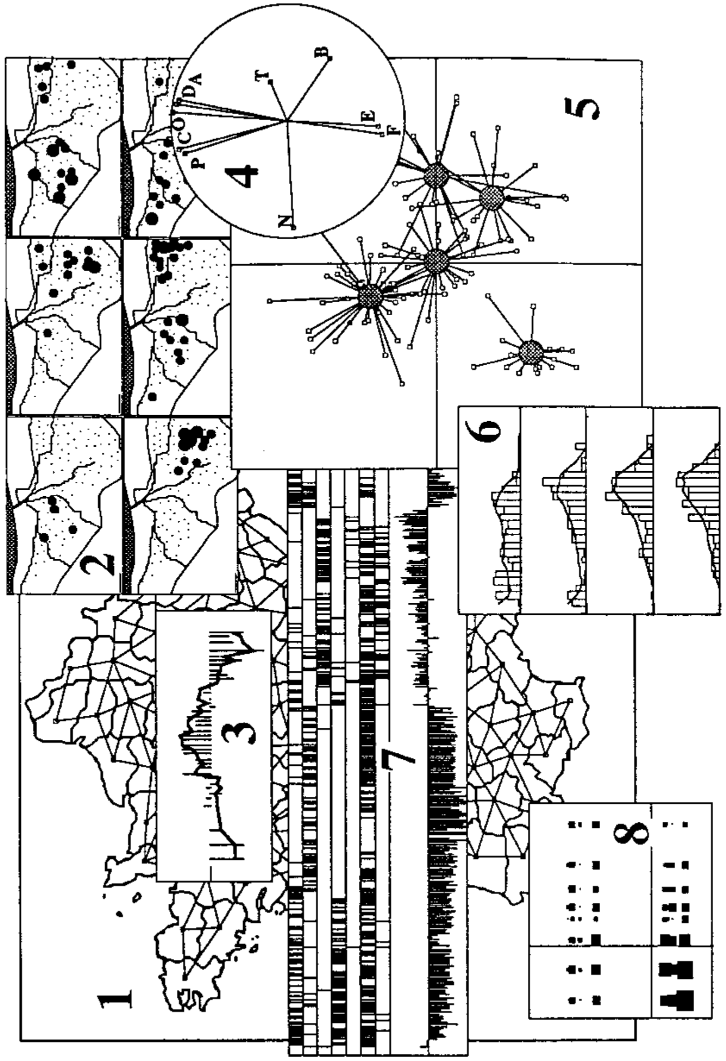


Figure 1: Illustration of some basic principles in scientific graphics. Labels represent the meaning of each point (4), lines symbolize the belonging to a class (5) or neighboring relationship (1). Size of pictures are proportional to a given quantity (2, 6, 8). Assembly of elementary graphs using collection (2, 6, 7), superimposition (1, 2, 3, 6) or both (6). All these graphs have been drawn on a Macintosh™ microcomputer using ADECO, GraphMu, and Claris MacDraw™ software.

Chambers *et al.* (1983) stress the importance of the criteria governing visual perception ( for example, it is easier to perceive the position on an axis than more complex aspects, like the size of an object), and associated mental processes (for example, additivity of visual effects).

Computer support for EDA, which was initially lacking (Tukey, 1977), or restricted to computing center users, is now widely available on microcomputers. On the Macintosh, we can quote Data Desk (Cornell University) and JMP (SAS Institute).

EDA approaches multivariate analysis in a very shallow way, and without concern for multidimensional methods. Chambers *et al.* (1983) make use of multiwindowing for the so-called "draftsman display" (projections on all the possible couples of axes) and its generalization to multidimensional data. This approach rapidly leads to an explosion of the number of possible graphs (see figure 2, and Chambers and Kleiner, 1982). Other graphical techniques met with the same difficulty when trying to approach multivariate data tables abreast: Chernoff faces (Chernoff, 1973; Wang, 1978), star diagrams, Andrew's curves (Andrews, 1972), etc. Several practical examples can be found in Everitt (1978), Wang (1978), or Wainer and Thissen (1981). In all cases, the discrimination, or the search for global trends, must be done visually. This is illusory when the number of variables and measures grows. Moreover, as pointed out by Gower and Digby (1981), these methods are dependent on subjective criteria which can be misleading (assigning variables to the elements of Chernoff faces, to the axes of star diagrams, to coefficients of Andrew's curves, etc.).

## 1.3. GRAPHS FOR MULTIVARIATE DATA ANALYSIS ON MICROCOMPUTERS.

To overcome the above problems, multidimensional data analysis methods (PCA: principal component analysis, CFA: correspondence factor analysis, MCA: multiple correspondence analysis) must be used, but this does not mean that we have to abandon our graphical tools. It is worth noting that the use of improved graphical techniques (as compared to mere factorial maps) in the field of multivariate analysis has not formed the subject of a synthesis, or of a theoretical settlement (which has been done for graphs in general, see for example Tufte (1983), or Bertin (1967)). It is however frequently met in scientific literature. Several examples may be found in the book from Barnett (1981): Gabriel (1981) draws ellipses in the plane of his biplot to summarize the set of points belonging to the same group. In the same paper, the author uses circles on a biplot with the Mahalanobis metric to represent the 5% thresholds of $T^2$ Hotelling tests between sample couples. In both cases, the introduction on the factor map of an information about the data set makes the interpretation of the results easier. A similar technique consists in drawing on the factor map the results of another analysis of the same data set (generally a hierarchical classification when the aim is to discriminate between sub-populations). Gower and Digby (1981, figure 6.22, p. 110) give an example of such a strategy by plotting on the factor plane of a procuste analysis the mean of each sub-population, linked to all the individuals which compose it by a "hedgedog" diagram. One can find in Everitt (1978) other examples: bidimensional oriented symbols plotted on the factor map, minimum spanning tree, automatic classification, etc.

In ecology, Owen (1990) uses 95% confidence ellipses in the factor plane of a detrended correpondence analysis (DCA) (Hill, 1979a; Gauch, 1982) to summarize the position and

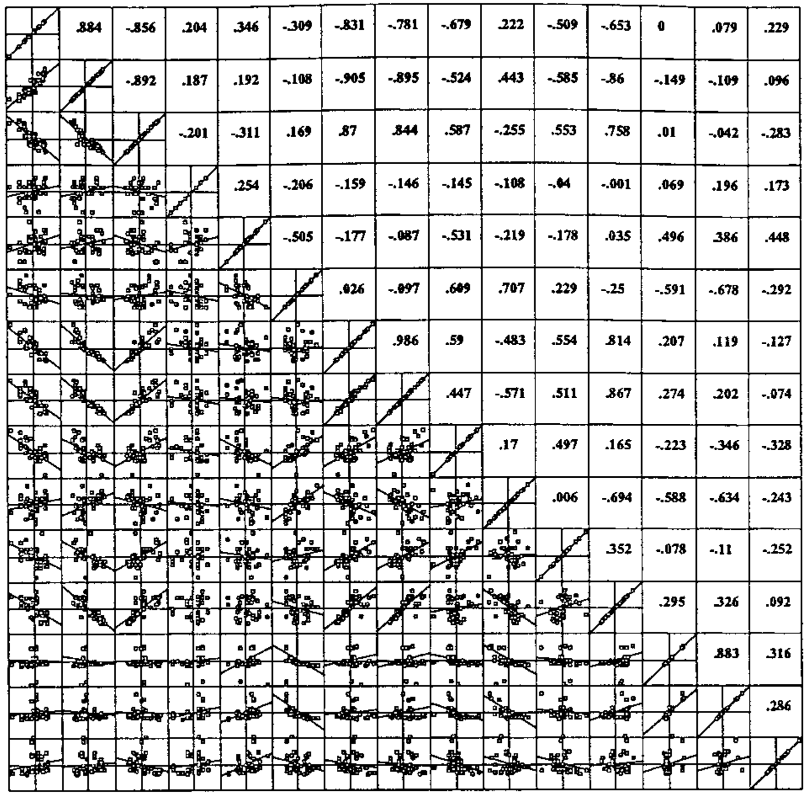


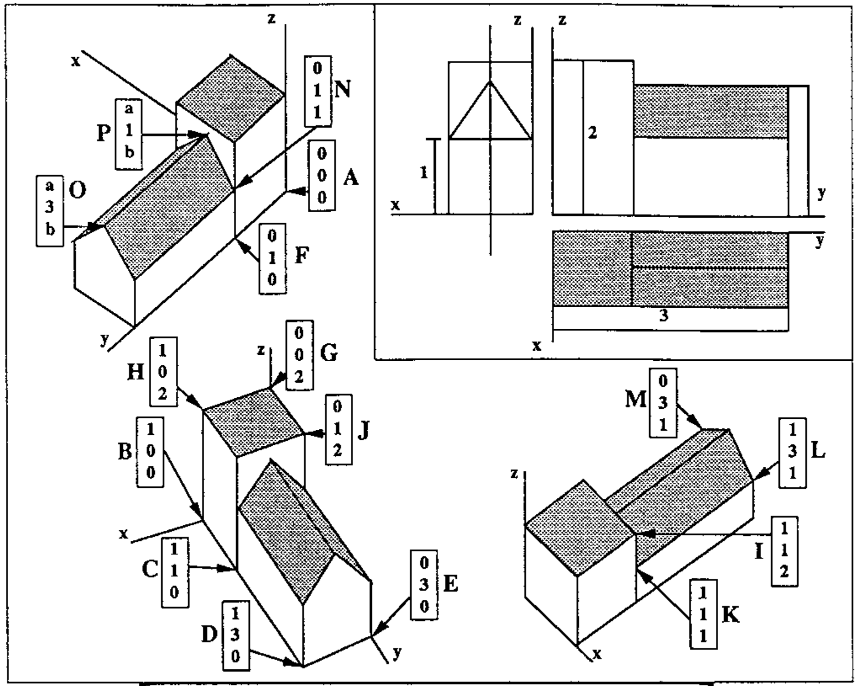
Figure 2: Numerical information vs. graphical information: representation of a correlation matrix. Upper triangle: correlation coefficients between each couple of variables. Lower triangle: scatter diagrams (each cloud is standardized: mean = 0 and variance = 1) for each couple of variables. Data are from Carrel *et al.* (1986), 15 physicochemical variables are measured at 39 sampling dates. Although the power of graphical representations is high, the need for numerical tools to summarize the information is obvious.

the dispersion of 141 mammalian species in 189 quadrats systematically di Texas, according to eight types of environment defined by a classificat (TWINSPAN) (Hill, 1979b; Gauch, 1982). Moreover, the author uses contoı draw on the geographical map of Texas the variations of the values of the first from DCA (the first factor reflects the global productivity and the second th seasonality). Conversely, the species richness is fitted with two-variables sec polynomials, and the results is represented as gray-level maps in the first plane.

The above examples are two kinds of graphical techniques which can alternatives to factor maps: representation of the values of factors as functions structures (graphical representation of a factor on a geographical map, or as a time, called "functional representations"), or, reciprocally, representation of da structure of data) on the factor plane. This approach, although being natural, e use of classical multidimensional software, which are limited to classical fa Moreover, to be really efficient, graphical tools for multivariate data analysi interactive: a graph must not be the result of several hours of work, whether i paper sheet or a computer console. The user should be able to use the "trial method to obtain, starting from the results of the statistical methods, the graph to their expression (for example choosing from factor map, functional represen diagram, hedgedog diagram, etc.). Lastly, the integration of computing and gra in the everyday desktop environment (word processor, spreadsheet, electronic makes much easier the task of writing reports, papers, and various manuscrip the results of data analyses must be presented. The availability of thes microcomputers widely spread in research laboratories allows the users to h elements necessary to the practical realization of these analyses. The M microcomputer fills a particular place in this field, because of its graphical use and the advanced integration between the different programs, allowing to easi and graphs. The ability to operate the pictures coming from a data analysi assemble them into an orderly graph (for example the correlation circle for va histogram of eigenvalues, and the functional representation of row coordina include them directly into a word processor program is extremely valuable and Macintosh™ and related microcomputers particularly suitable for the a environmental data.

## 1.4. ALGEBRAIC BASES FOR GRAPHICAL DATA ANALYSIS.

Graphs can be used to understand the relationships existing between data anal underlying algebraic bases. Most of the difficulties are already present in th three-dimensional space $\mathbb{R}^3$: figure 3 shows a simple object made of 16 points three coordinates. Starting from two angles a and b, elementary algebraic opera to define a new base of $\mathbb{R}^3$, (vectors U,V,W) and compute the coordinates of this new base (figure 4). By using only the last two elements of the new coord obtains the projection of the object onto the plane which is orthogonal to vector by angles a and b. The cartesian representation of the 16 points with the prec dimensional coordinates is an euclidean representation of the object. Thus operation of data analysis (figure 5) consists in projecting objects onto plane looking at these planes in front-view. Mathematically, these operations are ident when points are defined by p coordinates (i.e.; the measure of p variables, or oı



| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1/2 | 1/2 |
| 0 | 0 | 1 | 3 | 3 | 1 | 0 | 0 | 1 | 1 | 1 | 3 | 3 | 1 | 3 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | $\sqrt{3}$ | $\sqrt{3}$ |

Figure 3: Three-dimensional representation of an object made of 16 points in $R^3$. Points are labeled from A to P. The upper-right part of the figure gives the classical (top, front and side) views. The value of each component of the 16 coordinates is listed with the corresponding labels under the figure.

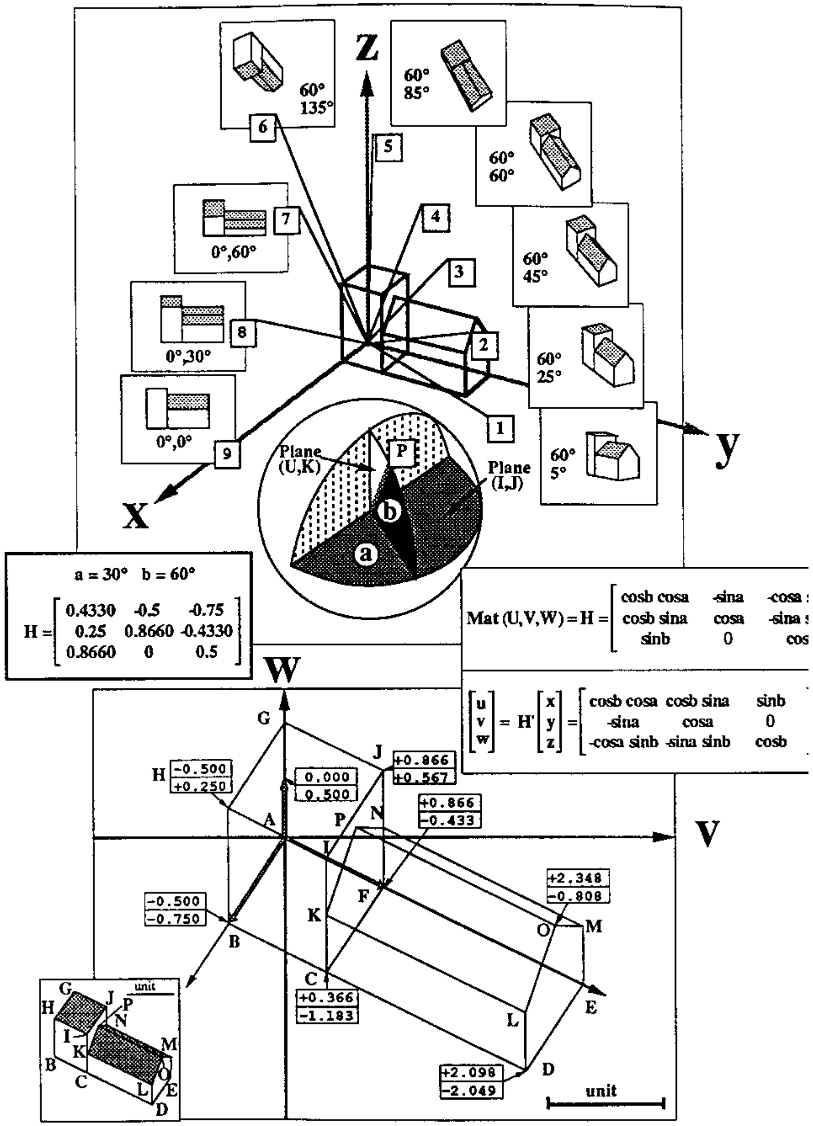

Z
X
y
60°
135°
60°
85°
60°
60°
60°
45°
60°
25°
60°
5°
0°,60°
0°,30°
0°,0°
1
2
3
4
5
6
7
8
9
Plane (U,K)
Plane (I,J)
P
a
b
a = 30° b = 60°
H = [0.4330 -0.5 -0.75; 0.25 0.8660 -0.4330; 0.8660 0 0.5]
Mat (U,V,W) = H = [cosb cosa -sina -cosa; cosb sina cosa -sina; sinb 0 cos]
[u v w] = H' [x y z] = [cosb cosa cosb sina sinb; -sina cosa 0; -cosa sinb -sina sinb cosb]
W
V
G
H
-0.500
+0.250
0.000
0.500
J
+0.866
+0.567
+0.866
-0.433
A
P
N
I
F
K
-0.500
-0.750
B
C
+0.366
-1.183
+2.348
-0.808
O
M
L
E
D
+2.098
-2.049
unit

Figure 4: Projection of a three-dimensional object on a plane.

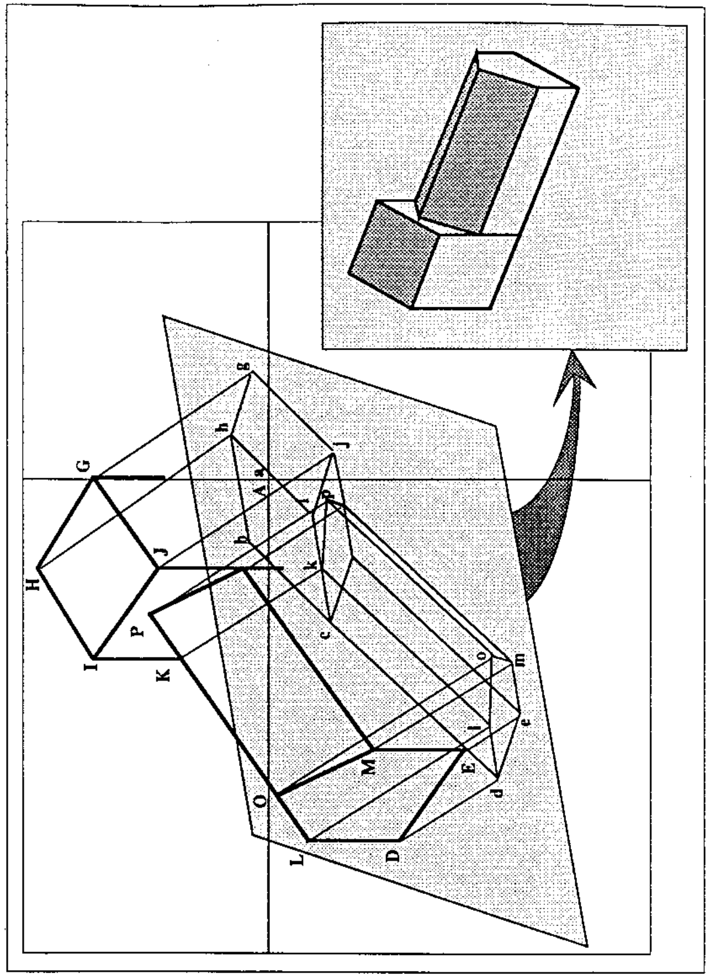


Figure 5: Geometrical view of the basic data analysis operation: a cloud of points in a vector space is projected onto a plane and the image of the cloud on this plane is interpreted.

data table). By this way, one can obtain a concrete image of an abstract object, which basis of the French school of data analysis.

## 1.5. DATA ANALYSIS PRINCIPLES.

The transition between the preceding algebraic bases and numerical data analysis is on a simple principle. Three measures (i.e.; three variables) define three values (x,y, therefore one point in $\mathbb{R}^3$. Repeated n times, these measures define a cloud of n po $\mathbb{R}^3$. There is an infinity of ways to project this cloud over a two-dimensional plar data analysis proposes particular planes, called "inertia planes", which have interesting properties: the mean of the distances between each couple of poi optimized. This is the basic feature of all the so-called "eigenvalue methods' particularly of PCA, CA and MCA.

The main conceptual difficulty comes from the fact that a data table having n (measures) and p columns (variables) can be considered either as a cloud of n po $\mathbb{R}^p$, or as a cloud of p points in $\mathbb{R}^n$ (in the example of figure 6, as a cloud of 21 po $\mathbb{R}^2$, or a cloud of two points in $\mathbb{R}^{21}$). Therefore, one can consider two kinds of maps, one which is the projection from $\mathbb{R}^p$ into $\mathbb{R}^2$ and the other from $\mathbb{R}^n$ intc Figure 7 explains this point of view, and allows to introduce the following comr centering the variables (i.e.; subtracting the mean from all the observed values) can b as a translation (figure 7A --> 7B) or as a projection (figure 7C). The variance can b as the dispersion of a cloud (figure 7A and 7B), or as the square of the length of a (figure 7C). The correlation can be seen as the elongation of a cloud (figure 7A an or as the angle between two vectors (angle between vectors X and Y or $X_0$ and figure 7C). These remarks are derived from the mathematical formalism of the c diagram (Fisher, 1915; Escoufier, 1987).

- - - - - - - - - - - - - - - - - -

Figure 6: The measures of three variables (temperature, ammoniac, nitrate) at 21 loc can be represented as a cloud of 21 points in $\mathbb{R}^3$. The three axes in this space are the variables. Data analysis defines a particular plane (ABCD) on which the cloud is prc (on the left), and onto which we can look to find out the characteristics of the cloud ( right). On this plane, the variables can be represented as the projection of the three a $\mathbb{R}^3$. The cloud may be then projected onto the horizontal axis (upper part of the g and represented on this axis with vertical bars proportional to the distance betwee: point and the plane of projection.

- - - - - - - - - - - - - - - - - -

Figure 7: A: n points (i.e.; n measures) in space $\mathbb{R}^2$ (raw data). Point number coordinates $(x_i,y_i)$ in the canonical base of $\mathbb{R}^2$. B: n points in space $\mathbb{R}^2$ (centered the cloud has only been translated so that its center has coordinates (0,0). C: two (i.e.; two variables, X and Y) in space $\mathbb{R}^{21}$; point X has coordinates (x1, x2, ..., x point Y has coordinates (y1, y2, ..., yn). $X_0$ and $Y_0$ are the centered variables (proj onto the subspace orthogonal to $\mathbb{1}_n$ ).

Thioulouse, J., J. Devillers, D. Chessel, and Y. Auda. 1991. Graphical techniques for multidimensional data analysis. Pages 153-205 in J. Devillers and W. Karcher, editors. Applied Multivariate Analysis in SAR and Environmental Studies. Kluwer Academic Publishers. http://pbil.univ-lyon1.fr/R/articles/arti067.pdf

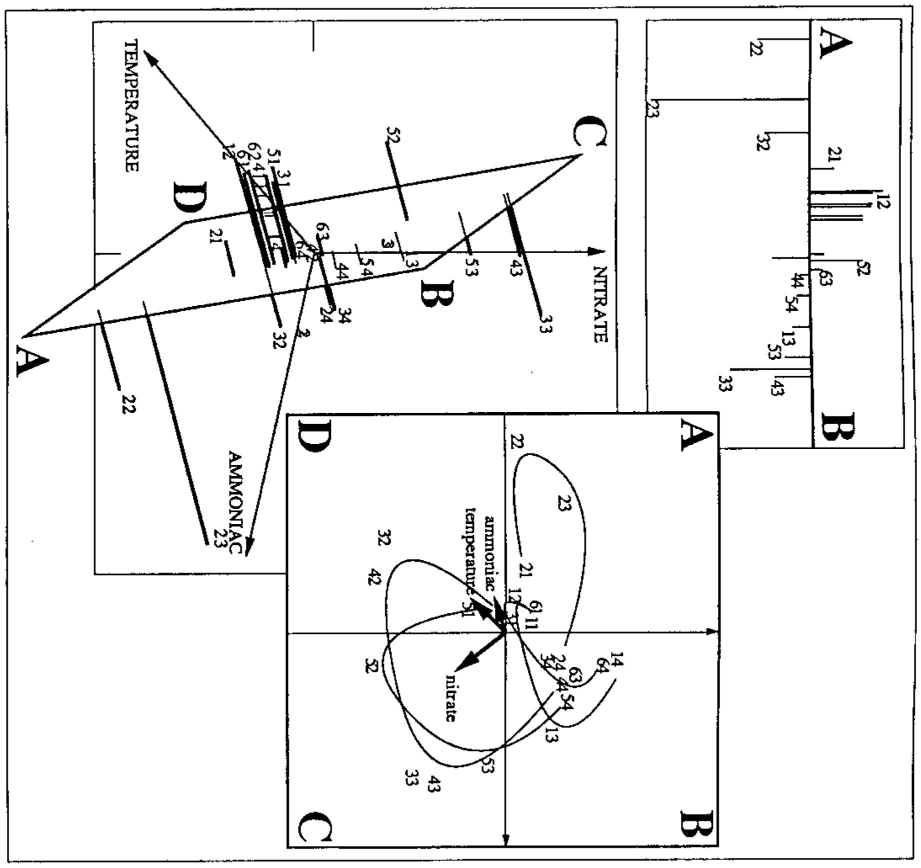


Figure 6

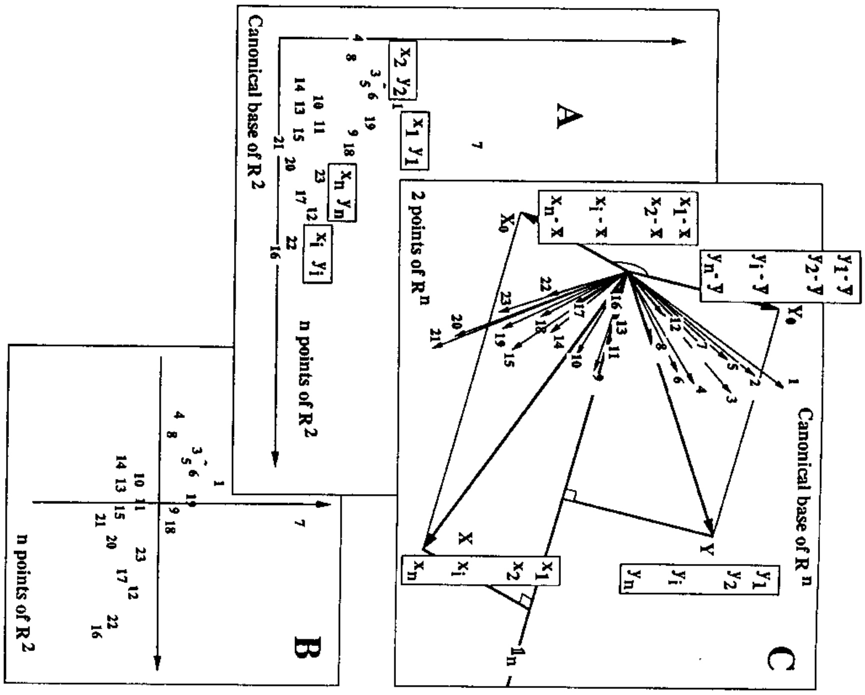


Figure 7

Thioulouse, J., J. Devillers, D. Chessel, and Y. Auda. 1991. Graphical techniques for multidimensional data analysis. Pages 153-205 in J. Devillers and W. Karcher, editors. Applied Multivariate Analysis in SAR and Environmental Studies. Kluwer Academic Publishers. http://pbil.univ-lyon1.fr/R/articles/arti067.pdf

## 2. Graphics and modeling of homogeneous tables.

The first goal of a multivariate approach, like PCA or CFA, is to condense the information of a data matrix on a graph. According to Ramsay (1989), the following model can be used to analyze the structure of a data table:

Data = Structural component + residual component + error of observation

These three elements must be separated from each other, and the approach combining multivariate methods and graphs leads to good results. This can be illustrated from the work of Slooff *et al.* (1983) and de Zwart and Slooff (1983) on the relative susceptibility of 15 taxonomical groups (table 1) to 11 chemicals (table 2).

Table 1. Identity of the organisms.

| N° | Species | Taxonomic level |
|---|---|---|
| 1 | *Photobacterium phosphoreum* | Bacteria |
| 2 | *Daphnia magna* | Crustacea |
| 3 | *Daphnia pulex* | Crustacea |
| 4 | *Daphnia cucullata* | Crustacea |
| 5 | *Aedes aegypti* | Insects |
| 6 | *Culex pipiens* | Insects |
| 7 | *Hydra oligactis* | Worms |
| 8 | *Lymnaea stagnalis* | Molluscs |
| 9 | *Leuciscus idus melanotus* | Fish |
| 10 | *Salmo gairdneri* | Fish |
| 11 | *Poecilia reticulata* | Fish |
| 12 | *Oryzias latipes* | Fish |
| 13 | *Pimephales promelas* | Fish |
| 14 | *Xenopus laevis* | Amphibia |
| 15 | *Ambystoma mexicanum* | Amphibia |

Table 2. Test chemicals.

| N° | Name |
|---|---|
| 1 | n-propanol |
| 2 | n-heptanol |
| 3 | Ethyl acetate |
| 4 | Acetone |
| 5 | Trichloroethylene |
| 6 | Allylamine |
| 7 | Aniline |
| 8 | Benzene |
| 9 | Pyridine |
| 10 | o-cresol |
| 11 | Salicylaldehyde |

From a statistical point of view, we have an homogeneous data matrix of 15 rows ( species) and 11 columns (i.e.; 11 chemicals) (table 3).

Table 3. Row data matrix (toxicities in log µmol/l).

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 5.49 | 2.24 | 4.82 | 5.7 | 2.95 | 2.46 | 3.72 | 3.48 | 4.43 | 2.15 | 2.07 |
| 5.02 | 2.75 | 3.83 | 5.43 | 2.85 | 2.83 | 0.84 | 3.71 | 4.14 | 1.94 | 1.68 |
| 4.7 | 2.62 | 3.47 | 5.18 | 2.53 | 2.77 | 0.03 | 3.59 | 3.86 | 1.95 | 1.65 |
| 4.99 | 2.86 | 3.27 | 5.12 | 2.64 | 2.69 | 0.86 | 3.68 | 4.49 | 2.18 | 1.65 |
| 4.86 | 3.14 | 3.6 | 5.41 | 2.56 | 3.32 | 3.22 | 3.41 | 3.22 | 2.87 | 2.12 |
| 4.9 | 3.02 | 4.65 | 5.47 | 2.62 | 3.48 | 3 | 2.96 | 2.92 | 2.63 | 2.65 |
| 5.05 | 3.14 | 4.19 | 5.37 | 2.76 | 2.49 | 3.64 | 2.64 | 4.16 | 2.84 | 1.76 |
| 5.03 | 2.54 | 4.1 | 5.08 | 2.63 | 1.94 | 3.93 | 3.47 | 3.65 | 3.17 | 1.73 |
| 4.91 | 2.44 | 3.49 | 5.23 | 3.21 | 2.96 | 2.72 | 3.23 | 3.39 | 2.22 | 1.43 |
| 4.73 | 2.57 | 3.47 | 5.11 | 2.5 | 2.42 | 2.66 | 2.86 | 3.85 | 2.08 | 1.04 |
| 5.05 | 2.74 | 3.38 | 5.22 | 3.14 | 2.32 | 3.03 | 3.73 | 4.24 | 2.55 | 1.63 |
| 4.99 | 2.62 | 3.15 | 5.39 | 3.31 | 2.45 | 3.25 | 3.51 | 4.29 | 2.58 | 1.54 |
| 4.92 | 2.47 | 3.49 | 5.41 | 2.55 | 1.57 | 2.84 | 3.03 | 3.16 | 2.5 | 1.54 |
| 4.82 | 2.58 | 3.31 | 5.62 | 2.53 | 1.94 | 3.78 | 3.39 | 4.25 | 2.55 | 1.8 |
| 4.82 | 2.65 | 3.22 | 5.54 | 2.56 | 1.5 | 3.67 | 3.68 | 4.08 | 2.57 | 1.79 |

It is possible to modify the raw data matrix (table 3) according to various procedı order to perform different types of Principal Component Analyses. Among the selected the following procedures:

- non centering (PCA on the matrix of the raw scalar products)
- column centering (PCA on covariance matrix)
- row centering ( PCA on covariance matrix of the transposed table)
- double centering (i.e.; row and column).

All these analyses can be carried out with a non centered PCA prograı transformations of the input table (general PCA from Lebart *et al.*, 1984). In ; following figures, the factor maps are represented in an orthonormal basis (i.e.; scaling is performed on the factor coordinates).

The first analysis dealing with the non centering procedure can be used when the dı homogeneous and expressed in the same unit (like here, in log µmol/l). It is perfı directly on table 3 without any transformation and leads to the graphical display of fi (F1 x F2 for columns (8A) and rows (8B)). Figure 8A shows an opposition on th axis between salicylaldehyde (compound 11) and the couple n-propanol (compound acetone (compound 4). The second axis underlines the particular behavior of a (compound 7). Figure 8B exhibits a strong cluster between *Daphnia magna* (org number 2), *Daphnia pulex* (organism number 3), and *Daphnia cucullata* (orgı number 4). Back to the raw data (table 3), one observes that this cluster is due tc sensitivity to aniline.

The second analysis, performed on table 4, leads to the graphical display of figure 9 F2 for columns (9A) and rows (9B)). From a statistical point of view, table 4 is obı by centering each column of table 3 (in MacMul, it can be computed by means ı "preparation" step in the case of a centered PCA). Figure 9A shows that a (compound number 7) is an outlier, having a high negative value on the first Allylamine (compound 6) is opposed to pyridine (compound 9) on the second axis. I

Thioulouse, J., J. Devillers, D. Chessel, and Y. Auda. 1991. Graphical techniques for multidimensional data analysis. Pages 153-205 in J. Devillers and W. Karcher, editors. Applied Multivariate Analysis in SAR and Environmental Studies. Kluwer Academic Publishers. http://pbil.univ-lyon1.fr/R/articles/arti067.pdf

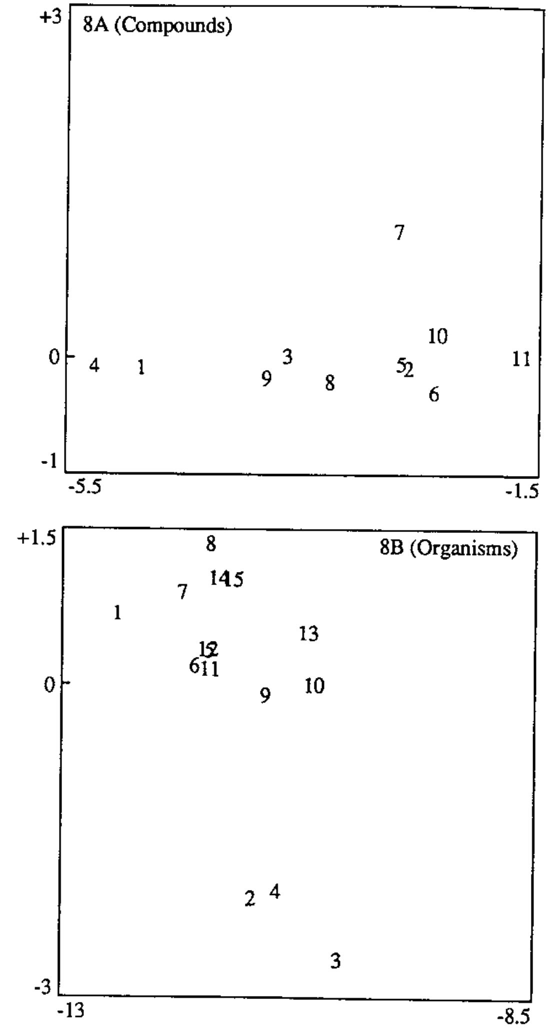


Figure 8: Factor maps of the first PCA (non centered analysis). 8A: Factor map of the 11 compounds. 8B: Factor map for the 15 organisms.

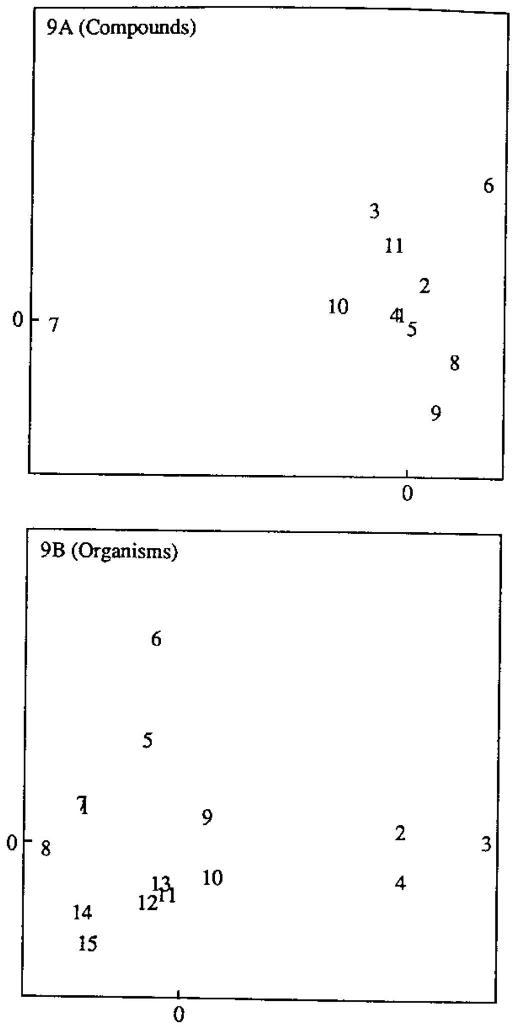


Figure 9: Factor maps of the second PCA (centering by compounds). 9A: Factor map of the 11 compounds. 9B: Factor map for the 15 organisms.

Table 4. Data matrix obtained from table 3 by column centering (i.e.; chemicals).

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.538 | -0.452 | 1.124 | 0.348 | 0.194 | -0.016 | 0.974 | 0.122 | 0.5547 | -0.302 | 0.3313 |
| 0.068 | 0.058 | 0.134 | 0.078 | 0.094 | 0.354 | -1.906 | 0.352 | 0.2647 | -0.512 | -0.0587 |
| -0.252 | -0.072 | -0.226 | -0.172 | -0.226 | 0.294 | -2.716 | 0.232 | -0.0153 | -0.502 | -0.0887 |
| 0.038 | 0.168 | -0.426 | -0.232 | -0.116 | 0.214 | -1.886 | 0.322 | 0.6147 | -0.272 | -0.0887 |
| -0.092 | 0.448 | -0.096 | 0.058 | -0.196 | 0.844 | 0.474 | 0.052 | -0.6553 | 0.418 | 0.3813 |
| -0.052 | 0.328 | 0.954 | 0.118 | -0.136 | 1.004 | 0.254 | -0.398 | -0.9553 | 0.178 | 0.9113 |
| 0.098 | 0.448 | 0.494 | 0.018 | 0.004 | 0.014 | 0.894 | -0.718 | 0.2847 | 0.388 | 0.0213 |
| 0.078 | -0.152 | 0.404 | -0.272 | -0.126 | -0.536 | 1.184 | 0.112 | -0.2253 | 0.718 | -0.0087 |
| -0.042 | -0.252 | -0.206 | -0.122 | 0.454 | 0.484 | -0.026 | -0.128 | -0.4853 | -0.232 | -0.3087 |
| -0.222 | -0.122 | -0.226 | -0.242 | -0.256 | -0.056 | -0.086 | -0.498 | -0.0253 | -0.372 | -0.6987 |
| 0.098 | 0.048 | -0.316 | -0.132 | 0.384 | -0.156 | 0.284 | 0.372 | 0.3647 | 0.098 | -0.1087 |
| 0.038 | -0.072 | -0.546 | 0.038 | 0.554 | -0.026 | 0.504 | 0.152 | 0.4147 | 0.128 | -0.1987 |
| -0.032 | -0.222 | -0.206 | 0.058 | -0.206 | -0.906 | 0.094 | -0.328 | -0.7153 | 0.048 | -0.1987 |
| -0.132 | -0.112 | -0.386 | 0.268 | -0.226 | -0.536 | 1.034 | 0.032 | 0.3747 | 0.098 | 0.0613 |
| -0.132 | -0.042 | -0.476 | 0.188 | -0.196 | -0.976 | 0.924 | 0.322 | 0.2047 | 0.118 | 0.0513 |

Table 5. Data matrix obtained from table 3 by row centering (i.e.; organisms).

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1.898 | -1.352 | 1.228 | 2.108 | -0.6418 | -1.132 | 0.1282 | -0.1118 | 0.8382 | -1.442 | -1.522 |
| 1.836 | -0.4336 | 0.6464 | 2.246 | -0.3336 | -0.3536 | -2.344 | 0.5264 | 0.9564 | -1.244 | -1.504 |
| 1.759 | -0.3209 | 0.5291 | 2.239 | -0.4109 | -0.1709 | -2.911 | 0.6491 | 0.9191 | -0.9909 | -1.291 |
| 1.86 | -0.27 | 0.14 | 1.99 | -0.49 | -0.44 | -2.27 | 0.55 | 1.36 | -0.95 | -1.48 |
| 1.43 | -0.29 | 0.17 | 1.98 | -0.87 | -0.11 | -0.21 | -0.02 | -0.21 | -0.56 | -1.31 |
| 1.418 | -0.4618 | 1.168 | 1.988 | -0.8618 | -0.0018 | -0.4818 | -0.5218 | -0.5618 | -0.8518 | -0.8318 |
| 1.592 | -0.3182 | 0.7318 | 1.912 | -0.6982 | -0.9682 | 0.1818 | -0.8182 | 0.7018 | -0.6182 | -1.698 |
| 1.642 | -0.8482 | 0.7118 | 1.692 | -0.7582 | -1.448 | 0.5418 | 0.0818 | 0.2618 | -0.2182 | -1.658 |
| 1.707 | -0.7627 | 0.2873 | 2.027 | 0.0073 | -0.2427 | -0.4827 | 0.0273 | 0.1873 | -0.9827 | -1.773 |
| 1.704 | -0.4564 | 0.4436 | 2.084 | -0.5264 | -0.6064 | -0.3664 | -0.1664 | 0.8236 | -0.9464 | -1.986 |
| 1.684 | -0.6264 | 0.0136 | 1.854 | -0.2264 | -1.046 | -0.3364 | 0.3636 | 0.8736 | -0.8164 | -1.736 |
| 1.619 | -0.7509 | -0.2209 | 2.019 | -0.0609 | -0.9209 | -0.1209 | 0.1391 | 0.9191 | -0.7909 | -1.831 |
| 1.876 | -0.5736 | 0.4464 | 2.366 | -0.4936 | -1.474 | -0.2036 | -0.0136 | 0.1164 | -0.5436 | -1.504 |
| 1.495 | -0.7445 | -0.0145 | 2.295 | -0.7945 | -1.385 | 0.4555 | 0.0655 | 0.9255 | -0.7745 | -1.525 |
| 1.54 | -0.63 | -0.06 | 2.26 | -0.72 | -1.78 | 0.39 | 0.4 | 0.8 | -0.71 | -1.49 |

9B reveals that the Daphnids (organisms number 2, 3, and 4) present a par ecotoxicological behavior. These three species have high positive coordinates on th axis. On the second axis, *Culex pipiens* (organism number 6) is opposed amphibians (organisms 14 and 15). A careful inspection of table 3 stresses the stru of the above figures. Aniline (chemical 7), mostly toxic towards *Daphnia magna* (org 2), *Daphnia pulex* (organism 3), and *Daphnia cucullata* (organism 4), explains the of these organisms, opposed to the other species. Allylamine (chemical 6) is the mos towards Amphibians (organisms 14 and 15) and the less towards *A. aegypti* (organi and *C. pipiens* (organism 6). The ecotoxicological behavior of pyridine (compoun opposed.

The third analysis, performed on table 5, leads to the graphical display of figure 10 F2 for columns (10A) and rows (10B)). The corresponding table (table 5) is obtain centering each row of table 3 (in MacMul, it can be computed by means "preparation" step in the case of a centered PCA on the transpose of table 3). The r are similar to those obtained with the first approach (see figures 8A and 8B).

The fourth analysis, performed on table 6, leads to the graphical display of figure 11 F2 for columns (11A) and rows (11B)). The results are similar to those obtained wi second approach (see figures 9A and 9B).

These different analyses (figures 8 to 11) show that more information can be extr from an homogeneous data matrix, by performing different centering procedures b PCA. Thus, the difference between figures 9 and 10 (i.e.; chemical centering organism centering) can be explained as follows:

- after centering by organisms, a "compound effect" remains in the data table, due differences of toxicity among the chemicals (i.e.; salicylaldehyde is much more towards all species than acetone and n-propanol). This is usually called a "size effe PCA, and is represented by the gradient of toxicity on the first axis of figure 10.

- after centering by chemicals , an "organism effect " remains in the data table, due t differences of sensitivity among species (i.e.; Daphnids are much more sensitive Molluscs to aniline and p-cresol and the converse is true for allylamine).

As the non-centered analysis is similar to the species-centered, we can conclude that i original data table, compound effect is predominant as compared to the species effect same conclusion can be derived from the similarity between the double centered an compound-centered analyses.

Confirmation of the above results can be obtained by means of graphical displays o data table itself. Figure 12 shows a collection of maps with circles and squares. O first map (figure 12A), the original data table was plotted with circles proportional t value of the raw toxicity (no centering and no standardization, chemicals in columns organisms in rows). The "compound effect" is obvious, low toxicity compounds appe columns of large circles, whereas highly toxic compounds appear as columns of s circles. This figure also underlines the fact that, would a "species effect" exist, completely hidden by the compound effect. Figure 12B is the same representation, bu the data table after centering by chemicals (this is the table on which the second PCA performed). The circles are proportional to positive values, and squares are proportion negative values. The compound effect has disappeared as expected, but the species e

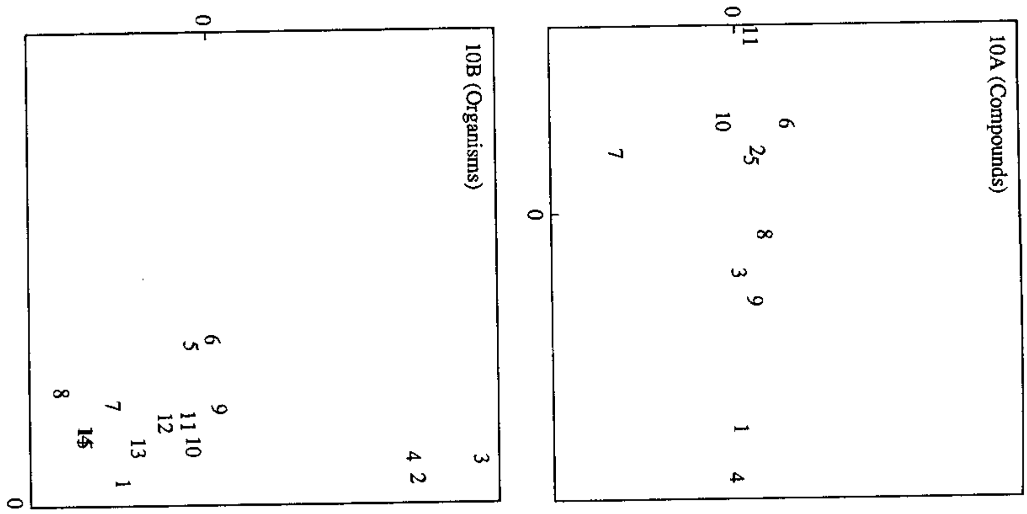


Figure 10: Factor maps of the third PCA (centering by organisms). 10A: Factor map of the 11 compounds. 10B: Factor map for the 15 organisms.

Table 6. Data matrix obtained from table 3 by double (row and column) centering.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.2275 | -0.7625 | 0.8135 | 0.0375 | -0.1165 | -0.3265 | 0.6635 | -0.1885 | 0.2441 | -0.6125 | 0.0208 |
| 0.1656 | 0.1556 | 0.2316 | 0.1756 | 0.1916 | 0.4516 | -1.808 | 0.4496 | 0.3623 | -0.4144 | 0.039 |
| 0.0884 | 0.2684 | 0.1144 | 0.1684 | 0.1144 | 0.6344 | -2.376 | 0.5724 | 0.325 | -0.1616 | 0.2517 |
| 0.1893 | 0.3193 | -0.2747 | -0.0807 | 0.0353 | 0.3653 | -1.735 | 0.4733 | 0.7659 | -0.1207 | 0.0626 |
| -0.2407 | 0.2993 | -0.2447 | -0.0907 | -0.3447 | 0.6953 | 0.3253 | -0.0967 | -0.8041 | 0.2693 | 0.2326 |
| -0.2525 | 0.1275 | 0.7535 | -0.0825 | -0.3365 | 0.8035 | 0.0535 | -0.5985 | -1.156 | -0.0225 | 0.7108 |
| -0.0789 | 0.2711 | 0.3171 | -0.1589 | -0.1729 | -0.1629 | 0.7171 | -0.8949 | 0.1078 | 0.2111 | -0.1556 |
| -0.0289 | -0.2589 | 0.2971 | -0.3789 | -0.2329 | -0.6429 | 1.077 | 0.0051 | -0.3322 | 0.6111 | -0.1156 |
| 0.0365 | -0.1735 | -0.1275 | -0.0435 | 0.5325 | 0.5625 | 0.0525 | -0.0495 | -0.4068 | -0.1535 | -0.2301 |
| 0.0329 | 0.1329 | 0.0289 | 0.0129 | -0.0011 | 0.1989 | 0.1689 | -0.2431 | 0.2296 | -0.1171 | -0.4438 |
| 0.0129 | -0.0371 | -0.4011 | -0.2171 | 0.2989 | -0.2411 | 0.1989 | 0.2869 | 0.2796 | 0.0129 | -0.1938 |
| -0.0516 | -0.1616 | -0.6356 | -0.0516 | 0.4644 | -0.1156 | 0.4144 | 0.0624 | 0.325 | 0.0384 | -0.2883 |
| 0.2056 | 0.0156 | 0.0316 | 0.2956 | 0.0316 | -0.6684 | 0.3316 | -0.0904 | -0.4777 | 0.2856 | 0.039 |
| -0.1753 | -0.1553 | -0.4293 | 0.2247 | -0.2693 | -0.5793 | 0.9907 | -0.0113 | 0.3314 | 0.0547 | 0.0181 |
| -0.1307 | -0.0407 | -0.4747 | 0.1893 | -0.1947 | -0.9747 | 0.9253 | 0.3233 | 0.2059 | 0.1193 | 0.0526 |

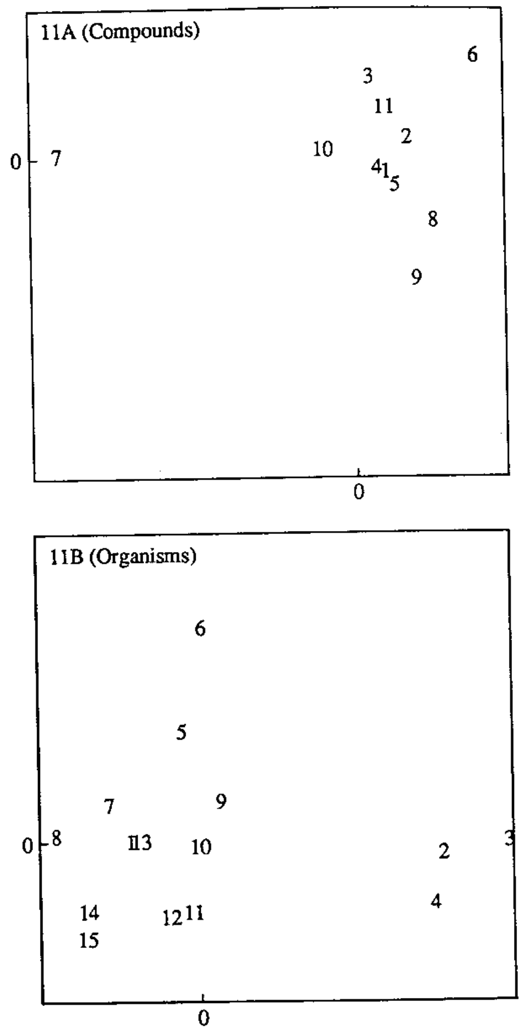


Figure 11: Factor maps of the first PCA (doubly centered analysis). 11A: Factor map of the 11 compounds. 11B: Factor map for the 15 organisms.

is still not very high. The only visible structure is the high sensitivity of the trout (*Sa gairdneri*, organism number 10) to most of the compounds (all the values are negat Figure 12C shows the table after double (rows and columns) centering. There is no r row or column effect, but interactions may still be present. This is clearly stressed b reorganization of the rows and columns which is performed in figure 12D: the rows columns on this figure are now positioned according to their coordinates on the first fa of the PCA (fourth analysis), but the values plotted are identical (i.e.; the doubly cent data table). The interaction between aniline and Daphnids is obvious. The next step find out whether other interactions subsist after removing the previous one, or not. can be achieved by computing the differences between the data set (doubly centered) the reconstitution of the data by the first factor of the PCA. The result is plotted in fi 12E, and shows another interaction between Insects and pyridine (as quoted abov figure 9). This procedure could be followed up by computing the difference betwee data set and the reconstitution of the data by the first two factors. The answer tc question "when should this procedure be stopped ?" is given in figure 12F: the decrea the following eigenvalues is slow and regular, thus indicating that the variability remai in the table is random, or may be attributed to observational error.

In addition to the possibilities offered by manipulations and graphical representatio the input table, we can extract still more information from data by using graphic improved factor maps. Thus, if we consider the results of the compound-centered PC good graphical technique consists in plotting the input data over the factorial map. way to achieve this purpose is to draw pictures on the factorial planes with s proportional to the toxicity. Their location on the graph is given by the coordinates o compound and species points. Squares and circles are used according to the sign o plotted values. The sign (+/-) can be recalled in the center of the picture.

On figure 13, we have represented a collection of 11 factorial maps for the species map per compound). On each map, the size of circles and squares is proportional to values of the corresponding column in table 4. The main advantage of this representati that it allows a direct interpretation of the factorial map without the necessity of loo back at the raw data matrix. Furthermore, one have a good view of the overall variat of the toxicity of each compound for the different species. For example, on figure 13 see that the cluster of Daphnids is due to their high sensitivity to aniline (large circles negative values), their same ecotoxicological behavior towards benzene (equal size pos squares), and so on.

Figure 14 is the reciprocal representation: we draw a collection of 15 factorial maps fo compounds (one map per species). Interpretation is similar to the above figure, bu comparison between species is more fruitful: one can observe for example the parti ecotoxicological behavior of *Photobacterium phosphoreum* and *Salmo gairdneri*.

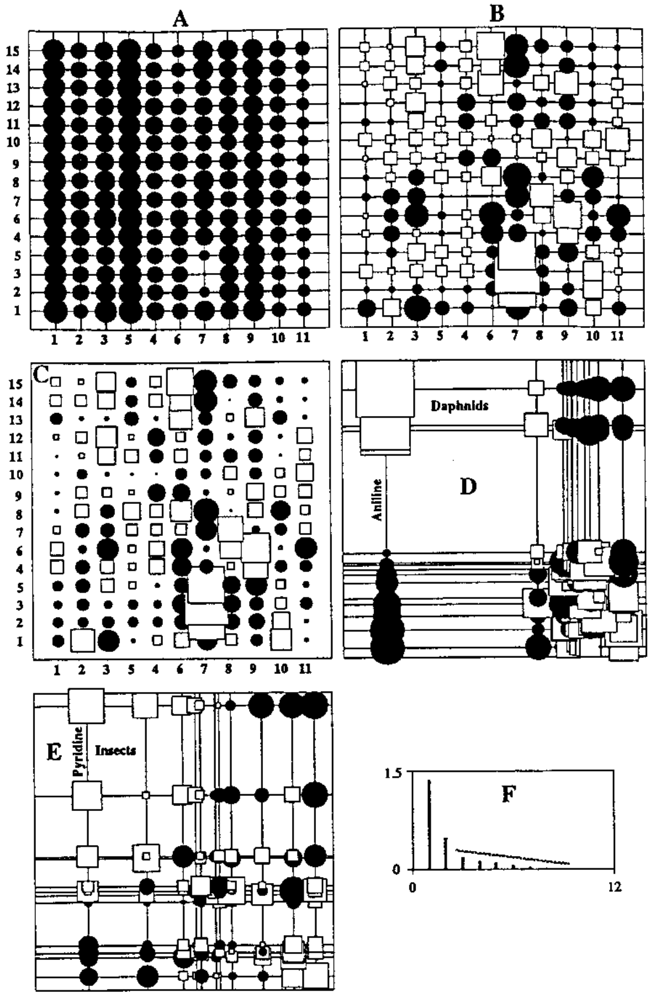


Figure 12: Graphical representations allowing to model the data table (15 organisms in rows, 11 compounds in columns). 12A: circles are proportional to the original data. 12B: centering by compound (circles are proportional to positive values, squares to negative). 12C: double centering. 12D: double centering with rows and columns positioned according to their factor scores. 12E: residuals after modeling by the first factor from PCA (rows and columns positioned according to their factor scores). 12F: eigenvalues.

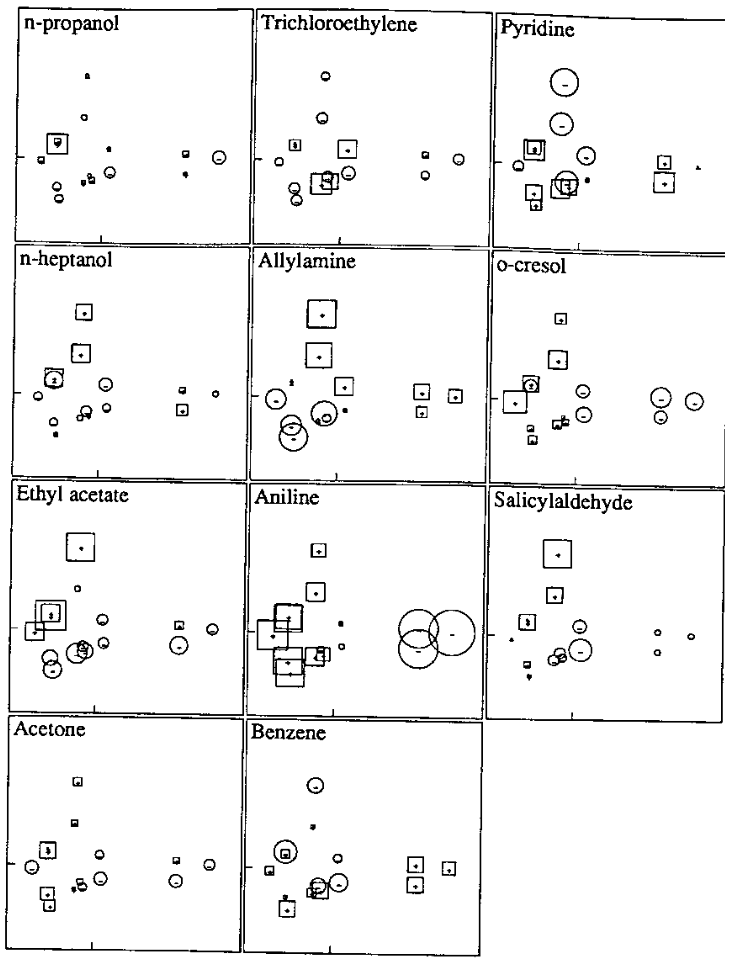


Figure 13: collection of 11 factorial maps (one for each compound). On each map, the centered value of the toxicity have been represented by a circle (negative values) or a square (positive values).

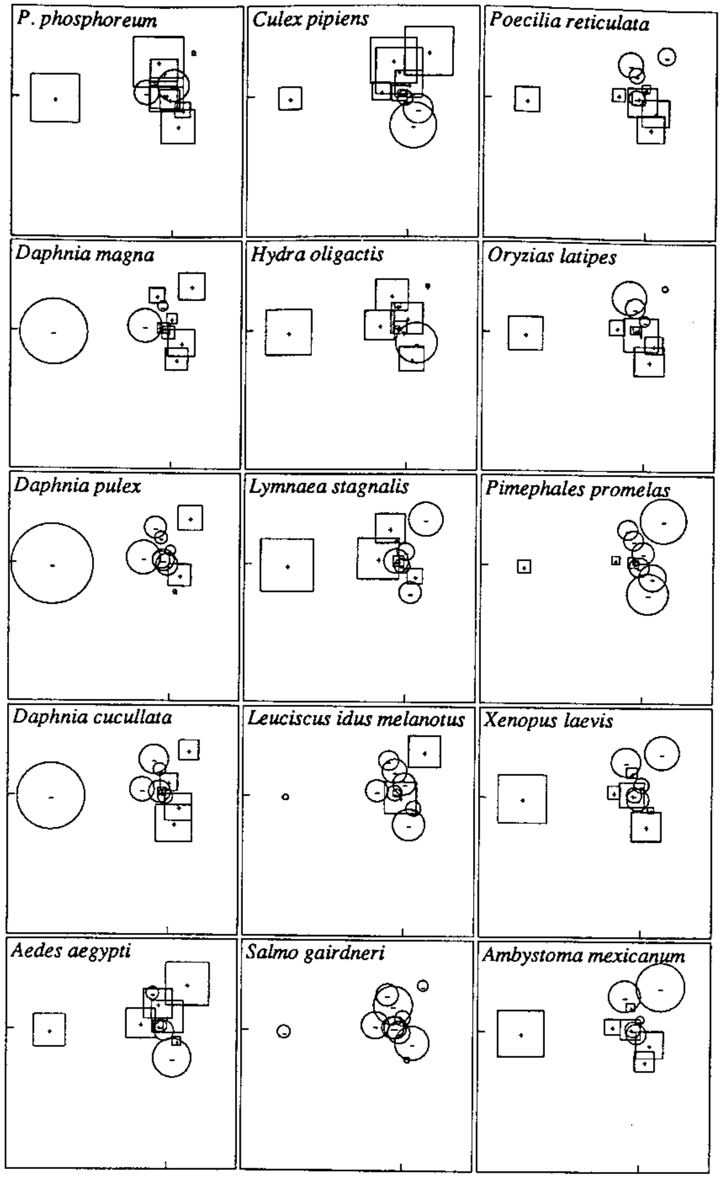


Figure 14: collection of 15 factorial maps (one for each organism). For caption, see figure 13.

## 3. Canonical graphs and duality in PCA.

In QSAR studies, the biological activity is traditionally related to the molecular des by means of multiple regression analyses (Dearden and Nicholson, 1986; Calam Vighi, 1987; Devillers *et al.*, 1988; Schultz, 1987; Nendza and Seydel, 1988; Ha al., 1989; Warne *et al.*, 1989). There are statistical drawbacks to using such a when for example, the molecular descriptors are too intercorrelated (Deville Chambon, 1989; Devillers *et al.*, 1989; Devillers and Lipnick, 1990). To overco problem it is possible to use PCA to define new independent variables which ar combinations of the original molecular descriptors. These variables, called pr components, can be introduced in regression analysis procedures. However, this s is powerful only if we can clearly interpret the components selected in the model. T of this paragraph is to introduce a new graphical method allowing to solve this probl The data used in this study were secured from Moulton and Schultz (1986). It conc para-substituted pyridines tested on *Tetrahymena pyriformis*. The toxicity value population growth inhibition) were expressed as the logarithm of the inverse of th concentration in moles/l (log BR). Each chemical was described by means of the fol six molecular descriptors: the corrected molar volume term of molar refractivity (M single bond fragment molecular connectivity index, $^1X^v{}_{sub}$ ; the hydrophobic para Pi; the para position Hammett electronic constant, σp; the field electronic parameter, the resonance electronic parameter, R. The values of the toxicity data and mol descriptors are listed in table 7.

Table 7. Molecular descriptors and biological response for 20 selected four-p pyridines.

| Derivative | MR | $^1X^v{}_{sub}$ | Pi | σp | F | R | log BR |
|---|---|---|---|---|---|---|---|
| H | 1.03 | 0 | 0 | 0 | 0 | 0 | -1.327 |
| $CH_3$ | 5.65 | 0.5 | 0.56 | -0.17 | -0.04 | -0.13 | -0.895 |
| $CH_2CH_3$ | 10.3 | 1.061 | 1.02 | -0.15 | -0.05 | -0.1 | -0.297 |
| Cl | 6.03 | 0.598 | 0.71 | 0.23 | 0.41 | -0.2 | -0.862 |
| Br | 8.88 | 0.35 | 0.86 | 0.23 | 0.44 | -0.17 | -0.31 |
| CN | 6.33 | 0.474 | -0.57 | 0.66 | 0.51 | 0.19 | -0.819 |
| $COCH_3$ | 11.18 | 0.704 | -0.55 | 0.5 | 0.32 | 0.2 | -0.835 |
| CHO | 6.88 | 0.525 | -0.65 | 0.42 | 0.31 | 0.13 | -0.159 |
| $COC_6H_5$ | 30.33 | 2.364 | 1.05 | 0.43 | 0.31 | 0.16 | -0.093 |
| $OCOCH_3$ | 12.47 | 0.816 | -0.64 | 0.31 | 0.41 | -0.07 | -0.814 |
| $NH_2$ | 5.47 | 0.289 | -1.23 | -0.66 | 0.02 | -0.68 | -0.439 |
| OH | 2.85 | 0.224 | -0.67 | -0.37 | 0.29 | -0.64 | -1.587 |
| $N(CH_3)_2$ | 15.55 | 0.816 | 0.18 | -0.83 | 0.1 | -0.92 | -0.635 |
| $CH_2OH$ | 7.19 | 0.57 | -1.03 | 0 | 0 | 0 | -1.329 |
| COOH | 6.93 | 0.678 | -0.32 | 0.45 | 0.33 | 0.15 | -1.386 |
| CHNOH | 10.28 | 0.458 | -0.38 | 0.1 | 0.25 | -0.13 | -0.547 |
| $CONH_2$ | 9.81 | 0.493 | -1.49 | 0.36 | 0.24 | 0.14 | -1.015 |
| $C_6H_5$ | 25.36 | 1.91 | 1.96 | -0.01 | 0.08 | -0.08 | 0.664 |
| $CH_2C_6H_5$ | 30.01 | 2.264 | 2.01 | -0.09 | -0.08 | -0.01 | 0.676 |
| $C(CH_3)_3$ | 19.62 | 1.5 | 1.98 | -0.2 | -0.07 | -0.13 | 0.164 |

A standardized PCA was first performed on the six molecular descriptors of table 7. The classical factorial maps are displayed in figure 15 (figure 15A: correlation circle; figure 15B: chemical map). The correlation circle shows that the first factor has a high correlation coefficient with Pi, MR, and $^1X^v{}_{sub}$, (i.e.; basically to the size of the molecules). The second factor is related to σp and R, and in a lower part to F, and can be interpreted as an electronic parameter. These results agree with those from Moulton and Schultz (1986). The chemical map confirms this interpretation, with the dichotomy between large and small molecules on the first axis, and the distribution of the chemicals on the second axis principally according to an electronic gradient.

In order to use these factors in regression analyses for predicting *Tetrahymena* toxicities, it is necessary to have an accurate idea of their relationships (which may be non linear) to each molecular descriptor. The best way to assess these relationships is to draw all the graphs representing the values of each molecular descriptor as a function of each factor. This method, called PCA canonical graph, is based on the property of components to maximize the correlation with the initial variables.

Figure 16 (A to F) is a collection of such graphs for the first factor and the six molecular descriptors under study (standardized values). Figure 16G shows the same relationship between the first factor and toxicity (log BR). The value of the linear correlation coefficient (r) and the corresponding significance probability (p) are listed in each elementary graph.

The same approach has been used for the second factor (figure 17, A to G). Figure 16 underlines positive linear correlations between the first principal component and the three following molecular descriptors: Pi, MR, and $^1X^v{}_{sub}$. This is not surprising if we consider the position of these molecular descriptors in the correlation circle (figure 15A). Figure 16G stresses the same type of relationship. Therefore, we can conclude that the first factor can be used as an explicative variable in regression analyses to summarize the physicochemical information encoded by Pi, MR, and $^1X^v{}_{sub}$. Figure 17 displays a strong correlation between the second component and σp, and significant correlations with F and R, but no relationship is found with the toxicity. As a conclusion, this factor is not relevant to predict the toxicity of the 20 pyridines on *Tetrahymena pyriformis*.

Figure 18 illustrates the geometrical point of view introduced in paragraph 1.5: the toxicity (log BR) is projected as an additional variable in the PCA. The corresponding geometric interpretation consists of the projection of the vector representing the toxicity (a vector of $\mathbb{R}^{20}$, with 20 coordinates) onto the plane defined by the first two principal components. On figure 18, the angles between vectors represent the correlation between the corresponding variables. The bundle of vectors corresponding to molecular descriptors Pi, MR, and $^1X^v{}_{sub}$ is very narrow and nearly parallel to the first principal component (and consequently nearly orthogonal to the second). The projection of the vector corresponding to toxicity falls among this bundle. The bundle of vectors corresponding to molecular descriptors F, σp and R is less narrow and is in opposite direction to the second principal component. This is an alternative way to see the relationships depicted in figures 16 and 17, and to state that the principal components maximize the sum of the squared correlations with all the variables.

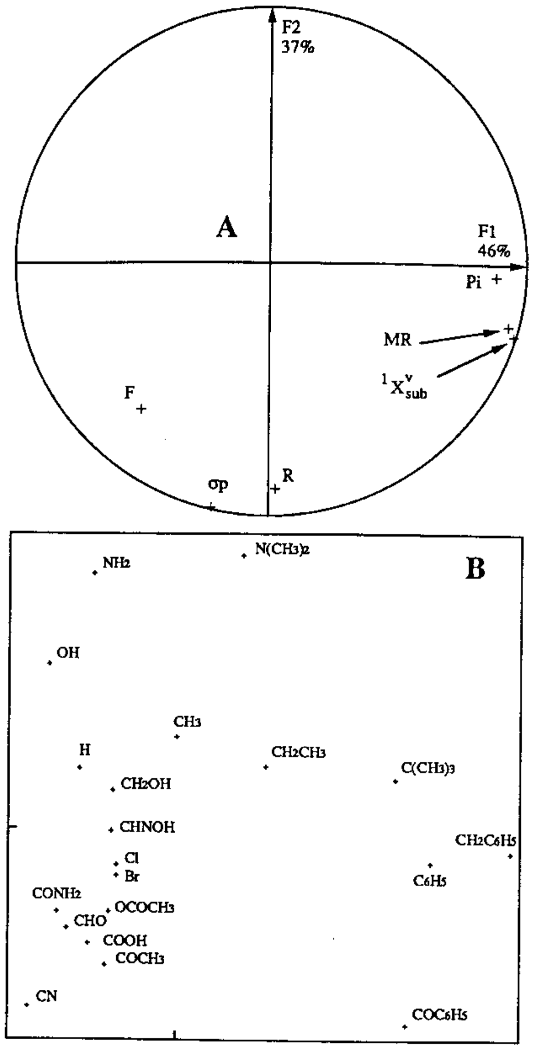


Figure 15: Results of the PCA on the molecular descriptors from table 10. 15A: correlation circle, 15B: chemical map.

Thioulouse, J., J. Devillers, D. Chessel, and Y. Auda. 1991. Graphical techniques for multidimensional data analysis. Pages 153-205 in J. Devillers and W. Karcher, editors. Applied Multivariate Analysis in SAR and Environmental Studies. Kluwer Academic Publishers. http://pbil.univ-lyon1.fr/R/articles/arti067.pdf

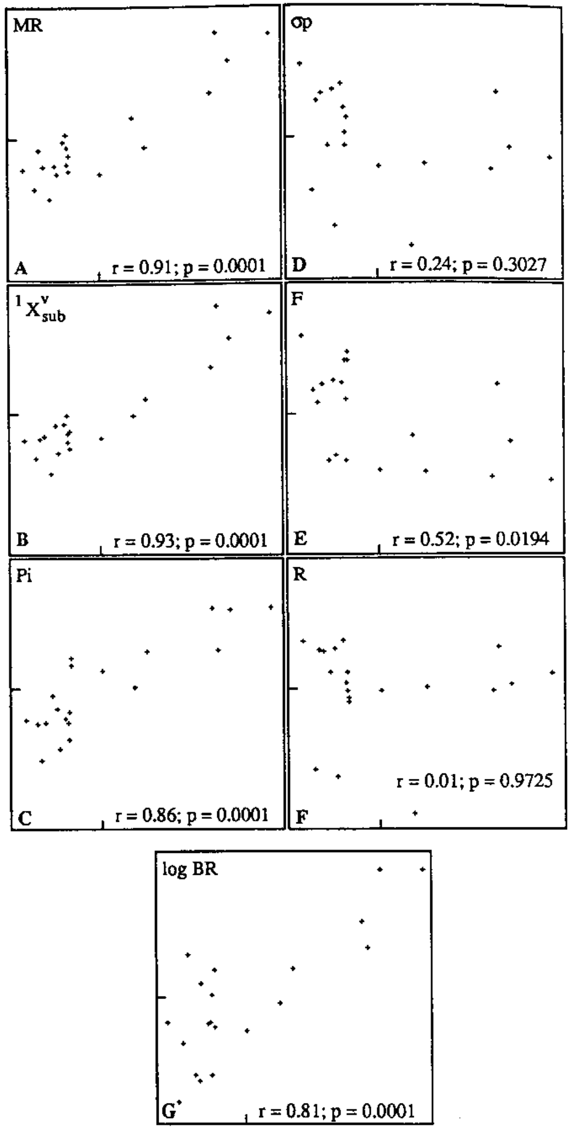


Figure 16: Canonical representation in PCA: the values of each standardized molecular descriptor (A to F) or toxicity (G) is plotted versus the first principal component.

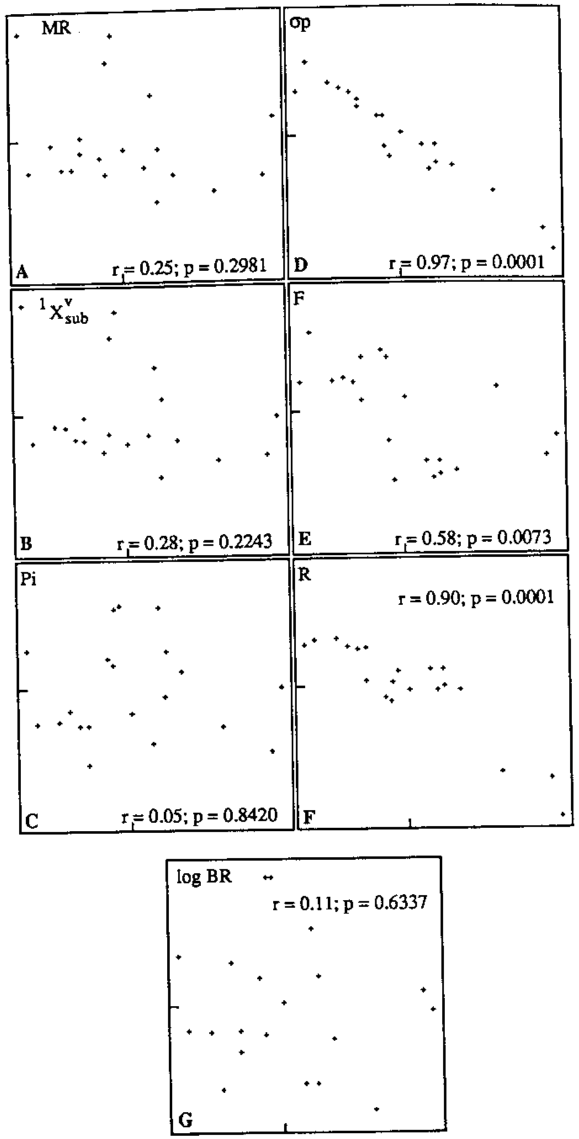


Figure 17: Canonical representation in PCA: the values of each standardized molecular descriptor (A to F) or toxicity (G) is plotted versus the second principal component.

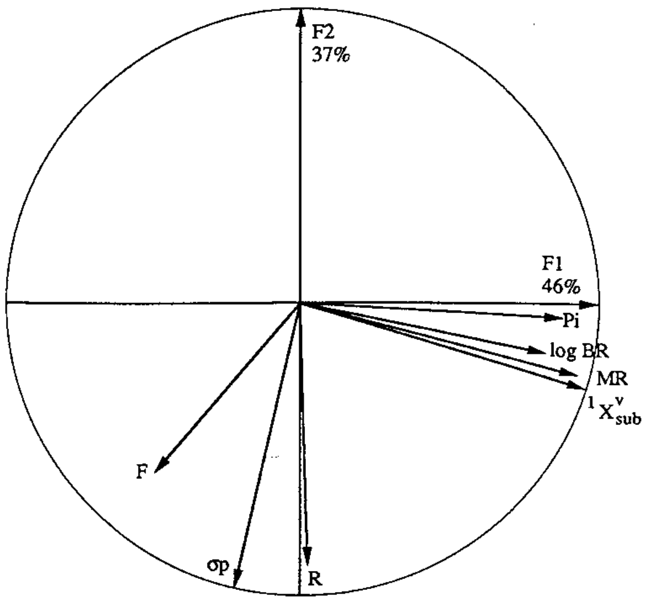


Figure 18: Correlation circle with toxicity (log BR) projected as additional element.

## 4. Discrimination, between-groups PCA, and graphical display.

Numerous structure-activity relationships (SAR) studies are based on discriminant ar (DA) (de Flora *et al.*, 1985; Benigni and Giuliani, 1987; Devillers *et al.*, 1987; Ens *al.*, 1988; Niemi *et al.*, 1987). This method is very powerful when the activiti boolean (e.g.; mutagenicity) and/or when they are not measured with accuracy. Hov the outputs of these analyses are often limited to numerical tables with a percent good classification, the discriminating power of the molecular descriptors and so o scope of this paragraph is principally to show the heuristic potency of the gra analysis in SAR. To reach this goal, a data matrix of 70 sweet (class 1), 17 tasteless 2), and 21 bitter (class 3) L-aspartyl dipeptides (L-Asp-NH-R), encoded by physicochemical descriptors (MR, L, $W_r$, $W_l$, $W_u$, $W_d$, and σ*, numbered from 1 t the maps), has been selected in the literature (Miyashita *et al.*, 1986). The molecular descriptors are listed in table 8. Table 9 gives the chemical stru configuration and taste quality of the dipeptides under study.

Table 8. Class of activity and molecular descriptors.

| no. | Class | MR | L | $W_r$ | $W_l$ | $W_u$ | $W_d$ | σ* |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 39.3 | 8.33 | 3.15 | 2.76 | 3.11 | 3.16 | 0.03 |
| 2 | 1 | 44 | 8.33 | 3.15 | 2.76 | 3.16 | 3.16 | -0.12 |
| 3 | 1 | 40.8 | 8.33 | 3.15 | 2.76 | 3.11 | 3.16 | 0.29 |
| 4 | 1 | 42.6 | 9.01 | 3.15 | 2.76 | 3.11 | 3.16 | 0.26 |
| 5 | 1 | 40.6 | 8.22 | 4.94 | 2.76 | 3.16 | 3.16 | -0.26 |
| 6 | 1 | 40.6 | 8.22 | 4.94 | 2.76 | 3.16 | 3.16 | -0.26 |
| 7 | 1 | 33.6 | 8.22 | 5.87 | 2.76 | 1.9 | 3.16 | -0.27 |
| 8 | 1 | 33.6 | 6.78 | 5.18 | 2.76 | 3.16 | 3.16 | -0.26 |
| 9 | 1 | 33.6 | 6.78 | 5.18 | 2.76 | 3.16 | 3.16 | -0.26 |
| 10 | 1 | 38 | 9.03 | 6.39 | 2.76 | 1.9 | 3.16 | -0.27 |
| 11 | 1 | 28.8 | 6.97 | 4.94 | 2.76 | 1.9 | 3.16 | -0.26 |
| 12 | 1 | 35.9 | 8.22 | 4.94 | 1.52 | 3.16 | 3.16 | -0.16 |
| 13 | 1 | 16.5 | 5.98 | 4.28 | 1.52 | 1.9 | 1.9 | 0.82 |
| 14 | 1 | 25.8 | 8.04 | 5.71 | 1.52 | 1.9 | 1.9 | 0.82 |
| 15 | 1 | 25.8 | 6.59 | 5.03 | 1.52 | 3.16 | 3.16 | 0.82 |
| 16 | 1 | 30.4 | 8.86 | 6.21 | 1.52 | 1.9 | 1.9 | 0.82 |
| 17 | 1 | 37.5 | 7.97 | 6.71 | 1.52 | 3.16 | 3.16 | 0.82 |
| 18 | 1 | 21.1 | 6.78 | 4.81 | 1.52 | 1.9 | 1.9 | 0.28 |
| 19 | 1 | 30.4 | 8.01 | 5.54 | 1.52 | 3.16 | 3.16 | 0.28 |
| 20 | 1 | 36 | 6.59 | 5.03 | 2.76 | 3.16 | 3.16 | 0.62 |
| 21 | 1 | 21.1 | 5.98 | 4.28 | 2.76 | 1.9 | 3.16 | 0.72 |
| 22 | 1 | 25.7 | 6.79 | 4.78 | 2.76 | 1.9 | 3.16 | 0.72 |
| 23 | 1 | 30.3 | 8.04 | 5.71 | 2.76 | 1.9 | 3.16 | 0.72 |
| 24 | 1 | 30.3 | 6.59 | 5.03 | 2.76 | 3.16 | 3.16 | 0.72 |
| 25 | 1 | 34.9 | 8.86 | 6.21 | 2.76 | 1.9 | 3.16 | 0.72 |
| 26 | 1 | 39.5 | 10.1 | 7.15 | 2.76 | 1.9 | 3.16 | 0.72 |
| 27 | 1 | 42.1 | 7.97 | 6.71 | 2.76 | 3.16 | 3.16 | 0.72 |
| 28 | 1 | 25.7 | 8.03 | 5.73 | 1.52 | 1.9 | 1.9 | 0.1 |
| 29 | 1 | 25.7 | 5.98 | 4.28 | 2.76 | 1.9 | 3.16 | 0.71 |

Table 8 (continued).

| no. | Class | MR | L | $W_r$ | $W_l$ | $W_u$ | $W_d$ | σ* |
|---|---|---|---|---|---|---|---|---|
| 30 | 1 | 44.3 | 6.59 | 5.03 | 2.76 | 3.16 | 3.16 | 0.59 |
| 31 | 1 | 35 | 8.04 | 5.71 | 2.76 | 1.9 | 3.16 | 0.71 |
| 32 | 1 | 35 | 6.59 | 5.03 | 2.76 | 3.16 | 3.16 | 0.71 |
| 33 | 1 | 30.4 | 6.17 | 4.42 | 3.42 | 1.9 | 4.29 | 0.69 |
| 34 | 1 | 39.7 | 8.04 | 5.71 | 3.63 | 1.9 | 3.99 | 0.7 |
| 35 | 1 | 39.7 | 6.59 | 5.03 | 3.63 | 3.16 | 3.99 | 0.7 |
| 36 | 1 | 44.3 | 6.59 | 5.03 | 3.63 | 3.16 | 3.99 | 0.6 |
| 37 | 1 | 39.7 | 8.04 | 5.71 | 3.49 | 1.9 | 4.41 | 0.69 |
| 38 | 1 | 35 | 6.97 | 4.94 | 3.42 | 1.9 | 4.29 | 0.66 |
| 39 | 1 | 40.7 | 6.97 | 4.94 | 3.42 | 1.9 | 4.29 | 0.66 |
| 40 | 1 | 36 | 6.59 | 5.03 | 3.49 | 3.16 | 4.41 | 0.7 |
| 41 | 1 | 37.5 | 6.97 | 4.94 | 3.42 | 2.52 | 4.29 | 0.98 |
| 42 | 1 | 37.5 | 6.97 | 4.94 | 3.42 | 1.9 | 4.29 | 0.98 |
| 43 | 1 | 44.3 | 9.03 | 6.39 | 3.42 | 1.9 | 4.29 | -0.1 |
| 44 | 1 | 46.5 | 8.33 | 3.15 | 3.42 | 3.11 | 4.29 | 0.9 |
| 45 | 1 | 46.8 | 8.22 | 4.94 | 3.42 | 3.16 | 4.29 | 0.66 |
| 46 | 1 | 49.8 | 9.05 | 3.15 | 3.42 | 3.11 | 4.29 | 0.87 |
| 47 | 1 | 54 | 9.01 | 3.15 | 3.42 | 3.11 | 4.29 | 0.87 |
| 48 | 1 | 41.2 | 8.04 | 5.51 | 3.42 | 3.16 | 4.29 | 1.06 |
| 49 | 1 | 32.3 | 6.85 | 5.56 | 3.42 | 3.16 | 4.29 | 1.06 |
| 50 | 1 | 22.6 | 5.98 | 4.28 | 2.76 | 1.9 | 3.16 | 1.03 |
| 51 | 1 | 27.2 | 6.79 | 4.78 | 2.76 | 1.9 | 3.16 | 1.03 |
| 52 | 1 | 31.8 | 8.04 | 5.71 | 2.76 | 1.9 | 3.16 | 1.03 |
| 53 | 1 | 31.8 | 6.59 | 5.03 | 2.76 | 3.16 | 3.16 | 1.03 |
| 54 | 1 | 36.5 | 8.86 | 6.21 | 2.76 | 1.9 | 3.16 | 1.03 |
| 55 | 1 | 36.5 | 8.04 | 5.51 | 2.76 | 3.16 | 3.16 | 1.03 |
| 56 | 1 | 43.7 | 7.97 | 6.71 | 2.76 | 3.16 | 3.16 | 1.03 |
| 57 | 1 | 36 | 6.36 | 4.69 | 3.42 | 1.9 | 4.29 | 1.1 |
| 58 | 1 | 36.6 | 5.98 | 4.28 | 2.76 | 1.9 | 3.77 | 1.04 |
| 59 | 1 | 37.6 | 8.04 | 5.71 | 2.76 | 1.9 | 3.77 | 0.98 |
| 60 | 1 | 37.6 | 6.59 | 5.03 | 2.76 | 3.16 | 3.77 | 0.98 |
| 61 | 1 | 28.3 | 5.98 | 4.28 | 3.55 | 1.9 | 3.16 | 0.98 |
| 62 | 1 | 37.6 | 8.04 | 5.71 | 3.55 | 1.9 | 3.16 | 0.98 |
| 63 | 1 | 37.6 | 6.59 | 5.03 | 3.55 | 3.16 | 3.16 | 0.98 |
| 64 | 1 | 48.9 | 10.1 | 7.22 | 3.42 | 1.9 | 4.29 | 0.83 |
| 65 | 1 | 44.3 | 8.88 | 6.61 | 3.42 | 1.9 | 4.29 | 0.84 |
| 66 | 1 | 36.6 | 8.03 | 5.73 | 3.42 | 1.9 | 4.29 | 0.91 |
| 67 | 1 | 43.7 | 7.5 | 6.03 | 3.42 | 4.05 | 4.29 | - 1.64 |
| 68 | 1 | 48.4 | 7.97 | 6.71 | 3.42 | 3.16 | 4.29 | 1.64 |
| 69 | 1 | 53 | 7.97 | 6.71 | 3.42 | 3.16 | 4.29 | 1.64 |
| 70 | 1 | 64.8 | 8.22 | 6.49 | 3.42 | 4.41 | 4.29 | 1.64 |
| 71 | 2 | 39.3 | 8.33 | 3.15 | 2.76 | 3.16 | 3.11 | 0.03 |
| 72 | 2 | 34.7 | 8.33 | 3.15 | 1.52 | 3.11 | 3.11 | 0.08 |
| 73 | 2 | 24.3 | 6.17 | 4.42 | 2.76 | 3.16 | 1.9 | -0.21 |
| 74 | 2 | 24.3 | 6.17 | 4.42 | 2.76 | 1.9 | 3.16 | -0.26 |
| 75 | 2 | 33.4 | 8.22 | 5.87 | 2.76 | 3.16 | 1.9 | -0.27 |

Table 8 (continued).

| no. | Class | MR | L | $W_r$ | $W_l$ | $W_u$ | $W_d$ | σ* |
|---|---|---|---|---|---|---|---|---|
| 76 | 2 | 26.4 | 5.98 | 4.28 | 2.76 | 3.16 | 1.9 | 0.71 |
| 77 | 2 | 43.4 | 8.33 | 3.15 | 2.54 | 3.11 | 3.41 | 0.73 |
| 78 | 2 | 48.3 | 8.33 | 5.2 | 3.42 | 4.29 | 1.9 | 0.91 |
| 79 | 2 | 47.3 | 8.33 | 5.2 | 3.42 | 4.29 | 1.9 | 0.87 |
| 80 | 2 | 55.1 | 10.3 | 5.2 | 3.42 | 2.87 | 4.29 | 0.99 |
| 81 | 2 | 60.2 | 10.3 | 6.23 | 3.42 | 2.87 | 4.29 | 0.99 |
| 82 | 2 | 22.6 | 5.98 | 4.28 | 2.76 | 3.16 | 1.9 | 1.03 |
| 83 | 2 | 38.6 | 7.42 | 4.98 | 3.42 | 4.29 | 1.9 | 0.07 |
| 84 | 2 | 33.6 | 6.17 | 4.42 | 3.42 | 2.52 | 4.29 | 1.31 |
| 85 | 2 | 48.3 | 9.01 | 3.15 | 3.42 | 4.29 | 3.11 | 0.87 |
| 86 | 2 | 28.3 | 5.98 | 4.28 | 3.42 | 3.16 | 1.9 | 0.98 |
| 87 | 2 | 45.2 | 9.01 | 3.15 | 2.54 | 3.11 | 3.41 | 0.7 |
| 88 | 3 | 52.8 | 11.1 | 4.56 | 2.76 | 3.42 | 3.16 | 0.08 |
| 89 | 3 | 28.9 | 6.17 | 4.21 | 2.76 | 3.16 | 3.16 | -0.26 |
| 90 | 3 | 28.9 | 6.17 | 4.21 | 2.76 | 3.16 | 3.16 | -0.26 |
| 91 | 3 | 21.7 | 5.98 | 4.28 | 2.76 | 3.16 | 1.9 | 0.72 |
| 92 | 3 | 31 | 5.98 | 4.28 | 3.63 | 3.99 | 1.9 | 0.7 |
| 93 | 3 | 36 | 6.79 | 4.78 | 3.49 | 4.41 | 1.9 | 0.69 |
| 94 | 3 | 35.1 | 6.17 | 4.21 | 3.42 | 3.16 | 4.29 | 0.66 |
| 95 | 3 | 45.3 | 6.59 | 5.03 | 3.7 | 3.16 | 4.41 | 0.66 |
| 96 | 3 | 35 | 6.17 | 4.42 | 3.42 | 3.16 | 4.29 | 0.66 |
| 97 | 3 | 49.8 | 9.03 | 6.39 | 3.42 | 1.9 | 4.29 | 0.66 |
| 98 | 3 | 40.9 | 8.33 | 3.15 | 2.56 | 3.11 | 3.45 | 1.13 |
| 99 | 3 | 48.2 | 8.33 | 3.15 | 3.49 | 3.11 | 4.42 | 0.84 |
| 100 | 3 | 52.6 | 8.33 | 3.15 | 3.44 | 3.11 | 4.32 | 0.84 |
| 101 | 3 | 46.5 | 8.33 | 3.15 | 3.42 | 4.29 | 3.11 | 0.9 |
| 102 | 3 | 47.4 | 8.04 | 5.51 | 3.42 | 3.16 | 4.29 | 1.12 |
| 103 | 3 | 40.4 | 8.13 | 5.78 | 3.42 | 1.9 | 4.29 | 0.83 |
| 104 | 3 | 56.6 | 8.67 | 4.58 | 3.97 | 4.29 | 1.9 | 0.82 |
| 105 | 3 | 56.6 | 8.67 | 4.58 | 3.97 | 1.9 | 4.29 | 0.82 |
| 106 | 3 | 42.7 | 9 | 3.15 | 2.56 | 3.11 | 3.45 | 1.1 |
| 107 | 3 | 50 | 9.01 | 3.15 | 3.49 | 3.11 | 4.42 | 0.81 |
| 108 | 3 | 56.5 | 9.01 | 3.15 | 3.44 | 3.11 | 4.32 | 0.81 |

Table 9: Chemical structure, configuration and taste quality of 108 *L*-aspartyl dipeptides (L-Asp-NH-R)

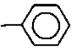
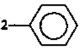
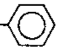
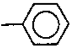
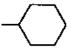
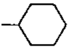
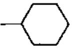
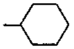
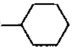
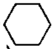
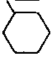
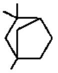
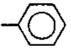
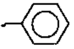
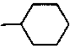
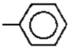
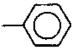
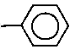
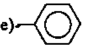
| no. NH-R | Config. | Taste | no. NH-R | Config. | Taste |
|---|---|---|---|---|---|
| 1 -CH(Me)-CH2—  | L | sweet | 21 Ala-OMe | D | sweet |
| 2 -C(Me)2-CH2—  |  | sweet | 22 Ala-OEt | D | sweet |
| 3 -CH(CH2OH)-CH2—  | L | sweet | 23 Ala-OPr $^n$ | D | sweet |
| 4 -CH(CH2OH)CH2-(p-OH)—  | L | sweet | 24 Ala-OPr $^i$ | D | sweet |
| 5 -CH(Me)-CH2-—  | L | sweet | 25 Ala-OBu $^n$ | D | sweet |
| 6 -CH(Me)-CH2-—  | D | sweet | 26 Ala-Pe $^n$ | D | sweet |
| 7 -CH(Me)-Pe$^n$ | L | sweet | 27 Ala-O—  | D | sweet |
| 8 -CH(Me)-Pe $^i$ | L | sweet | 28 γ-Abu-OMe |  | sweet |
| 9 -CH(Me)-Pe $^i$ | D | sweet | 29 Abu-OMe | D | sweet |
| 10 -CH(Me)-He$^n$ | L | sweet | 30 (α-Et)-Abu-OPr $^i$ |  | sweet |
| 11 -CH(Me)-Bu$^n$ | L | sweet | 31 Abu-OPr $^n$ | D | sweet |
| 12 -CH2-CH2-—  |  | sweet | 32 Abu-OPr $^i$ | D | sweet |
| 13 Gly-OMe |  | sweet | 33 NVa-OMe | L | sweet |
| 14 Gly-OPr $^n$ |  | sweet | 34 Val-OPr $^n$ | D | sweet |
| 15 Gly-OPr $^i$ |  | sweet | 35 Val-OPr $^i$ | D | sweet |
| 16 Gly-OBu $^n$ |  | sweet | 36 (α-Me)-Val-OPr $^i$ | D | sweet |
| 17 Gly-O—  |  | sweet | 37 NVa-OPr $^n$ | D | sweet |
| 18 β-Ala-OMe |  | sweet | 38 NLe-OMe | L | sweet |
| 19 β-Ala-OPr $^i$ |  | sweet | 39 NLe-OEt | L | sweet |
| 20 (α-Me)Ala-OPr $^i$ |  | sweet | 40 ILe-OPr $^i$ | D | sweet |

Table 9 (continued)

| no. NH-R | Config. | Taste | no. NH-R | Config. | T |
|---|---|---|---|---|---|
| 41 HyNle-OMe (erythro) | L | sweet | 61 *a*Thr-OMe | D | sv |
| 42 HyNle-OMe (threo) | L | sweet | 62 *a*Thr-OPr $^n$ | D | sv |
| 43 Cap-OMe | L | sweet | 63 *a*Thr-OPr $^i$ | D | sv |
| 44 Phe-OMe | L | sweet | 64 Lys(Ac)-OMe | L | sv |
| 45 HPhe-OMe | L | sweet | 65 Orn(Ac)-OMe | L | sv |
| 46 (*p*-NH2)-Phe-OMe | L | sweet | 66 Glu-(OMe)2 | L | sv |
| 47 Tyr-OMe | L | sweet | 67 Ama(O )-OMe |  | sv |
| 48 Ser(But$^i$)-OMe | L | sweet | 68 Ama(O  )-OMe |  | sv |
| 49 Ser(Bu$^t$)-OMe | L | sweet | 69 Ama(O  )-OMe |  | sv |
| 50 Ser-OMe | D | sweet | 70 μma(O  )-OMe |  | sv |
| 51 Ser-OEt | D | sweet | 71 -CH(Me)-CH2—  | D | ta |
| 52 Ser-OPr $^n$ | D | sweet | 72 -CH2-CH2—  |  | ta |
| 53 Ser-OPr $^i$ | D | sweet | 73 -CH(Me)-Pr $^n$ | D | ta |
| 54 Ser-OBu $^n$ | D | sweet | 74 -CH(Me)-Pr $^n$ | L | ta |
| 55 Ser-OBu $^i$ | D | sweet | 75 -CH(Me)-Pe$^n$ | D | ta |
| 56 Ser-O—  | D | sweet | 76 Abu-OMe | L | ta |
| 57 Met-OMe | L | sweet | 77 Phe-NH2 | L | ta |
| 58 Thr-OMe | D | sweet | 78 -CH(COOMe)-CH2-(*m*-OH)-—  | D | ta |
| 59 Thr-OPr $^n$ | D | sweet | 79 -CH(COOMe)-CH2-(*o*-OH)—  | D | ta |
| 60 Thr-OPr $^i$ | D | sweet | 80 -CH(COOMe)-CH2-(*m*-OH, *p*-OMe)—  | L | ta |

Thioulouse, J., J. Devillers, D. Chessel, and Y. Auda. 1991. Graphical techniques for multidimensional data analysis. Pages 153-205 in J. Devillers and W. Karcher, editors. Applied Multivariate Analysis in SAR and Environmental Studies. Kluwer Academic Publishers. http://pbil.univ-lyon1.fr/R/articles/arti067.pdf

Table 9 (continued)

| no. NH-R | Config. | Taste | no. NH-R | Config. | Taste |
|---|---|---|---|---|---|
| 81 -CH(COOMe)-CH2-(*m*, *p*-OMe)-  | L | tasteless | 95 Leu-OPr $^{i}$ | D | bitter |
| 82 Ser-OMe | L | tasteless | 96 ILeu-OMe | L | bitter |
| 83 Met-OMe | D | tasteless | 97 Cap-OEt | L | bitter |
| 84 Cys(Me(O$_2$))-OMe | L | tasteless | 98 Phe | L | bitter |
| 85 Tyr-OMe | D | tasteless | 99 Phe-NHMe | L | bitter |
| 86 Thr-OMe | L | tasteless | 100 Phe-N(Me)$_2$ | L | bitter |
| 87 Tyr-NH$_2$ | L | tasteless | 101 Phe-OMe | D | bitter |
| 88 -CH(Me)-CH$_2$- -NH-SO$_2$-CH$_3$ | L | bitter | 102 Thr(COPr $^{i}$)-OMe | L | bitter |
| 89 -CH(Me)-Bu $^{i}$ | D | bitter | 103 Lys-OMe | L | bitter |
| 90 -CH(Me)-Bu $^{i}$ | L | bitter | 104 Trp-OMe | L | bitter |
| 91 Ala-OMe | L | bitter | 105 Trp-OMe | D | bitter |
| 92 Val-OMe | L | bitter | 106 Tyr | L | bitter |
| 93 NVal-OEt | L | bitter | 107 Tyr-NHMe | L | bitter |
| 94 Leu-OMe | L | bitter | 108 Tyr-N(Me)$_2$ | L | bitter |

Other abbreviations used (Ariyoshi, 1976) : Pr $^{n}$, *n*-propyl; Pr $^{i}$, *i*-propyl; Bu $^{n}$, *n*-butyl; But $^{i}$, *i*-butyroyl; Pe $^{n}$, *n*-pentyl; Pe $^{i}$, *i*-pentyl; He $^{n}$, *n*-hexyl; Bu $^{t}$, *tert*-butyl; Cap, capryline = α-aminooctanoic acid; *a*Thr, allothreonine; HyNle, β-hydroxynorleucine; HPhe, β-cyclohexyl-α-alanine.

The classical display of the results obtained with discriminant analysis is shown in fig 19. It presents the factorial maps (F1xF2) for the three classes of activity (19A, 1' 19C), and for the three groups superimposed (19D). One can already notice that discrimination between classes is not accurate. From a SAR point of view, it is interesi to note the three groups of outliers in figure 19A (chemicals number 1 to 4; 14, 16, and 44, 46, and 47). To improve this display, and find out which molecular descriptor has best discriminating power, it is possible to look at the distribution of each molec descriptor, and compare it to the distribution of the discriminant functions built by analysis, as well as those computed from other analyses. Figure 20 summarizes th information: the Gauss curves are used to model the distribution of each molec descriptor. The correlation ratio for each descriptor is listed in each graph. It is easy to that only descriptors $W_r$, $W_u$, and $W_d$ provide a discrimination between sweet and sweet compounds. The curves for discriminant functions show that only the first fa has a good discriminating power, as opposed to plain PCA.

We recently showed (Dolédec and Chessel, 1987) the interest of within groups between groups PCA in the statistical analysis of physicochemical data. For compari purposes, the "between groups" method has been used on the data matrix under st (table 8). The distribution of the first two factors, modeled with Gauss curves in figure are very close from those of the discriminant functions. Thus, it was not useful to perf the discriminant analysis, especially if we consider that a simple PCA program allow carry out these two particular PCA: the "between" analysis consists of a PCA on the t of the means per class (table 10), followed by a projection of the rows of the standardi table as additional elements, and the within analysis consists of a PCA on the differe between the actual values and the means per classes.

Table 10. Means of molecular descriptors (columns) for the three taste classes (rows).

| Class | MR | L | $W_r$ | $W_l$ | $W_u$ | $W_d$ | σ* |
|---|---|---|---|---|---|---|---|
| 1 | -0.1526 | -0.0607 | 0.2995 | -0.1622 | -0.0312 | 0.1193 | 0.0403 |
| 2 | 0.0130 | 0.1006 | -0.4733 | -0.0086 | 0.6395 | -0.7408 | -0.2193 |
| 3 | 0.4983 | 0.1208 | -0.6151 | 0.5478 | 0.5230 | 0.2022 | 0.0433 |

A confirmation of these interpretations can be readily obtained by plotting the value each molecular descriptor for all chemicals. Figure 21 displays the values of table 8 a standardization. For $W_r$, the difference between sweet and non-sweet is obvious: alr all sweet chemicals have positive values, and non-sweet negative ones. Moreover, outliers detected in figure 19A are clearly visible: chemicals number 1, 2, 3, 4, 44, 46, 47 have highly negative values whereas they are sweet. For molecular descriptor numb ($W_u$), it is interesting to note that extreme values (1.90 or 4.29 in table 8) ha discriminating power (sweet compounds have low values and tasteless or bitter have l values) while mean values (3.16 in table 8) have not (sweet, tasteless and b compounds may have intermediate values).

Miyashita *et al.* (1986) used the SIMCA method (Wold and Sjöström, 1977) to fi structure-taste relationship on the 108 compounds under study (tables 8 and 9). SIMCA method is another variation of PCA, and their results are very close from tho the above PCA. Therefore we can state that all the above multivariate analyses are not to clearly separate the three classes of activity. The only undoubted discrimina concerns sweet and non-sweet.

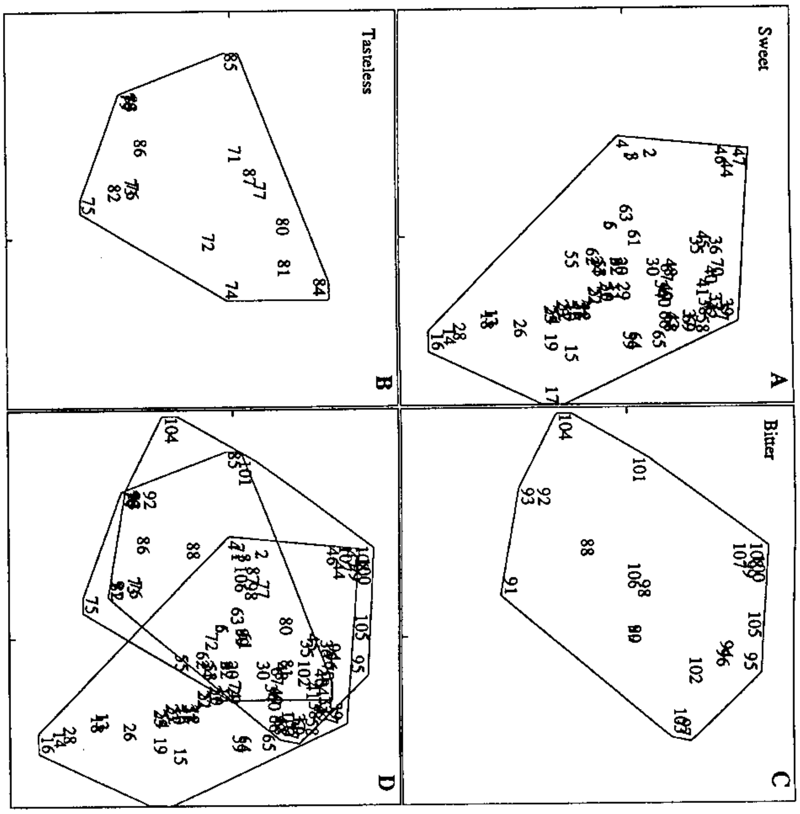


Figure 19: Factor map of the discriminant analysis. The cloud corresponding to the three taste classes have been represented separately (A, B, and C), and superimposed (D).

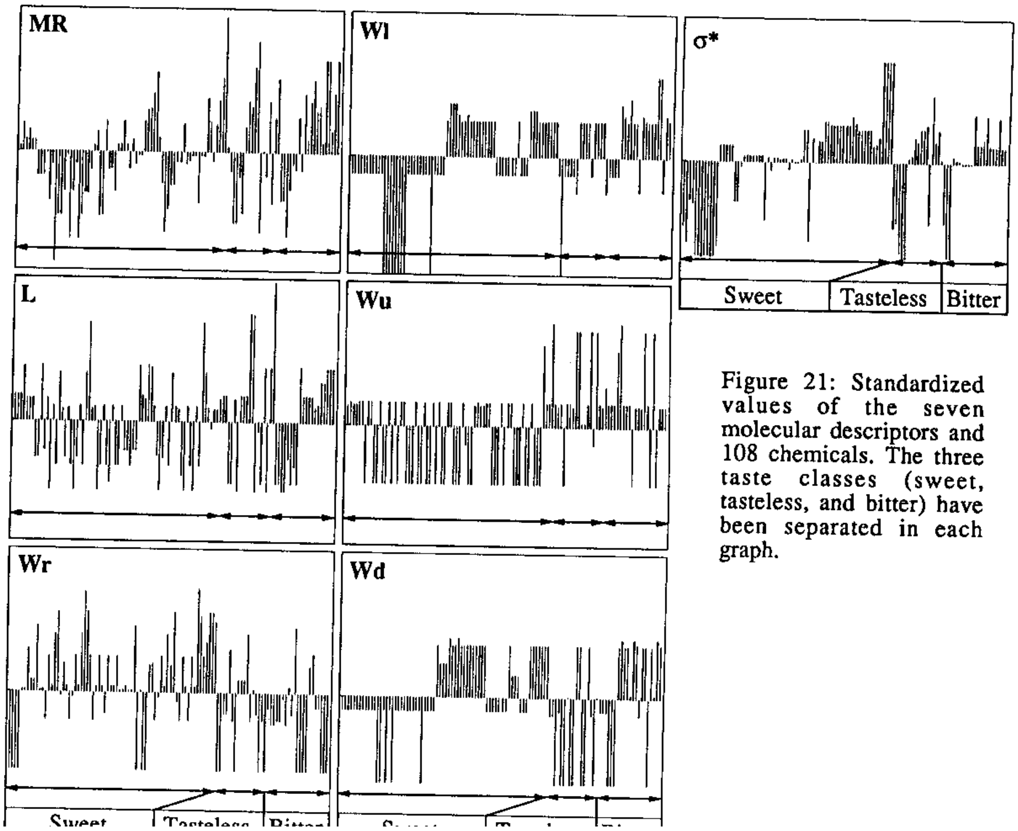


Figure 21: Standardized values of the seven molecular descriptors and 108 chemicals. The three taste classes (sweet, tasteless, and bitter) have been separated in each graph.

## 5. Graphical display of three-ways table analyses.

Several data analysis methods have been developed to study three-way tables. Among them, the French school of data analysis has presented the STATIS method (L'Hermier des plantes, 1976; Escoufier, 1980; Lavit, 1988). Three-mode principal component analysis (Tucker, 1966; Kroonenberg, 1983, 1989) is another approach of this question. Partial triadic analysis (Thioulouse and Chessel, 1987), derived from triadic analysis (Jaffrenou, 1978), is suitable for quantitative variables, when there is no missing data. The work of Jenkins and Buikema (1990) on the response of a winter plankton food web to simazine offers a good example of three-way table on which partial triadic analysis can be applied. The data set consists of four physicochemical parameters (temperature, pH, dissolved oxygen, and nitrate concentration) measured four times with an interval of seven days between measurements, in four microcosms experimentally contaminated by three concentrations of simazine plus one control. Data are listed in table 11.

Table 11. Physicochemical parameters affected by simazine and time (in days).

| Days | D0 | D7 | D14 | D21 | Simazine (mg/l) |
|---|---|---|---|---|---|
| Temp.(°C) | 9.10 | 2.20 | 7.70 | 9.90 | 0 |
| pH | 7.60 | 7.20 | 7.70 | 8.40 | 0 |
| $O_2$ (% Sat.) | 85.70 | 95.50 | 87.30 | 87.30 | 0 |
| $NO_3$ (mg/l) | 0.24 | 0.06 | 0.05 | 0.55 | 0 |
| Temp.(°C) | 9.30 | 1.40 | 7.50 | 10.00 | 0.1 |
| pH | 7.60 | 7.60 | 7.70 | 8.30 | 0.1 |
| $O_2$ (% Sat.) | 87.00 | 93.70 | 89.00 | 84.00 | 0.1 |
| $NO_3$ (mg/l) | 0.25 | 0.07 | 0.05 | 0.59 | 0.1 |
| Temp.(°C) | 9.00 | 1.40 | 7.00 | 9.80 | 0.5 |
| pH | 7.50 | 7.30 | 7.50 | 7.70 | 0.5 |
| $O_2$ (% Sat.) | 84.70 | 90.00 | 68.00 | 68.70 | 0.5 |
| $NO_3$ (mg/l) | 0.25 | 0.18 | 0.15 | 0.70 | 0.5 |
| Temp.(°C) | 9.10 | 1.70 | 7.50 | 10.00 | 1.0 |
| pH | 7.50 | 7.40 | 7.40 | 7.30 | 1.0 |
| $O_2$ (% Sat.) | 85.30 | 79.50 | 69.00 | 54.50 | 1.0 |
| $NO_3$ (mg/l) | 0.27 | 0.23 | 0.23 | 0.97 | 1.0 |

Figure 22 shows the different steps of this analysis. More details can be found in Thioulouse and Chessel (1987), and an example of use in Cadet and Thioulouse (1989). The first step of triadic analysis is called "Global analysis", and gives an overall insight of the between-dates structure. The second, called "Compromise" is aimed at finding an average table, as close as possible from all tables. The last is called "Close-up analysis" and gives a detailed description of each item (i.e.; concentrations and variables) at each date.

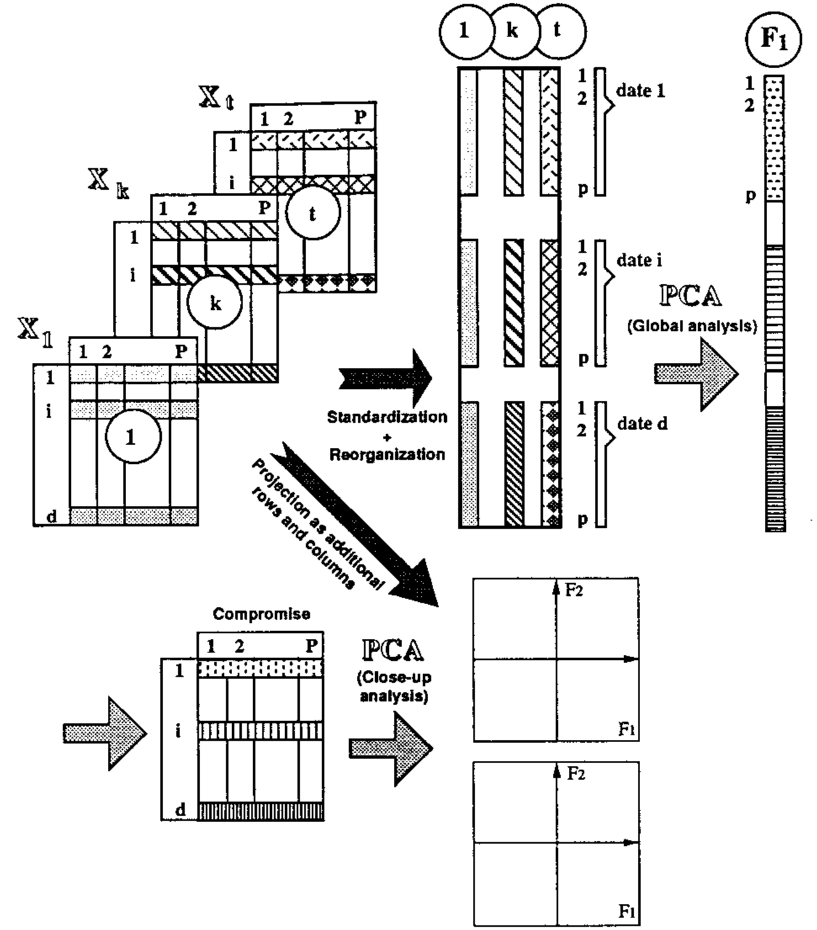


Figure 22: General principle of partial triadic analysis. The series of tables (X1, ..., Xk, ..., Xt) is standardized (each variable is transformed to have zero mean and unit variance) and is merged into one table after reorganization. Each table thus becomes one column of the new table. A (non) centered PCA is performed on this table (global analysis) and the first factor (F1) is reorganized to obtain a new table having the same structure (rows and columns) as one of the initial tables. This new table is called the compromise. A second (non) centered PCA is performed on this new table, and all the rows and columns of all the initial tables are projected as additional elements in this analysis (close-up analysis).

Thioulouse, J., J. Devillers, D. Chessel, and Y. Auda. 1991. Graphical techniques for multidimensional data analysis. Pages 153-205 in J. Devillers and W. Karcher, editors. Applied Multivariate Analysis in SAR and Environmental Studies. Kluwer Academic Publishers. http://pbil.univ-lyon1.fr/R/articles/arti067.pdf

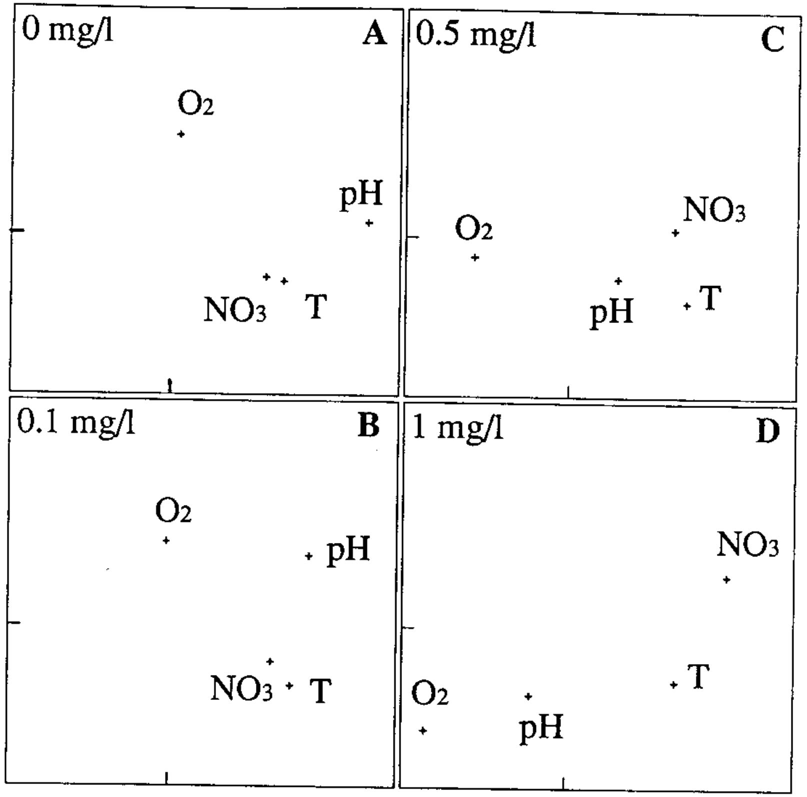


Figure 23: Factor maps of the rows of global analysis (the four maps correspond to the four concentrations).

Figure 23 represents the graph of the rows for the global analysis. The four maps correspond to the four concentrations, and in each map the four variables are displayed. The main feature is the opposition between the two following couples of concentrations: 0 - 0.1 mg/l and 0.5 - 1.0 mg/l (see the position of the four variables in figures 23A - 23B and in figures 23C - 23D). In figures 23C and 23D, the opposition between oxygen and nitrate is due to the decrease of the percentage of oxygen and the increase of nitrate concentration (see table 11).

The selection mechanism of the rows in the GraphMu program (Thioulouse, 1989, 1990) allows to easily redraw figure 23 with a reorganization of the points according to the variables instead of the concentrations (figure 24). This new presentation allows to stress that the temperature (figure 24A) is not influenced by the concentration of simazine while pH, $O_2$, and $NO_3$ are. Indeed, the aspect of the trajectory of pH (figure 24B) and oxygen (figure 24C) points reveals that these physicochemical parameters decrease when the concentrations of simazine increase. The converse is true for nitrate (figure 24D).

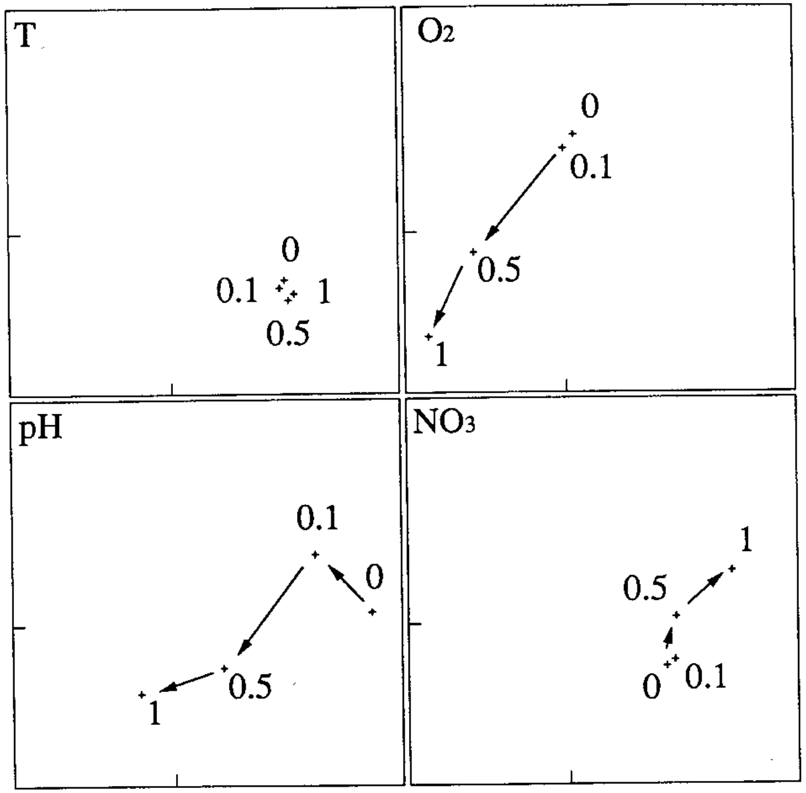


Figure 24: Factor maps of the rows of global analysis (the four maps correspond to the four variables).

Figure 25 represents the graph of the columns for the global analysis. It clearly shows that there is an evolution of the structure of the physicochemical parameters with time. The opposition between day seven and day 21 can be related to the fact that the concentration of nitrate increases during the experiment, whereas the dissolved oxygen decreases.

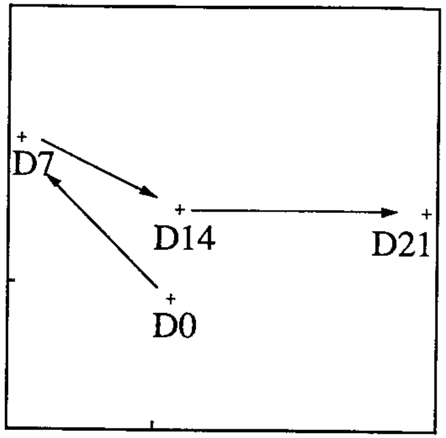


Figure 25: Factor maps of the columns of global analysis (i.e.; the four dates).

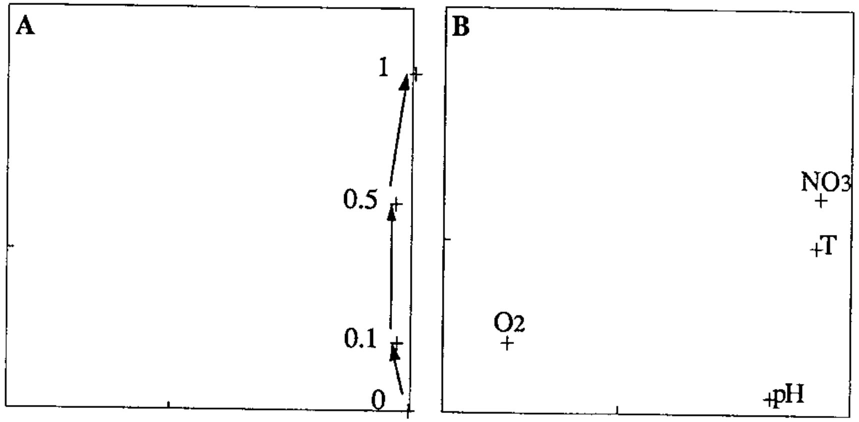


Figure 26: Compromise analysis. 26A: representation of the four concentrations. 26B: representation of the four physicochemical variables.

Figures 26A (concentrations) and 26B (variables) represent the compromise analysis. They do not give more information due to the simplicity of the structure of the example, but this compromise is necessary for the last step of partial triadic analysis. Indeed, t close-up analysis merely consists of a projection of the rows and columns of the init table (table 11, after reorganization and standardization) into this compromise analys Figures 27 and 28 confirm the results obtained with figures 24 and 25 but stress t evolution of the time structure as a function of the simazine concentrations (figure 27) a reciprocally (figure 28).

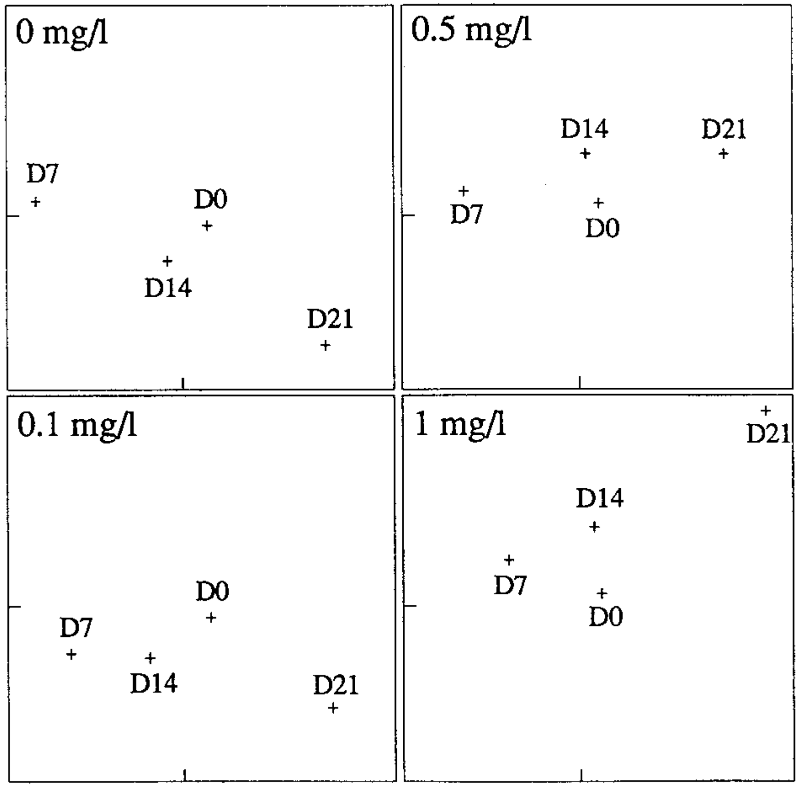


Figure 27: Close-up analysis. Representation of the four dates for each concentration.

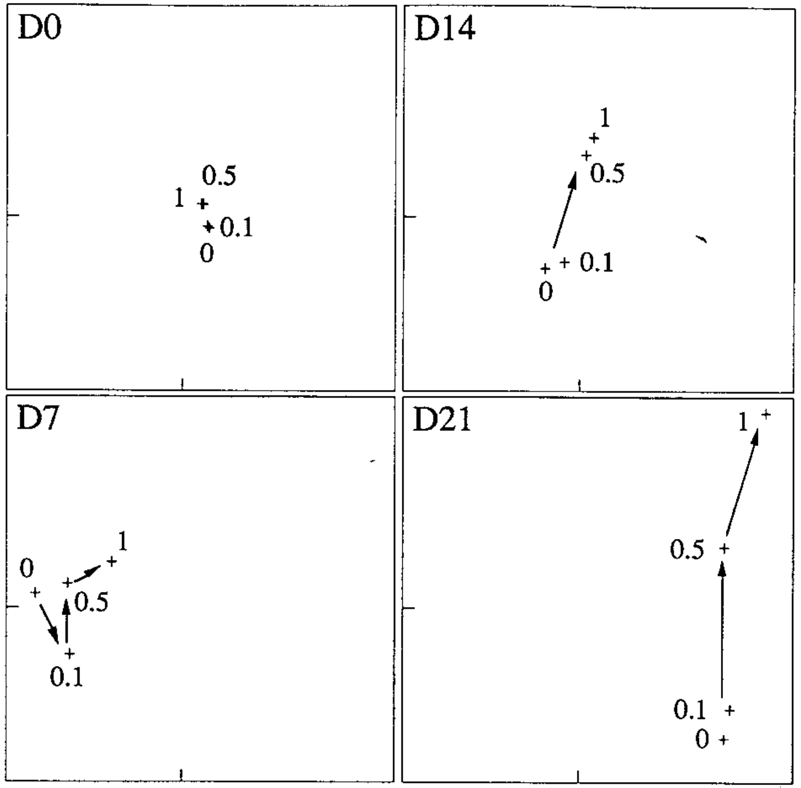


Figure 28: Close-up analysis. Representation of the four concentrations at each date.

These results agree with those from Jenkins and Buikema (1990), but the graphical approach gives a clearer view of the interactions between the different parameters involved in the contamination process by the herbicide under study.

## 6. Conclusion.

Multivariate methods are required to analyze large data sets. In this context, pictures a graphs play a key role in making easier for biologists the understanding of statisti methods and the interpretation of the results obtained by these methods. Howev classical factor maps are not necessarily the best way to represent the characterist stressed by analyses. Our examples show that factor coordinates can be considered useful numerical codes, which can be used to draw various kinds of graphical displays.

PCA coordinates of homogeneous toxicity tables provide an ordination which can be us to represent the original values (figure 12D), but they also provide a model to reconstit the data set, compute residuals from this model, and draw it in the same way (figure 12 Conversely, the toxicity of compounds can be projected onto the factor map, to ma easier the interpretation of the factors and the characteristics of each line and colu (organisms and compounds) of the table (figures 13 and 14).

The canonical graphical display for representing the correlations between fac coordinates and original data values is shown in figures 16 and 17. It clearly shows wh variables are well correlated to each factor. From another point of view, the projection the predicted variable into the correlation circle (figure 18) also shows its degree correlation with the factors.

Gauss curves are helpful to model the distribution of factor scores, and compare it to t distribution of original variables: this feature is particularly interesting for discrimin analysis, as it allows to detect which variables and factors have strong discriminati power (figure 20). Representation of original data (after centering or standardization) w vertical bars (figure 21) also helps in understanding the structuration of the data set, a detecting abnormal values.

Lastly, collections of factor maps can be used to analyze data sets with complex structu and particularly multi-way tables. The criterion used to build the collection is ve important, and figures 23 and 24 show that the same coordinates can stress differ structures according to the way they are grouped.

Graphics and multivariate data analysis require a large amount of computation, and th routine application in practice would be quite impossible without the aid of power computer-based interactive graphical display systems like MacMul, GraphMu, a ADECO. Information on these systems can be obtained from the authors.

## 7. Software availability.

MacMul and GraphMu are freely available on the Internet network by anonymous FTP biomol.univ-lyon1.fr (134.214.100.42). They are also available in several archive serve in the info-mac archive by anonymous FTP to sumex.stanford.edu (and also wuarchive.wustl.edu), or by e-mail on BITNET to the file serve FILESERV@IRLEARN in Europe, or LISTSERV@RICEVM1 in the USA, or BITFTP (BITFTP@PUCC). One can also find them in the statlib server (send a 'he mail to the server, at statlib@lib.stat.cmu.edu)

Thioulouse, J., J. Devillers, D. Chessel, and Y. Auda. 1991. Graphical techniques for multidimensional data analysis. Pages 153-205 in J. Devillers and W. Karcher, editors. Applied Multivariate Analysis in SAR and Environmental Studies. Kluwer Academic Publishers. http://pbil.univ-lyon1.fr/R/articles/arti067.pdf

## 8. References.

Andrews D. F. (1972). Biometrics 28, 125.

Ariyoshi Y. (1976). Kagaku Sosetsu, N°14 (Chemistry of Taste and Smell), The Chemical Society of Japan, Japan Scientific Society Press, Tokyo, p. 85.

Auda Y. (1983). Rôle des méthodes graphiques en analyse des données: application au dépouillement des enquêtes écologiques. Thèse de Troisième cycle, Université Lyon 1.

Barnett V. (Ed.) (1981). Interpreting Multivariate Data. John Wiley & Sons, New York.

Benigni R. and Giuliani A. (1987). Molecular Toxicol. 1, 143.

Bertin J. (Ed.) (1967). Sémiologie graphique. Les diagrammes, les réseaux, les cartes. Mouton et Gauthier-Villars, Paris.

Cadet P. and Thioulouse J. (1989). Revue Nématol., 12, 35.

Calamari D. and Vighi M. (1987). Quantitative Structure Activity Relationships in Ecotoxicology: Value and Limitations. Commission of the European Communities, Contract N° 86-B6602-11-001-N, p. 106.

Carrel G., Barthélémy D., Auda Y., and Chessel D. (1986). Acta Oecologica Oecol. Gener. 7, 189.

Chambers J.M., Cleveland W.S., Kleiner B., and Tukey P.A. (Eds.) (1983). Graphical Methods for Data Analysis, Duxbury Press, Boston.

Chambers J.M. and Kleiner B. (1982). Graphical Techniques for Multivariate Data and for Clustering. In, Krishnaiah P.R. and Kanal L.N. (Eds.) Handbook of Statistics, Vol. 2, North-Holland Publishing Company, Amsterdam, pp. 209-244.

Chernoff H. (1973). J. Am. Statist. Ass. 68, 361.

Dearden J.C. and Nicholson R.M. (1986). Pestic. Sci. 17, 305.

De Flora S., Koch R., Strobel K., and Nagel M. (1985). Toxicol. Environ. Chem. 10, 157.

Devillers J. and Chambon P. (1989). J. Fr. Hydrol. 20, 185-192.

Devillers J. and Lipnick R.L. (1990). Practical applications of regression analysis in environmental QSAR studies. In, Karcher W. and Devillers J. (Eds.). Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology, Kluwer Academic Publishers, Dordrecht, pp. 129-143.

Devillers J., Zakarya D., Chambon P., and Chastrette M. (1987). A model based on autocorrelation molecular descriptors for predicting mutagenicity of organic pollutants. In the proceedings of the International Symposium: Cellular Impact in Ecotoxicology", May 18-20, 1987, Lyon, France.

Devillers J., Zakarya D., and Chastrette M. (1988). Chemosphere 17, 1531.

Devillers J., Zakarya D., Chastrette M., and Doré J.C. (1989). Biomed. Environ. Sci. 385.

De Zwart D. and Slooff W. (1983). Aquatic Toxicol. 4, 129.

Dolédec S. and Chessel D. (1987). Acta Oecologica Oecol. Gener. 8, 403.

Enslein K., Blake B.W., Tuzzeo T.M., Borgstedt H.H., Hart J.B., and Salem H. (198 In Vitro Toxicol. 2, 1.

Escoufier Y. (1980). L'analyse conjointe de plusieurs matrices de données. In, Jolivet *et al.* (Eds.). Biométrie et Temps.

Escoufier Y. (1987). The duality diagram: a means for better practical applications. Legendre P. and Legendre L. (Eds.) Developments in Numerical Ecology. NATO A Series G (Ecological Sciences) Springer Verlag, Berlin pp. 139-156.

Everitt B.S. (Ed.) (1978). Graphical Techniques for Multivariate Data, Heinema Educational Books Ltd, London.

Fisher R.A. (1915). Biometrika 10, 507.

Gabriel K.R. (1981). Biplot display of multivariate matrices for inspection of data a diagnosis. In, Barnett V. (Ed.). Interpreting Multivariate Data, John Wiley & Sons, N York, pp. 147-173.

Gauch H.G. (Ed.) (1982). Multivariate analysis in community ecology. Cambrid University Press, Cambridge, England.

Gower J.C. and Digby P.G.N. (1981). Expressing complex relationships in t dimensions. In, Barnett V. (Ed.). Interpreting Multivariate Data, John Wiley & So New York, pp. 83-118.

Hansch C., Kim D., Leo A.J., Novellino E., Silipo C., and Vittoria A. (1989). CRC C Rev. Toxicol. 19, 185.

Hill M.O. (1979a). DECORANA - a FORTRAN program for detrended corresponden analysis and reciprocal averaging. Cornell University, Ithaca, New-York, USA.

Hill M.O. (1979b). TWINSPAN - a FORTRAN program for arranging multivariate d in an ordered two-way table by classification of the individuals and attributes. Corn University, Ithaca, New-York, USA.

Jaffrenou P.A. (1978). Sur l'Analyse des Familles Finies de Variables Vectorielles. Bas Algébriques et Application à la Description Statistique. Thesis Université Lyon I.

Jenkins D.G. and Buikema A.L. (1990). Environ. Toxicol. Chem. 9, 693.

Kroonenberg P.M. (Ed.) (1983). Three-mode Principal Component Analysis: Theory and Applications. D.S.W.O. Press, Leiden, NL.

Kroonenberg P.M. (1989). Acta Oecologica Oecol. Gener. 10, 245.

Lavit C. (Ed.) (1988). Analyse Conjointe des Tableaux Quantitatifs, Masson, Paris.

Lebart L., Morineau A. and Warwick K.M. (Eds.) (1984). Multivariate Descriptive Statistical Analysis. Correspondence Analysis and Related Techniques for Large Matrices. John Wiley & Sons, N.Y.

L'Hermier des Plantes H. (1976). Structuration des tableaux à trois imdices de la statistique. Théorie et application dEune méthode d'analyse conjointe. Thesis Université Montpellier II.

Miyashita Y., Takahashi Y., Takayama C., Sumi K., Nakatsuka K., Ohkubo T., Abe H., and Sasaki S. (1986). J. Med. Chem. 29, 906.

Moulton M.P. and Schultz T.W. (1986). Chemosphere 15, 59.

Morgenthaler S. and Tukey W. (1989). The next future of data analysis. In, Diday E. (Ed.). Data Analysis, Learning Symbolic and Numeric Knowledge, Nova Science Publishers, New York, pp. 1-12.

Nendza M. and Seydel J.K. (1988). Quant. Struct. Act. Relat. 7, 165.

Niemi G.J., Veith G.D., Regal R.R., and Vaishnav D.D. (1987). Environ. Toxicol. Chem. 6, 515.

Owen J. G. (1990). Ecology 71, 1823.

Ramsay J.O. (1989). The data analysis of vector-valued functions. In, Diday E. (Ed.). Data analysis, learning symbolic and numeric knowledge, Nova Science Publishers, New York, pp. 233-245.

Schultz T.W. (1987). Bull. Environ. Contam. Toxicol. 38, 994.

Slooff W., Canton J.H., and Hermens J.L.M. (1983). Aquatic Toxicol. 4, 113.

Thioulouse J. (1989). Computer Applications in the Biosciences, 5, 287.

Thioulouse J. (1990). Computers and Geosciences, 16, 1235.

Thioulouse J. and Chessel D. (1987). Acta Oecologica Oecol. Gener. 8, 463.

Tucker L.R. (1966). Psychometrika 31, 279.

Tufte E.R. (Ed.) (1983). The Visual Display of Quantitative Information, Graphic Press, Cheshire.

Tukey J. (Ed.) (1977). Exploratory Data Analysis, Addison-Wesley Publishing Company, Reading.

Tukey P.A. and Tukey J.W. (1981a). Graphical display of data sets in three or more dimensions. Preparation; prechosen sequences of views. In, Barnett V. (Ed.). Interpreting Multivariate Data, John Wiley & Sons, New York, pp. 189-213.

Tukey P.A. and Tukey J.W. (1981b). Graphical display of data sets in three or more dimensions. Data-driven view selection; agglomeration and sharpening. In, Barnett V. (Ed.). Interpreting Multivariate Data, John Wiley & Sons, New York, pp. 215-243.

Tukey P.A. and Tukey J.W. (1981c). Graphical display of data sets in three or more dimensions. Summarization; smoothing; supplemented views. In, Barnett V. (Ed.). Interpreting Multivariate Data, John Wiley & Sons, New York, pp. 245-275.

Wainer H. and Thissen D. (1981). Ann. Rev. Psychol. 32, 191.

Wang P.C.C. (Ed.) (1978). Graphical Representation of Multivariate Data, Academic Press, London.

Warne M.ST.J., Connell D.W., Hawker D.W., and Schüürmann G. (1989). Ecotoxicol Environ. Safety 17, 133.

Wold S. and Sjöström M. (1977). SIMCA, a method for analysing chemical data in term: of similarity and analogy. In, Kowalski B. (Ed.). Chemometrics, Theory and Application American Chemical Society Symposium Series N° 52.