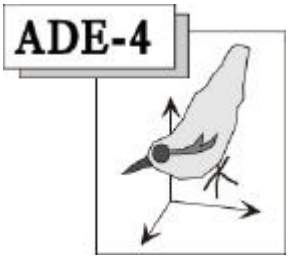


Fiche thématique 5.E

Mesures de la corrélation entre tableaux

Résumé

La fiche invite à utiliser les mesures de corrélation entre tableaux. Le problème est illustré par les données publiées par B. Statzner & Coll (1997, Reproductive traits, habitat use and templet theory: a synthesis of world-wide data on aquatic insects. Freshwater Biology : 38, 109-135) et divers tableaux de fréquences alléliques. Les coefficients de corrélation entre tableaux utilisés sont décrits dans les travaux de Lazraq & Coll. (1992, Mesures de liaison vectorielle et généralisation de l'analyse canonique. Revue de Statistique Appliquée : 39, 23-35) et de Kiers & Coll. (1994, Generalized canonical analysis based on optimizing matrix correlations and a relation with IDIOSCAL. Computational Statistics and Data Analysis : 18, 331-340).

Plan

D. Chessel

Introduction

Dans un article récent B. Statzner [1] pose clairement la question de la corrélation entre tableaux. Les données intégralement reproduites et documentées permettent une discussion méthodologique sur la signification du passage de la corrélation entre variables à la corrélation entre tableaux.

Les données sont présentées sous la forme de deux tableaux de variables floues. Chaque ligne est une espèce. Les colonnes sont des modalités regroupées par variables. Par exemple le taxon 43 (*Ephemera vulgata* L., EPHEMEROPTERA, Ephemeridae) pour le trait biologique “taille des femelles” présente le profil :

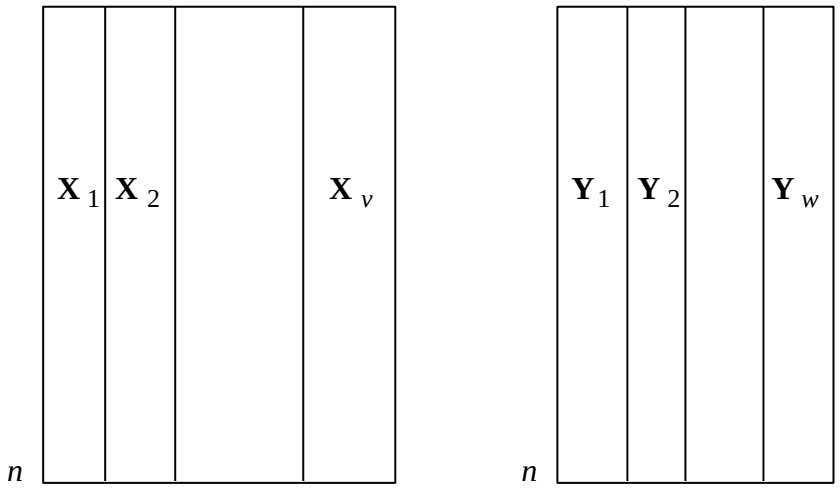
1	2	3	4	5	6	7
0	**0**	**2**	**2**	**0**	**0**	**0**

Ceci signifie que cette taille est dans la classe]10 mm, 15 mm] dans 50% des cas et dans la classe]15 mm, 20 mm] dans les autres cas. Le premier tableau est formé de la juxtaposition de 10 traits biologiques comptant au total 41 modalités.

Dans le second tableau, à la ligne 43 on trouve pour le trait écologique “lieu de ponte” le profil :

1	2	3	4	5	6	7
0	**1**	**2**	**0**	**0**	**0**	**0**

Ceci signifie que l’espèce dépose ses œufs une fois sur trois à la surface d’une eau stagnante et deux fois sur trois à la surface d’une eau peu courante. Le premier tableau est formé de la juxtaposition de 7 traits écologiques comptant au total 34 modalités. Il y a 131 espèces d’insectes aquatiques ainsi décrites et une structure de données du type :



Le sommaire de l’article indique :

> *1. Using world-wide data on the reproductive biology of 131 species (in eight orders) of aquatic insects, we used multivariate analyses to examine: (i)*

relationships among reproductive traits determining life cycle, fecundity, morphology, behaviour and physiology; (ii) relationships among traits determining spatial and temporal habitat characteristics at different scales; and (iii) the relationship between reproductive and habitat-use traits. This provided a test of predictions of the habitat templet concept on trends of species traits along gradients of habitat heterogeneity.

2. The major trends observed in the relationships among reproductive traits were that larger females had larger eggs, which were more vulnerable to perturbations such as droughts and often laid in cocoons. In addition, they laid the eggs in larger numbers of smaller clutches than smaller females. Other traits (e.g. egg number or incubation time) did not show clear trends.

3. Females that deposited eggs at sites of low local temporal heterogeneity (within plants) used, at the same time, gross habitats of high temporal heterogeneity (temporary waters). In contrast, traits in habitat use did hardly differ along well-known gradients of temporal heterogeneity along running waters (from source to estuary). The number of habitat units used by ovipositing females generally increased with the spatial scale considered, i.e. most species oviposited in a single small habitat unit but in several gross habitats.

4. A significant ($P<0.01$*) relationship between traits in reproduction and habitat use demonstrated that habitat acted as a templet for reproductive strategies. This relationship was dominated by larger females having larger, unattached eggs which were more vulnerable to droughts and were oviposited in temporally more stable small-scale habitats (within wood or macrophytes, or within cocoons spun by the female) but more unstable large-scale habitats (primarily temporary waters). Thus, only on the small habitat scale did some of our observations correspond to the predictions of the habitat templet concept (e.g. larger size or higher vulnerability in more stable habitats). However, many species had traits in reproduction that did not show trends as predicted by the concept.*

5. This and other recent studies of the relationships between traits of freshwater organisms and the heterogeneity of their habitats have shown that habitat acts as a templet for species life history traits. However, many of the details observed in these studies did not correspond to predictions of the templet concept because of trade-offs among the traits and scale problems in the description of habitat heterogeneity. Therefore, future studies should focus on groups of organisms that are as similar as possible in the trade-offs among their species traits and on the potential relationships of habitat heterogeneity across multiple scales.

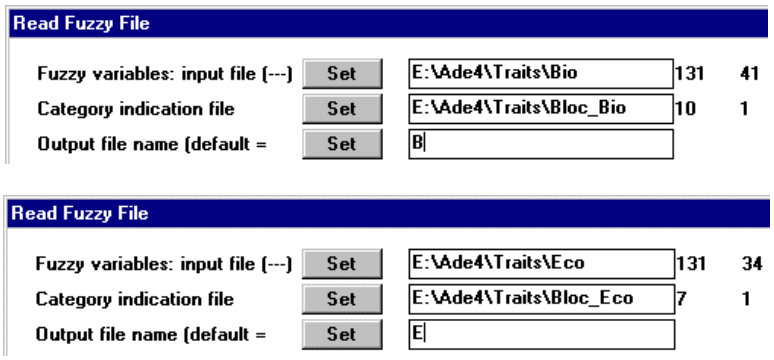
L'analyse est conduite par une analyse de co-inertie [2] entre deux ACM sur variables floues [3]. On peut se demander si l'introduction des méthodes *K*-tableaux modifie sensiblement l'opinion qu'on peut avoir sur les données. En particulier est-il possible de mesurer

directement la corrélation entre deux traits biologiques, deux traits écologiques et éventuellement sélectionner des associations de traits biologiques et écologiques exprimant une ou plusieurs formes de relations entre les deux types d'information ?

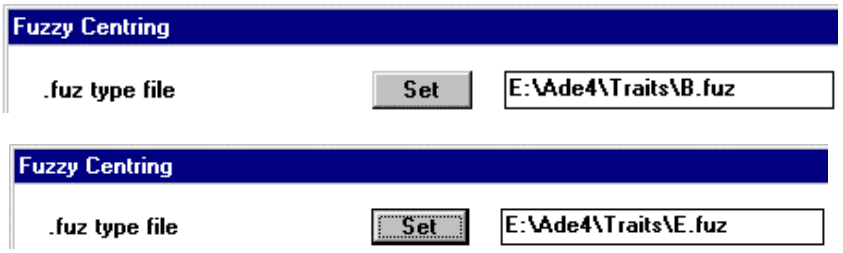
Il est assez étonnant que l'emploi des indices de corrélations entre tableaux ne soit pas établi en biologie. Dans un tableau de fréquences alléliques (populations en lignes, bloc d'allèles par locus en colonnes), un tableau écologique (stations en lignes, blocs d'espèces par groupe en colonnes), un tableau d'usage du code génétique (gènes en lignes, blocs de codons par acides aminés en colonnes) comme dans un tableau de traits on a d'abord besoin d'une mesure de redondance entre tableaux comme l'usage de plusieurs variables demandait d'abord une mesure de la covariation.

Les coefficients RV

Les fichiers (carte Traits de la pile de données) sont lus (FuzzyVar: Read Fuzzy File) :



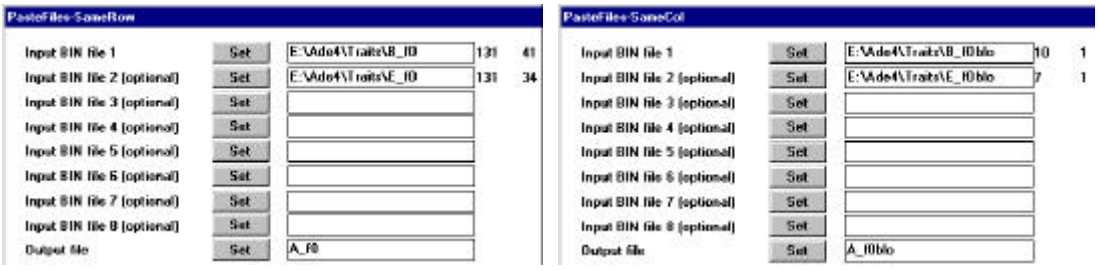
Les tableaux sont simplement centrés :



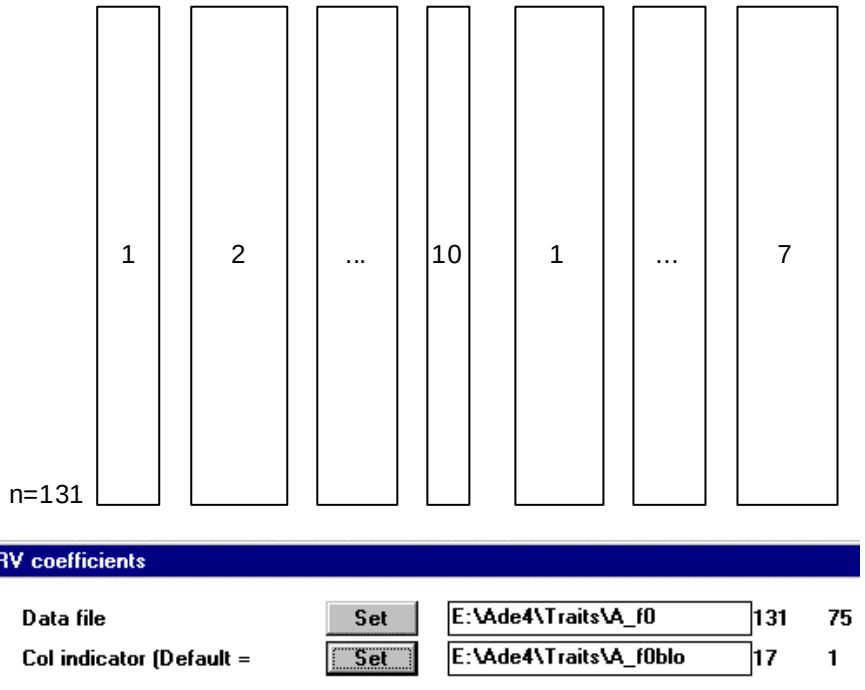

```
Centring of a fuzzy array
Input file: E:\Ade4\Traits\B.fuz for access to file E:\Ade4\Traits\B
Row number: 131, column number: 41
Uniform row weights

Missing data: 120
File E:\Ade4\Traits\B_f0mm contains the table mm = mean for missing data
It has 131 rows and 41 columns (categories)

File E:\Ade4\Traits\B_fO contains the centred table
It has 131 rows and 41 columns (categories)

File E:\Ade4\Traits\B_fOpl contains the row weights
It has 131 rows and 1 column
```

```
File E:\Ade4\Traits\B_fOblo contains the column block indicator
It has 10 rows and 1 column

Centring of a fuzzy array
Input file: E:\Ade4\Traits\E.fuz for access to file E:\Ade4\Traits\E
Row number: 131, column number: 34
Uniform row weights

Missing data: 104
File E:\Ade4\Traits\E_f0mm contains the table mm = mean for missing data
It has 131 rows and 34 columns (categories)

File E:\Ade4\Traits\E_fO contains the centred table
It has 131 rows and 34 columns (categories)

File E:\Ade4\Traits\E_fOpl contains the row weights
It has 131 rows and 1 column

File E:\Ade4\Traits\E_fOblo contains the column block indicator
It has 7 rows and 1 column
```

Pour simplifier la discussion, ils sont assemblés :



On a donc la configuration :

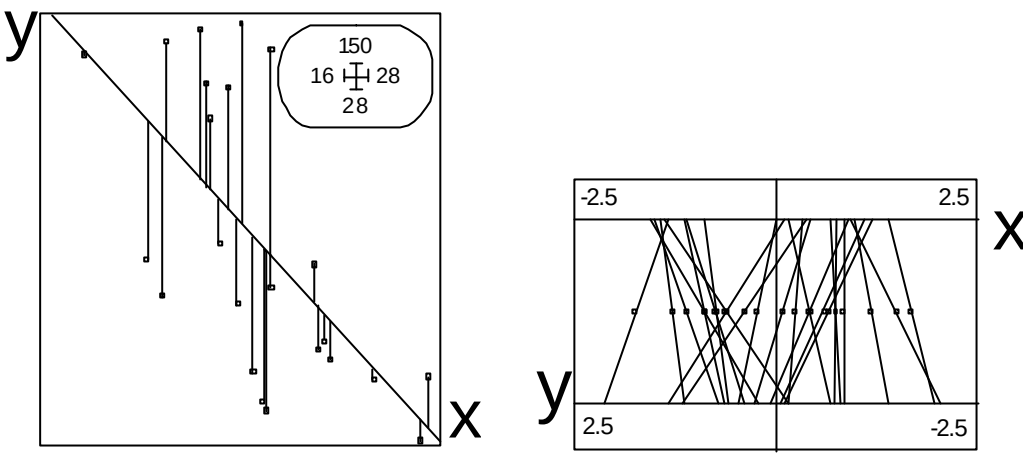



```
Input file: E:\Ade4\Traits\A_f0
    -> Rows: 131, columns: 75
    -> 17 blocks: 7/6/6/3/3/3/3/4/3/3/7/6/4/8/2/4/3/
```

```
RV coefficients Escoufier 1973
---------------------- Correlation matrix -------------------

[  1] 1000
[  2]  202 1000
[  3]   71  102 1000
[  4]  147   45   36 1000
...
[ 16]   62   70   16   49   14   19   40   38   13    7   63
        42   71   44    2 1000
[ 17]   23   28   23    9   48   27   53    8   46   21   43
         5   42   13    1   28 1000
------------------------------------------------------------
E:\Ade4\Traits\A_f0_RV2 is a binary file with 17 rows and 17 columns
Content: Coefficients RV
```

Fem_Size	Egg_length	Egg-number	Generations	Oviposition	Incubation	Egg_shape	Egg_attach	Clutch_struc	Clutch_numbe	Oviposi_site	Substrat_eggs	Egg_deposition	Gross_habitat	Saturation	Time_day	Season	
1000	202	71	147	53	25	133	90	59	35	56	50	112	49	23	62	23	Fem_Size
202	1000	102	45	43	31	264	103	62	97	168	123	153	86	45	70	28	Egg_length
71	102	1000	36	24	28	52	43	33	39	46	47	49	64	26	16	23	Egg-number
147	45	36	1000	83	45	36	9	25	4	34	14	49	27		49	9	Generations
53	43	24	83	1000	92	15	7	33	5	35	17	16	40	22	14	48	Oviposition
25	31	28	45	92	1000	6	67	29	43	47	45	15	71	2	19	27	Incubation
133	264	52	36	15	6	1000	106	26	161	85	113	90	55	3	40	53	Egg_shape
90	103	43	9	7	67	106	1000	91	104	68	58	154	42	47	38	8	Egg_attach
59	62	33	25	33	29	26	91	1000	146	98	19	177	41	9	13	46	Clutch_struc
35	97	39	4	5	43	161	104	146	1000	31	8	86	26	23	7	21	Clutch_number
56	168	46	34	35	47	85	68	98	31	1000	196	214	137	110	63	43	Oviposi_site
50	123	47	14	17	45	113	58	19	8	196	1000	120	100	18	42	5	Substrat_eggs
112	153	49	49	16	15	90	154	177	86	214	120	1000	30	52	71	42	Egg_deposition
49	86	64	27	40	71	55	42	41	26	137	100	30	1000	13	44	13	Gross_habitat
23	45	26		22	2	3	47	9	23	110	18	52	13	1000	2	1	Saturation
62	70	16	49	14	19	40	38	13	7	63	42	71	44	2	1000	28	Time_day
23	28	23	9	48	27	53	8	46	21	43	5	42	13	1	28	1000	Season

A chaque couple de deux tableaux on a associé une mesure de corrélation (RV) comprise entre 0 et 1 éditée dans Excel en entier (x 1000). On peut se demander pourquoi la valeur n'est pas, comme dans le cas d'une corrélation entre variables dans l'intervalle $[-1,+1]$. L'indice mesure la ressemblance de deux typologies et non pas le lien entre les valeurs. Le RV vaut exactement le carré de la corrélation. Par exemple, pour deux variables corrélées négativement, le carré de corrélation est représenté par la droite de régression et exprime le pourcentage de variance expliquée :

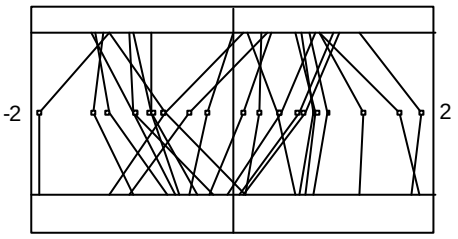


Si on normalise les deux variables et si on change le signe de la seconde, il exprime la ressemblance entre les deux ordinations par le principe (pour des variables normalisées) :

$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i - y_i\right)^2 = 2 - r$$

On voit aussi (ci-dessus, à droite) le principe issu de l'ACP des deux variables. Au milieu, on a la variable $\mathbf{z} = \frac{\mathbf{x}+\mathbf{y}}{2}$ (après normalisation) qui maximise $cor^2(\mathbf{z},\mathbf{x}) + cor^2(\mathbf{z},\mathbf{y})$.

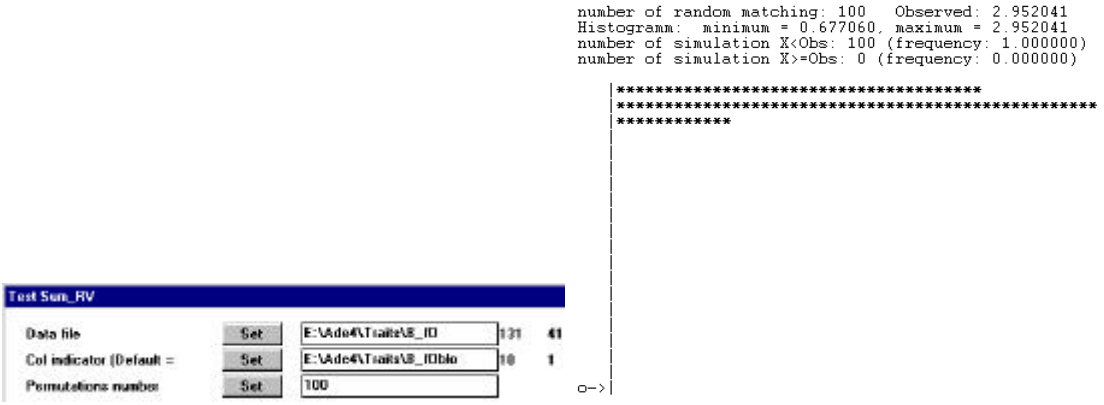
Ou encore, suivant l'analyse canonique, il existe une variable de variance 1 (ci-dessous, au milieu) qui présente avec chacune des deux une corrélation (pour deux variables corrélées négativement) de $\frac{1-r}{2}$:

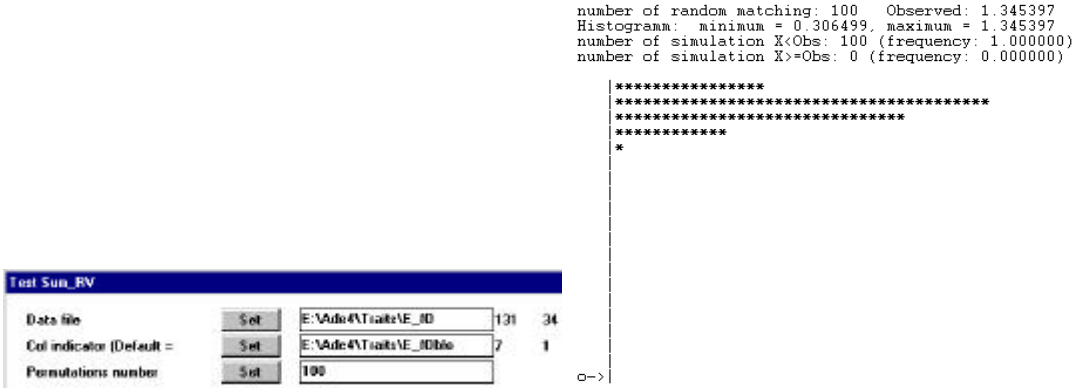


Dans le premier point de vue, on s'intéresse aux valeurs et on prédit y par x. Dans les autres on s'intéresse aux structures et on montre comment elles se ressemblent. Dés qu'il y a plus d'une variable de chaque côté, les points de vue deviennent distincts. Le RV utilise alors la co-inertie des tableaux pris deux à deux.

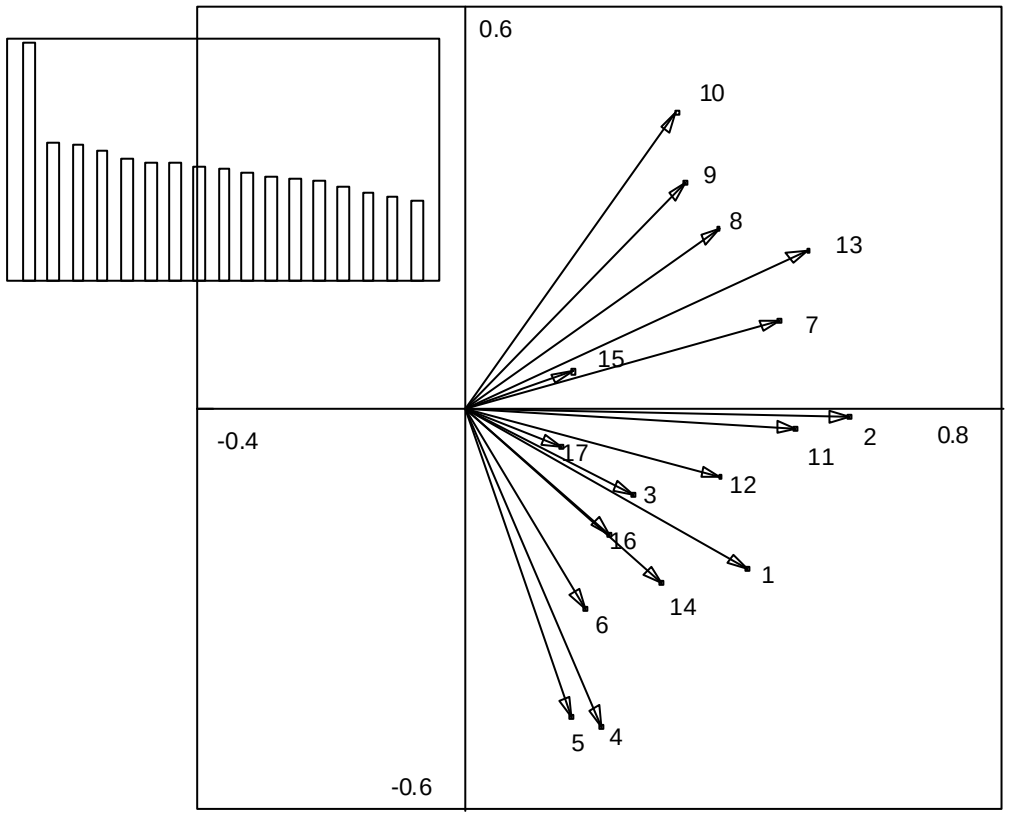
$$RV(\mathbf{X},\mathbf{Y}) = \frac{Trace\left(\frac{1}{n}\mathbf{X}\mathbf{X}^t\frac{1}{n}\mathbf{Y}\mathbf{Y}^t\right)}{\sqrt{Trace\left(\frac{1}{n}\mathbf{X}\mathbf{X}^t\frac{1}{n}\mathbf{X}\mathbf{X}^t\right)Trace\left(\frac{1}{n}\mathbf{Y}\mathbf{Y}^t\frac{1}{n}\mathbf{Y}\mathbf{Y}^t\right)}}$$

Les RV observés sont faibles. Ils le sont tellement qu'on peut se demander s'ils ne sont pas tous nuls (statistiquement). Le test de permutations de Canonical: Test Sum_RV répond négativement :





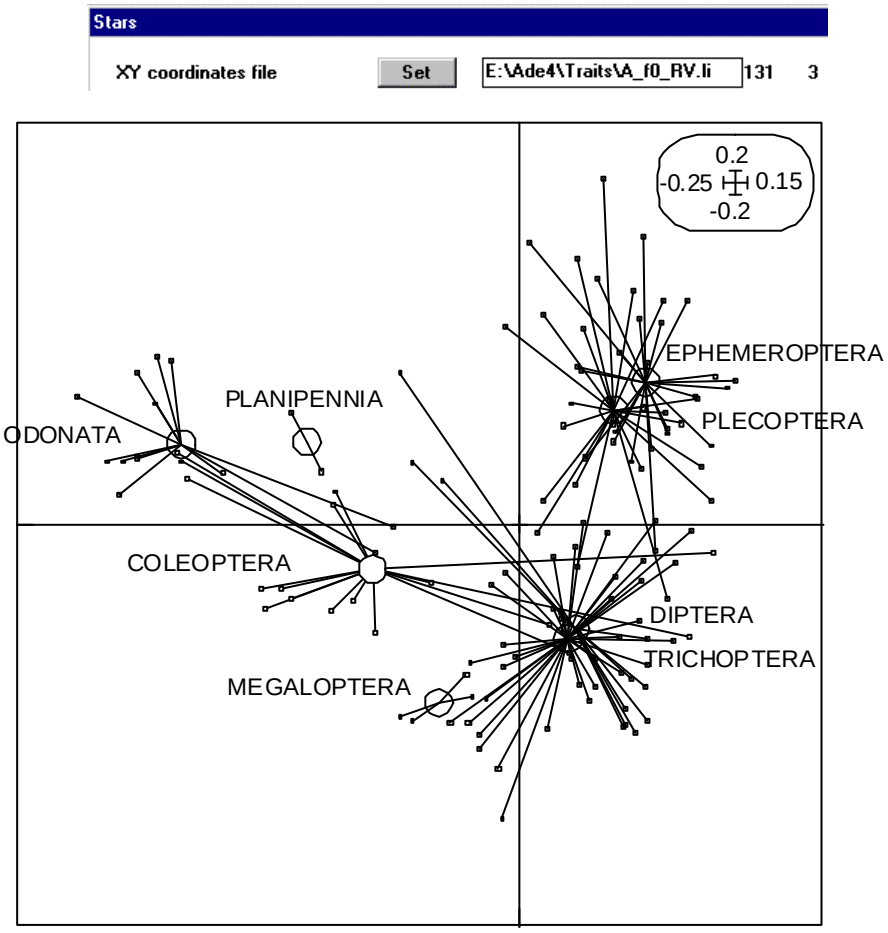
Les RV sont significativement non tous nuls. Cependant le compromis de STATIS est franchement faible :

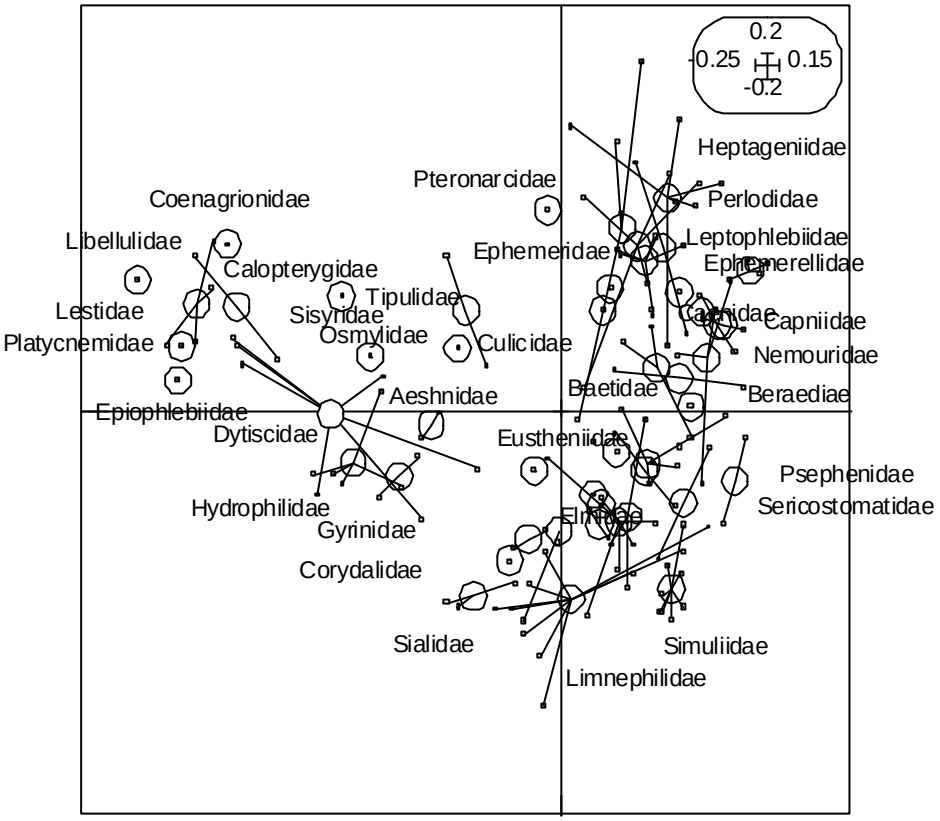


On lit sur ce graphe la géométrie du nuage des opérateurs d'inertie par tableaux. Certes le premier axe (horizontal) contient un minimum de signification, mais ce qui frappe le plus, c'est une absence globale de corrélation entre typologies induites par les traits biologiques et écologiques. On retrouve l'essentiel du résumé des auteurs, par exemple :

> *The major trends observed in the relationships among reproductive traits were that larger females had larger eggs* [Corrélation 1-2, RV=0.202] , *which were more vulnerable to perturbations such as droughts and often laid in cocoons* [RV 2-7 = 0.264]. *In addition, they laid the eggs in larger numbers of smaller clutches than smaller females* [RV 2-10 = 0.161]. *Other traits (e.g. egg number* [RV <= 0.1] *or incubation time* [RV <= 0.07]*) did not show clear trends ... A significant (*P*<0.01) relationship between traits in reproduction and habitat use demonstrated that habitat acted as a templet for reproductive strategies.*

L'essentiel est cependant dans l'absence de liens, comme si était optimisée la multiplicité des combinaisons possibles tant dans les stratégies de reproduction que dans les stratégies écologiques et qu'entre les deux le nombre de combinaisons s'accroisse encore. Il y a des corrélations : elles sont d'abord le résultat de la phylogénie. Parce que les espèces d'un même genre, d'une même famille et d'un même ordre se ressemblent, on obtient un minimum de cohérence. Ceci se voit parfaitement sur le typologie de synthèse :





Le compromis décrit une structure parfaitement identifiable. Pourtant, aucun des tableaux d'origine ne joue un rôle particulier, son origine est diffuse :

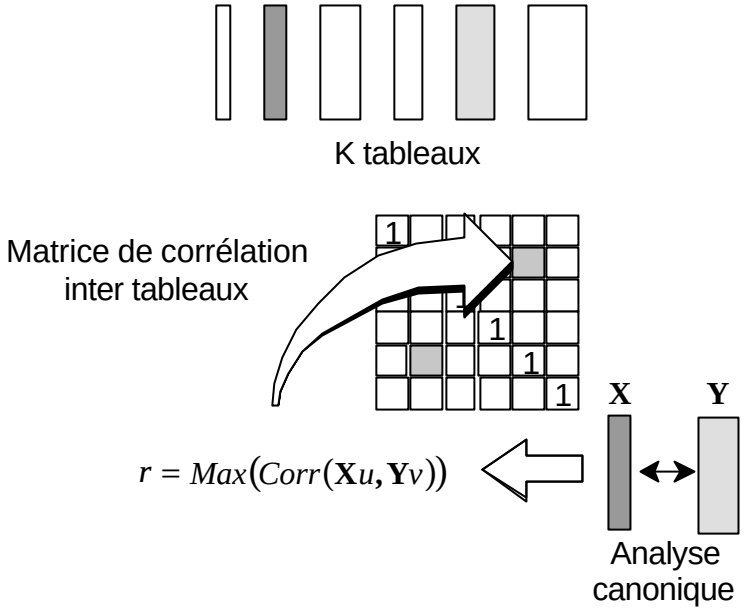
```
Cols = column number of each table
Weights = Weights of operators in the consensus
HS norm  = Norm (Hilbert-Schmidt) of operators
RVcons = RV (Wk, Sum akWk) 1/1000
RV1/2 = sqrt(RV (Wk, Sum akWk))  1/1000
Dist2 = squared distance = 2(1-RVcons)
```

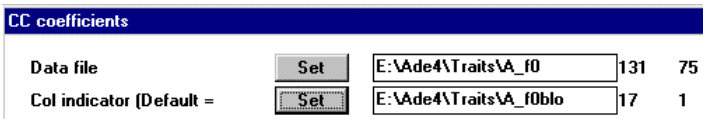
Number	Cols	Weights	NS norm	RVcons	RV1/2	Dist2
1	7	2.922e-01	2.311e-01	**419**	647	1.163e+00
2	6	3.997e-01	2.897e-01	**573**	757	8.545e-01
3	6	1.738e-01	2.012e-01	**249**	499	1.502e+00
4	3	1.415e-01	1.482e-01	**203**	450	1.594e+00
5	3	1.093e-01	3.432e-01	**157**	396	1.687e+00
6	3	1.236e-01	2.355e-01	**177**	421	1.646e+00
7	3	3.264e-01	3.532e-01	**468**	684	1.065e+00
8	4	2.636e-01	2.884e-01	**378**	615	1.245e+00
9	3	2.287e-01	3.461e-01	**328**	572	1.345e+00
10	3	2.187e-01	3.574e-01	**313**	560	1.373e+00
11	7	3.445e-01	2.403e-01	**494**	703	1.013e+00
12	6	2.654e-01	2.447e-01	**380**	617	1.239e+00
13	4	3.554e-01	3.708e-01	**509**	714	9.816e-01
14	8	2.033e-01	1.593e-01	**291**	540	1.417e+00
15	2	1.109e-01	1.420e-01	**159**	399	1.682e+00
16	4	1.495e-01	1.243e-01	**214**	463	1.572e+00
17	3	9.890e-02	1.937e-01	**142**	376	1.717e+00

Le point de vue des auteurs n'est pas contredit. Il est exprimé d'une autre manière.

Les corrélations canoniques

Une autre manière de mesurer la corrélation entre deux tableaux vient de l'analyse canonique :



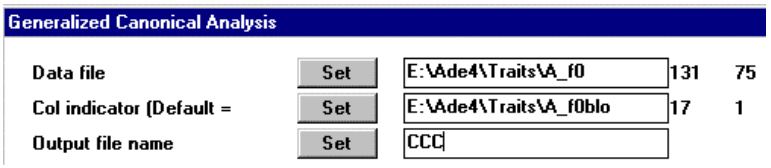



```
Input file: E:\Ade4\Traits\A_f0
    -> Rows: 131, columns: 75
    -> 17 blocs: 7/6/6/3/3/3/3/4/3/3/7/6/4/8/2/4/3/

Block:  1  Dim: 131 -  7 Rank:   6
Block:  2  Dim: 131 -  6 Rank:   5
...
Block: 16  Dim: 131 -  4 Rank:   3
Block: 17  Dim: 131 -  3 Rank:   2
Canonical correlation coefficients
Hotelling 1936
---------------------- Correlation matrix -------------------

[  1] 1000
[  2]  496 1000
[  3]  149  257 1000
...
[ 16]  335  208   83  100  101   65  125  116   28   44  158
       116  207  101    3 1000
[ 17]   82   73   65   15   82   52   68   37   72   46  114
        40   74   74    4   73 1000
------------------------------------------------------------
cancor squared
---------------------- Correlation matrix -------------------
[  1] 1000
[  2]  246 1000
...
[ 16]  112   43    7   10   10    4   16   13    1    2   25
```

```
         13   43   10    0 1000
[ 17]     7    5    4    0    7    3    5    1    5    2   13
          2    5    5    0    5 1000
------------------------------------------------------------
E:\Ade4\Traits\A_f0_CC2 is a binary file with 17 rows and 17 columns
Content: Squared canonical correlation (lambda1)
```

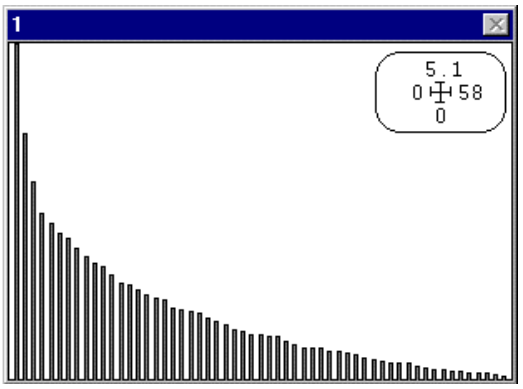
Fem_Size	Egg_length	Egg-number	Generations	Oviposition	Incubation	Egg_shape	Egg_attach	Clutch_struc	Clutch_numbe	Oviposi_site	Substrat_eggs	Egg_deposition	Gross_habitat	Saturation	Time_day	Season	
1000	246	22	113	20	18	154	63	37	28	20	40	243	30	5	112	7	Fem_Size
246	1000	66	8	14	8	360	44	13	28	211	76	161	84	12	43	5	Egg_length
22	66	1000	4	2	7	10	24	9	7	21	24	12	38	4	7	4	Egg-number
113	8	4	1000	24	4	11		1	1	10	3	6	4		10		Generations
20	14	2	24	1000	24	1	1	6	1	10	1	1	18	1	10	7	Oviposition
18	8	7	4	24	1000		18	1	3	12	4	1	26		4	3	Incubation
154	360	10	11	1		1000	55	4	60	82	60	108	39		16	5	Egg_shape
63	44	24		1	18	55	1000	23	23	50	43	113	39	3	13	1	Egg_attach
37	13	9	1	6	1	4	23	1000	58	32	7	61	30		1	5	Clutch_struc
28	28	7	1	1	3	60	23	58	1000	10	2	59	17	3	2	2	Clutch_number
20	211	21	10	10	12	82	50	32	10	1000	178	216	265	44	25	13	Oviposi_site
40	76	24	3	1	4	60	43	7	2	178	1000	88	60	5	13	2	Substrat_eggs
243	161	12	6	1	1	108	113	61	59	216	88	1000	27	7	43	5	Egg_deposition
30	84	38	4	18	26	39	39	30	17	265	60	27	1000	3	10	5	Gross_habitat
5	12	4		1			3		3	44	5	7	3	1000			Saturation
112	43	7	10	10	4	16	13	1	2	25	13	43	10		1000	5	Time_day
7	5	4		7	3	5	1	5	2	13	2	5	5		5	1000	Season

Le résultat est fondamentalement le même. On peut comparer les RV et les carrés de corrélation canonique. L'analyse canonique généralisée est possible (le nombre des individus 131 est grand par rapport aux nombres de variables de chaque tableau) :

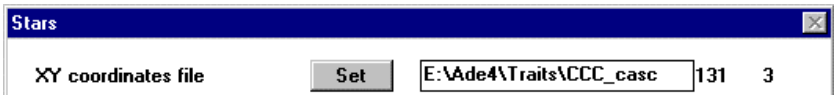



```
Input file: E:\Ade4\Traits\A_f0
    -> Rows: 131, columns: 75
    -> 17 blocks: 7/6/6/3/3/3/3/4/3/3/7/6/4/8/2/4/3/

Block:  1  Dim: 131-  7 Rank:   6
Block:  2  Dim: 131-  6 Rank:   5
...
Block: 16  Dim: 131-  4 Rank:   3
Block: 17  Dim: 131-  3 Rank:   2

Num  Eigenval.|Num. Eigenval.|Num. Eigenval.|Num. Eigenval.|
001  5.091e+00|002  3.746e+00|003  3.021e+00|004  2.542e+00|
005  2.396e+00|006  2.230e+00|007  2.174e+00|008  2.021e+00|
009  1.881e+00|010  1.792e+00|011  1.742e+00|012  1.603e+00|
...
```

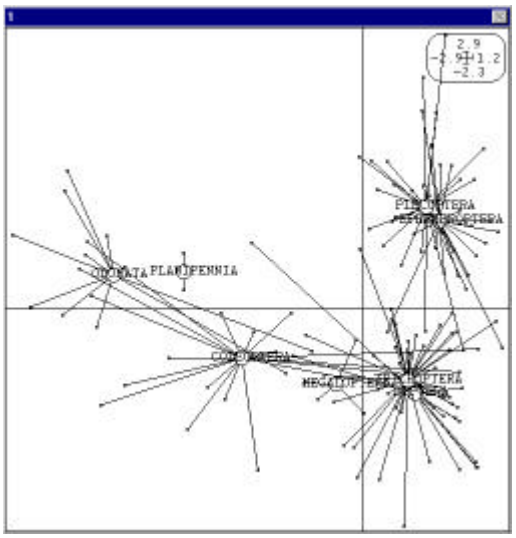
Le dépouillement des premières dimensions est valide :




```
Eigenvalues in file CCC_vpro - Rows: 58 -Col: 1
Canonical scores in file CCC_casc - Rows: 131 -Col: 3
File :CCC_casc
|Col.|   Mini    |   Maxi    |
|----|-----------|-----------|
|   1|-2.836e+00| 1.142e+00|
|   2|-2.246e+00| 2.810e+00|
|   3|-1.984e+00| 3.053e+00|
|----|-----------|-----------|
```



On reconnaît la même typologie :



Les associations taille des femelles (1), taille des œufs (2), forme des œufs (7) et lieu de pontes (13) puis groupement des œufs (9) et mode de dépôt (14) est à nouveau mises en évidence :

```
Eigenvalue ratios 1/10000
|----------|------|------|------|
|          |fac  1|fac  2|fac  3|
|----------|------|------|------|
|  Tab  1  |  1196|   366|  1463|
|  Tab  2  |  1546|   414|  1528|
|  Tab  3  |   273|   811|   826|
|  Tab  4  |   100|   387|   548|
|  Tab  5  |    49|   416|   988|
```

```
|  Tab  6  |    15|   571|   227|
|  Tab  7  |  1341|   242|   664|
|  Tab  8  |   776|   923|   281|
|  Tab  9  |   278|  1254|   342|
|  Tab 10  |   466|   442|   174|
|  Tab 11  |   863|   768|   769|
|  Tab 12  |   691|   629|   667|
|  Tab 13  |  1150|  1376|   407|
|  Tab 14  |   698|   603|   492|
|  Tab 15  |     4|   153|   255|
|  Tab 16  |   505|   145|   169|
|  Tab 17  |    50|   499|   201|
|----------|------|------|------|
|----------|------|------|------|
|  Total   | 10000| 10000| 10000|
|----------|------|------|------|
```

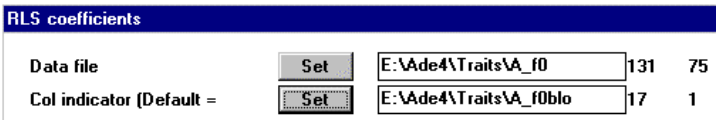
Rien ne contredit l'interprétation des auteurs qui aurait pu se faire à partir d'une analyse canonique généralisée. Il est remarquable de voir ici, que STATIS qui est en accord avec l'analyse factorielle multiple et l'analyse de co-inertie multiple dans leur domaine de validité [4] ne s'éloigne pas d'une analyse canonique généralisée [5] dans un autre cadre. Les deux méthodes donnent directement des indications sur les tableaux (qui sont ici des variables au sens expérimental) plutôt que sur les modalités et simplifient l'interprétation.

Les coefficients RLS

Ils sont définis par :

$$RLS(\mathbf{X},\mathbf{Y}) = \frac{Trace\left(\left(\frac{1}{n}\mathbf{X}\mathbf{X}^t\frac{1}{n}\mathbf{Y}\mathbf{Y}^t\right)^{1/2}\right)}{\sqrt{Trace\left(\frac{1}{n}\mathbf{X}\mathbf{X}^t\right)Trace\left(\frac{1}{n}\mathbf{Y}\mathbf{Y}^t\right)}}$$

Cet indice est attribué à Lingoes & Schönemann [6] par Lazraq & Coll. [7] et étudié par Kiers & Coll. (1994) [8]. On a :

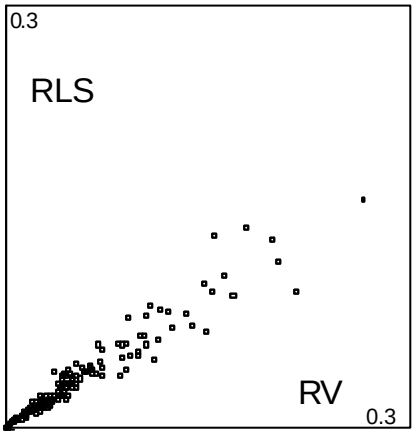


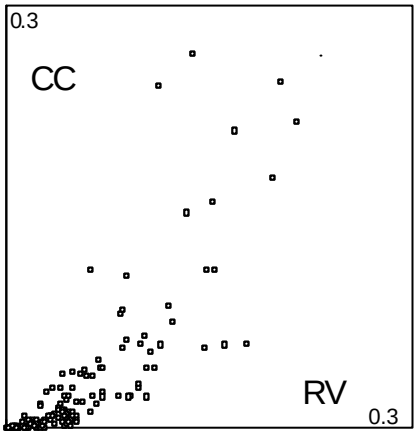

```
Input file: E:\Ade4\Traits\A_f0
    -> Rows: 131, columns: 75
    -> 17 blocs: 7/6/6/3/3/3/3/4/3/3/7/6/4/8/2/4/3/

RLS coefficients
Lingoes & Schonemann 1974
Lazraq, Cleroux & Kiers 1992
Kiers, Cleroux & Ten Berge 1994
----------------------- Correlation matrix -------------------

[  1] 1000
```

```
[  2]  344 1000
...
[ 16]  213  218  109  186  101  125  179  164  104   76  196
       193  208  166   35 1000
[ 17]  125  142  120   87  197  147  214   80  177  128  153
        67  172   99   35  142 1000
------------------------------------------------------------
E:\Ade4\Traits\A_f0_LS1 is a binary file with 17 rows and 17 columns
Content: Coefficients RLS

E:\Ade4\Traits\A_f0_LS.dist is a binary file with 17 rows and 17 columns
Content: Among array distances = sqrt(2(1-r))


RLS matrix is positive - Trace =  1.700e+01

PCO - RLS matrix diagonalization - Trace =  1.700e+01
Num  Eigenval.|Num. Eigenval.|Num. Eigenval.|Num. Eigenval.|
001  3.954e+00|002  1.292e+00|003  1.156e+00|004  1.102e+00|
005  9.784e-01|006  9.536e-01|007  9.246e-01|008  8.503e-01|
009  8.026e-01|010  7.709e-01|011  7.069e-01|012  6.826e-01|
013  6.700e-01|014  6.224e-01|015  5.862e-01|016  5.206e-01|
017  4.271e-01|

E:\Ade4\Traits\A_f0_LS.divp is a binary file with 17 rows and 1 columns
Content: Eigenvalues

E:\Ade4\Traits\A_f0_LS.dico is a binary file with 17 rows and 6 columns
Content: Array coordinates

File :E:\Ade4\Traits\A_f0_LS.dico
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1| 2.880e-01| 6.243e-01|
|   2|-5.090e-01| 4.003e-01|
|   3|-4.878e-01| 4.479e-01|
|   4|-5.290e-01| 4.848e-01|
|   5|-4.014e-01| 5.376e-01|
|   6|-4.097e-01| 4.278e-01|
|----|----------|----------|

RLS squared
---------------------- Correlation matrix -------------------

[  1] 1000
[  2]  118 1000
...
[ 16]   45   47   12   35   10   16   32   27   11    6   38
        37   43   28    1 1000
[ 17]   16   20   14    8   39   22   46    6   32   16   24
         4   30   10    1   20 1000
------------------------------------------------------------
E:\Ade4\Traits\A_f0_LS2 is a binary file with 17 rows and 17 columns
Content: RLS squared
```

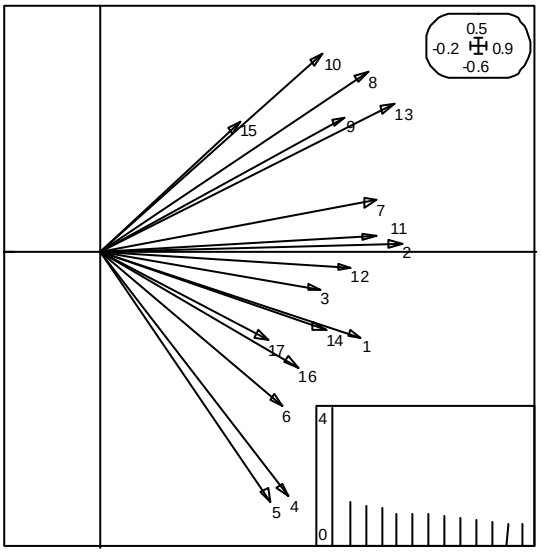
On compare encore les carrés des RLS et les RV :

Fem_Size	Egg_length	Egg-number	Generations	Oviposition	Incubation	Egg_shape	Egg_attach	Clutch_struc	Clutch_numbe	Oviposi_site	Substrat_eggs	Egg_deposition	Gross_habitat	Saturation	Time_day	Season	
1000	118	56	69	30	16	82	42	40	22	43	35	62	35	11	45	16	Fem_Size
118	1000	65	30	26	22	162	60	43	52	94	71	98	51	21	47	20	Egg_length
56	65	1000	24	17	15	38	40	23	22	38	42	33	42	13	12	14	Egg-number
69	30	24	1000	62	32	40	7	19	4	21	12	29	15		35	8	Generations
30	26	17	62	1000	52	8	6	18	6	19	14	9	24	16	10	39	Oviposition
16	22	15	32	52	1000	6	39	19	32	25	29	10	38	1	16	22	Incubation
82	162	38	40	8	6	1000	88	25	109	38	84	62	36	2	32	46	Egg_shape
42	60	40	7	6	39	88	1000	78	81	59	40	136	34	33	27	6	Egg_attach
40	43	23	19	18	19	25	78	1000	104	56	20	142	29	7	11	32	Clutch_struc
22	52	22	4	6	32	109	81	104	1000	16	6	59	19	20	6	16	Clutch_number
43	94	38	21	19	25	38	59	56	16	1000	134	96	73	49	38	24	Oviposi_site
35	71	42	12	14	29	84	40	20	6	134	1000	83	67	13	37	4	Substrat_eggs
62	98	33	29	9	10	62	136	142	59	96	83	1000	22	33	43	30	Egg_deposition
35	51	42	15	24	38	36	34	29	19	73	67	22	1000	6	28	10	Gross_habitat
11	21	13		16	1	2	33	7	20	49	13	33	6	1000	1	1	Saturation
45	47	12	35	10	16	32	27	11	6	38	37	43	28	1	1000	20	Time_day
16	20	14	8	39	22	46	6	32	16	24	4	30	10	1	20	1000	Season





Les carrés de coefficients RLS sont sensiblement proportionnels au RV, la liaison étant meilleure qu'avec les carrés des corrélations canoniques de nature très différente. C'est pourquoi l'image euclidienne associée à ces est sensiblement celle de l'inter-structure de STATIS :



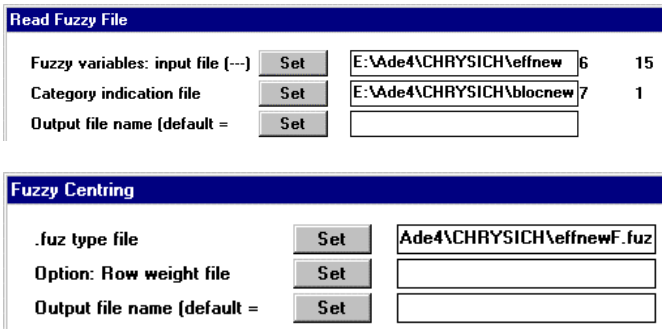
Toutes ces remarques montrent qu'avec ADE-4, il est possible d'étendre la notion de matrice de corrélation entre variables à celle de matrices de corrélation entre tableaux sans grand effort.

Corrélations entre locus

Les tableaux de fréquences alléliques supportent une telle approche. En effet, ils s'agit de variables floues définie par la distribution de fréquences des allèles d'un locus donné dans une population. Prenons l'exemple de la carte Chrysich [9]. Enlevons les colonnes associées à des allèles non représentés et les locus sans variabilité :

26	0	0	28	0	28	0	0	24	10	0	0	30	0	28	20	0	32	0	0	30	0	16	0
28	0	0	32	0	32	0	0	24	18	0	0	32	0	32	30	0	34	0	0	32	2	22	0
89	5	0	92	4	92	0	0	73	23	0	0	96	0	96	60	0	84	0	16	62	0	60	0
12	0	0	12	0	12	0	0	12	0	0	0	12	0	12	12	0	12	0	2	10	0	12	0
57	5	0	66	2	64	0	0	46	8	0	0	66	0	66	44	0	62	0	32	28	8	52	0
28	0	28	0	28	0	0	1	10	3	0	0	28	0	28	22	0	1	11	12	0	24	0	0
26	0	0	28	0	28	0	24	10	32	0	0	30	0	16									
28	0	0	32	0	32	0	24	18	34	0	0	32	2	22									
89	5	0	92	4	92	0	73	23	84	0	16	62	0	60									
12	0	0	12	0	12	0	12	0	12	0	2	10	0	12									
57	5	0	66	2	64	0	46	8	62	0	32	28	8	52									
28	0	28	0	28	0	1	10	3	1	11	12	0	24	0									

On obtient des données du type 2/2/2/3/2/2/2. Passer en binaire le résultat et centrer :



Les RV sont donnés par Canonical: RV coefficients:

```
---------------------- Correlation matrix -------------------
[  1] 1000
[  2]   94 1000
[  3]   72  998 1000
[  4]   51   37   36 1000
[  5]   94 1000  998   37 1000
[  6]   27  746  770   60  746 1000
[  7]   54  980  981   41  980  804 1000
-------------------------------------------------------------
```

Les carrés de RLS sont donnés par Canonical: RLS coefficients :

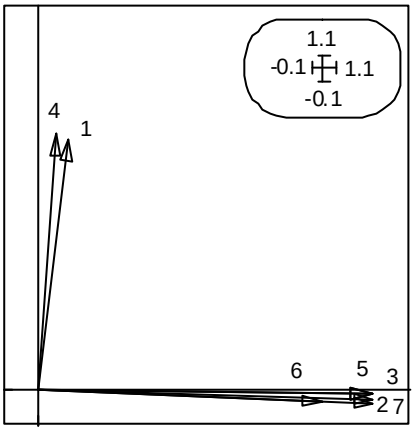
```
---------------------- Correlation matrix -------------------
[  1] 1000
[  2]   94 1000
[  3]   72  998 1000
```

```
[  4]   50   36   35 1000
[  5]   94 1000  998   36 1000
[  6]   27  746  770   58  746 1000
[  7]   54  980  981   40  980  804 1000
------------------------------------------------------------
```

Les carrés de corrélation canonique sont donnés par Canonical: CC coefficients :

```
---------------------- Correlation matrix ------------------
[  1] 1000
[  2]    9 1000
[  3]    5  996 1000
[  4]   18 1000  996 1000
[  5]    9 1000  996 1000 1000
[  6]    1  556  594  664  556 1000
[  7]    3  961  962  962  961  647 1000
------------------------------------------------------------
```

Les deux premières familles sont quasiment identiques, la troisième est différente. Les conditions numériques sont strictement défavorables à l'usage de l'analyse canonique (les tableaux 4 et 2 réunissent 5 colonnes pour 6 lignes). Chacun des tableaux induit une structure de rang 1 (à l'exception du quatrième) et RV et RLS sont alors confondus. Il y a deux compromis :

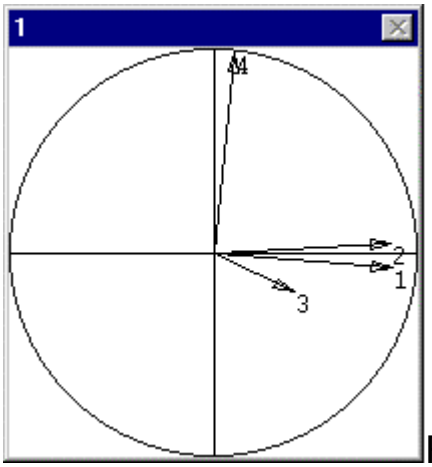


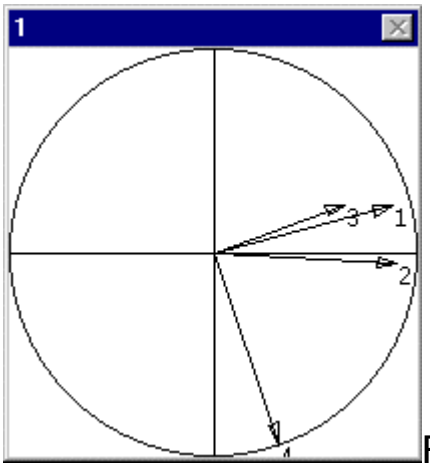
Les locus 2, 3, 5, 6 et 7 isolent la population 6 (Niger), les locus 1 et 4 ne participent pas à cet isolement et ont une autre signification (ou n'ont pas de signification). Même sur un exemple très simple comme celui-ci on voit que tous les gènes ne disent pas tous la même chose. Pour des tableaux à locus très polymorphes, cette pratique peut être fort utile.

Pour les données de la carte Chevaine **[10]**, les conditions numériques sont différentes. On a 25 groupes pour des locus peu polymorphes. On trouve :

```
RV coefficients Escoufier 1973
---------------------- Correlation matrix ------------------
[  1] 1000
[  2]  610 1000
[  3]  171  136 1000
[  4]    4   76    7 1000
------------------------------------------------------------
RLS squared
---------------------- Correlation matrix ------------------
[  1] 1000
[  2]  577 1000
```

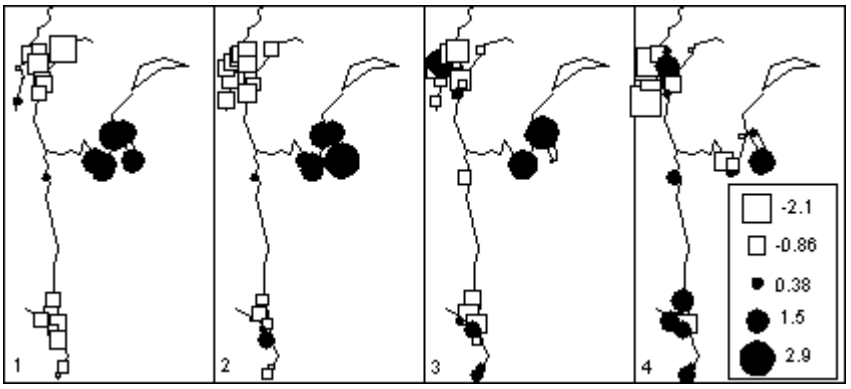
```
[  3]  162  136 1000
[  4]    4   76    7 1000
------------------------------------------------------------
cancor squared
---------------------- Correlation matrix ------------------
[  1] 1000
[  2]  402 1000
[  3]   81   19 1000
[  4]    5    6    0 1000
------------------------------------------------------------
```



RV


RLS

Les deux premiers donne une typologie de locus identique (seuls deux d'entre eux représentent une information cohérence) tandis que l'analyse canonique généralisée montre cette cohérence :

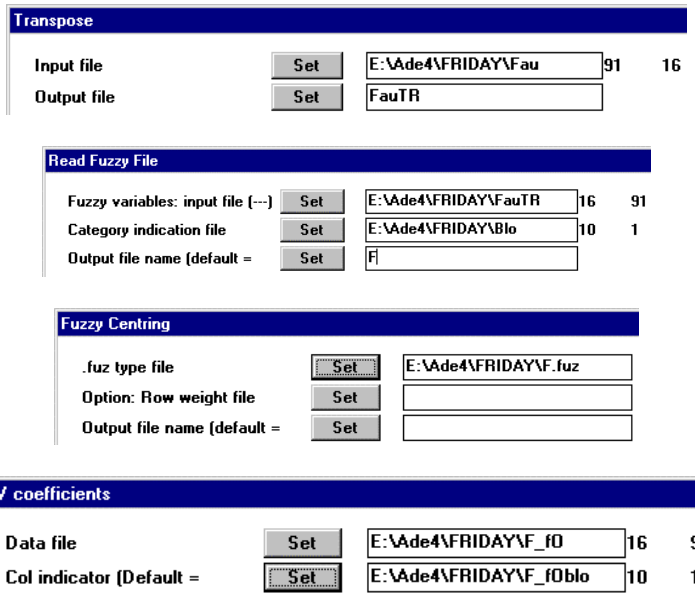


Pour chaque tableau k, on a constitué un score de variance unité $\mathbf{z}_k$ par combinaison des variables de ce tableau et une variable normée $\mathbf{z}$ de référence. Ces scores optimise la quantité :

$$\sum_{k=1}^{K} cor^2\left(\mathbf{z}_k, \mathbf{z}\right)$$

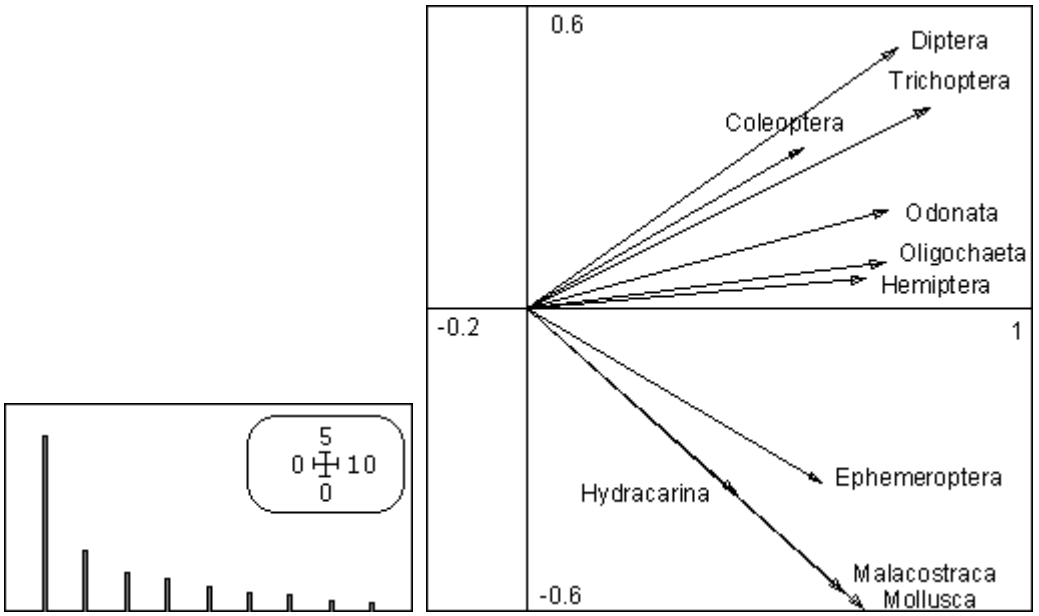
Les deux premiers locus fabriquent le même score, le troisième s'en approche imparfaitement et le quatrième fait autre chose. Tous les tableaux n'ont pas la même fonction.

Sur les données de L. Friday, que nous avons beaucoup utilisé [11], on obtient rapidement :

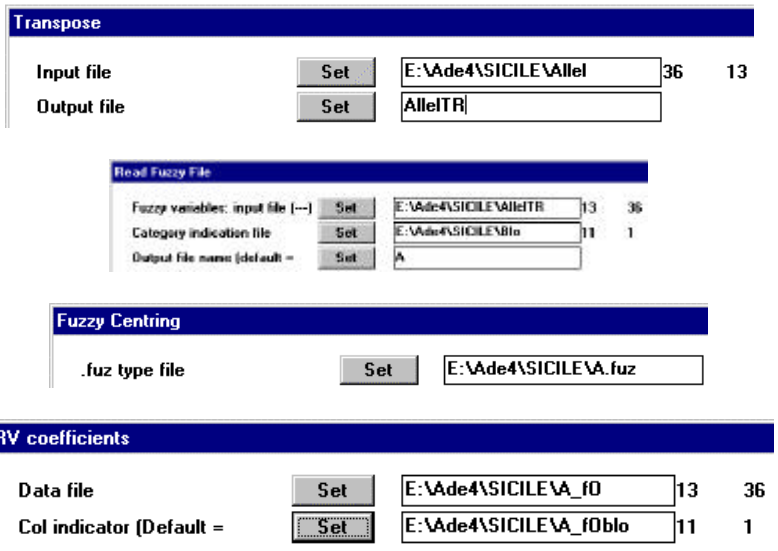



```
---------------------- Correlation matrix ------------------
[  1] 1000
[  2]  329 1000
[  3]  517  563 1000
[  4]  254  377  340 1000
[  5]  283  326  497  220 1000
[  6]  407  645  729  229  464 1000
[  7]  273  129  212  116  178  147 1000
[  8]  327  370  271  523  206  191  235 1000
[  9]  360  381  292  458  202  224  483  676 1000
[ 10]  543  367  551  329  232  514  237  341  368 1000
-----------------------------------------------------------------
```

Les RV sont importants. Il est possible qu'il y ait deux compromis :

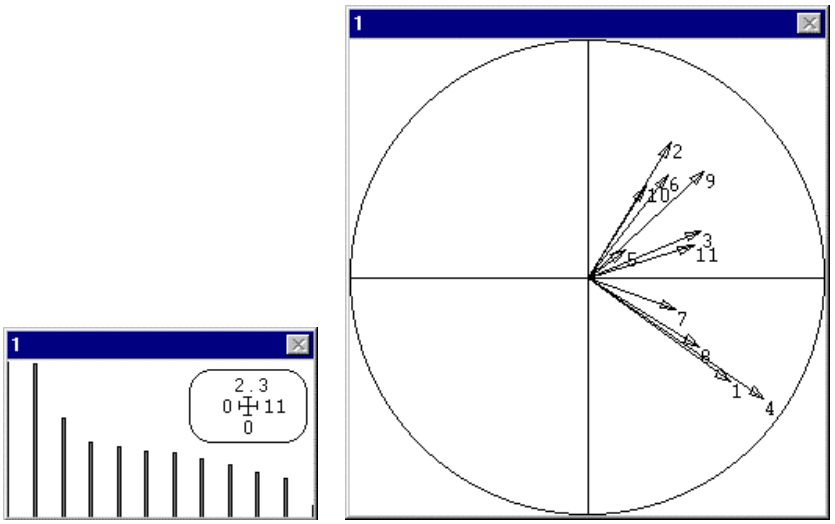


Une analyse K-tableaux approfondie est utile. Une dernière illustration porte sur la carte Sicile **12** :

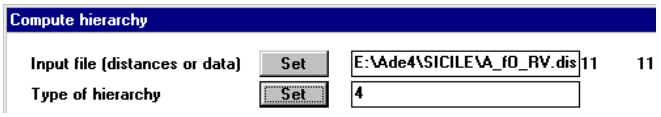



```
Input file: E:\Ade4\SICILE\A_f0
    -> Rows: 13, columns: 36
    -> 11 blocks: 3/4/4/5/3/2/2/3/3/2/5/
RV coefficients Escoufier 1973
---------------------- Correlation matrix ------------------
[  1] 1000
[  2]   71 1000
[  3]  203   59 1000
[  4]  627   26   78 1000
[  5]   45   15   92   18 1000
[  6]   11  282  189  128   65 1000
[  7]  117    8  199  184   59    0 1000
[  8]   61   24   91  462   41   53  144 1000
[  9]   72  237  274  113   26  125  148  105 1000
[ 10]   13  149   18   38   88   69    4   82  227 1000
[ 11]  112  198  149  271   37   63   42   39  172   49 1000
------------------------------------------------------------
```

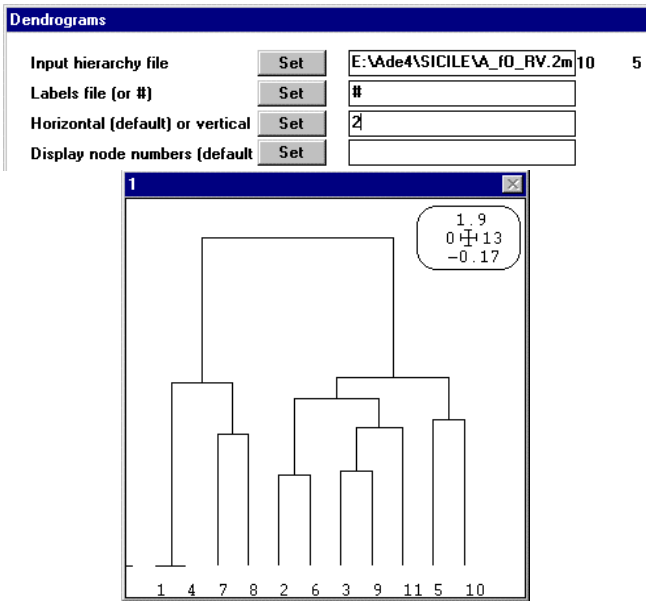
Les RV sont ici très variables et la situation n'est pas simple :



Le coefficient RV définit une distance entre tableaux dont on a ci-dessus la représentation euclidienne. On peut alors faire une classification des locus :




```
Clusters: Compute hierarchy
Data file: E:\Ade4\SICILE\A_fO_RV.dist
Number of rows: 11, columns: 11
Output file: E:\Ade4\SICILE\A_fO_RV.2mha
Number of rows: 10, columns: 5
Hierarchy algorithm used : second order moment (Ward's method)
```



On est passé de la typologie de variables (les allèles) à une typologie de tableaux (les locus). De manière générale, le plus utile des coefficients de corrélation est le RV. Il est central. Dans le situation « beaucoup d'individus pour peu de variables par tableau » il est en compétition avec le carré de corrélation canonique qui invite à l'analyse canonique généralisée. Dans la situation « peu d'individus pour beaucoup de variables par tableau » il est proche du RLS. Rappelons que RV et RLS se réfère à l'analyse de co-inertie de deux tableaux respectivement par :

$$RLS(\mathbf{X},\mathbf{Y}) = \frac{\sum_{k=1}^{r} Cov(\mathbf{X}\mathbf{u}_k, \mathbf{Y}\mathbf{v}_k)}{\sqrt{\sum_{i=1}^{r_X} l_i\left(\frac{1}{n}\mathbf{X}^t\mathbf{X}\right)\sum_{j=1}^{r_Y} l_j\left(\frac{1}{n}\mathbf{Y}^t\mathbf{Y}\right)}}$$

$$RV(\mathbf{X},\mathbf{Y}) = \frac{\sum_{k=1}^{r} Cov^2(\mathbf{X}\mathbf{u}_k, \mathbf{Y}\mathbf{v}_k)}{\sqrt{\sum_{i=1}^{r_X} l_i^2\left(\frac{1}{n}\mathbf{X}^t\mathbf{X}\right)\sum_{j=1}^{r_Y} l_j^2\left(\frac{1}{n}\mathbf{Y}^t\mathbf{Y}\right)}}$$

Le RV envoie à STATIS tandis que le RLS est plus proche de l'analyse de co-inertie multiple elle-même proche de l'analyse factorielle multiple. Kiers a montré que les matrices de RLS ne sont pas nécessairement semi-définie positive (voir la documentation du module Canonical).

Elles le sont souvent en pratique. Mais l'équivalent inter-tableaux de la corrélation inter-variables est le RV d'Escoufier. La simple mesure de cette corrélation lorsqu'on a des données structurées en multiples tableaux semblent un préalable indispensable pour orienter l'analyse.

Références

1 Statzner, B., Hoppenhaus, K., Arens, M.-F. & Richoux, Ph. (1997) Reproductive traits, habitat use and templet theory: a synthesis of world-wide data on aquatic insects. *Freshwater Biology* : 38, 109-135.

2 Doledec, S. & Chessel, D. (1994) Co-inertia analysis: an alternative method for studying species-environment relationships. *Freshwater Biology* : 31, 277-294.

3 Chevenet, F., Dolédec, S. & Chessel, D. (1994) A fuzzy coding approach for the analysis of long-term ecological data. *Freshwater Biology* : 31, 295-309.

4 Chessel, D. & Hanafi, M. (1996) Analyses de la co-inertie de K nuages de points. *Revue de Statistique Appliquée* : 44, 35-60.

5 Carrol, J.D. (1968) A generalization of canonical correlation analysis to three or more sets of variables. *Proceeding of the 76th Convention of the American Psychological Association* : 3, 227-228.

6 Lingoes, J.C. & Schönemann, P.H. (1974) Alternative measures of fit for the Schönemann-Carrol matrix fitting algorithm. *Psychometrika* : 39, 423-427.

7 Lazraq, A., Cléroux, R. & Kiers, H.A.L. (1992) Mesures de liaison vectorielle et généralisation de l'analyse canonique. *Revue de Statistique Appliquée* : 39, 23-35

8 Kiers, H.A.L, Cléroux, R. & Ten Berge, M.F. (1994) Generalized analysis based on optimizing matrix correlations and a relation with IDIOSCAL. *Computational Statistics and Data Analysis* : 18, 331-340.

9 Agnese, J.F. (1989) *Différenciation génétique de plusieurs espèces de Siluriformes ouest-africains ayant un intérêt pour la pèche et l'aquaculture*. Thèse de Doctorat, Université des Sciences et Techniques du Languedoc, Montpellier. 1-194.

10 Guinand, B., Bouvet, Y. & Brohon, B. (1996) Spatial aspects of genetic differentiation of the European chub in the Rhone River basin. *Journal of Fish Biology* : 49, 714-726.

11 Friday, L.E. (1987) The diversity of macroinvertebrate and macrophyte communities in ponds. *Freshwater Biology* : 18, 87-104.

12 Pigliucci, M. & Barbujani, G. (1991) Geographical patterns of gene frequencies in Italian populations of Ornithogalum montanum (Liliaceae). *Genetical research, Cambridge* : 58, 95-104.