

Individus et informations supplémentaires

Résumé

La fiche décrit l'usage des individus supplémentaires dans des circonstances variées. En particulier, cette pratique est étendue aux analyses inter et intra-classes et aux analyses de co-inertie. Une illustration porte sur l'analyse de co-inertie de deux analyses inter-classes. On distingue ce qui relève directement de la projection euclidienne de vecteurs quelconques sur des plans principaux de ce qui touche à l'illustration des plans factoriels par de l'information externe aux tableaux de données.

Plan

1 — La question	2
2 — Individus supplémentaires	3
2.1 — Individus supplémentaires en ACP	4
2.2 — Individus supplémentaires en ACP inter-classes...	5
2.3 — Classes supplémentaires en intra-classes.....	8
3 — Équivalents en analyse des correspondances	11
4 — Lignes supplémentaires en co-inertie	14
5 — Couplage d'inter-classes	17
6 — Information supplémentaire	20
Références	24

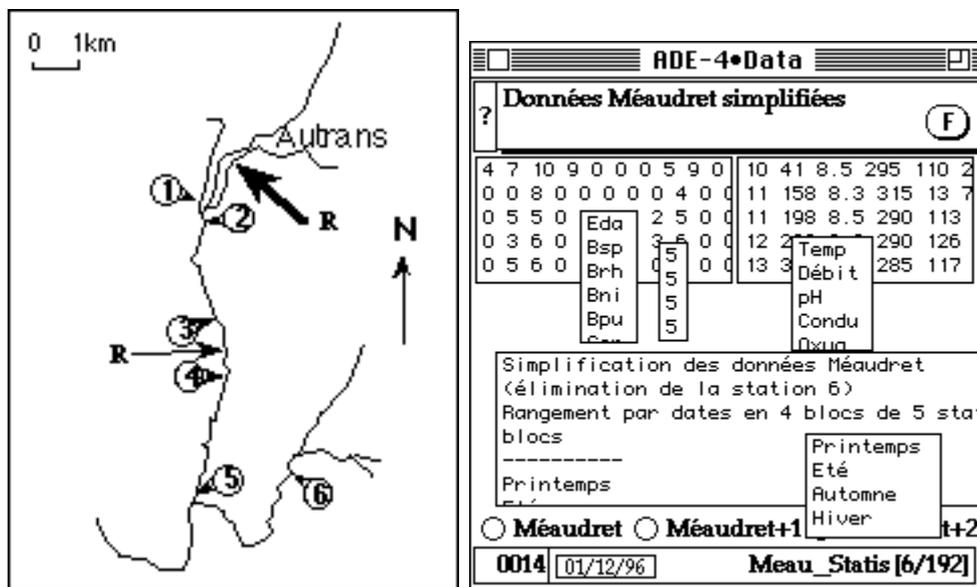
D. Chessel

1 — La question

Les conversations sur Internet sont souvent intéressantes. Un utilisateur averti dit :

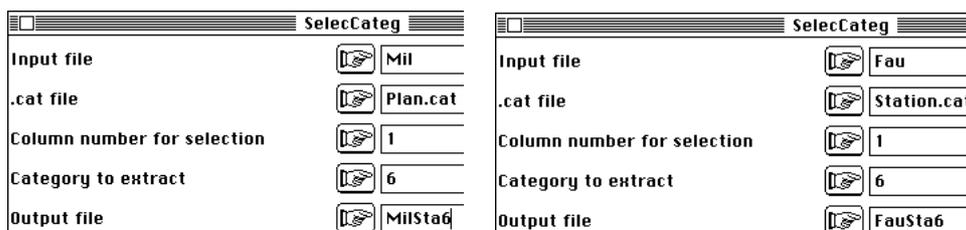
J'ai une petite question sur les lignes supplémentaires dans une AFC inter. En plus des lignes du tableau initial (qu'on obtient dans le fichier .bels), je veux projeter de vrais lignes supplémentaires dans une AFC inter. Comment faut-il faire ? Je pense qu'il faut faire la transformation initiale (pij/pi.p.j)-1 en utilisant les marges colonnes de l'AFC inter et les marges lignes des lignes supplémentaires, mais je me demandais si il n'y avait pas un truc ou une option spéciale dans un des modules qui permette de le faire automatiquement plutôt qu'à la main ?

Pour faire une réponse utile, on peut construire la situation expérimentale suivante, à partir des (inépuisables) données des cartes Méaudret de la pile ADE-4•Data.

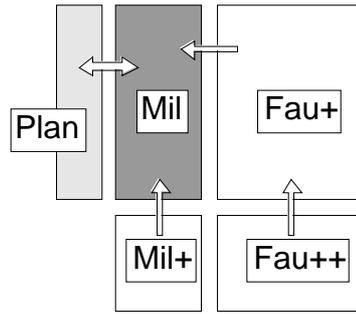


On voit bien, sur la carte des stations, que 5 d'entre elles sont dans un gradient amont-aval, la sixième servant de référence sur une autre rivière. C'est le cas typique de la présence d'individus supplémentaires. Sur la carte meau_Stat6 qui illustre la méthode STATIS sur les X, cette station est éliminée. Créer un dossier de travail contenant ces données.

A partir des autres cartes de la série, extraire l'information correspondant à la station supplémentaire :



On se retrouve avec une situation fort simple qui pose la question du supplémentaire de multiples façons :



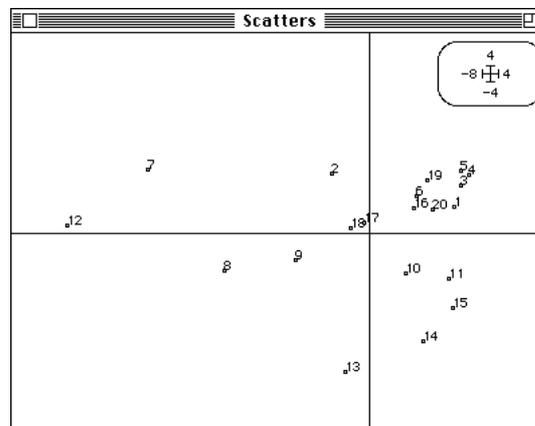
2 — Individus supplémentaires

Plusieurs auteurs ont déjà souligné que le terme supplémentaire s'applique souvent de manière abusive à tout ce qui ne fait pas partie du tableau des données alors qu'on devrait bien réserver le terme projection en individus supplémentaires à une opération géométrique précise.

Faisons l'ACP normée du tableau Mil de référence :

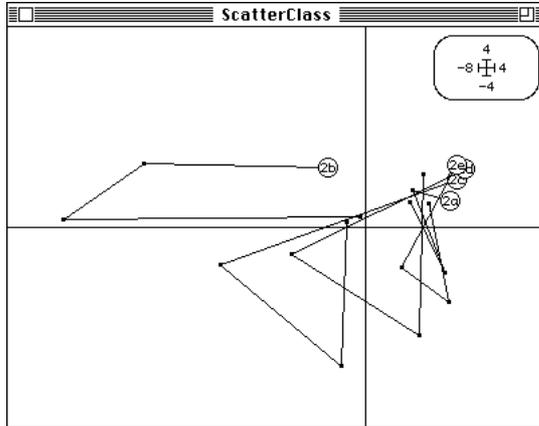
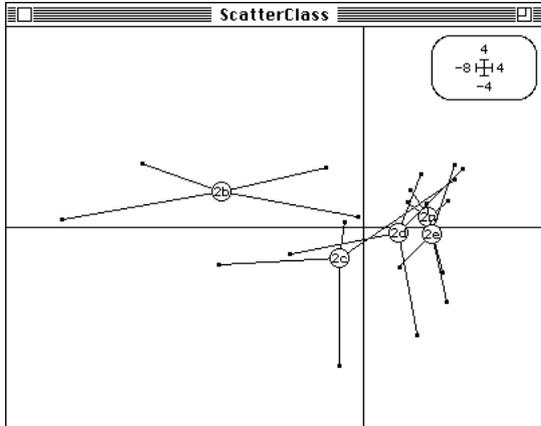
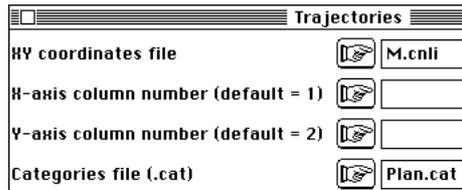
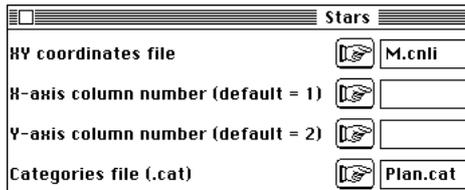


Nous conserverons 2 axes pour simplifier l'illustration.

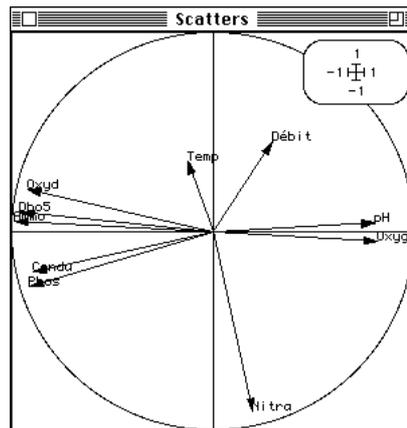
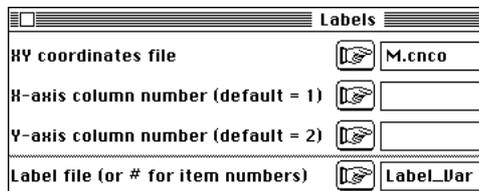


Dès que cette carte est tracée, on cherche à voir ce qu'elle ne contient pas explicitement, c'est à dire la disposition spatiale et temporelle des points :





a, b, c, d et e sont les labels des stations (a est la référence amont, b la station polluée, c, d, et e les stations aval progressivement restaurées. Cela correspond à l'interprétation par les variables :



On a représentée sur la carte des stations de l'information extérieure. L'opération est légitime (on a une idée rapidement de la synergie spatio-temporelle) mais n'est en rien une projection en individus supplémentaires.

2.1 — Individus supplémentaires en ACP

Pour représenter la station 6 on peut projeter les points correspondants sur le plan des axes principaux du nuage. Les lignes du fichier MilSta6 sont des vecteurs de \mathbb{R}^{10} , l'espace dans lequel on trouve les 20 vecteurs lignes du tableau M. Les données dans le tableau M ont été normalisées, ce qui signifie qu'au vecteur

$$(x_{i1}, x_{i2}, \dots, x_{ip})$$

on a substitué le vecteur :

$$\frac{x_{i1} - m(1)}{\sqrt{\text{var}(1)}}, \frac{x_{i2} - m(2)}{\sqrt{\text{var}(2)}}, \dots, \frac{x_{ip} - m(p)}{\sqrt{\text{var}(p)}}$$

C'est le vecteur transformé qui est projeté et pour les individus supplémentaires la même transformation s'impose. Elle est implicite dans l'option DDUtil : Supplementary rows :

Supplementary rows			
Initial analysis (up)		M.cnvp	10 2
Supplementary row file		MilSta6	4 10
Output file name		MS	

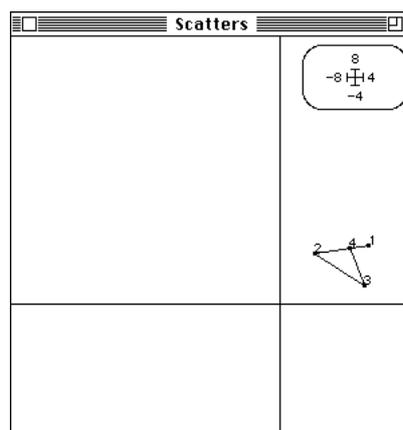
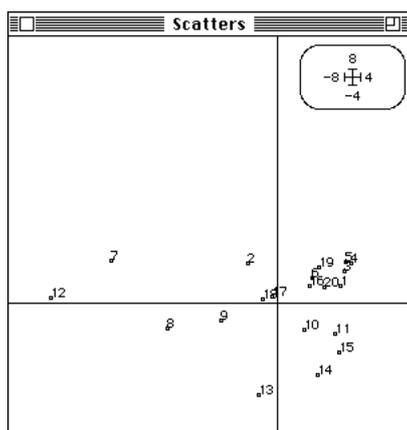
File MS contains the coordinates of the projected rows
It has 4 rows and 2 columns

File :MS

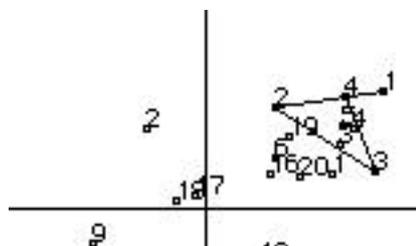
Col.	Mini	Maxi
1	1.041e+00	2.669e+00
2	5.472e-01	1.730e+00

Labels	
HV coordinates file	M.cnl
H-axis column number (default = 1)	
V-axis column number (default = 2)	
Label file (or # for item numbers)	#

Trajectories	
HV coordinates file	MS
H-axis column number (default = 1)	
V-axis column number (default = 2)	
Label file (or # for item numbers)	#

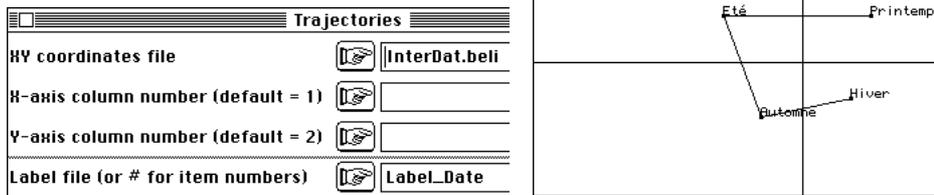


Comme il est possible de copier-coller entre deux figures, la station 6 vient prendre position comme témoin non pollué :

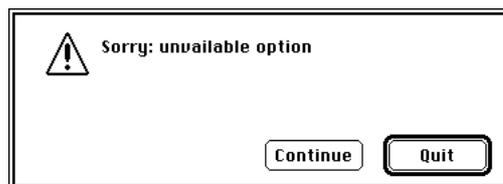


Cette opération est bien connue mais la question posée est beaucoup plus subtile : peut-on l'étendre en ACP inter ou intra-classes ?

2.2 — Individus supplémentaires en ACP inter-classes



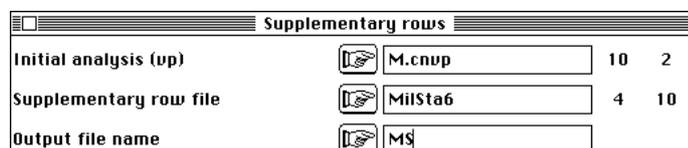
Sur l'axe 2, une composante saisonnière indépendante est représentée par la seule variable température. Que faire ici de la station 6 ? Répondre suppose qu'on ait d'abord identifié les opérations exécutées dans l'inter-classe. Le tableau normalisé est moyenné par classe (date) ce qui donne un tableau de centre de gravité 4-10. Les 4 points de \mathbb{R}^{10} définissent un plan principal, sur lequel ils sont projetés (fichier ---.beli). L'analyse elle-même contient une projection en individus supplémentaires des lignes du tableau normalisés (fichier ---.bels). Il serait donc logique de projeter en lignes supplémentaires les mêmes points que les lignes supplémentaires de l'analyse initiale :



Manifestement le coup n'était pas prévu. Les options disponibles sont repérées par le type de l'analyse c'est à dire les deux premiers caractères de l'extension des noms de fichiers, soit cn (ACP normée), cp (ACP centrée) et fc (AFC). Il est dommage que nous ne puissions récupérer le tableau des vecteurs :

$$\frac{x_{i1} - m(1)}{\sqrt{\text{var}(1)}}, \frac{x_{i2} - m(2)}{\sqrt{\text{var}(2)}}, \dots, \frac{x_{ip} - m(p)}{\sqrt{\text{var}(p)}}$$

Puisqu'il est calculé dans l'option DDUtil : Supplementary rows il est regrettable de ne pouvoir en disposer. C'est réparé avec la nouvelle version :



Title of the analysis: M.cnta
Number of rows: 20, columns: 10

Supplementary row after transformation in : MS_tab
Rows :4 Columns: 10

L'opération prend un sens car le module contient l'option DDUtil : Row projections qui projette des lignes après une analyse sans modification. Sur l'analyse de type be on peut donc projeter des vecteurs lignes modifiés par ailleurs :

Row projections		
Initial analysis (vp)	<input type="text" value="InterDat.beup"/>	4 2
Supplementary row file	<input type="text" value="MS_tab"/>	4 10
Output file name	<input type="text" value="InterS"/>	

Title of the analysis: InterDat.beta
 Number of rows: 4, columns: 10

File InterS contains the coordinates of the projected rows
 It has 4 rows and 2 columns

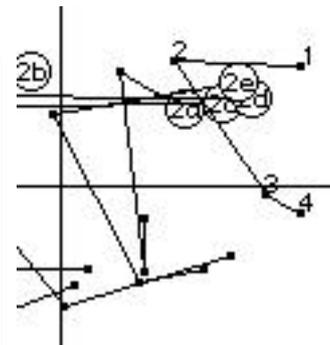
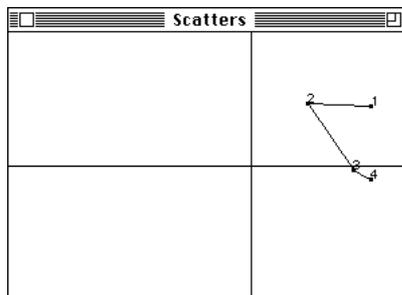
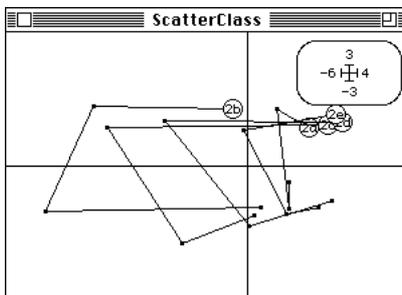
File :InterS

Col.	Mini	Maxi
1	1.403e+00	2.974e+00
2	-3.136e-01	1.386e+00

D'où les représentations :

Trajectories	
HV coordinates file	<input type="text" value="InterDat.bels"/>
H-axis column number (default = 1)	<input type="text"/>
V-axis column number (default = 2)	<input type="text"/>
Categories file (.cat)	<input type="text" value="Plan.cat"/>

Trajectories	
HV coordinates file	<input type="text" value="InterS"/>
H-axis column number (default = 1)	<input type="text"/>
V-axis column number (default = 2)	<input type="text"/>
Label file (or # for item numbers)	<input type="text" value="#"/>



puisqu'on peut aussi copier-coller entre deux fenêtres de deux modules.

2.3 — Classes supplémentaires en intra-classes

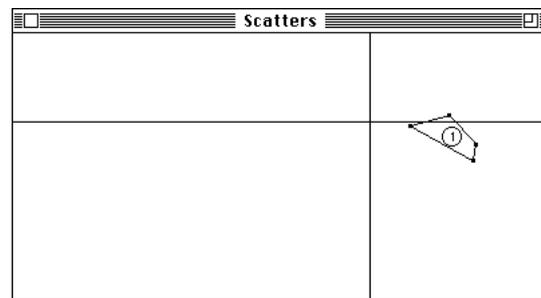
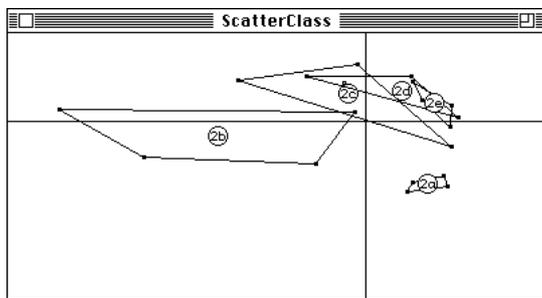
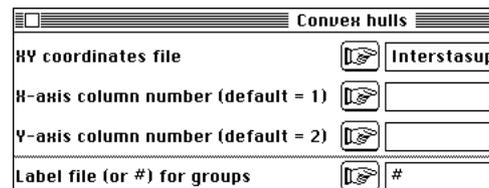
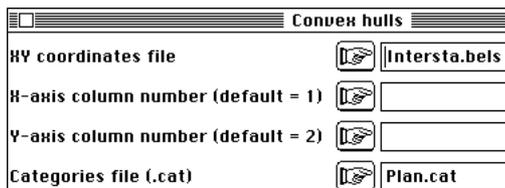
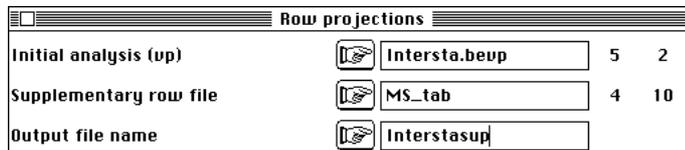
Continuons la réflexion ouverte par la question posée. Nous venons de faire une ACP inter-dates et la station supplémentaire ajoutait un point supplémentaire dans chaque classe. Mais si nous faisons l'analyse inter-stations la station supplémentaire ajoutera une classe. Pour savoir si la projection en individus supplémentaires est possible et valide, là encore il convient de revenir à la procédure interne. Pour l'inter-classe, il n'y a pas de difficulté nouvelle.

Associer la partition par stations et exécuter l'intra-classes :

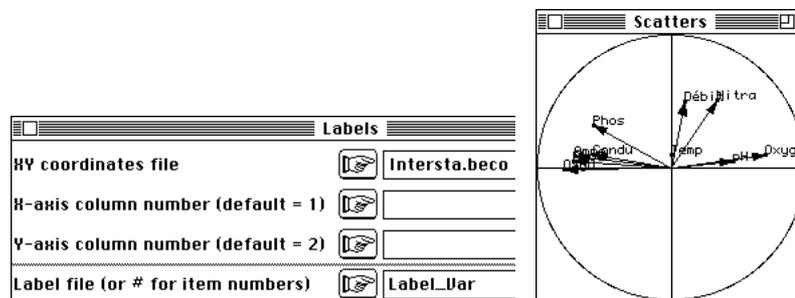
Initialize: LinkPrep		
Statistical triplet	<input type="text" value="M.cnta"/>	20 10
Categories file (.cat)	<input type="text" value="Plan.cat"/>	
Selected variable (default=1)	<input type="text" value="2"/>	
Output file name	<input type="text" value="Intersta"/>	



Projeter les lignes supplémentaires :

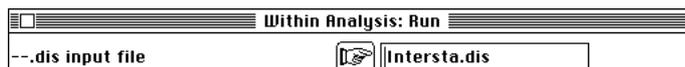


avec :

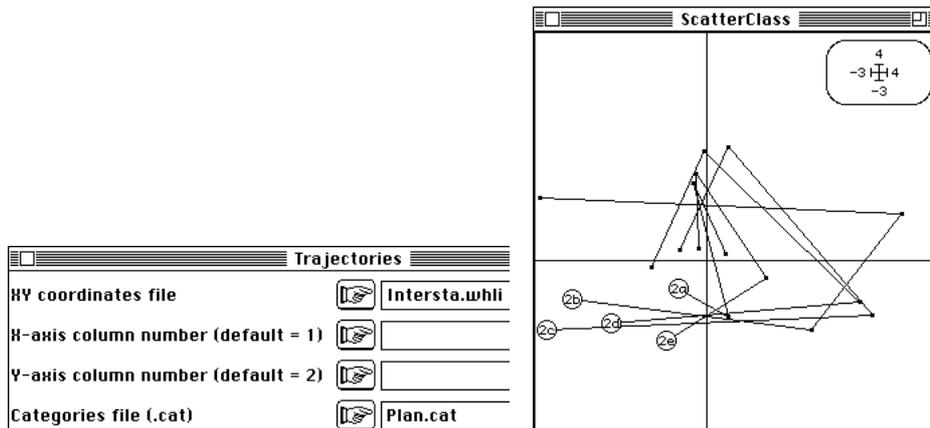


Dans l'intra-classe, par contre, le tableau normalisé est à nouveau centré par classes et le nouveau nuage de 20 points de \mathbb{R}^{10} définit des axes principaux sur lequel sont projetés les lignes recentrées (---.whli) mais aussi les lignes du tableau normalisé (---.whls). Donc, il est concevable de projeter les lignes supplémentaires de l'ACP initiale, ce qui donne la procédure suivante.

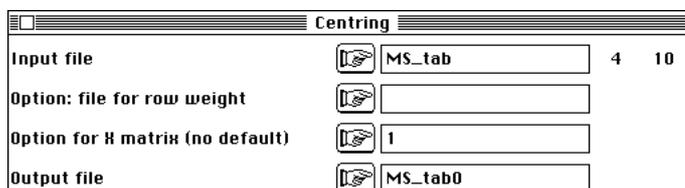
Exécuter l'intra-classes :



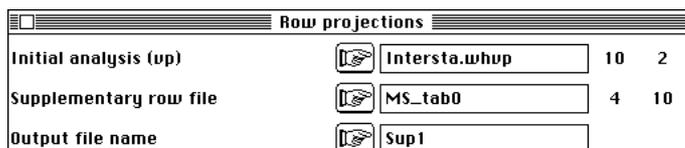
Sur le plan de l'intra-classe on a les nuages recentrés par stations :



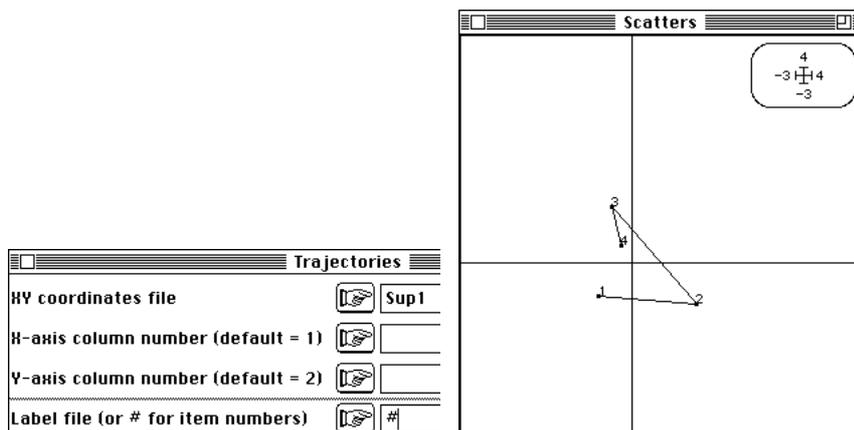
Si on veut la station 6 en supplémentaire sur cette carte, il nous faut le tableau recentré du tableau normalisé par rapport au tableau analysé. Ce n'est pas compliqué :



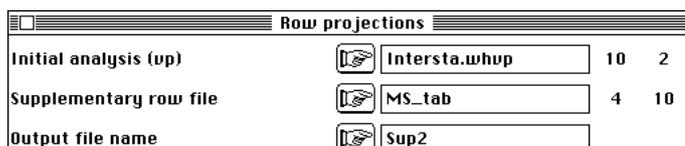
Projeter les lignes supplémentaires :

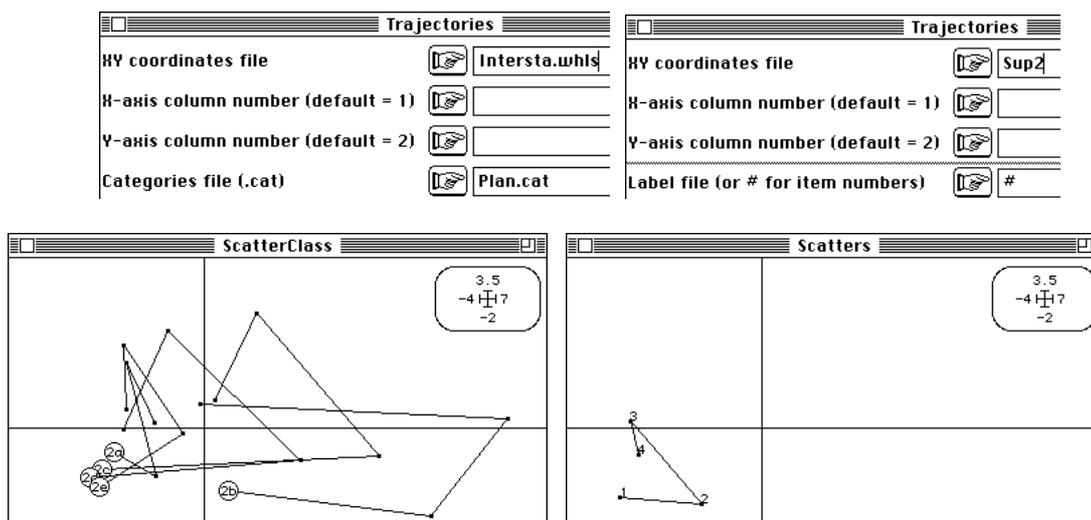


Vérifier la cohérence (les témoins ont des variations très voisines) :



Lorsqu'on utilise les projections des points non recentrés, l'équivalent demande :





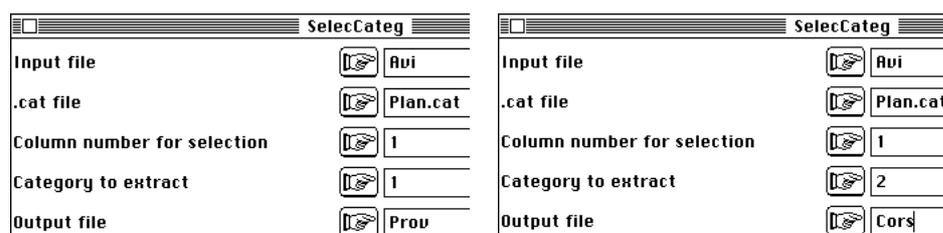
On retiendra donc qu'en inter et intra-classes, la projection des individus supplémentaires est aisée. Seul un minimum d'attention est requis pour associer les nuages de lignes actifs et passifs de manière à ce que les superpositions associent des tableaux ayant subi des traitements cohérents.

3 — Équivalents en analyse des correspondances

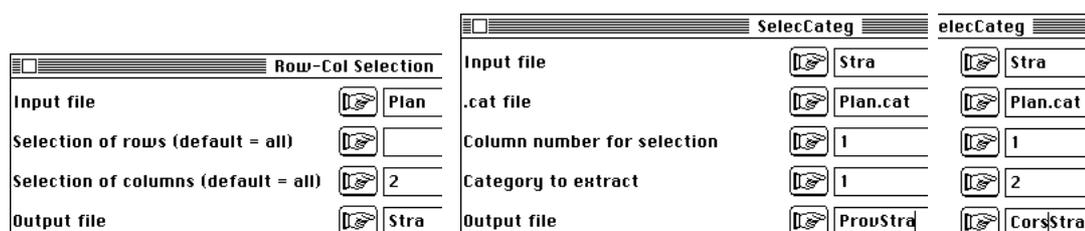
Pour illustrer le fonctionnement de pratiques identiques en AFC, utilisons la carte Provence-Corse de la pile ADE-4•Data. Le tableau faunistique a 302 lignes et 60 colonnes. Lire le fichier associé au plan d'expérience :



Séparer le tableau en deux parties :

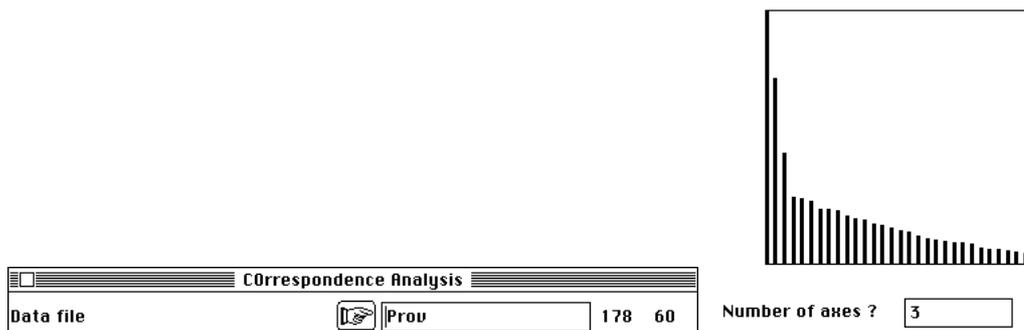


Sélectionner la seconde variable du fichier Plan et séparer en deux parties :

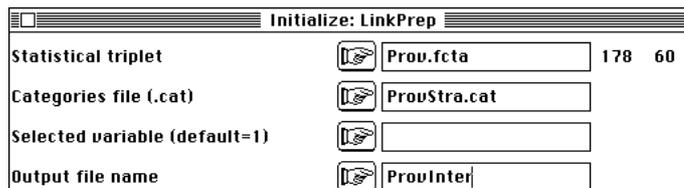


On veut projeter en individus supplémentaires les relevés du second tableau sur l'inter-classe du premier en analyse des correspondances :

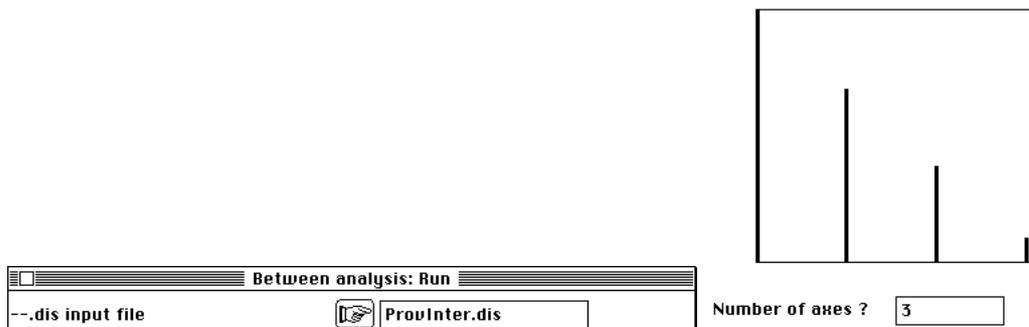
Faire l'analyse initiale :



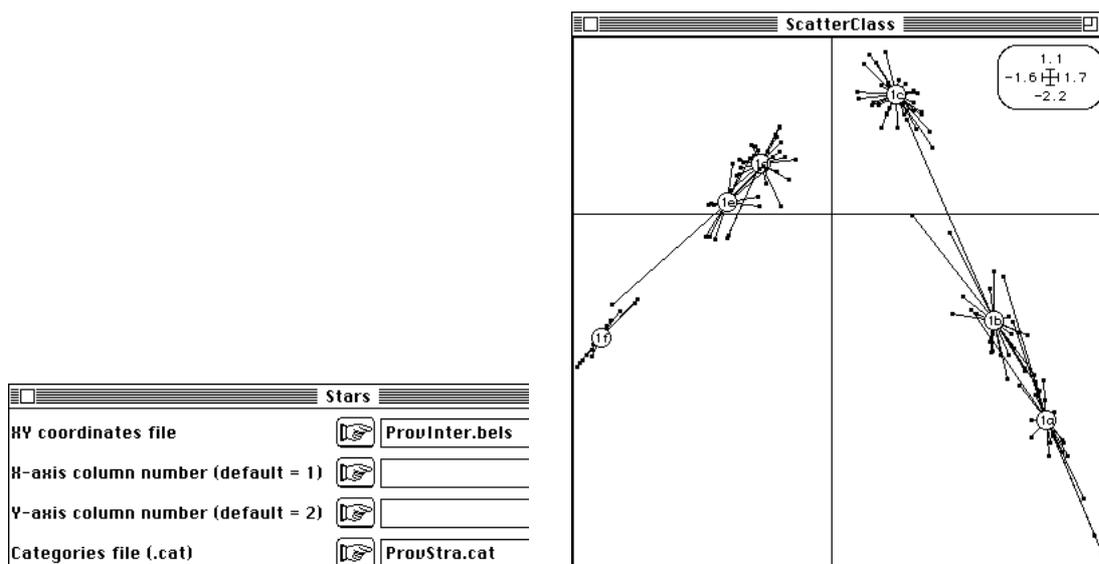
Coupler avec la variable de stratification (architecture de la végétation) :



Exécuter l'analyse inter-classe :



Tracer le plan de l'inter-classe :



Le plan 1-3 mérite aussi d'être examiné.

Projeter en individus supplémentaires les lignes de l'autre tableau :

Supplementary rows			
Initial analysis (up)		Prov.fcup	60 2
Supplementary row file		Cors	124 60
Output file name		CorSup	

Title of the analysis: Prov.fcta
 Number of rows: 178, columns: 60

Supplementary row after transformation in : CorSup_tab
 Rows :124 Columns: 60

File CorSup contains the coordinates of the projected rows
 It has 124 rows and 3 columns

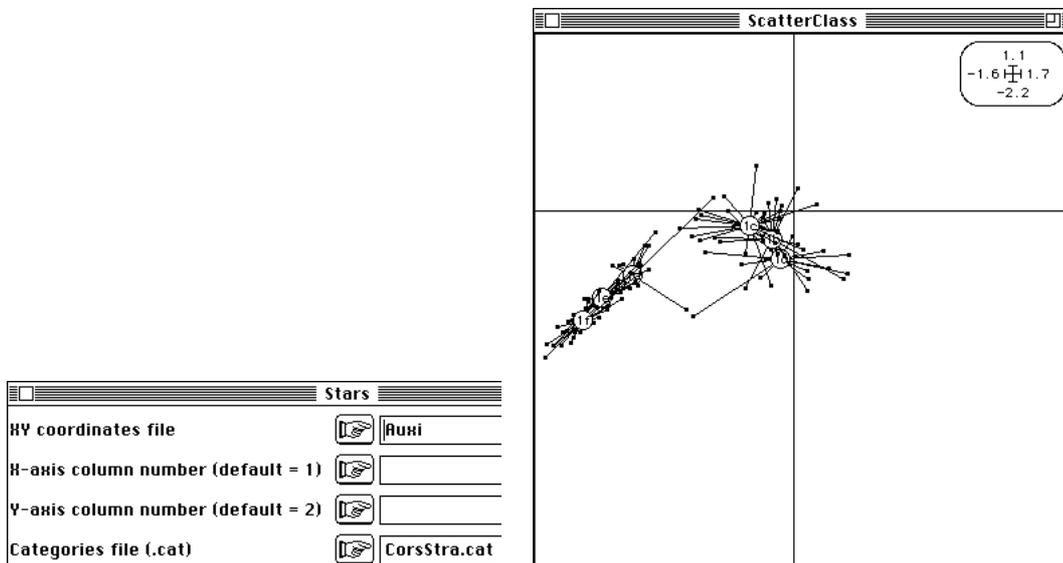
File :CorSup

Col.	Mini	Maxi
1	-1.525e+00	3.463e-01
2	-9.929e-01	2.215e-01
3	-9.422e-01	8.967e-01

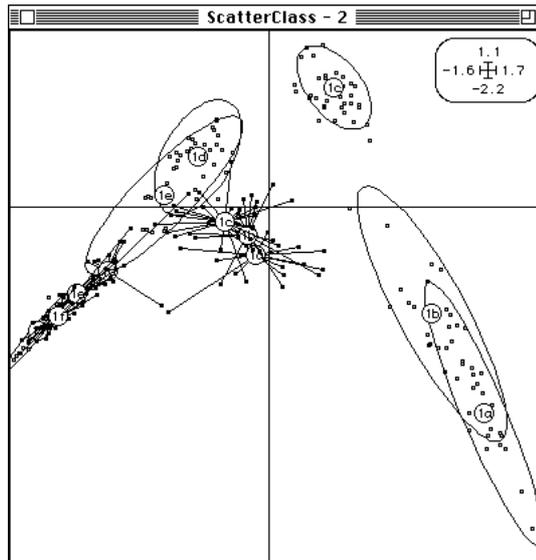
On utilise le fichier _Tab pour faire de la projection dans l'inter-classe :

Row projections			
Initial analysis (up)		ProvInter.beup	6 2
Supplementary row file		CorSup_tab	124 60
Output file name		Ruxi	

On refait alors la carte inter-strate des relevés de Corse en individus supplémentaires sur la carte inter-strate des relevés de Provence :



L'opération est fort significative au plan biologique. La convergence des cortèges faunistiques en milieu forestier ¹ positionne les relevés forestiers des deux zones au même endroit :



Mais, en n'utilisant que les espèces des milieux ouverts de Provence d'une part et en intégrant les différences de structure des communautés en milieu insulaire ² d'autre part, les milieux ouverts de Corse ne prennent pas la place qu'on attend. On peut donc retenir de cet exemple que la représentation d'individus supplémentaires en inter-classe après une ACP (centrée ou normée) et une AFC inter-classe ne pose pas de problème.

4 — Lignes supplémentaires en co-inertie

Profitons de l'occasion pour étendre la notion d'individus supplémentaires à l'analyse de co-inertie. Reprendre le dossier Meau_Statis utilisé au paragraphe 2. Cette opération distingue radicalement l'analyse de co-inertie ³ de l'analyse canonique des correspondances ⁴ dans le cas où il y a des stations supplémentaires dans les deux tableaux.

Faisons l'ACP normée du premier tableau :

Correlation matrix PCA	
Matrix input file	M 20 10

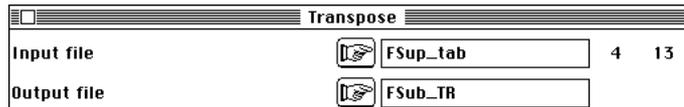
Préparons le tableau des individus supplémentaires de cette analyse :

Supplementary rows	
Initial analysis (vp)	M.cnvp 10 2
Supplementary row file	MilSta6 4 10
Output file name	MSup

Faisons l'ACP centrée du second tableau :

Covariance matrix PCA	
Matrix input file	F 20 13

Préparons le tableau des individus supplémentaires associés :



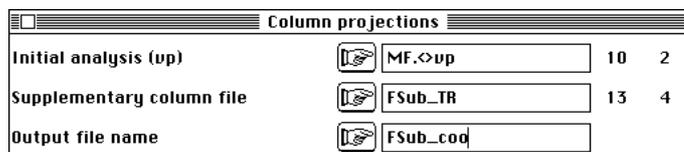
File MF.<>pl contains the weight of each row
It has 13 rows and 1 column

File MF.<>pc contains the weight of each column
It has 10 rows and 1 column

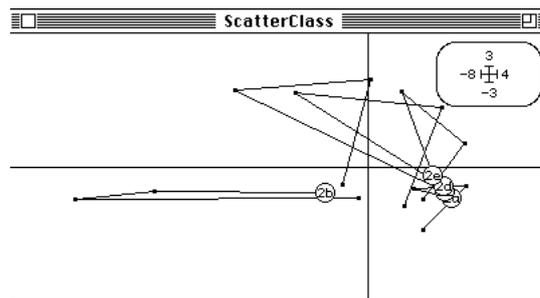
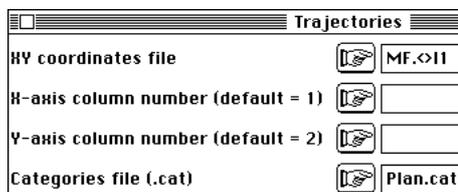
File MF.<>ma contains the name of the input statistical triplet
1 ----> M.cnta
2 ----> F.cpta

total inertia: 101.180710

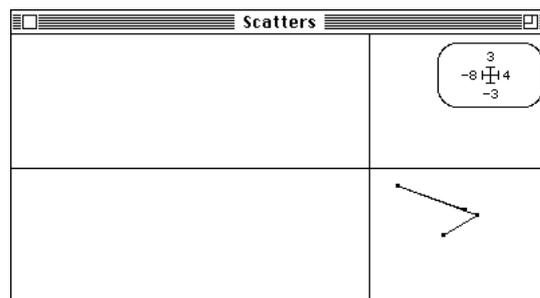
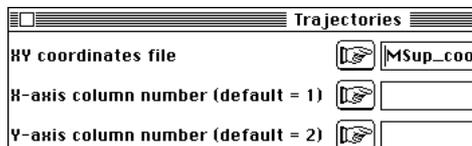
Les relevés de milieu sur le plan de co-inertie sont projetés :



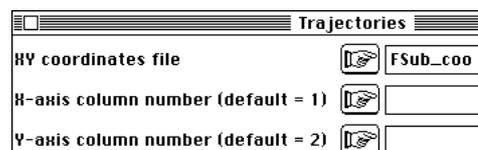
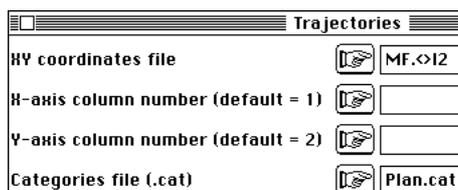
On peut alors représenter les relevés supplémentaires dans les deux espaces de référence. Pour les relevés de milieu :



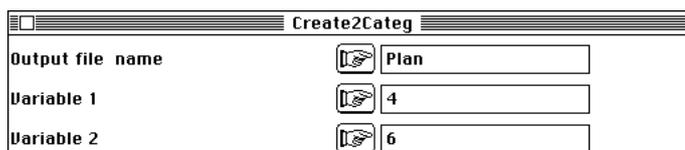
et pour la station supplémentaire :



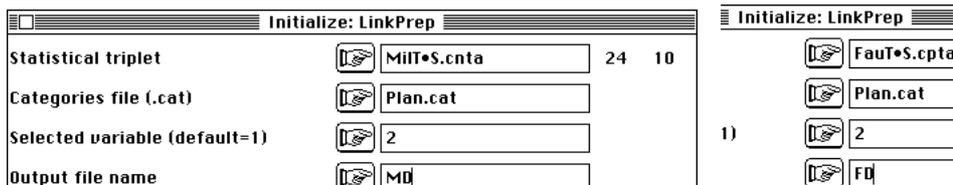
De même dans le tableau faunistique et pour la station supplémentaire :



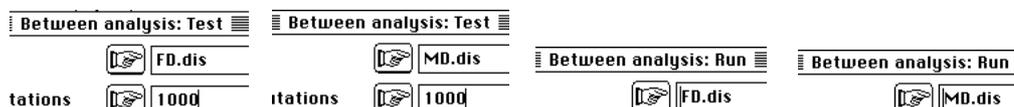
Établir le plan d'observation :



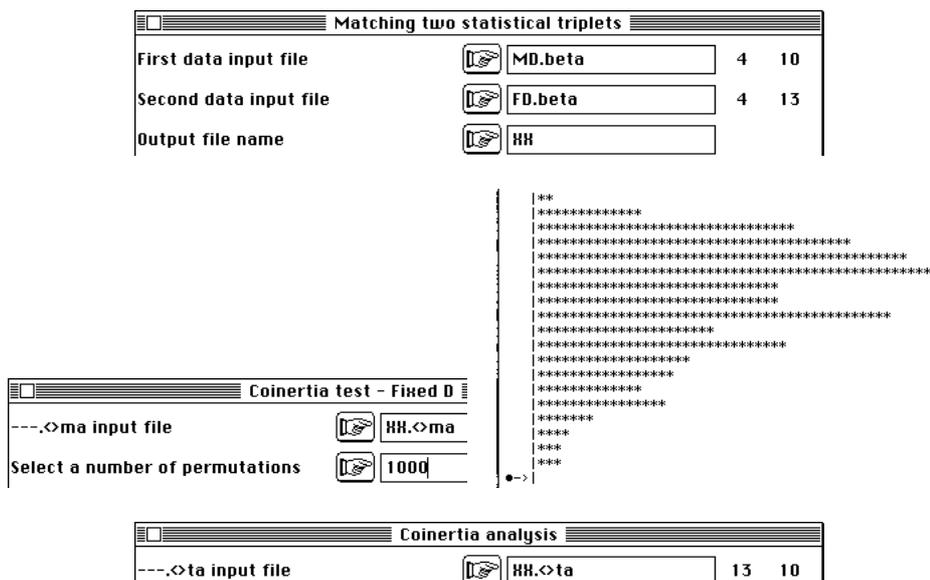
Exécuter l'ACP normée du tableau MilT•S et l'ACP centrée du tableau FauT•S. Préparer les deux analyse inter-classes associées à la partition des tableaux en 6 stations (inter-stations) :



Vérifier que les deux inter-classes sont très significatives et exécuter :



Coupler les deux inter-classes et vérifier la signification du couple:

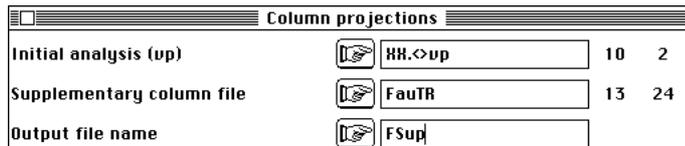
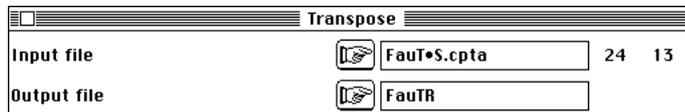


Noter que dans le tableau croisé les espèces sont en lignes et les variables de milieu sont en colonnes. En couplant les deux inter-classes, les lignes des tableaux sont les stations (moyennes sur les 4 dates).

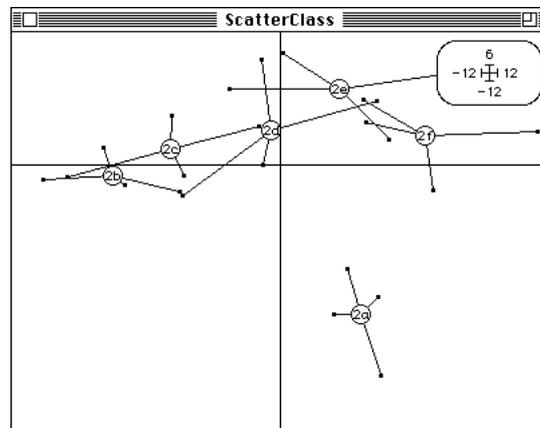
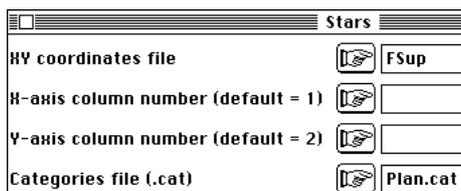
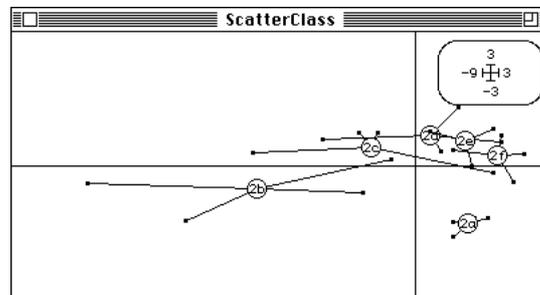
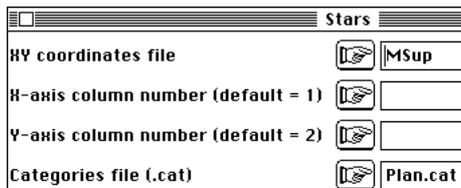
Pour obtenir les points station-date, il faut projeter sur les axes de co-inertie les tableaux d'origine. Comme dans le cas précédent, les lignes du tableau de milieu sont des lignes supplémentaires :



Pour les lignes du tableau faunistique qui deviennent des colonnes supplémentaires du tableau croisé, il faut d'abord transposer :



D'où les figures associées



La sixième station retrouve sa place particulière.

Cette pratique est décrite dans ⁶. On peut l'étendre au couplage d'intra-classes : la méthode STATICO (fiche thématique 9-1) est pour cet objectif préférable.

On fera, pour finir un exercice qui montre que les vecteurs projetés en supplémentaire ne sont pas forcément des individus. Par exemple, on veut projeter sur les axes de co-inertie inter-classes les axes principaux de l'ACP normée du tableau de milieu. Ce sont des vecteurs de \mathbb{R}^{10} dont les composantes sont les scores des variables dits de variance unité :



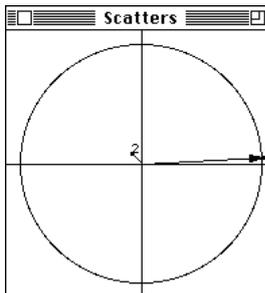
Ils sont donc en colonnes dans MilT•S.cnc1. En transposant, on obtient deux vecteurs de \mathbb{R}^{10} normés pour la métrique canonique qui se projette sur le plan de co-inertie par :

Transpose	
Input file	MilT•S.cnc1 10 2
Output file	FauT

Row projections	
Initial analysis (up)	HH.◇up 10 2
Supplementary row file	FauT 2 10
Output file name	FauSup

La représentation de la projection se fait simplement avec :

Labels	
XY coordinates file	FauSup
X-axis column number (default = 1)	
Y-axis column number (default = 2)	
Label file (or # for item numbers)	#
Draw vectors from origin (yes = 1)	1
Draw unit circle (yes = 1)	1

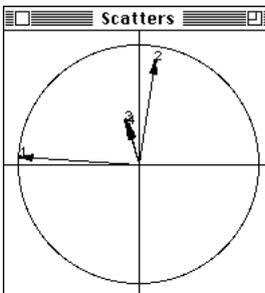


Pour le tableau faunistique la transposition n'est pas utile :

Column projections	
Initial analysis (up)	HH.◇up 10 2
Supplementary column file	FauT•S.cpc1 13 4
Output file name	Fau2Sup

Add normed scores
FauT•S.cpcp

Labels	
XY coordinates file	Fau2Sup
X-axis column number (default = 1)	
Y-axis column number (default = 2)	
Label file (or # for item numbers)	#
Draw vectors from origin (yes = 1)	1
Draw unit circle (yes = 1)	1



6 — Information supplémentaire

Il n'est pas toujours pertinent de privilégier la projection en individus supplémentaire au sens euclidien par rapport à l'expression directe par un procédé quelconque d'information supplémentaire. C'est vrai pour les variables qualitatives et les variables faunistiques (abondance taxonomique). En effet, l'averaging (représentation des moyennes) est en général plus explicite et plus simple à expliquer.

Prenons par exemple, le jeu de données de la carte Light_Trap⁷. Les variables météorologiques (11) donnent une ACP de dépouillement facile :

Correlation matrix PCA

Matrix input file

Labels

HV coordinates file

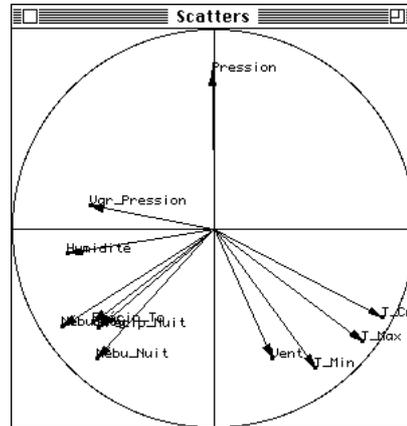
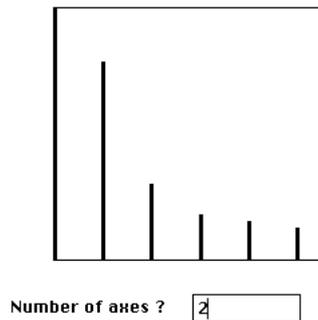
H-axis column number (default = 1)

Y-axis column number (default = 2)

Label file (or # for item numbers)

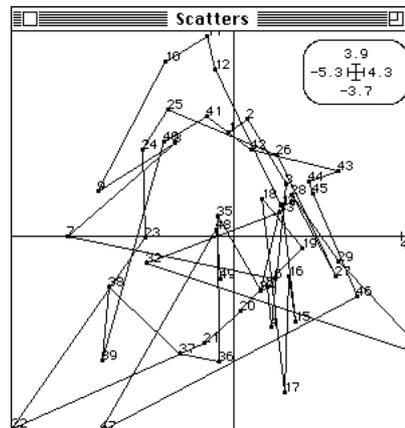
Draw vectors from origin (yes = 1)

Draw unit circle (yes = 1)



Trajectories

HV coordinates file 49 2



On note la succession haute pression (beau temps) puis fortes températures puis précipitations (orages d'été) caractéristiques du temps estival de la région. La question porte sur l'influence des variables météorologique sur l'abondance des piégeages lumineux. Le tableau faunistique a 17 espèces (variables).

On peut chercher à représenter l'information sur l'abondance des taxons dans le cadre de ce plan factoriel très explicite. La notion d'objets supplémentaires porte cette fois sur les variables :

Supplementary columns

Initial analysis (up) 11 2

Supplementary column file 49 17

Output file name

Add. col. 1 Mean= 0.06122 Var.= 0.05748
 Add. col. 2 Mean= 0.06122 Var.= 0.05748
 ...

Add. col. 16 Mean= 0.1837 Var.= 0.2316
 Add. col. 17 Mean= 5.939 Var.= 81.32

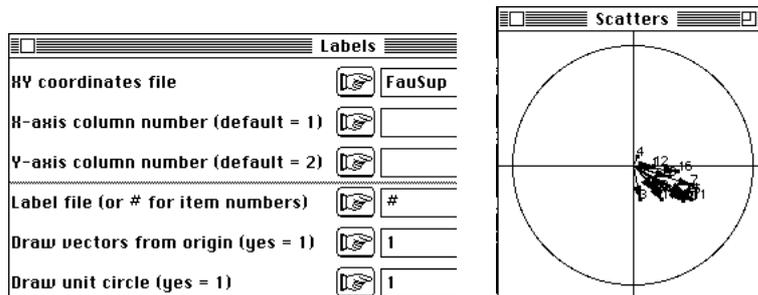
Title of the analysis: Meteo.cnta
 Number of rows: 49, columns: 11

File FauSup contains the coordinates of the projected columns
 It has 17 rows and 2 columns

File :FauSup

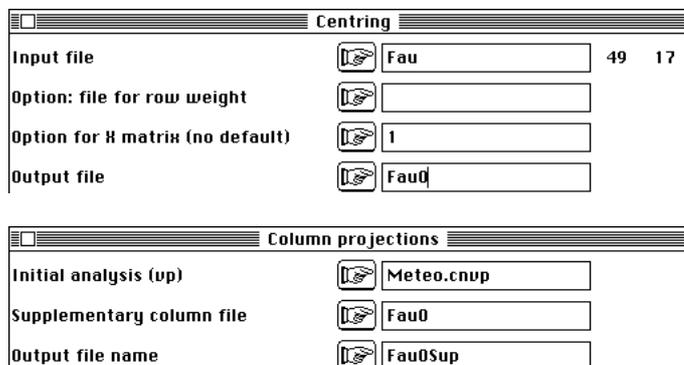
Col.	Mini	Maxi
1	4.268e-02	4.837e-01
2	-2.882e-01	7.055e-02

Les colonnes de Fau, comme dans l'analyse initiale, sont normalisées.

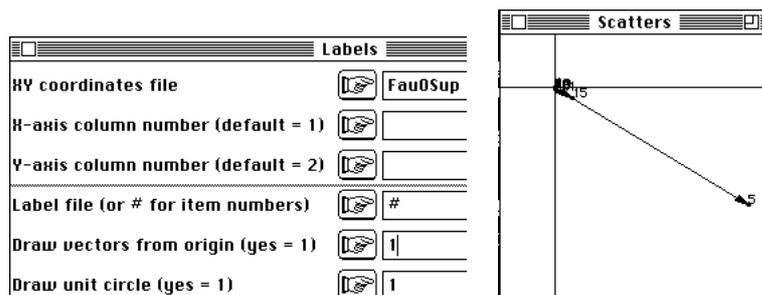


Les projections des variables supplémentaires normalisées (vecteurs de norme 1) donnent des coordonnées qui sont des coefficients de corrélation avec les coordonnées factorielles. Ces corrélation sont faibles mais de même signe.

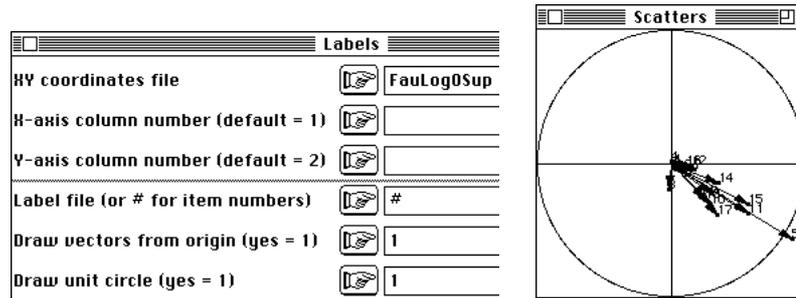
On peut centrer le tableau Fau et projeter ses colonnes :



Les coordonnées des projections sont des covariances et l'image intègre la variabilité des abondances :

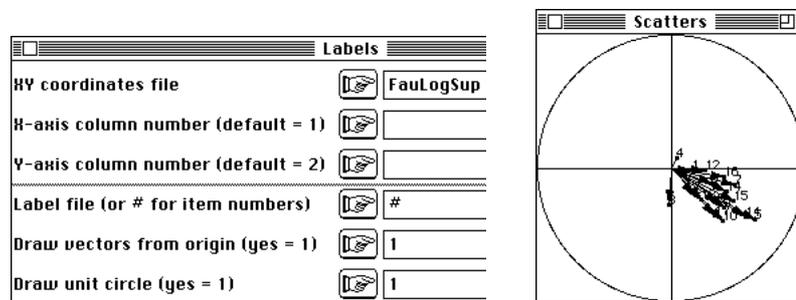


La variance de l'espèce est trop forte. On peut recommencer après transformation en $\text{Log}(x+1)$:

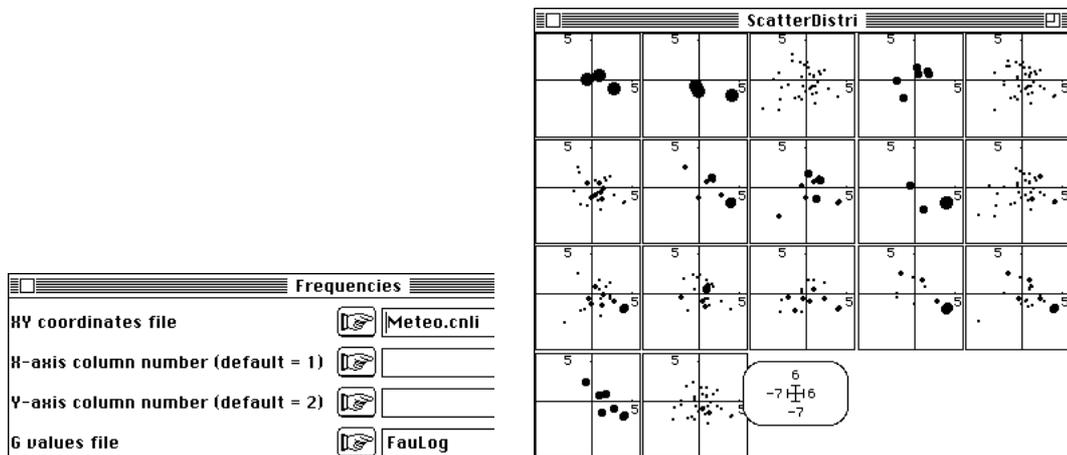


La présence du cercle unité sert de repère mais la longueur des vecteurs projetés n'étant plus égale à 1, les points ne tombent pas forcément dans ce cercle.

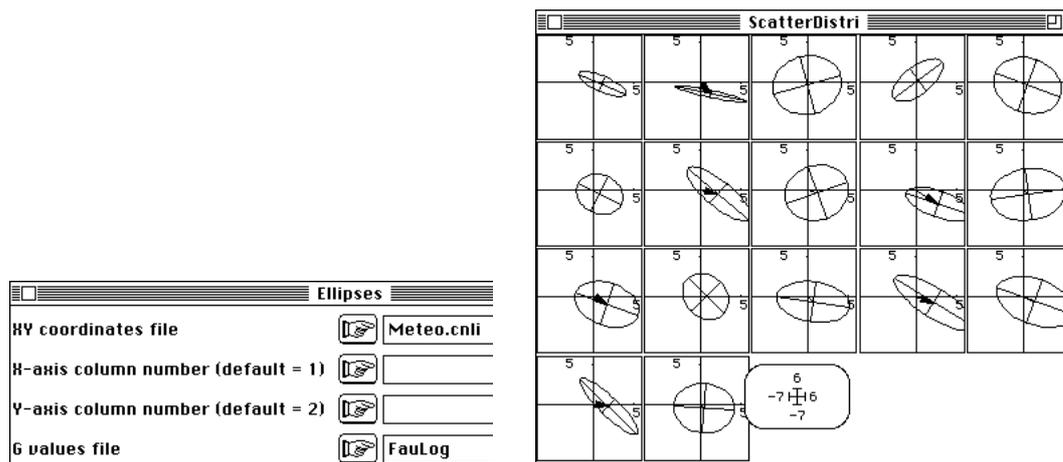
On peut utiliser les variables normalisées après changement de variables :



L'abondance d'un taxon peut être considérée comme une distribution de fréquence sur l'ensemble des points du plan factoriel. On passe de la représentation de variables supplémentaire à celle d'information supplémentaire :



La représentation des moyennes s'impose :



On pourra aussi superposer les moyennes des positions des espèces, représenter l'abondance des espèces sur les plans de co-inertie, ... Ici domine l'idée d'une combinaison de variables météorologiques ayant une influence commune sur les émergences de tous les taxons. Notons enfin qu'il arrive souvent que de véritables projections euclidiennes soient également des représentation par moyennes de distribution et que les notions d'individus supplémentaires et d'information supplémentaire se confondent.

Quoiqu'il en soit la graphique appliquée à la statistique multidimensionnelle est un moyen d'expression. Cela suppose quelques libertés dans les choix et la référence à un comportement "conforme à la règle" peut être le signe d'une certaine absence d'imagination. Ce n'est évidemment pas une raison pour faire n'importe quoi.

Références

- ¹ Blondel, J. & Farre, H. (1988) The convergent trajectories of bird communities along ecological successions in european forests. *Œcologia (Berlin)* : 75, 83-93.
- ² Blondel, J., Chessel, D. & Frochet, B. (1988) Niche expansion and density compensation of island birds in mediterranean habitats. A case study from comparison of two ecological successions. *Ecology* : 69, 6, 1899-1917.
- ³ Dolédec, S. & Chessel, D. (1994) Co-inertia analysis: an alternative method for studying species-environment relationships. *Freshwater Biology* : 31, 277-294.
- ⁴ Ter Braak, C.J.F. (1986) Canonical correspondence analysis : a new eigenvector technique for multivariate direct gradient analysis. *Ecology* : 69, 69-77.
- ⁵ Franquet, E. & Chessel, D. (1994) Approche statistique des composantes spatiales et temporelles de la relation faune-milieu. *C.R. Acad. Sci. Paris, Sciences de la vie* : 317, 202-206.
- ⁶ Franquet, E., Dolédec, S. & Chessel, D. (1995) Using multivariate analyses for separating spatial and temporal effects within species-environment relationships. *Hydrobiologia* : 300-301, 425-431.
- ⁷ Usseglio-Polatera, P. & Auda, Y. (1987) Influence des facteurs météorologiques sur les résultats de piégeage lumineux. *Annales de Limnologie* : 23, 1, 65-79.