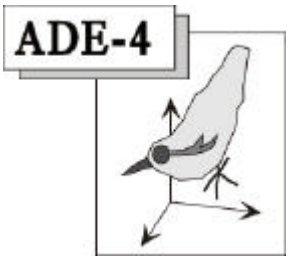


Fiche thématique 2.D

Les tableaux de fréquences alléliques

Résumé

La fiche répond à une question posée plusieurs fois sur le forum Adelist « Que faire avec un tableau de fréquences alléliques ? » La connexion faite par B. Guinand avec l'analyse des correspondances floues est contestable dans son fondement. On montre qu'un tableau de fréquences alléliques dérive simplement d'un tableau de génotypes par une moyenne en absence de données manquantes. L'inertie inter-classe est directement reliée à la mesure de fixation du Fst. On introduit un codage des données manquantes pour respecter cette propriété en utilisant les propositions en œuvre dans le logiciel GENETIX (http://www.univ-montp2.fr/~genetix/genetix.htm). L'ACP inter-classe et l'AFC inter-classe de ces tableaux est appropriée. Elles permettent une représentation optimale des individus et des populations.

Plan

D. Chessel & D. Laloë

Introduction : l'indice de fixation

Les tableaux de fréquences alléliques ont la forme des tableaux de variables floues. C'est cette apparence qui permet à Guinand [1] de rebaptiser l'ACF (Analyse des Correspondances Floues) [2] en CRT-MCA (Constant Row Total - Multiple Correspondence Analysis) sans y apporter la moindre modification (MCA: Fuzzy Correspondence Analysis). L'auteur associe l'ACF à la génétique par le biais d'une similitude entre la notion de rapport de corrélation (pourcentage de variance expliquée d'une variable quantitative par une variable qualitative) et celle d'Indice de fixation (F_{st}) rapport de la variance des fréquences d'un allèle entre populations sur le maximum possible. Le rapport de corrélation de l'AFC d'un tableau de fréquences alléliques sur l'axe *j* devient alors le « F_{st} par locus/axe *j* » dans [3].

Cette association mérite d'être reprise à partir des définitions. Considérons un tableau de fréquences alléliques portant sur un seul locus. Les lignes de ce tableau sont référés à des groupes d'individus ayant une propriété commune : ils appartiennent à une même catégorie (race, génération, population, biotope). On peut utiliser le terme général de groupe. Dans chaque groupe on utilise une partie des individus. Ces individus fournissent un ou deux gènes (suivant le succès des manipulations, le nombre de chromosomes, ...) et chaque gène est classé dans une des formes possibles (allèle). Le groupe est caractérisé par l'ensemble des fréquences géniques.
Supposons que le tableau porte sur *g* groupes et que le locus étudié présente *p* allèles. Il s'écrit accompagné des effectifs ayant permis de calculer les fréquences :

$$\mathbf{Y} = \begin{bmatrix} p_{11} & \cdots & p_{1j} & \cdots & p_{1m} \\ p_{i1} & \cdots & p_{ij} & \cdots & p_{im} \\ p_{g1} & \cdots & p_{gi} & \cdots & p_{gm} \end{bmatrix} \quad \mathbf{e} = \begin{bmatrix} e_1 \\ e_i \\ e_g \end{bmatrix}$$

Avec l'utilisation de *v* loci, on a alors des données formées de l'assemblage de *v* tableaux du type :

$$\mathbf{X}^k = \begin{bmatrix} p^k_{11} & \cdots & p^k_{1j} & \cdots & p^k_{1m_k} \\ p^k_{i1} & \cdots & p^k_{ij} & \cdots & p^k_{im_k} \\ p^k_{g1} & \cdots & p^k_{gi} & \cdots & p^k_{gm_k} \end{bmatrix} \quad \mathbf{e}^k = \begin{bmatrix} e^k_1 \\ e^k_i \\ e^k_g \end{bmatrix}$$

On a donc un K-tableau accompagné d'un tableau de pondération :

$$\mathbf{X} = \left[\mathbf{X}^1, \cdots \mathbf{X}^k, \cdots, \mathbf{X}^v \right] \quad \mathbf{e} = \left[\mathbf{e}^1, \cdots \mathbf{e}^k, \cdots, \mathbf{e}^v \right]$$

Ces données pourraient dériver d'une expérience parfaite dans laquelle le groupe *k* comporte n_k individus ayant fourni exactement $2n_k$ gènes dont on aurait identifié les allèles de tous les loci étudiés. Les fréquences alléliques dérivent alors des fréquences génotypiques. A l'autre extrême, chacun des tableaux pourrait être constitué à partir d'un échantillon extrait du groupe sans que l'on trouve, dans deux tableaux différents, deux individus communs. La diversité extrême des matériels et des méthodes et le rôle de l'interprétation dans la lecture des résultats individuels ne permet pas de repérer la forme exacte des données d'origine, le tableau des fréquences alléliques étant un dérivé des observations de base.

Nous considérons donc, dans un premier temps, un tableau de fréquences de la forme :

$$\mathbf{X} = \left[\mathbf{X}^1, \cdots \mathbf{X}^k, \cdots, \mathbf{X}^v \right]$$

où le poids de chaque groupe est pour tous les loci est uniformément $1/g$ puis nous examinerons l'utilisation des données qui, en amont, conduisent à ce type de résumé.

L'indice de fixation est une propriété de chacun des sous-tableaux. On peut reprendre la définition et l'explicitation du calcul dans [4] (Chapitre 3, Box B, p.164). L'indice de fixation est une mesure de la variabilité génétique entre groupes définie par un locus. Le tableau $\mathbf{Y}$ est typiquement un table d'ACP centrée. Les individus sont les groupes, les variables $\mathbf{a}_j$ sont les allèles. On calcule les moyennes et la variance des variables :

$$\text{moy}\left(\mathbf{a}_j\right) = \overline{p}_j = \frac{1}{g}\sum\nolimits_{i=1}^{g} p_{ij} \qquad \text{var}\left(\mathbf{a}_j\right) = s_j^2 = \frac{1}{g}\sum\nolimits_{i=1}^{g}\left(p_{ij} - \overline{p}_j\right)^2$$

La variabilité génétique entre groupes est l'inertie totale du tableau centré, soit exactement :

$$Iner = \sum\nolimits_{i=1}^{g} s_j^2$$

Il serait logique de pondérer la fréquence de chaque groupe par le nombre de mesures faite dans ce groupe. On aurait alors :

$$\textit{Effectif total} : e = \sum\nolimits_{i=1}^{g} e_i \;\; \textit{Poids d'un groupe} : \; f_i = \frac{e_i}{e}$$

$$\textit{Fréquences moyennes} : \text{moy}\left(\mathbf{a}_j\right) = \overline{p}_j = \sum\nolimits_{i=1}^{g} f_i\, p_{ij}$$

$$\textit{Variance d'un allèle} : \text{var}\left(\mathbf{a}_j\right) = s_j^2 = \sum\nolimits_{i=1}^{g} f_i \left(p_{ij} - \overline{p}_j\right)^2$$

$$\textit{Variance totale} : \; Iner = \sum\nolimits_{i=1}^{g} s_j^2$$

En utilisant cette remarque, on simplifie la définition de l'indice de fixation. En effet, on peut considérer les variables indicatrices $\mathbf{I}_j$ des allèles (voir [5], Chapitre 5, p.143) qui codent les gènes de l'échantillon 1 si l'allèle porté est j et 0 sinon. Ces variables portent sur $e = e_1 + \cdots + e_g$ éléments. Elles sont disjonctives complètes, c'est-à-dire que sur chaque ligne du tableau $\mathbf{I} = \left[\mathbf{I}_1, \cdots, \mathbf{I}_p\right]$ on trouve des valeurs 0 à l'exception d'une valeur 1 et une seule. Chaque gène de chaque échantillon compte pour $1/e$. Les moyennes et les variances des indicatrices sont :

$$\text{moy}\left(\mathbf{I}_j\right) = \frac{1}{e}\sum\nolimits_{k=1}^{e} \mathbf{I}_{kj} = p_j$$

$$\text{var}\left(\mathbf{I}_j\right) = \frac{1}{e}\sum\nolimits_{k=1}^{e}\left(\mathbf{I}_{kj} - p_j\right)^2 = p_j\left(1 - p_j\right)^2 + \left(1 - p_j\right)\left(0 - p_j\right)^2 = p_j\left(1 - p_j\right)$$

La moyenne pondérée des fréquences alléliques par groupe est alors exactement la fréquence totale :

$$p_j = \overline{p}_j$$

La moyenne conditionnelle de l'indicatrice pour le groupe k est :

$$\text{moy}_{/i}\left(\mathbf{I}_j\right) = \frac{1}{e}\sum\nolimits_{\substack{k=1 \\ k \in groupe k}}^{e} \mathbf{I}_{kj} = p_{ij}$$

La variance conditionnelle pour le groupe k est :

$$\text{var}_{/i}\left(\mathbf{I}_j\right) = \frac{1}{e}\sum\nolimits_{\substack{k=1 \\ k \in groupe k}}^{e}\left(\mathbf{I}_{kj} - p_{ij}\right)^2 = p_{ij}\left(1 - p_{ij}\right)$$

L'équation de l'analyse de la variance donne (*b* pour variance inter-classe -*between*- et *w* pour variance intra-classe -*within*-) :

$$\mathrm{var}\left(\mathbf{I}_j\right) = p_j\left(1-p_j\right) = b\left(\mathbf{I}_j\right) + w\left(\mathbf{I}_j\right)$$

On a alors :

$$b\left(\mathbf{I}_j\right) = \sum_{i=1}^{g} f_i \left(\mathrm{moy}_{/i}\left(\mathbf{I}_j\right) - \mathrm{moy}\left(\mathbf{I}_j\right)\right)^2 = s_j^2$$

$$w\left(\mathbf{I}_j\right) = \sum_{i=1}^{g} f_i \,\mathrm{var}_{/i}\left(\mathbf{I}_j\right) = \sum_{i=1}^{g} f_i\, p_{ij}\left(1-p_{ij}\right)$$

D'où l'inertie totale du tableau des indicatrices (variabilité génétique du locus) :

$$Iner_{tot} = \sum_{j=1}^{p} p_j\left(1-p_j\right) = 1-\sum_{j=1}^{p} p_j^{\,2}$$

l'inertie inter-classe du tableau des indicatrices (variabilité génétique inter-groupe) :

$$Iner_b = \sum_{j=1}^{p} s_j^{\,2}$$

et l'inertie intra-classe du tableau des indicatrices (variabilité génétique intra-groupe) :

$$Iner_w = \sum_{j=1}^{p}\sum_{i=1}^{g} f_i\, p_{ij}\left(1-p_{ij}\right) = \sum_{i=1}^{g} f_i\left(1-\sum_{j=1}^{p} p_{ij}^{\,2}\right)$$

En écologie, on dirait « la diversité totale des relevés ($Iner_{tot}$ = indice de Simpson) se décompose en moyenne des diversités internes des relevés (diversité alpha = $Iner_w$) + variabilité de la composition taxonomique entre relevés (diversité beta = $Iner_b$) ». Ce point de vue est exprimé dans **6**.

En génétique, le même calcul a une autre signification. Pour un allèle multiple, dans une grande sous-population *i* avec accouplement aléatoire, $2\,p_{ij}\left(1-p_{ij}\right)$ est le taux d'hétérozygotes pour l'allèle et $1-\sum_{j=1}^{p} p_{ij}^{\,2}$ est le taux d'hétérozygotes pour le locus étudié. En moyenne ($f_i = 1/g$) sur l'ensemble des populations ce taux vaut $Iner_w$. Vue ici comme variance intra-classe, elle ne peut dépasser la variance totale ($1-\sum_{j=1}^{p} p_j^{\,2}$) qui est le taux d'hétérozygotes théorique dans la population totale. En fonction de la fixation ($p_{ij} = 1$) ou de la perte ($p_{ij} = 0$) d'allèles dans les sous-populations, la variabilité intra-groupe diminue et on appelle indice de fixation la quantité :

$$F_{st} = \frac{Iner_{tot} - Iner_w}{Iner_{tot}} = \frac{\sum_{j=1}^{p} s_j^{\,2}}{1-\sum_{j=1}^{p} \overline{p}_j^{\,2}}$$

Lorsque chaque sous-population est identique à la population totale, donc à chacune des autres, l'indice est nul. Lorsque chaque sous-population a un allèle fixé, la variabilité intra-groupe est nulle et la variabilité totale est exclusivement inter-groupe. L'indice de fixation vaut 1.

L'indice de fixation nous indique donc que la variabilité dans un tableau de fréquences alléliques se mesure par une somme de variances (inertie) à laquelle on donne un sens par référence à une inertie de référence. Il est une propriété d'un locus et donc est associé à la description de la variabilité inter-groupe sur un seul locus. Il nous permet de comparer la valeur typologique de deux loci ou plus dans une stratégie multi-tableaux. Peux t'on faire un parallèle avec un rapport de corrélation intervenant dans une ACF ?

En reprenant les notations qui précèdent, prenons le locus du tableau **Y**. Le rapport de corrélation suppose l'existence d'un score numérique. Il y a deux types de scores. Le premier est associé aux groupes. Notons $\mathbf{x} = \left(x_1, \cdots, x_g \right)$ un tel score. Il a une moyenne $\text{moy}\left(\mathbf{x} \right) = \frac{1}{g} \sum_{i=1}^{g} x_i$ que l'on peut supposer nulle et une variance $\text{var}\left(\mathbf{x} \right) = \frac{1}{e} \sum_{i=1}^{e} \left(x_i - \text{moy}\left(\mathbf{x} \right) \right)^2$ qu'on peut supposer égale à 1 (il suffit de normaliser le score pour qu'il ait ces propriétés). Le score étant connu, on calcule pour chaque allèle :

$$moy_j\left(\mathbf{x} \right) = \frac{\sum_{i=1}^{g} p_{ij} x_i}{\sum_{i=1}^{g} p_{ij}}$$

Ces moyennes forment un score des allèles, qui a alors une variance qui est inférieure ou égale à la variance de départ C'est un rapport de corrélation (variance des moyennes conditionnelles rapportée à la variance de départ).

Le second est associé aux allèles. Notons $\mathbf{y} = \left(y_1, \cdots, y_p \right)$ un tel score. Il a une moyenne $\text{moy}\left(\mathbf{y} \right) = \sum_{i=1}^{g} p_i y_i$ que l'on peut supposer nulle et une variance $\text{var}\left(\mathbf{y} \right) = \sum_{i=1}^{p} \left(p_i y_i - \text{moy}\left(\mathbf{y} \right) \right)^2$ qu'on peut supposer égale à 1 (il suffit de normaliser le score pour qu'il ait ces propriétés). Le score étant connu, on calcule pour chaque groupe :

$$moy_i\left(\mathbf{y} \right) = \sum_{j=1}^{p} p_{ij} y_j$$

Ces moyennes forment un score des groupes, qui a alors une variance qui est inférieure ou égale à la variance de départ C'est un rapport de corrélation (variance des moyennes conditionnelles rapportée à la variance de départ).

Ces opérations d'averaging, dans un sens comme dans l'autre, qui se réfère à la notion d'ordination en écologie, n'ont pas de signification expérimentale en génétique. Cette statistique est un pourcentage de variance comme le F_{st} mais les rapports de corrélation de l'ACF que la répartition des gènes portant un même allèle entre populations alors que la génétique s'intéresse à la répartition des allèles dans une population de gènes . Le F_{st} est une mesure originale de la variabilité qu'il convient d'introduire en analyse de données plutôt que d'importer en génétique une statistique ressemblante qui n'a pas de justification ni mathématique ni expérimentale.

Analyses simples sur les fréquences alléliques

Ce préambule n'implique pas, par ailleurs, que l'ACF est inappropriée aux tableaux étudiés. Il ne fait que souligner qu'une similitude ne peut servir de base, encore moins justifier des assertions du type « A value [correlations ratio] is considered significant if its correlation ratio is twice the eigenvalue associated with a factorial axis » ([3] Table III p. 720) ou « because the eigenvalues are equivalent to F_{st} the equation for the estimate of gene flow (*Nm) :*

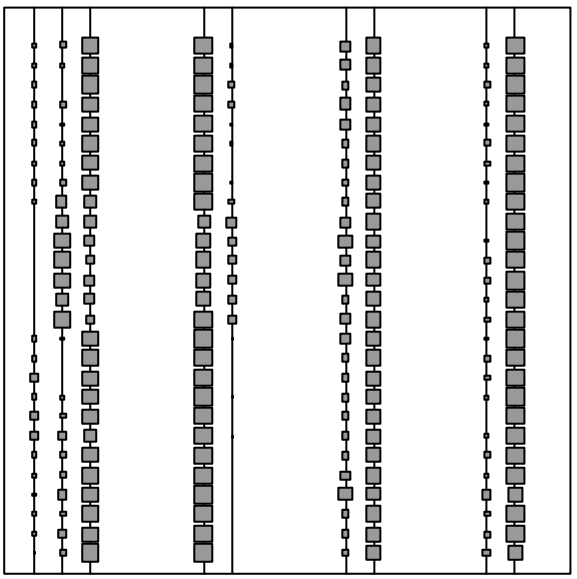
$$Nm \approx 0.25\left((1/F_{st})-1\right)$$

[may] be replaced by :

$$Nm \approx 0.25\left((1/l)-1\right)$$ » ([3] p. 723)

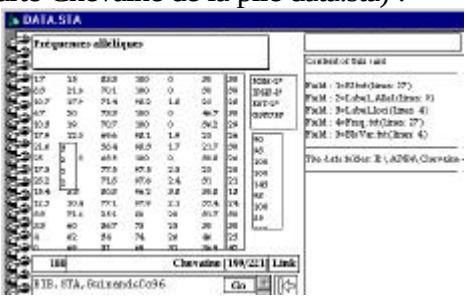
Le raisonnement s'écrit : le F_{st} est un pourcentage de variance, le rapport de corrélation est un pourcentage de variance, donc le rapport de corrélation est un F_{st}. La valeur propre est une moyenne de rapports de corrélation, donc la valeur propre est un F_{st}. Il s'agit ici de métaphore (*figure de rhétorique par laquelle on transpose la signification propre d'un mot à une autre signification qui ne lui convient qu'en vertu d'une comparaison sous-entendue, Petit Larousse).*

Pour être plus utile, on peut se demander quelles analyses de base peuvent être employées sur un tableau de fréquences alléliques. Reprenons le tableau de [3] p. 716. Dans 27 stations, on a les fréquences alléliques sur 4 loci polymorphes pour des échantillons de chevaines. On peut représenter les données brutes (Tables: Fuzzy Variables) :

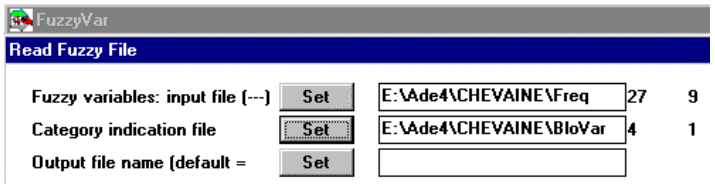


On reconnaît les individus (lignes), les modalités (colonnes) et les variables (paquet de colonnes), ici les échantillons, les allèles et les loci.. Le numérateur de l'indice de fixation étant la somme des variances des fréquences par allèle, on peut penser faire une ACP floue (PCA: Fuzzy PCA). La procédure est :

1) Implanter les données (carte Chevaine de la pile data.sta) :

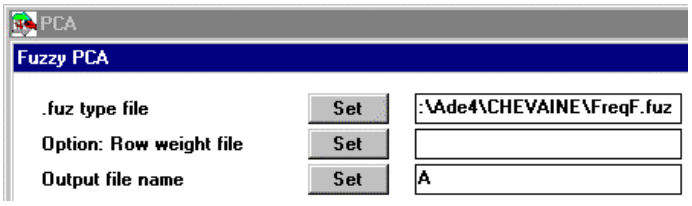


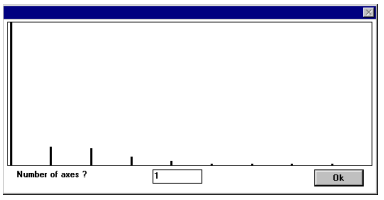
2) Lire le fichier :




```
Description of categories:
-----------------------------------------
Variable number 1 has 3 categories
-----------------------------------------
[   1]  Category:    1    Freq.:       0.11
[   2]  Category:    2    Freq.:       0.24
[   3]  Category:    3    Freq.:       0.65

Variable number 2 has 2 categories
-----------------------------------------
[   4]  Category:    1    Freq.:      0.901
[   5]  Category:    2    Freq.:     0.0986

Variable number 3 has 2 categories
-----------------------------------------
[   6]  Category:    1    Freq.:      0.322
[   7]  Category:    2    Freq.:      0.678

Variable number 4 has 2 categories
-----------------------------------------
[   8]  Category:    1    Freq.:      0.114
[   9]  Category:    2    Freq.:      0.886
-----------------------------------------
```

3) Exécuter l'analyse :



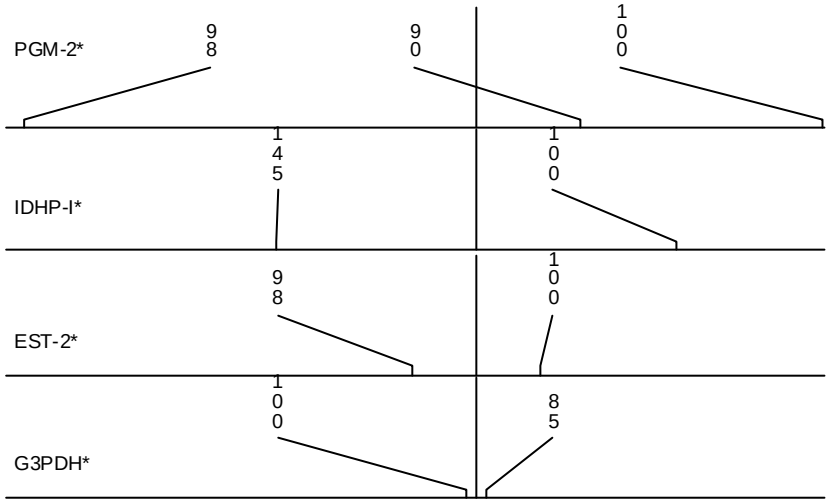



```
fp/FuzzyPCA: PCA on fuzzy table
Input file: E:\Ade4\CHEVAINE\FreqF.fuz for access to file E:\Ade4\CHEVAINE\FreqF
Row number: 27, column number: 9
Uniform row weights

File A.cppl contains the row weights
It has 27 rows and 1 column

File A.cppc contains the column weights Diag(Unif1,...UnifV)
It has 9 rows and 1 column

File A contains the raw table
It has 27 rows and 9 columns (categories)

File A.cpta contains the centred table
It has 27 rows and 9 columns (categories)
```

Cette analyse utilise le triplet

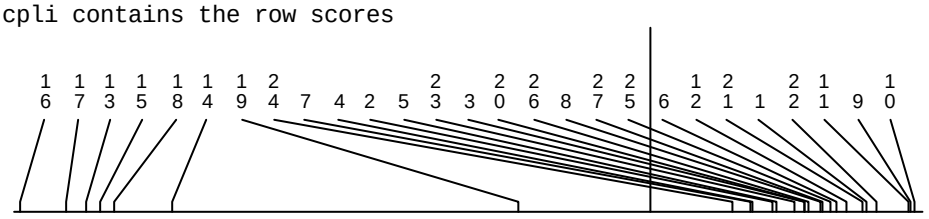
$$\mathbf{X} = \left[p_{ij}^{k} - \overline{p}_{j}^{k} \right] \mathbf{D}_p = Diag\left(\underbrace{\frac{1}{m_1}, \cdots, \frac{1}{m_1}}_{m_1 \ fois}, \underbrace{\frac{1}{m_2}, \cdots, \frac{1}{m_2}}_{m_2 \ fois}, \cdots, \underbrace{\frac{1}{m_v}, \cdots, \frac{1}{m_v}}_{m_v \ fois} \right) \mathbf{D}_p = Diag\left(\frac{1}{g}, \cdots, \frac{1}{g} \right)$$

```
--------------------------------------------------
DiagoRC: General program for two diagonal inner product analysis
Input file: A.cpta
--- Number of rows: 27, columns: 9
----------------------
Total inertia: 0.0487198
----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +3.7243E-02 +0.7644 +0.7644   |02   +4.5295E-03 +0.0930 +0.8574 |
03   +4.1779E-03 +0.0858 +0.9432   |04   +1.9885E-03 +0.0408 +0.9840 |
05   +7.8070E-04 +0.0160 +1.0000   |06   +0.0000E+00 +0.0000 +1.0000 |
07   +0.0000E+00 +0.0000 +1.0000   |08   +0.0000E+00 +0.0000 +1.0000 |
09   +0.0000E+00 +0.0000 +1.0000

File A.cpvp contains the eigenvalues and relative inertia for each axis
File A.cpco contains the column scores
```



Graph1D: Labels sur A.cpco

File A.cpli contains the row scores

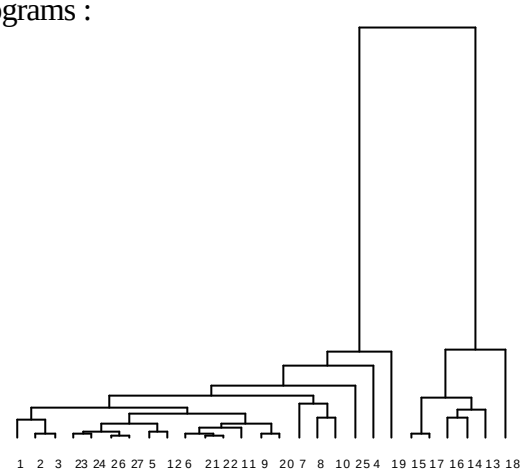


Graph1D: Labels

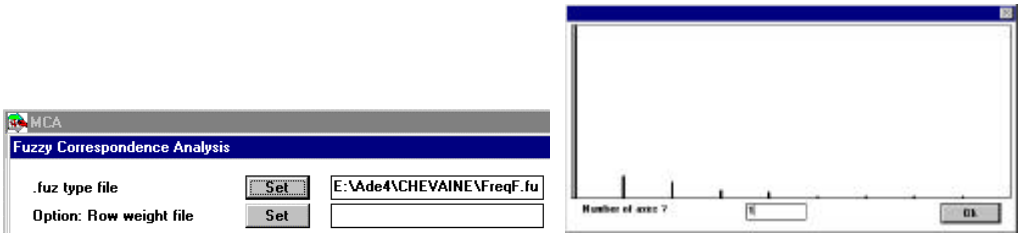
```
--------------------------------------------------
Total inertia: 4.8720e-02 - Number of axes: 1
Contribution of fuzzy variable to total inertia
|Num  | Inertia  |  Ratio   | Max=Total|    FST   |
|-----|----------|----------|----------|----------|
|    1|2.9337e-02|      6022|5.0830e-01|      1731|
|    2|1.2271e-02|      2519|1.7780e-01|      1380|
|    3|4.6273e-03|       950|4.3692e-01|       212|
|    4|2.4848e-03|       510|2.0235e-01|       246|
|-----|----------|----------|----------|----------|
|Sum  |4.8720e-02|     10000|
--------------------------------------------------
Contribution of fuzzy variable to eigenvalues (in 1/10000)
```

```
|Num   |  Fac  1  |
|------|----------|
|    1 |      7157|
|    2 |      2575|
|    3 |       264|
|    4 |         5|
|------|----------|
|Sum   |     10000|
-----------------------------------------------------
```

La structure très simple est visible sur le tableau de données et entièrement exprimée par les coordonnées sur un axe (séparation du Haut-Rhône). On retrouve ce résultat par une classification sur une matrice de distances génétiques (enchaîner DMAUtil: Genetic distance, option distance de Nei, DMAUtil: ToClusters, Clusters: Compute hierarchy option UGPMA et Dendrograms: Dendrograms :



Avec une AFC floue, on a :




```
fl/FuzzyMCA: Multiple correspondence analysis on fuzzy table
Input file: E:\Ade4\CHEVAINE\FreqF.fuz for access to file E:\Ade4\CHEVAINE\FreqF
Row number: 27, column number: 9
Uniform row weights

File E:\Ade4\CHEVAINE\FreqF.flpl contains the row weights
It has 27 rows and 1 column

File E:\Ade4\CHEVAINE\FreqF.flta contains the table processed by MCA
It has 27 rows and 9 columns (categories)

File E:\Ade4\CHEVAINE\FreqF.flpc contains the column weights (1/V)*DM
It has 9 rows and 1 column
```

Cette analyse utilise le triplet

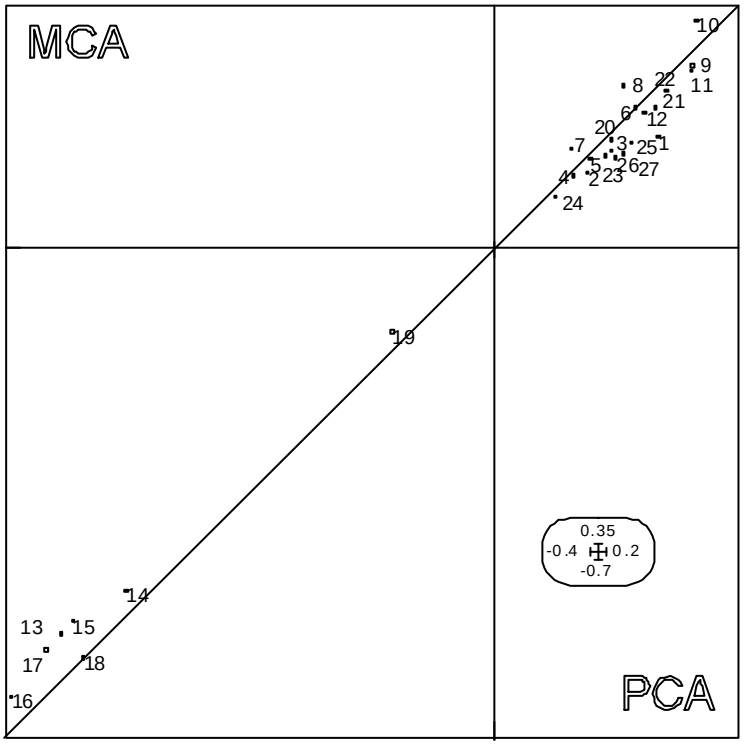
$$\mathbf{X} = \left[\frac{p_{ij}^k}{\overline{p}_j^k} - 1 \right] \quad \mathbf{D}_p = \frac{1}{v} Diag\left(\overline{p}_j^k \right) \quad \mathbf{D}_p = Diag\left(\frac{1}{g}, \cdots, \frac{1}{g} \right)$$

équivalent à :

$$\mathbf{X} = \left[p_{ij}^k - \overline{p}_j^k \right] \quad \mathbf{D}_p = \frac{1}{v} Diag\left(\frac{1}{\overline{p}_j^k} \right) \quad \mathbf{D}_p = Diag\left(\frac{1}{g}, \cdots, \frac{1}{g} \right)$$

```
--------------------------------------------------
DiagoRC: General program for two diagonal inner product analysis
Input file: E:\Ade4\CHEVAINE\FreqF.flta
--- Number of rows: 27, columns: 9
----------------------
Total inertia: 0.123066
----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +9.6403E-02 +0.7833 +0.7833   |02   +1.1904E-02 +0.0967 +0.8801 |
03   +8.4026E-03 +0.0683 +0.9483   |04   +3.7027E-03 +0.0301 +0.9784 |
05   +2.6538E-03 +0.0216 +1.0000   |06   +0.0000E+00 +0.0000 +1.0000 |
07   +0.0000E+00 +0.0000 +1.0000   |08   +0.0000E+00 +0.0000 +1.0000 |
09   +0.0000E+00 +0.0000 +1.0000

File E:\Ade4\CHEVAINE\FreqF.flvp contains the eigenvalues and relative inertia
for each axis
File E:\Ade4\CHEVAINE\FreqF.flco contains the column scores
File E:\Ade4\CHEVAINE\FreqF.flli contains the row scores
```

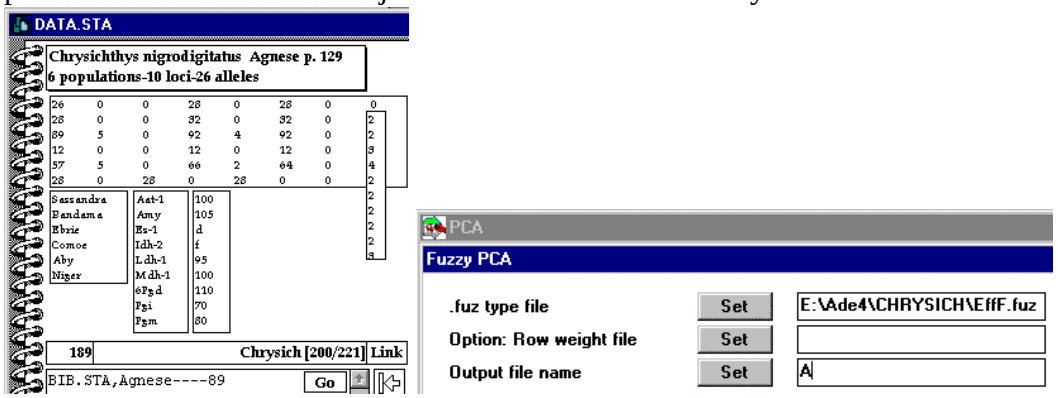
Les deux coordonnées expriment la même structure :



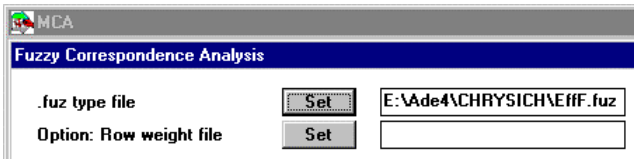

```
--------------------------------------------------
CorRatioFCA: Correlation ratios after a FCA
Title of the analysis: E:\Ade4\CHEVAINE\FreqF.fl
Number of rows: 27, columns: 4
Variable : 1
> Categ=   1 Weight=  0.110   0.498
> Categ=   2 Weight=  0.240  -0.914
```

```
> Categ=   3 Weight=  0.650   0.253
----------------------> r=    0.270
Variable : 2
> Categ=   1 Weight=  0.901   0.110
> Categ=   2 Weight=  0.099  -1.009
----------------------> r=    0.111
Variable : 3
> Categ=   1 Weight=  0.322  -0.097
> Categ=   2 Weight=  0.678   0.046
----------------------> r=    0.005
Variable : 4
> Categ=   1 Weight=  0.114   0.031
> Categ=   2 Weight=  0.886  -0.004
----------------------> r=    0.000

File E:\Ade4\CHEVAINE\FreqF.flrc contains the correlation ratios between
each categorical variable and each factorial score
It has 4 rows and 1 columns
--------------------------------------------------
```

Les ordres de grandeur entre indices de fixation et rapport de corrélation sont respectés. On peut se demander s'il en est toujours ainsi. Utiliser la carte Chrysich [7] :

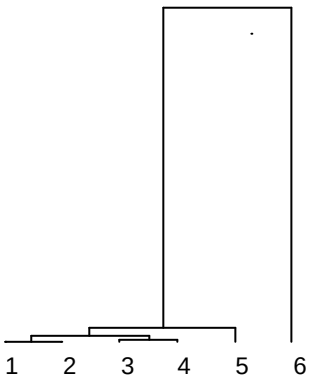



```
Total inertia: 0.567017
------------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +5.3214E-01 +0.9385 +0.9385   |02   +2.8136E-02 +0.0496 +0.9881 |
03   +6.0683E-03 +0.0107 +0.9988   |04   +6.6114E-04 +0.0012 +1.0000 |
05   +9.1427E-06 +0.0000 +1.0000   |06   +0.0000E+00 +0.0000 +1.0000 |
--------------------------------------------------------
Total inertia: 5.6702e-01 - Number of axes: 1
Contribution of fuzzy variable to total inertia
|Num  | Inertia  |  Ratio   | Max=Total|    FST   |
|-----|----------|----------|----------|----------|
|    1|1.0579e-03|        19|4.3625e-02|       485|
|    2|1.3889e-01|      2449|2.7778e-01|     10000|
|    3|9.0126e-02|      1589|2.9348e-01|      9213|
|    4|8.8965e-03|       157|3.6235e-01|       982|
|    5|6.1630e-33|         0|0.0000e+00|       000|
|    6|6.1630e-33|         0|0.0000e+00|       000|
|    7|6.1630e-33|         0|0.0000e+00|       000|
|    8|1.1671e-01|      2058|2.5888e-01|      9016|
|    9|1.2490e-01|      2203|4.3340e-01|      5764|
|   10|8.6445e-02|      1525|3.2332e-01|      8021|
|-----|----------|----------|----------|----------|
|Sum  |5.6702e-01|     10000|
```

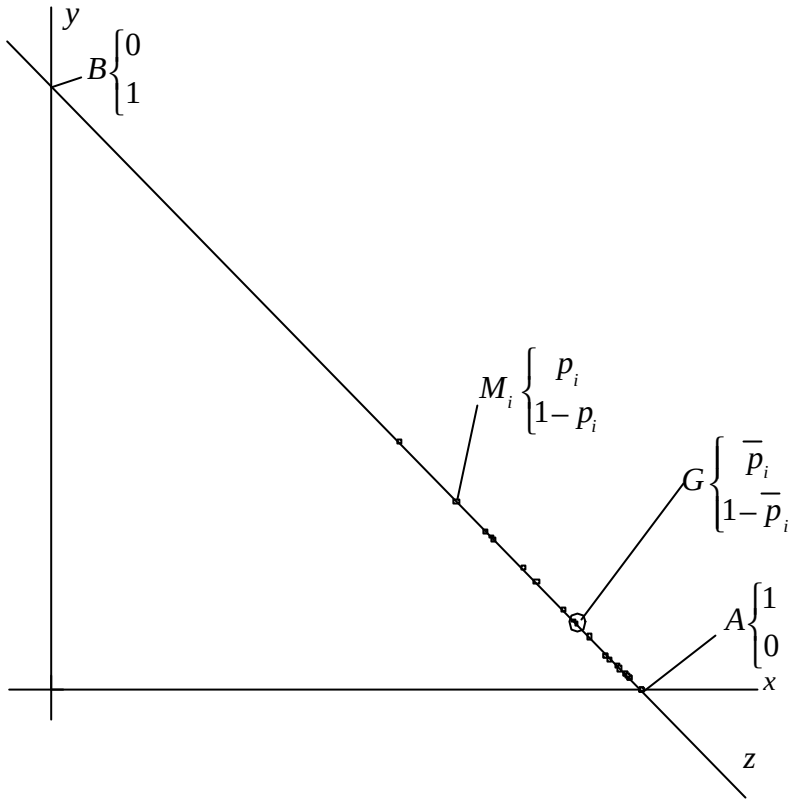



```
----------------------
Total inertia: 0.441023
----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +4.1316E-01 +0.9368 +0.9368   |02   +1.9736E-02 +0.0448 +0.9816 |
03   +7.2069E-03 +0.0163 +0.9979   |04   +9.0939E-04 +0.0021 +1.0000 |
05   +9.5556E-06 +0.0000 +1.0000   |06   +0.0000E+00 +0.0000 +1.0000 |
...
File E:\Ade4\CHRYSICH\EffF.flrc contains the correlation ratios between
each categorical variable and each factorial score
It has 10 rows and 1 columns
-----------------------------------------------
Binary input file: E:\Ade4\CHRYSICH\EffF.flrc - 10 rows, 1 cols.
          RC  |    FST   |
   1 |      26|       485|
   2 |    9926|     10000|
   3 |    9171|      9213|
   4 |     598|       982|
   5 |       0|         0|
   6 |       0|         0|
   7 |       0|         0|
   8 |    8949|      9016|
   9 |    4696|      5764|
  10 |    7950|      8021|
```

La similitude est encore très apparente. Cet exemple partage avec le précédent deux propriétés. La majorité des loci présente peu d'allèles (entre 2 et 4) d'une part, la structure est simple avec un seul facteur d'autre part. Les deux systèmes de coordonnées reproduisent la partition (Distances de Nei, lien UGPMA) :



Quand il y a deux allèles, l'indice de fixation pour un seul locus est exactement égal au rapport de corrélation. En effet, on a :



L'inertie du nuage vaut :

$$I_T = \mathrm{var}(x_i) + \mathrm{var}(y_i) = \frac{1}{g}\sum_{i=1}^{g}\left(p_i - \overline{p}_i\right)^2 + \frac{1}{g}\sum_{i=1}^{g}\left(q_i - \overline{q}_i\right)^2 = \frac{2}{g}\sum_{i=1}^{g}\left(p_i - \overline{p}_i\right)^2$$

De plus :

$$1 - \overline{p}_i^{\,2} - \left(1 - \overline{p}_i\right)^2 = 2\overline{p}_i\left(1 - \overline{p}_i\right)$$

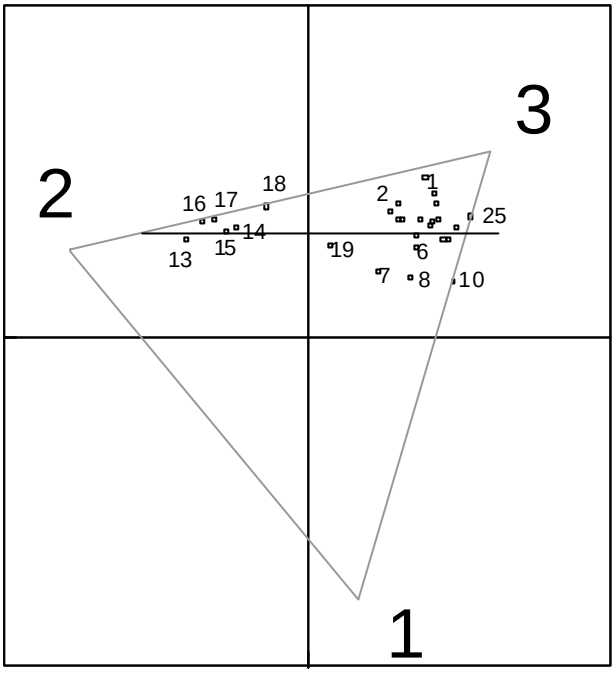
D'où :

$$F_{st} = \frac{\frac{2}{g}\sum_{i=1}^{g}\left(p_i - \overline{p}_i\right)^2}{2\overline{p}_i\left(1 - \overline{p}_i\right)}$$

L'axe de l'analyse de ce tableau est évidemment l'axe sur le quel se trouve les points. Chaque point sur l'axe z est au centre de gravité des points *A* et *B* muni des poids p_i et $1 - p_i$. Le rapport de corrélation est le rapport de la variance des coordonnées sur z sur la variance des coordonnées de *A* et *B* munis des poids $\overline{p}_i$ et $1 - \overline{p}_i$, soit :

$$r = \frac{\frac{1}{g}\sum_{i=1}^{g}\left(\sqrt{2}\left(p_i - \overline{p}_i\right)\right)^2}{\overline{p}_i\left(\sqrt{2}\left(1 - \overline{p}_i\right)\right)^2 + \left(1 - \overline{p}_i\right)\left(\sqrt{2}\left(0 - \overline{p}_i\right)\right)^2} = F_{st}$$

Cette propriété est encore vraie pour un locus à 3 allèles si les points sont alignés et approximativement vraie si les points sont approximativement alignés. Par exemple pour le locus à 3 allèles du tableau Chevaine on a :

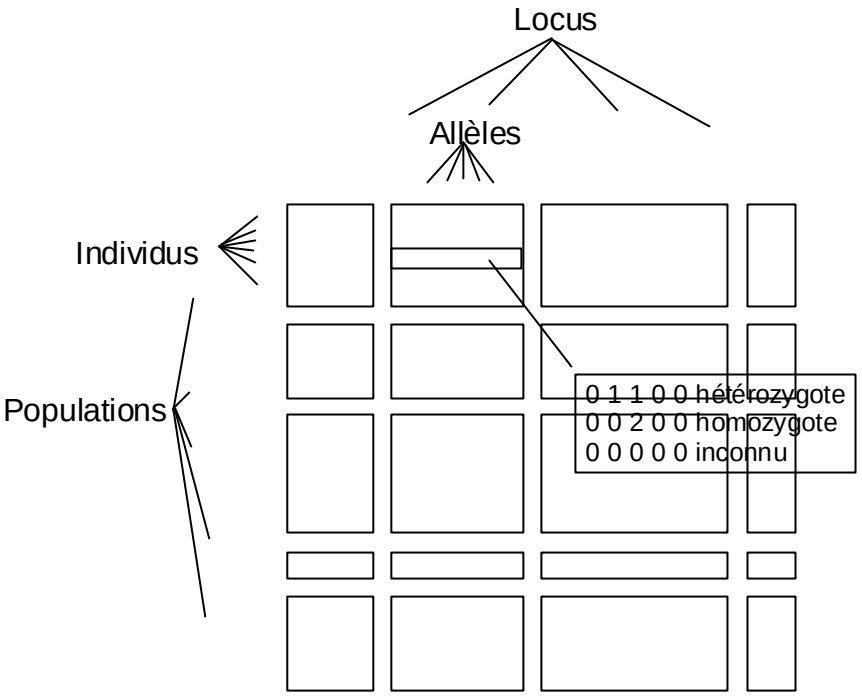


Il faut enfin que les ordinations par locus soient concordantes, ce qui se produit si le taux d'inertie sur l'axe 1 est très élévée. C'est la cas des exemples qui précèdent. On reconnaît ce-dessus la typologie globale induite par le locus seul. Seules des structures simples comme des partitions ou des gradients spatiaux (avec des loci peu polymorphes) induisant une forte prédominance d'un axe permettent de trouver une similitude entre valeurs des indices de fixation et rapports de corrélation de l'ACF. Dans les cas plus compliqués cette association n'a pas lieu d'être. Par contre l'inertie du nuage des points pour un locus vu dans une ACP en pourcentage (PCA: After row % transformation PCA) est exactement dans tous les cas le numérateur du F_{st} du locus correspondant. Les analyses simples qui permettent d'aborder les tableaux de fréquences alléliques sont donc des ACP quand on travaille en fréquence et des AFC quand on travaille en dénombrements. Ces prémisses étant posées, on va s'intéresser à une question de fond. Comment mesurer la cohérence des gènes, leur valeur typologique, la covariation de plusieurs gènes polymorphes. Il s'agit de problèmes de statistiques multi-tableaux.

Des données brutes aux tableaux de fréquence

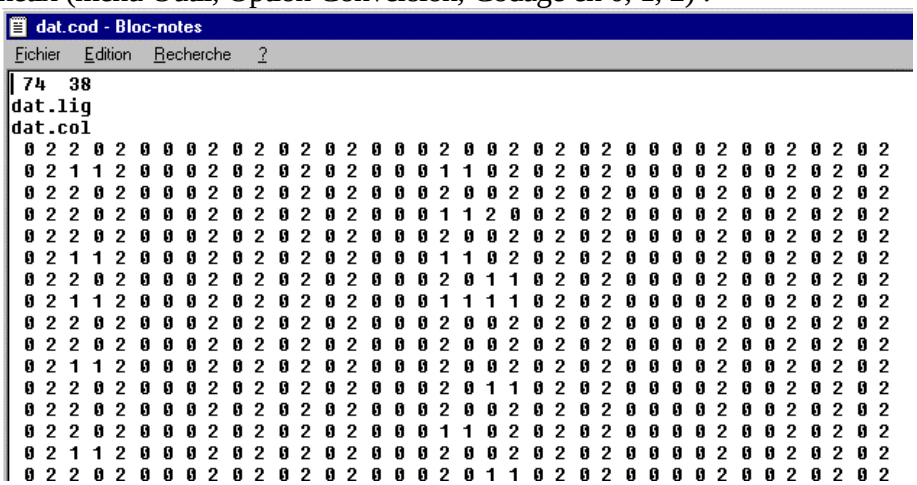
Du point de vue de l'analyse des données, on cherchera à conserver la notion d'individu. Si la fréquence allélique est une caractéristique du groupe, il s'agit d'une moyenne. La typologie de population par le biais de moyennes multidimensionelles renvoie à la notion d'analyse inter-classe. Voir la variabilité intra-classe et la variabilité inter-classe ne peut se faire qu'à partir des données brutes.

Ces données ont une forme particulière résumée dans la figure :

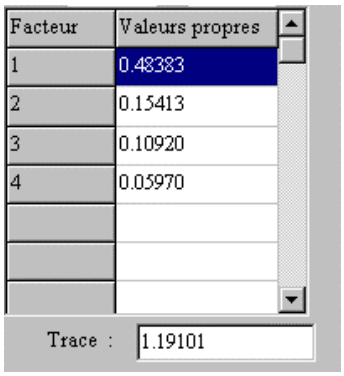


Le logiciel GENETIX [8] propose de faire l'analyse des correspondances de ce tableau. Nous utilisons l'exemple proposé dans la documentation du logiciel [9]. Ces données sont dans la carte Genetix (voir FuzzyVar: From_GENETIX) :

Le tableau Data est un tableau de variables floues. Un individu hétérozygote est codé 0 ½0 ½ 0.., un individu hétérozygote est codé 000010..., une donnée manquante est codée 000000... Dans Genetix (menu Outil, Option Conversion, Codage en 0, 1, 2) :

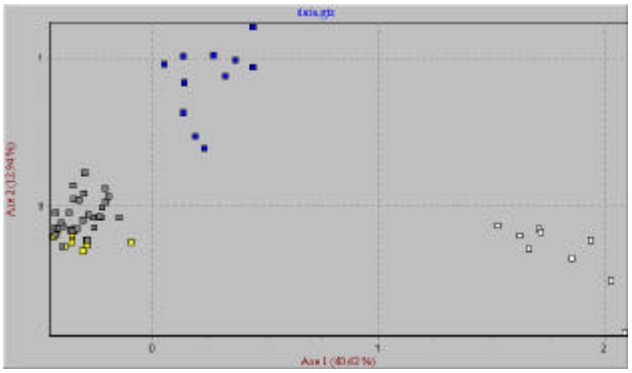


L'aption AFC2D de Genetix affiche :



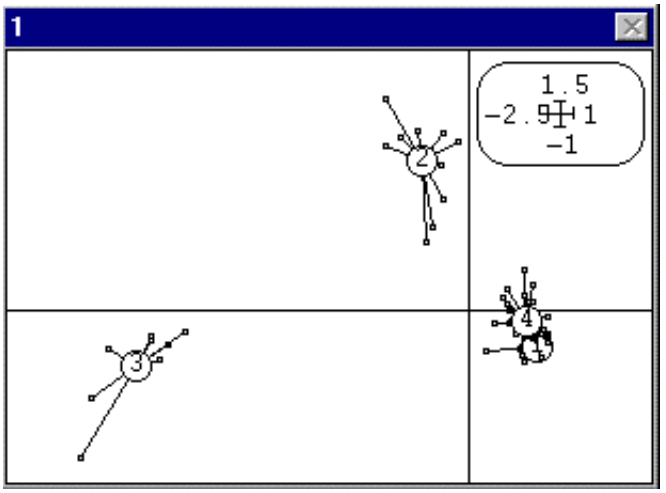
Facteur	Valeurs propres
1	0.48383
2	0.15413
3	0.10920
4	0.05970

Trace : 1.19101

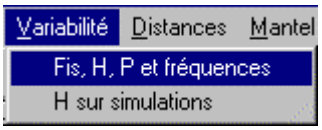


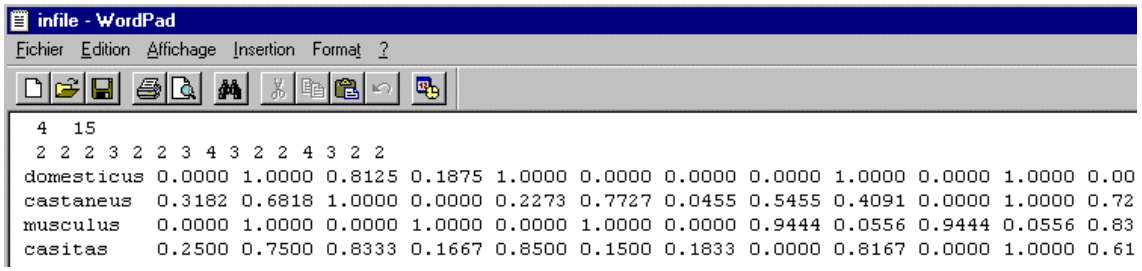
On peut obtenir la même chose par :

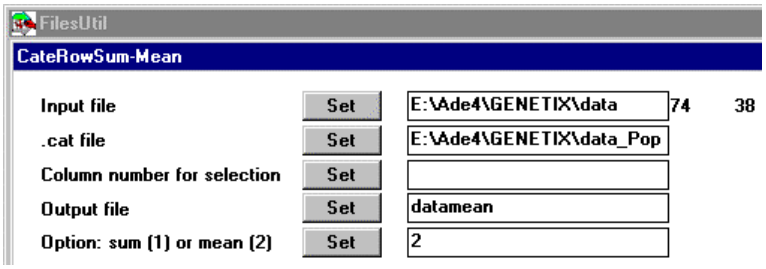

```
-----------------------
Total inertia:  1.19101
-----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +4.8383E-01 +0.4062 +0.4062   |02   +1.5413E-01 +0.1294 +0.5356 |
03   +1.0920E-01 +0.0917 +0.6273   |04   +5.9696E-02 +0.0501 +0.6775 |
```

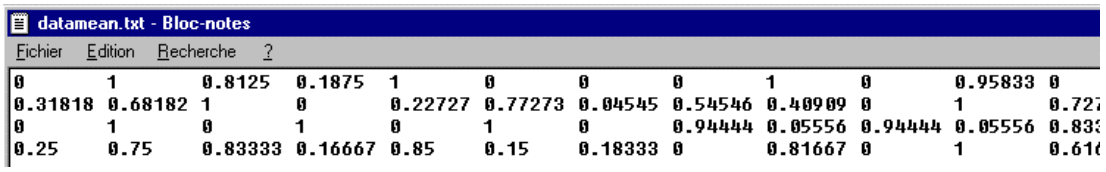


Ce faisant, on n'a tenu compte, ni de la structure en paquets de lignes (populations) ni de la structure en paquets de colonnes (locus). Genetix donne le tableau de fréquences alléliques dans l'option :



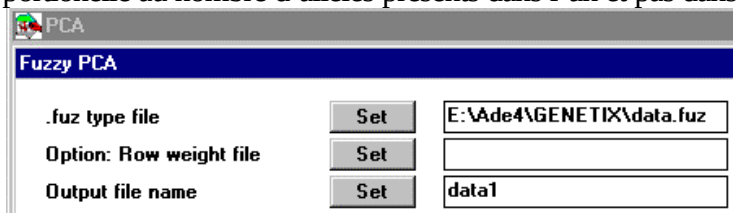


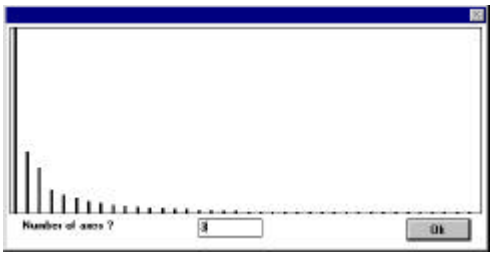




C'est presque la même chose (la différence vient des données manquantes dont nous reparlerons). L'analyse d'un tableau de fréquences alléliques est très proche (ou strictement identique si les données sont complètes) d'une analyse inter-classe (populations). Faire l'analyse simple du tableau des génotypes, c'est ne pas centrer l'analyse sur la typologie des groupes. Faire l'analyse du tableau des fréquences alléliques, c'est faire la typologie des groupes sans voir la diversité interne. Faire une analyse inter-classe est un bon compromis.

Dans l'ACP du tableau flou des génotypes, on a une propriété simple. La distance entre deux individus est proportionelle au nombre d'allèles présents dans l'un et pas dans l'autre.



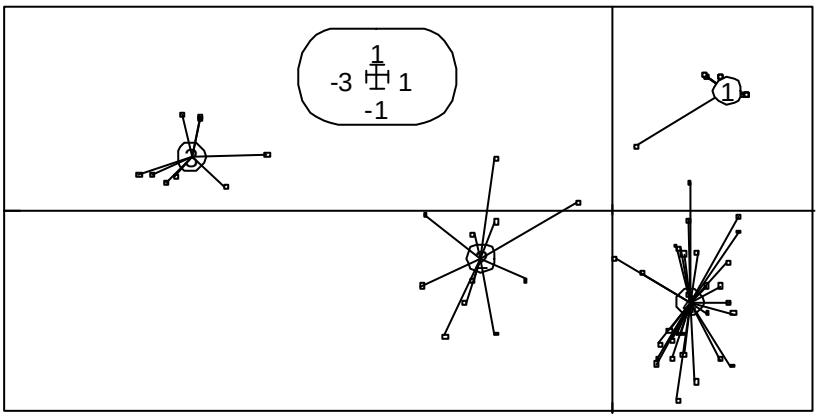



```
fp/FuzzyPCA: PCA on fuzzy table
Input file: E:\Ade4\GENETIX\data.fuz for access to file E:\Ade4\GENETIX\data
Row number: 74, column number: 38
Uniform row weights

File data1.cppl contains the row weights
```

```
It has 74 rows and 1 column

File data1.cppc contains the column weights Diag(Unif1,...UnifV)
It has 38 rows and 1 column

File data1 contains the raw table
It has 74 rows and 38 columns (categories)
```

Dans le nouveau tableau de données, les données manquantes sont remplacées par le profil total et la moyenne par population est exactement le profil de fréquences alléliques.

```
File data1.cpta contains the centred table
It has 74 rows and 38 columns (categories)

File data1.cpma contains
----- number of rows: 74
----- number of variables: 15
----- number of categories: 38
----- variable number of each category (vector of 38 values)

-----------------------------------------------------
DiagoRC: General program for two diagonal inner product analysis
Input file: data1.cpta
--- Number of rows: 74, columns: 38
-----------------------
Total inertia:  1.71489
-----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +7.8288E-01 +0.4565 +0.4565   |02   +2.5728E-01 +0.1500 +0.6065 |
03   +1.9094E-01 +0.1113 +0.7179   |04   +9.5537E-02 +0.0557 +0.7736 |
...

File data1.cpvp contains the eigenvalues and relative inertia for each axis
--- It has 38 rows and 2 columns
```

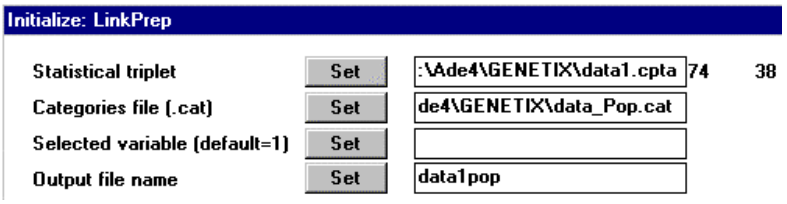


Les quatre populations sont bien séparées et la contribution des locus à la structure globale est équilibrée.

```
-----------------------------------------------------
Contribution of fuzzy variable to eigenvalues (in 1/10000)
|Num  |  Fac  1  |  Fac  2  |  Fac  3  |
|-----|----------|----------|----------|
|    1|         1|      1170|       801|
|    2|      1123|       548|      3458|
```

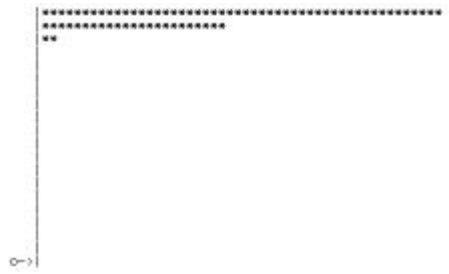
```
|    3|      2816|       171|      1135|
|    4|      1575|       378|       107|
|    5|      1609|       262|      1389|
|    6|      1004|      5583|         2|
|    7|      1492|      1488|       749|
|    8|       693|       165|      1050|
|    9|       343|        41|       680|
|   10|       472|      6908|      3563|
|   11|      2144|         7|      1431|
|   12|       913|        71|       937|
|   13|      1074|        21|        32|
|   14|       102|        69|       317|
|   15|      1789|       267|      1498|
|-----|----------|----------|----------|
|Sum  |     10000|     10000|     10000|
-------------------------------------------------
```

L'analyse inter-classe associée :




```
Between and Within-class inertia
Groups are defined by column 1 of file E:\Ade4\GENETIX\data_Pop
Input statistical triplet: table E:\Ade4\GENETIX\data1.cpta
total inertia:  1.715e+00
Between-class inertia  1.137e+00 (ratio:  6.632e-01)
Within-class inertia  5.775e-01 (ratio:  3.368e-01)
```

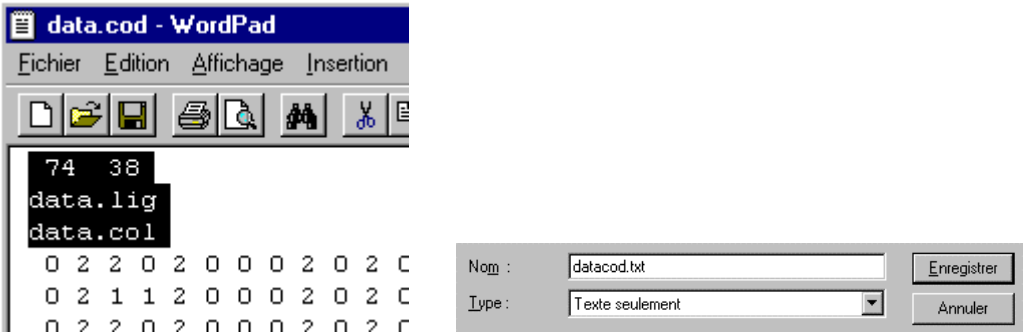
donne 2/3 d'inertie inter-populations et 1/3 d'inertie intra-populations. Le test de Monte-Carlo est évidemment très significatif :

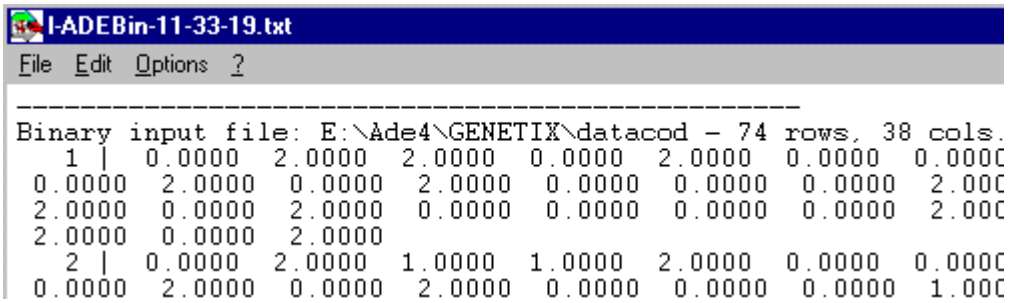


et la carte de l'analyse inter-classe est très voisine de celle de l'analyse simple. On obtient des résultats voisins en AFC et AFC inter-classe. On retiendra de ces premiers essais qu'une ACP ou une AFC floue sur tableaux de fréquences alléliques est une ACP ou une AFC inter-classe qui s'ignore. Vérifions.

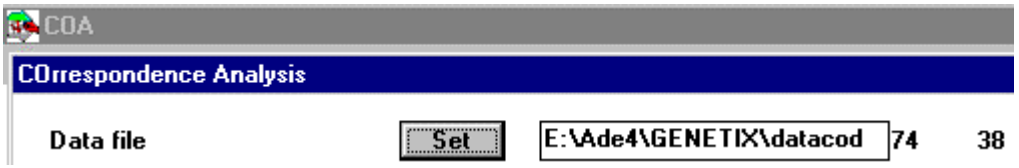
Fréquences alléliques et inter-classe

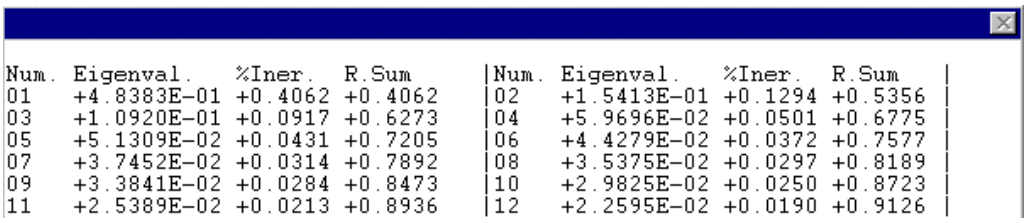
Reprenons les données directement dans Genetix. Le fichier dat.cod (codage 0, 1, 2) est passé en binaire. On enlève l'en-tête du fichier :





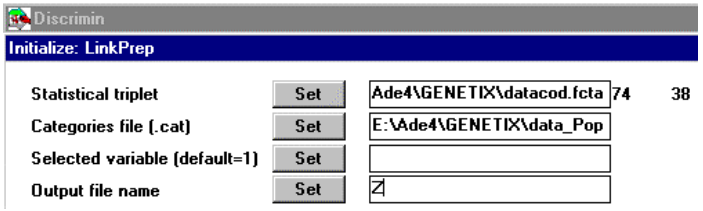
On fait son AFC :





Num.	Eigenval.	%Iner.	R.Sum	Num.	Eigenval.	%Iner.	R.Sum
01	+4.8383E-01	+0.4062	+0.4062	02	+1.5413E-01	+0.1294	+0.5356
03	+1.0920E-01	+0.0917	+0.6273	04	+5.9696E-02	+0.0501	+0.6775
05	+5.1309E-02	+0.0431	+0.7205	06	+4.4279E-02	+0.0372	+0.7577
07	+3.7452E-02	+0.0314	+0.7892	08	+3.5375E-02	+0.0297	+0.8189
09	+3.3841E-02	+0.0284	+0.8473	10	+2.9825E-02	+0.0250	+0.8723
11	+2.5389E-02	+0.0213	+0.8936	12	+2.2595E-02	+0.0190	+0.9126

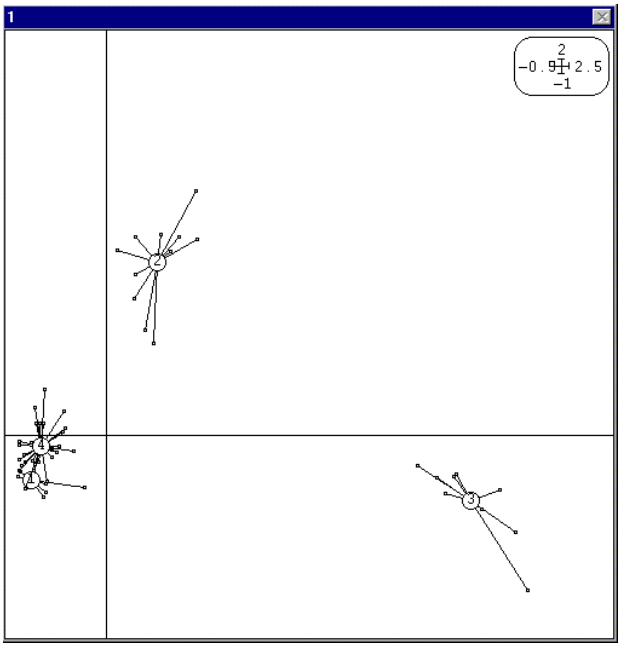
On couple cette analyse avec l'indicatrice des populations :



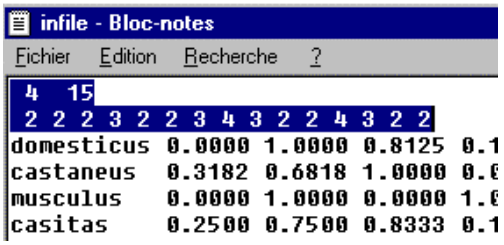
between-class inertia 6.886e-01 (ratio: 5.782e-01)

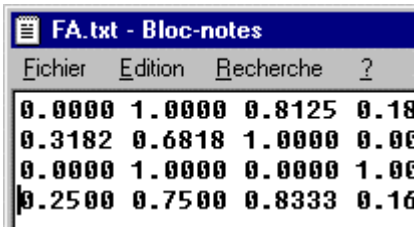
Num.	Eigenval.	R.Iner.	R.Sum	Num.	Eigenval.	R.Iner.	R.Sum
01	+4.7385E-01	+0.6881	+0.6881	02	+1.3695E-01	+0.1989	+0.8870

```
03   +7.7793E-02 +0.1130 +1.0000   |04   +0.0000E+00 +0.0000 +1.0000 |
```

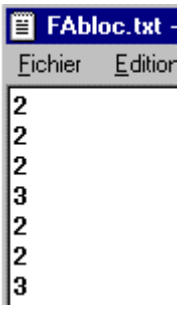


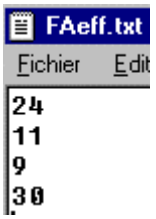
On reprend ensuite le tableau des fréquences alléliques qu'on passe en binaire :



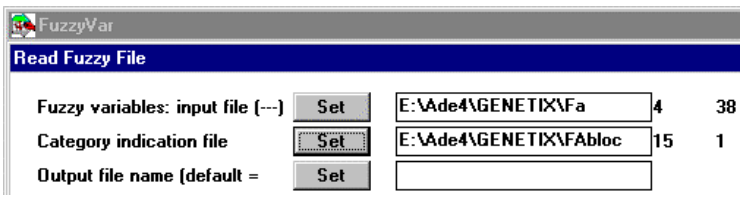


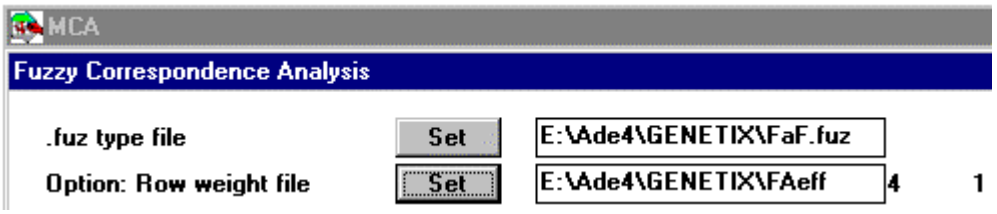
On conserve les effectifs des allèles par locus (Fabloc.txt) et les effectifs par populations (Faeff.txt) et on passe en binaire :





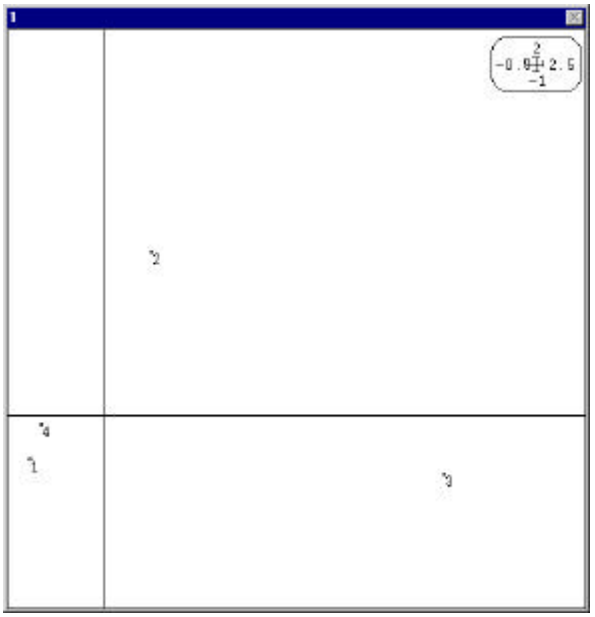
On lit les fichiers :



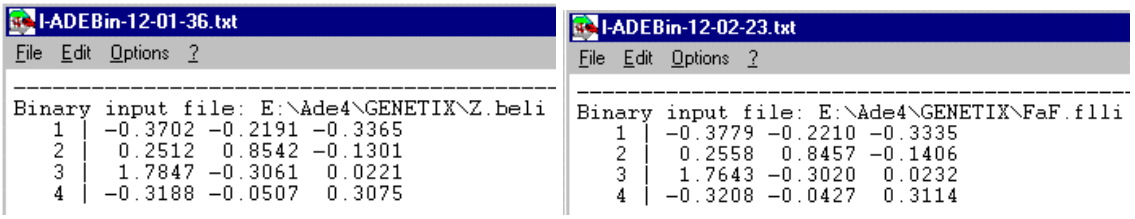



```
----------------------
Total inertia: 0.688747
----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +4.7635E-01 +0.6916 +0.6916   |02   +1.3400E-01 +0.1946 +0.8862 |
03   +7.8399E-02 +0.1138 +1.0000   |04   +0.0000E+00 +0.0000 +1.0000 |
```

Les valeurs propres sont très voisines et on a :



Ou encore :



I-ADEBin-12-01-36.txt

File Edit Options ?

```
-----------------------------------------
Binary input file: E:\Ade4\GENETIX\Z.beli
   1 | -0.3702 -0.2191 -0.3365
   2 |  0.2512  0.8542 -0.1301
   3 |  1.7847 -0.3061  0.0221
   4 | -0.3188 -0.0507  0.3075
```

I-ADEBin-12-02-23.txt

File Edit Options ?

```
-----------------------------------------
Binary input file: E:\Ade4\GENETIX\FaF.flli
   1 | -0.3779 -0.2210 -0.3335
   2 |  0.2558  0.8457 -0.1406
   3 |  1.7643 -0.3020  0.0232
   4 | -0.3208 -0.0427  0.3114
```

L'identité n'est pas parfaite à cause des données manquantes. Elle n'est obtenu qu'en codant une mesure manquante par le profil moyen des individus renseignés. Par exemple, pour l'individu 24, les locus 5, 6, 7 et 12 ne sont pas renseignés. Il faudrait coder :

```
1     0.0000  1.0000
2     0.5000  0.5000
3     1.0000  0.0000
4     0.0000  0.0000  1.0000
5     0.1164  0.8836
6     0.4658  0.5342
```

```
7     0.3750  0.5347  0.0903
8     0.0000  0.0000  0.0000  1.0000
9     0.0000  1.0000  0.0000
10    1.0000  0.0000
11    1.0000  0.0000
12    0.0616  0.0890  0.0616  0.7877
13    0.0000  0.0000  1.0000
14    0.0000  1.0000
15    0.0000  1.0000
```

Hors ce détail numérique, l'analyse d'un tableau de fréquences alléliques est d'abord une analyse inter-classe. Pour voir au mieux les individus (comme dans l'AFC de Genetix) et les populations (comme dans l'AFC floue de Guinand) les analyses inter-classes s'imposent. Ce détail pose cependant un problème particulier. On attribue à un individu, sur une donnée manquante, le profil moyen. Si le tableau est destiné à une inter-classe, il serait plus judicieux de lui attribuer le profil moyen de tous les individus connus du même groupe. Cette opération est assurée dans l'utilitaire Genetic Missing Data de FuzzyVar.

Références

1 Guinand, B. (1996) Use of a multivariate model using allele frequency distributions to analyse patterns of genetic differentiation among populations. *Biological Journal of the Linnean Society* : 58, 173-195.

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

3 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.

4 Hartl, D.L. (1980) *Principles of population genetics*. Sinauer Associates, Sunderland, Massachussetts. 1-488.

5 Weir, B.S. (1990) *Genetic data analysis*. Sinauer Associates, Sunderland, Massachussetts. 377 pp.

6 Lande, R. (1996) Statistics and partioning of species diversity, and similarity among multiple communities. *Oikos* : 76, 5-13.

7 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.

8 Belkhir (K.) GENETIX 4.0 Logiciel sous Windows™ pour la génétique des populations du Laboratoire Génome et Populations, CNRS UPR 9060, Université de Montpellier II,

Montpellier (France), Université Montpellier 2, Place E. Bataillon, 34095 Montpellier cedex 05, France. e-mail : Genetix@crit.univ-montp2.fr, tél. : (33) 67 14 38 87, fax : (33) 67 14 45 54. Freeware à http://www.univ-montp2.fr/~genetix/genetix.htm.

9 Orth, A., Adama, T., Din, W. & Bonhomme, F. (1998) Hybridation naturelle entre deux sous espèces de souris domestique Mus musculus domesticus et Mus musculus castaneus près de Lake Casitas (Californie). *Genome* : 41, 104-110.