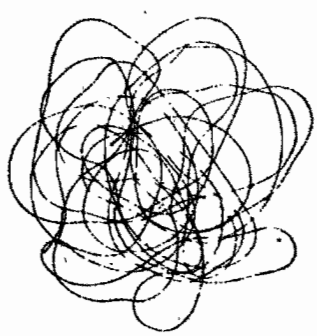
R. RENE CHAUME

METHODES DE CLASSIFICATION STATISTIQUES ET ANALYSE DES ENSEMBLES DE POPULATIONS



OFFICE DE LA RECHERCHE SCIENENTIFIQUE ET TECHNIQUE OUTRE-MER

CENTRE D'ADIOPODOUMÉ – CÔTE D'IVOIRE

B. P. 20 – ABIDJAN

Avril 1972

OFFICE DE LA RECHERCHE SCIENTIFIQUE ET TECHNIQUE OUTRE-MER

CENTRE D'ADIOPODOUME

Laboratoire de Génétique

METHODES DE CLASSIFICATION STATISTIQUES ET ANALYSE DES ENSEMBLES DE POPULATIONS.

par

Régine RENE-CHAUME

Mars 1972

METHODES DE CLASSIFICATION STATISTIQUES ET ANALYSE DES ENSEMBLES DE POPULATIONS.

I - GENERALITES

1. Les problèmes de classification
2. Historique
3. Domaine d'application

II- LES DONNEES

1. Choix des caractères
2. Collecte des données
3. Codification des données

III - REDUCTION DES DONNEES

1. Indices de similarité

   1.1. Indices simples

   1.2. Indice de similarité et probabilité
   - Couple ordonné pour la similarité
   - Définition de la Prob. d'un couple
   - Similarité = Complément Prob.
   - Originalité caractères rares
   - Exemple de calcul de l'indice de GOODALL

   1.3. Indice de SMIRNOV
   - Description
   - Exemple - K6 - matrice de SMIRNOV

2. Distances
   - d_{jk} moyenne des modules
   - Δ_{jk} distance enclidienne
   - D^2_{jk} Mahalamobis
   - χ^2 distance distributionnelle

3. Coefficents de correlation
   - r_{jk}
   - ρ de SPEARMANN

IV - FORMATION DES GROUPES

1. Sans analyse statistique

1.1. Ombrages différentiels.

1.2. Dendrogrammes.

1.3. Méthode nodale.

1.4. Arbre de longueur minimum

2. Analyses statistiques

Analyses des variances et de covariance.
matrices W résiduelle B inter individu

2.1. Indice de proximités

2.2. Distance généralisée de MAHALANOBIS

2.3. Analyse en composantes principales.

231. à partir de B

232. Représentation géométrique

233. Variables canoniques

234. Composantes varimax.

3. Analyse factorielle des correspondances.

définition de la distance du χ^2

I - GENERALITES.

1. Les problèmes de classification.

Dans bien des domaines il est important de pouvoir décrire un individu et surtout de savoir le situer relativement à d'autres. C'est ainsi le cas de l'Amélioration des Plantes où l'on cherche à déterminer la valeur propre d'une population et sa valeur relative par rapport à des populations voisines.

L'étude du polymorphisme des populations de P. maximum d'Afrique de l'Est nous a conduit à passer en revue les différentes méthodes de taxonomie numérique.

Dans tout problème de classification on se trouve en présence d'un tableau à double entrée (individus par caractères par exemple). Chaque colonne étant la description d'un individu au travers des différents caractères.

Le but de la classification est le regroupement des individus (c'est à dire des colonnes) les plus ressemblants.

La taxonomie numérique est l'étude théorique des méthodes de classification "c'est l'évaluation numérique des affinités ou similitude entre les unités taxonomiques et la mise en ordre de ces unités en taxons sur la base de leurs affinités""c'est la définition de SOKAL et SNEATH.

Un autre problème de classification peut se présenter les groupes étant constitués, dans quel groupe placer un individu donné, c'est plutôt un problème de classement.

2. Historique.

Depuis longtemps, ces problèmes ont attirés l'attention des scientifiques.

Sans remonter jusqu'à ARISTOTE, on peut mentionner des travaux de LINNE en 1735. L'aspect quantitatif de ces problèmes n'est abordé que 2 siècles plus tard ; FISHER 1936 en introduisant la notion de fonction discriminante, donne le départ à l'étude des méthodes d'analyse multivariable.

RAO (1948-1952) semble être le premier à s'intéresser aux deux problèmes de classification.

Plus récemment GOOD 1965 et KENDALL 1966 précisent la distinction entre classification et discrimination. De ce désintérêt des statiticiens pour le premier problème a résulté indirectement une éclosion en ordre dispersé, de méthodes fort nombreuses et plus ou moins acceptables connues notamment sous le nom de "Cluster analysis".

L'utilisation de plus en plus courante des moyens modernes de calcul et de traitement de l'information a accentué ce mouvement.

SOKAL, SNEATH et MICHENER sont à la base de cette dernière tendance.

3. Domaines d'application.

Le champ d'application des méthodes de classification est très important.

Mises à part la botanique et la zoologie qui furent toujours concernées par ces méthodes.

On trouve des applications en océanographie, sociologie, pédologie, écologie.

En psychologie, en médecine, en pédagogie les publications sont nombreuses.

En météorologie et climatologie, pour la définition et la classification des types de temps et même les domaines plus littéraires, comme la documentation, linguistique, archéologie, anthropologie sont concernés par ces méthodes.

II - LES DONNEES.

A la base du problème se trouve un tableau à double entrée :

	individus
caractères	

Pour les caractères quantitatifs le tableau comprend les différentes valeurs numériques observées. Pour les caractères qualitatifs le tableau est un ensemble de signes conventionnels 0 ou 1, + ou -, ou pour les caractères à plusieurs états des codes correspondant à des couleurs, des aspects etc.

1. Le choix des caractères à prendre en considération est propre à chaque matériel étudié (Fig. 1)

2. La collecte des données demande certaines précautions ; elle doit par exemple être homogène dans l'espace et dans le temps ; ne pas varier d'un observateur à l'autre; (par exemple pour les caractères qualitatifs la collecte des données doit être faite par le même observateur) (Fig. 2).

3. La codification des données.
C'est un problème général à tous les domaines. Chaque fois qu'il s'agit de mettre sous forme numérique des données qualitatives.

Pour les caractères qualitatifs ordonnés logiquement le code sera un nombre entier, par exemple 0 1 2 pour les couleurs blanc rose rouge.

Pour les caractères non ordonnés on aura souvent intérêt à les remplacer par plusieurs caractères à 2 variantes (absence - présence). Par exemple chez P. maximum le port de la 3e feuille.

Dressée	Cassée	Retombante
0	1	2

sera transformé en 3 caractères

A dressée ou non
B cassée ou non
C retombante ou non.

Ce qui donne le tableau suivant :

	A	B	C
0	1	0	0
1	0	1	0
2	0	0	1

Pour les données quantitatives il est parfois utile de les transformer pour normaliser les distributions.

ou linéariser des régressions

ou encore pour remplacer les variables

par des variables réduites ou non correlées.

III - REDUCTION DES DONNEES.

Du tableau I x J, individus par caractères, on doit passer à un tableau I x I , individus par individus.

Ce tableau I x I peut-être de plusieurs natures: une matrice d'indices de similarités (ou coefficients d'association), une matrice de coefficients de correlation ou bien une matrice de distances.

1. Indices de similarité.

Le problème de la définition d'un indice de similarité biologiquement valable est complexe, et les solutions données par les auteurs consultés sont nombreuses.

Il s'est avéré qu'un nombre de plus en plus grand de taxonomistes font appel aux ordinateurs pour le traitement de leurs données. L'importance des calculs necessités par les différentes méthodes montre que la machine électro-mécanique devient insuffisante.

SOKAL et SNEATH (1963) donnent leurs principes de base de la taxonomie numérique.

"1. La taxonomie idéale est celle dans laquelle le taxon a le plus grand contenu d'information et qui est basée sur le nombre de caractères le plus grand possible.

2. A priori chaque caractère est de poids égal dans la création naturelle des taxa.

3. Toute similarité (ou affinité) entre deux entités quelqonques est une fonction de la similarité de beaucoup de caractères pour lesquels on les compare.

4. Des taxa distincts peuvent être construits en raison des corrélations des divers caractères qu'on étudie.

5. La taxonomie numérique comme nous la concevons est donc strictement une science empirique.

6. L'affinité est estimée indépendamment des considérations phylogénétiques.

Sur le point 5 tous les auteurs ne sont pas d'accord. GOODALL (1964-1966 a - 1966 b - 1966 c - 1966 d - 1966 e - 1968) ne considère pas la taxonomie numérique comme une science empirique mais raisonne en termes probabilistes. Nous verrons cet aspect du problème plus loin.

1.1. Indices simples

- Caractères binaires

k \ j	+	−	
+	n_{jk}	n_{jk}	$n\ K$
−	n_{jk}	n_{jk}	$n\ k$
	n_j	n_j	n

$m = n_{JK} + n_{jk}$ coïncidences

$u = n_{jk} + n_{jk}$ non coïncidences

$n = m + u$

$= n\ k + n_{\ k} = n_j + n_j$ etc...

Le coefficient de JACCARD

$S_J = {}^{n_{JK}} /_{nJK + u}$ le plus simple

compris entre 0 et 1.

Le coefficient d'association SOKAL MICHENER

$S_{SM} = \frac{n}{n} = \frac{m}{m+u}$ compris entre 0 et 1

Le coefficient de ROGERS et TANIMOTO

$S_{RT} = \frac{m}{m + 2u} = \frac{m}{n + u}$ compris entre 0 et 1

- Caractères à plusieurs états.

Indice de ROGERS et TANIMOTO

	A				B		C			D				
	1	2	3	4	1	2	1	2	3	1	2	3	4	5
individu 1	-	+	-	-	-	+	-	-	+	-	+	-	-	-
individu 2	-	+	-	-	+	-	-	-	+	-	-	-	-	+

C'est le rapport du nombre de caractère ayant même état pour les deux variétés au nombre d'états représentés par les deux variétés pour tous les caractères.

Cet indice se ramène à S_{RT} dans le cas de caractères binaires.

1.2. Indice de similarité et probabilité.

Dans le problème qui nous préoccupe les variables décrivant les individus sont des caractères qualitatifs à deux ou plusieurs états et des caractères non métriques mais ordonnés à deux ou plusieurs classes. Ce sont des caractères morphologiques sujets à des processus aléatoires puisqu'ils dépendent du génotype de l'organisme lui-même fonction des effets aléatoires de mutation et de recombinaison. Partant de ce fait GOODALL a construit un indice de similarité basé directement sur la théorie des probabilités.

Pour chaque caractère i et pour chaque couple d'états (i_k, i_l), la similarité est déterminée de telle manière qu'elle engendre une relation d'ordre permettant de ranger les couples d'états. Etant donné deux couples $(i_k\ i_l)$ et $(i_m\ i_n)$ ou bien $(i_k\ i_l) < (i_m\ i_n)$ ou bien $(i_k\ i_l) \geqslant (i_m\ i_n)$ pour la similarité.

Soit $(i_m\ i_n)$ un couple pris au hasard

on définit Prob $\left\{ (i_m\ i_n) \geqslant (i_k\ i_l) \right\} = P_{(i_k\ i_l)}$

La similarité entre i_k et i_l est alors

$$S\ (i_k\ i_l) = 1 - P_{(i_k\ i_l)}.$$

L'indice de similarité pour chaque couple est donc défini comme le complément de la probabilité qu'un couple pris au hasard ait une similarité supérieure ou égale à celle du couple en question. Ces probabilités sont combinées en faisant l'hypothèse que les valeurs prises par les différents caractères dans le même individu sont indépendants. Le calcul imposé par la méthode exacte devient très rapidement impraticable même pour un calculateur électronique lorsque ce nombre de caractères et le nombre d'états par caractère augmentent. Une amélioration a été, malgré toute les approximations qu'on peut faire le calcul ne peut-être exécuter que sur ordinateur.

Il est évident que cet indice est original par rapport à ceux couramment utilisés en taxonomie numérique. Il s'applique à toutes les catégories de caractères, qualitatifs, ordonnés ou quantitatifs. Les concordances qui sont les moins probables donnent une meilleure preuve de ressemblance donc une contribution plus forte dans la similarité entre deux individus. Cet indice donne donc aux caractéristiques rares le poids accru que les taxonomistes tendent à préférer. De plus un indice de déviation défini par GOODALL 1966, ouvre la voie à un processus complet de taxonomie numérique contrôlé par des niveaux de signification. A moins qu'on ait une information extérieure sur les fréquences des caractères dans la population, elles doivent être estimées à partir des individus considérés pris comme un échantillon de la population. Cet indice dépend donc du contexte dans lequel le couple d'individus est considéré ce n'est pas un indice absolu.

D'autres auteurs tout en établissant des indices empiriquement se rapprochent de l'idée de GOODALL par le fait qu'ils pensent qu'il est souhaitable de donner plus d'importance aux états rares des caractères. C'est à dire un état d'un caractère représenté dans 10 % de la population aura plus de poids dans la similarité entre deux individus qu'un état représenté dans 90 % de la population. Pour SMIRNOV les poids accordés aux caractères sont fonction de la fréquence de ceux-ci dans la population.

Indice de GOODALL : exemple

Soit un caractère qualitatif prenant les états 1 2 3 4
soit m un échantillon d'une population

supposons, pour fixer les idées, m = 15

et $f_1 = 2 \quad f_2 = 4 \quad f_3 = 4 \quad f_4 = 5$

on sait que si $i \neq j \quad S_{ij} = 0$

si $i = j \quad S_{ij} \quad 0$

on a de plus $f_i < f_k \Rightarrow S_{ii} > S_{kk}$

et $P_{ij} = 1$ pour $i \neq j$

P_{ii} = Prob { un couple pris au hasard $\geqslant$ ii } = pour la similarité

$$\frac{\text{nb de couples} \geqslant ii}{\text{nb de couples total}} = \frac{\sum_{k \in E \{ k : f_k \leq f_i \}} \text{nb de couples kk}}{\text{nb de couples total}}$$

les relations d'ordre entre couple sont :

(11) > (22) = (33) > (44)

$$P_{22} = \frac{\sum_{k \in (1, 2, 3)} \frac{1}{2} f_k (f_k - 1)}{\frac{m(m-1)}{2}} = \frac{2 + 12 + 12}{210} = 0.12$$

$S_{22} = 0.88.$

Pour le problème qui nous préoccupe l'indice de SMIRNOV est apparu comme plus conforme aux données. Pour se rendre compte de l'information supplémentaire apportée par la pondération des caractères un indice simple de ROGERS et TANIMOTO est essayé sur les mêmes données.

1.3. Indice de SMIRNOV

a) Description

L'indice de SMIRNOV se formule de la façon suivante:

S = nombre de clones

E_i = nombre de clones présentant le caractère E à l'état i.

e_i = nombre de clones ne possédant pas le caractère E l'état i.

On a : $E_i + e_i = s$ et $\sum_i E_i = s \; i \in (1, e)$ pour l'état E.

La pondération se fait de la façon suivante :

pour un caractère E et un état i considérés le poids accordé W_{Ei} sera :

* $\frac{e_i}{E_i}$ si les deux clones sont à l'état i pour le caractère E.

** $\frac{E_i}{e_i}$ si les deux clones ne sont pas à l'état i pour le caractère E.

*** -1 si l'un des clones est à l'état i, l'autre ne l'étant pas pour le caractère E.

On fait ensuite la moyenne des poids pour tous les états pour le caractère E.

$$\bar{W}_E = \frac{1}{e} \sum_{i=1}^{e} W_{Ei}$$

Si on considère enfin tous les caractères E^1 E^2 E^3.. E^{α}.. E^m ;

$$\bar{W}_{E^{\alpha}} = \sum_{i=1}^{e^{\alpha}} W_{Ei}^{\alpha}$$

L'indice de similarité entre les deux clones f et g sera :

$$t_{f,g} = \frac{1}{n} \sum_{\alpha=1}^{m} \bar{W}_{E^{\alpha}} ; \quad n = \sum_{\alpha=1}^{m} e^{\alpha}$$

$$t_{f,g} = \frac{1}{n} \sum_{\alpha=1}^{m} \left[\sum_{i=1}^{e^{\alpha}} W_{Ei}^{\alpha} \right]$$

Cet indice peut prendre des valeurs positives ou négatives supérieures ou inférieures à 1 en valeur absolue.

Une étude des propriétés statistiques de l'I.S. a montré que la moyenne de cet indice est nulle et que la variance est

$$\sigma_s^2 = \frac{1}{n} + \frac{1}{n^2} \sum_{\alpha} \sum_{i \neq j} \frac{E_i^{\alpha} E_j^{\alpha}}{e_i^{\alpha} e_j^{\alpha}}$$

qui se ramène à $\frac{1}{m}$ pour les caractères binaires.

m = nb de caractères.

Exemple de matrice des indices de SMIRNOV : fig. 3.

2. Distances.

distance moyenne.

$$d_{jk} = \frac{1}{n} \sum_{i=1}^{n} \left| X_{ij} - X_{ik} \right|$$

distance enclidienne

$$\Delta_{jk} = \left(\sum_{i=1}^{n} (X_{ij} - X_{ik})^2 \right)^{1/2}$$

distance généralisée de MAHALANOBIS

$$D^2_{jk} = \sum_{\alpha} \sum_{\beta} (\bar{X}_{\alpha j} - \bar{X}_{\alpha k})^2 (\bar{X}_{\beta j} - \bar{X}_{\beta k}) \lambda^{\alpha\beta}$$

où $(\lambda^{\alpha\beta})$ est la matrice inverse de $(\lambda_{\alpha\beta})$,

matrice des corrélations résiduelles, entre caractères pris deux à deux.

Par exemple pour deux caractères, Δ_{jk} est la distance réelle entre individus dans une représentation plane ;

$$\Delta^2_{jk} = (X_{1j} - X_{1k})^2 + (X_{2j} - X_{2k})^2$$

si ρ est le coefficient de corrélation résiduel entre les caractères 1 et 2 alors on a :

$$D^2_{jk} = \frac{1}{1 - \rho^2} \left((\bar{X}_{1j} - \bar{X}_{1k})^2 - 2\rho (\bar{X}_{1j} - \bar{X}_{1k}) (\bar{X}_{2j} - \bar{X}_{2k}) + (\bar{X}_{2j} - \bar{X}_{2k})^2 \right)$$

3. Coefficients de corrélation

$$r_{jk} = \frac{\sum_{i=1}^{n} (X_{ij} - \bar{X}_j) (X_{ik} - \bar{X}_k)}{\sqrt{\sum_{i=1}^{n} (X_{ij} - \bar{X}_j)^2 \sum_{i=1}^{n} (X_{ik} - \bar{X}_k)^2}}$$

Coefficient de corrélation de rang de SPEARMANN.

$$\rho = 1 - \frac{6 \sum d_i^2}{n^3 - n.}$$

IV - FORMATION DES GROUPES.

1. Sans analyse statistique.

1.1. Ombrages différentiels.

L'étendue de la mesure de la ressemblance est partagée en intervalles à chacun desquels on fait correspondre un ombrage plus ou moins foncé selon la grandeur de la similarité ; l'ombrage le plus foncé correspondant aux valeurs de la similarité les plus grandes. On intervertit alors lignes et colonnes pour que les carreaux sombres soient le plus proche possible de la diagonale (fig. 4).

1.2. Dendrogrammes.

La matrice de similarité étant calculée, le problème qui reste à résoudre est celui de la formation des groupes. La construction du dendrogramme ne conduit pas effectivement à la formation des groupes mais à l'établissement d'une structure hiérarchisée des différents individus. Des groupes pourraient être formés si on pouvait déterminer le niveau de signification de l'indice utilisé.

L'établissement du dendrogramme peut-être fait par deux méthodes qui peuvent l'une et l'autre être pondérées WEIGHTED, (unweighted) variable-groupe method et weighted (unweighted) pair group method. Dans la première méthode tous les groupements possibles deux à deux sont faits à chaque cycle. Dans la seconde seul le groupement des deux individus les plus ressemblants est effectué à chaque cycle. La matrice de similarité entre les groupes ainsi formés peut se faire avec ou sans pondération. Dans le premier cas on utilise la méthode de SPEARMAN pour le calcul du coefficient de corrélation entre groupe.

$$r_{qQ} = \frac{\square qQ}{\sqrt{q + 2\,\Delta\, q}\ \sqrt{Q + 2\Delta Q.}}$$

où r_{qQ} est le coefficient de corrélation entre les groupes q et Q

qQ est la somme de tous les coefficients de corrélation entre les membres d'un groupe avec l'autre.

Δq est la somme de tous les coefficients de corrélation du premier groupe, ΔQ du deuxième groupe.

q est le nombre d'unités du premier groupe Q celui du second.

Dans le second cas on utilise la moyenne arithmétique des coefficients entre les membres d'un groupe avec l'autre. La méthode avec pondération présente l'inconvénient dans certains cas de donner des "renversements". C'est à dire que l'on peut trouver des coefficients entre groupe supérieurs à tous les coefficients à l'intérieur des deux groupes, ce qui rend difficile la construction du dendrogramme.

Exemple :

A partir de la matrice de la fig. 1 on peut construire un dendrogramme par la méthode la plus simple c'est à dire, sans pondération avec regroupement par deux. La détermination du dendrogramme se fait par étapes successives. Nous allons voir en détail la deuxième étape.

La matrice obtenue à la fin de la première étape est la suivante :

	K 107	A	B	C	K 117
K 107		- 23	+ 8	- 26	- 25
A	- 23		- 36	- 14	- 8
B	+ 8	- 36		- 34	- 23
C	- 26	- 14	- 34		- 29
K 117	- 25	+ 8	- 23	- 29	

A, B et C sont des groupes formés. Les indices soulignés sont les plus grands de chaque colonne.

Dans la première colonne la liaison la plus forte est celle de K 107 avec B, on regarde alors la colonne B, la liaison la plus forte est celle de B avec K 107 le groupe (B, K 107), qu'on appelle D, est formée au niveau + 8. L'ordre dans lequel on examine les colonnes est sans importance. Dans la quatrième colonne, la liaison la plus forte est celle de C avec A, mais si on regarde la colonne A la liaison la plus forte n'est pas celle de A avec C le groupe A, C n'est pas formé.

Le même raisonnement conduit à la formation du groupe (A, K 117), qu'on appelle E, au niveau + 8. On forme une nouvelle matrice avec D, E, C, en prenant les moyennes arithmétiques.

$$I_{D,E} = \frac{1}{4} (I_{B,A} + I_{B,K117} + I_{K107,A} + I_{K107,\ K117})$$

on obtient :

	D	E	C
D		- 27	- 30
E	- 27		- 21
C	- 30	- 21	

On recommence à partir de cette nouvelle matrice. On obtient le dendrogramme fig. 5.

1.3. Méthode nodale de ROGERS et TANIMOTO.

Cette méthode mise au point pour l'indice de ROGERS et TANIMOTO peut s'appliquer à n'importe quel coefficient d'association ou indice de similarité.

Soit S_{ij} l'indice de similarité entre les clones i et j. On appelle indice de "typicalité" le nombre Ri d'unités taxonomiques ayant au moins un état commun avec la variété i. (Nombre de $S_{ij} \neq 0$ pour i donné). Soit t le nombre de taxa ; Ri $\leq$ t - 1. Le taxon dont l'indice de typicalité Ri est le plus élevé est le plus "typique".

On calcule $H_i = \prod_{j=1}^{t}$ Si j pour $j \neq i$ et Sij > 0.

Le taxon i ayant la plus grande valeur Ri et la plus grande valeur H_i est considéré comme le plus "typique".

C'est le centroïde du système lorsque les coefficients sont en logarithme. Le taxon le plus typique est appelé premier noeud de l'étude et autour on forme un groupe avec tous les taxa de forts coefficients de similarité avec lui. Le second noeud choisi est le taxon suivant dans l'ordre de typicalité. Le rayon de groupe entourant le premier noeud doit être tel qu'il ne doit pas englober le second noeud.

On introduit maintenant la distance dij = $- \log_2$ Sij. comprise entre 0 et l'infini (de Sij = 0 à Sij = 1).

$$-\log_2 H_i = \sum_{j=1}^{j=t} dij = \sum_{j=1}^{j=t} (-\log_2 Sij) \quad j \neq i \; Sij > 0.$$

$d_{1,2}$ est la distance entre le premier et le second noeud. Toute unité dont la distance au premier noeud est inférieure à $d_{1,2}$ est lié au premier noeud. On forme ainsi un groupe.

Après la détermination de ce premier groupe on enlève tous ses éléments de l'étude et on recommence pour les groupes secondaires.

Exemple d'application de la méthode nodale.

La matrice des indices est transformée en matrice de distances en posant $dij = -\log_2 \left(\frac{Sij + k}{k'} \right)$

(avec k et k' tel que $0 < \frac{Sij + k}{k'} < 1$).

Les sommes des distances par colonne sont calculées. L'individu de la colonne dont la somme des distances est la plus petite est retenu comme premier centroïde (1) le second est l'individu dont la somme est immédiatement inférieure (2)

Soit $d_{1,2}$ la distance entre les deux premiers centroïdes.

Tous les individus dont la distance à (1) est inférieure à $d_{1,2}$f forment un groupe autour de (1) à conditions que leurs distances deux à deux soient inférieures à $d_{1,2}$.

Le tableau de la fig. 6 montre que K 107 est le premier centroïde, K 109 est le second - $d_{1,2}$ = 1,22
donc le premier groupe est K 107, K 112, K 114 la distance intragroupe est 0,70.

Fig. 6 - Tableau des $d_{ij} = -\log_2 \left(\frac{S_{ij} + 1{,}00}{2}\right)$

	K 107	K 109	K 110	K 112	K 114	K 115	K 116	K 117
K 107		1.22	1.58	0.89	0.88	1.56	1.34	1.41
K 109	1.22		0.66	1.69	1.71	1.22	1.14	0.98
K 110	1.58	0.66		1.67	1.56	1.29	1.24	0.82
K 112	0.89	1.69	1.67		0.32	1.32	1.43	1.62
K 114	0.88	1.71	1.56	0.32		1.81	1.98	1.17
K 115	1.56	1.22	1.29	1.32	1.81		0.28	1.45
K 116	1.34	1.14	1.24	1.43	1.98	0.28		1.54
K 117	1.41	0.98	0.82	1.62	1.62	1.45	1.54	
Total	7.88	8.62	8.82	8.94	9.88	8.93	8.95	8.99

Fig. 7 - Tableau de d_{ij}

	K 109	K 110	K 115	K 116	K 117
K 109		0.66	1.22	1.14	0.98
K 110	0.66		1.29	1.24	0.82
K 115	1.22	1.29		0.28	1.45
K 116	1.14	1.24	0.28		1.54
K 117	0.98	0.82	1.45	1.54	
Total	4.00	4.01	4.24	4.20	4.79

Fig. 8 -

	K 110	K 115	K 116	K 117
K 110		1.29	1.24	0.82
K 115	1.29		0.28	1.45
K 116	1.24	0.28		1.54
K 117	0.82	1.45	1.54	
Total	3.35	3.02	3.06	3.81

La fig. 7 montre que :

K 109 est effectivement le second centroïde.

K 110 est le troisième.

la distance entre K 109 et K 110 est 0.66 - K 109 se trouve isolé.

Si on supprime K 109 (fig. 8) le troisième centroïde est K 115 et non K 110.

On revient donc à la matrice précédente en groupant autour de K 109, les clones dont la distance à K 109 est inférieure à d (K 109, K 115) = 1,22.

Ce qui forme le groupe

(K 109 - K 110 - K 117 - K 116)

mais d (K 116, K 110) = 1,24 $>$ 1,22

d (K 116, K 117) = 1,54 $>$ 1,22

le clone K 116 n'est pas à intégrer dans ce groupe.

Le second groupe est donc K 109 - K 110 - K 117 sa distance intra-groupe est 0.82.

Le troisième est K 115 - K 116.

La fig. 9 montre la représentation plane des groupes

1.4. Arbre de longueur minimum .

On suppose n points dans un espace multidimensionnel. Un arbre raccordant ces points est un ensemble de segments joignant les points deux à deux tel que :

1 - Il n'existe pas de boucle

2 - Chaque point est visité au moins une fois.

La représentation plane de l'arbre est approximative mais fournit une image des rapprochements. fig. 10.

La détermination se fait par ordinateur ; un programme existe au laboratoire de statistique de Monsieur le Professeur BENZECRI.

2. Analyses statistiques proprement dites.

Les I individus sont comparés dans un essai en "blocs entièrement randomisés avec répétitions par bloc". Sur cet essai J caractères quantitatifs sont mesurés. A partir des J analyses de variances on peut construire un indice de proximité (PERNES 1968). A partir des J analyses de variance et des J $(\frac{J-1}{2})$ analyses de covariance, on obtient la matrice W des variances et covariances résiduelles, et la matrice B des variances et covariances entre individus.

Ce sont deux matrices carrées symétriques. En divisant chaque terme par le produit des écart-types (résiduels ou entre individus) on obtient deux nouvelles matrices ; à partir de W la matrice des corrélations résiduelles, à partir de B la matrice des corrélations entre individus. La première conduit à l'analyse des distances généralisées de MAHALANNOBIS, la seconde à l'analyse en composantes principales.

1.2. Indice de proximité.

Les I analyses de variances aboutissent à l'établissement de différences significatives entre individus (test de TUKEY), il est possible de calculer une 1ère distance entre 2 individus pour chaque caractère - et d'additionner toutes ces distances pour obtenir un indice généralisé. Cet indice de distance dij ($i \in I$, $j \in J$) sera transformé en Pij = M - dij où $M = \underset{ij}{\text{Max}}\ dij$

en divisant Pij par M - on obtient une valeur comprise entre 0 (dij = M) et 1 (dij = 0). A partir de cette matrice I x I des - Pij /M toutes les méthodes de regroupement taxonomiques sont applicables (par exemple dendrogrammes ou méthode nodale).

les dij peuvent servir à l'établissement d'un arbre de longueur minimum.

Exemple - (PERNES, COMBES, 1968).

L'analyse de variance donne pour un caractère le classement des clones suivant :

40 34 36 | 3 56 | 15 23 21 | 14 | 13 25 | 10 | 6 4 | 52

Les accolades réunissent les clones non significativement différents.

La distance entre deux clones, pour ce caractère, sera

0 si les deux clones ne sont pas différents

1 si les deux clones sont différents et qu'il n'existe entre eux aucun clone à la fois différents de l'un et de l'autre

2 si les clones sont séparés par un clone différent de l'un et de l'autre

n si les clones sont séparés par (n - 1) clones différents de l'un et de l'autre.

Fig. 11 - Tableau des distances.

	40	34	36	3	56	15	.....
40							
34	0						
36	0	0					
3	1	1	1				
56	1	1	1	0			
15	2	2	2	1	0		
.							
.							
.							

On obtient dij en cumulant les distances pour tous les caractères.

Cette méthode a permis d'aboutir à des conclusions biologiques très intéressantes.

Deux essais ont été faits sur clones issus de graines apomictiques (reproduction asexuée) et sur clones issus de boutures (reproduction végétative au sens stricte).

Les deux dendrogrammes (fig. 12) obtenus sont différents et montrent que l'expression du génotype a subi des modifications autoentretenus par multiplication végétative.

Après multiplication par graines les clones se regroupent par localisation géographique. Sans passage par graines le regroupement est climatique.

Par exemple le clone 56 originaire de Grand-Lahou est dans le groupe Abidjan dans l'essai avec passage par graines ; mais le milieu climatique de Grand-Lahou est plus proche de Sassandra que d'Abidjan (déficit hydrique cumulé supérieur à 400 mm contre 200 pour Abidjan ; Saison sèche plus longue de deux mois).

2.2. Distances généralisées D^2 de MAHALANOBIS

Introduites en 1936 par MAHALANOBIS cette distance n'est applicable qu'à des mesures de distributions normales.

WILKS d'abord, RAO et FISHER ensuite, généralisent l'analyse de variance à une seule variable au cas de variables multiples.

Les D^2 se calculent à partir de la matrice W des corrélations résiduelles.

$$D^2 = (\bar{x}_1 - \bar{x}_2)' \quad S^{-1} \quad (\bar{x}_1 - \bar{x}_2)$$

$\bar{x}_1$ désigne le vecteur, à J dimensions, des mesures de l'individu 1.

On retrouve une matrice I x I de distances à partir de laquelle on peut construire des regroupements par dendrogramme, par la méthode nodale, par l'ALM. Ou par la méthode des constellations qui repose sur les 2 critères suivants :

a) tout élément dont l'adjonction à un groupe n'augmente pas considérablement la distance intra-groupe appartient au groupe. Inversement tout accroissement important de la distance intra-groupe rejette l'élément.

b) une valeur D^2 supérieure à $\frac{n_1 + n_2}{n_1 \quad n_2} \chi^2_I (0,05)$ empêche l'appartenance des 2 individus à un même groupe à moins qu'il ne puisse être relié par une chaine de D^2 tous inférieures au seuil. (PERNES 1965).

(seuil de signification de la variable T^2 de comparaison deux à deux de vecteurs moyennes où n_1 et n_2 sont les effectifs sur lesquels sont évalués les deux vecteurs moyennes.)

Exemple : (PERNES 1965).

Les clones de Panicum maximum sont comparés aux essais blocs. Les D^2 sont calculés le seuil obtenu est 1,57.

Fig.

Groupe	D^2	Nb termes	D^2 moyen intra-groupe
6 - 4	1,03	1	1,03
6 - 4 - 52	3,90	3	1,30
6 - 4 - 52 - 14	17,55	6	2,93

Pour le troisième groupe le D^2 moyen intra-groupe 2,93 est supérieur au seuil d'une part, et d'autre part, il existe un saut non négligeable lorsqu'on ajoute le clone 14 au second groupe.

2.3. Analyse en composantes principales.

2.3.1. Le but de l'analyse en composantes principales est l'attribution des regroupements à des variables biologiquement significatives, obtenus par combinaisons linéaires de variables initiales.

Cette étude se fait à partir de la matrice des corrélations B' issue de B. Le calcul consiste en la recherche des vecteurs propres de cette matrice des corrélations inter-individus. La première direction propre correspond à la valeur propre la plus grande.

2.3.2. A la matrice B' correspond une forme quadratique : $Q(x) = (X - \mu)' (B')^{-1} (x - \mu)$ qui est l'équation d'un ellipsoïde dans un espace à p dimensions.

La recherche des directions propres et celle des axes de l'ellipsoïde ne font qu'un. Un changement d'axes, qui n'est autre qu'une rotation, donne de nouveaux axes parallèles à ceux de l'ellipsoïde. Dans ce nouveau système l'équation de l'ellipsoïde devient : $Q'(Y) = (Y - \lambda)' D^{-1} (Y - \lambda)$

D^{-1} est une matrice diagonale.

D et B' sont équivalentes puisqu'elles correspondent au même ellipsoïde.
Les vecteurs propres sont orthogonaux et indépendants, ils forment une base de l'espace (fig. 15).

Exemple d'analyse en composantes principales.

(PERNES et COMBES 1970)

L'analyse en composantes principales des types II de Panicum maximum de Côte d'Ivoire indistinguable qualitativement cultivés de façon homogène à Adiapo Doumé, a montré une différenciation quantitatives finement structurée suivant les gradients géographiques naturels de la Côte d'Ivoire (Fig. 16).

2.3.3. Utilisation des variables canoniques

C'est une méthode améliorée de la précédente, elle fait intervenir les deux matrices W et B pour, en quelque sorte, purifier la matrice B des corrélations résiduelles.

Le problème à résoudre est de trouver t fonctions linéaires ($t < p$) des variables qui maximisent la somme de toutes les valeurs possible de D^2 entre individus.

On montre que la meilleure fonction linéaire est la plus grande racine de

$$|B - \lambda W| = 0$$

2.3.4. Méthode des composantes varimax

Cette méthode permet d'identifier les groupes en diminuant les distances intra-groupes et en augmentant les distances inter-groupes. Chaque composante identifie un groupe.
Le calcul est fait par ordinateur. fig. 17.

Exemple : Thèse de PERNES (1972).

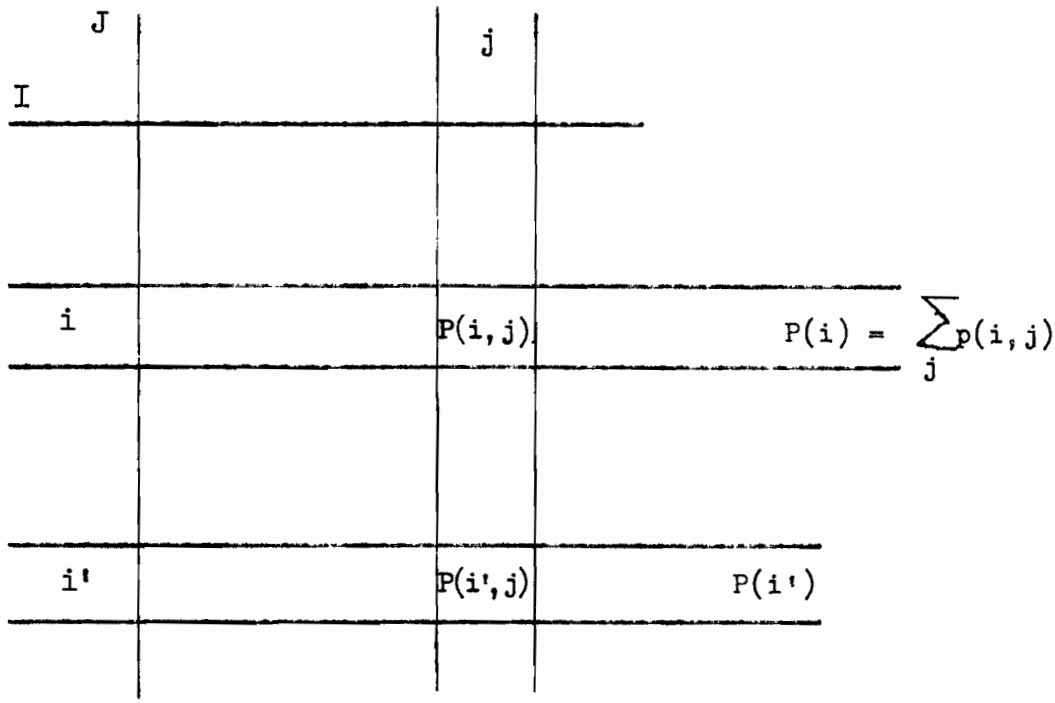
3. Analyse factorielle des correspondances.

Le principe est le même que celui de l'analyse en composantes principales ; mais la forme quadratique utilisée en guise de matrice de corrélations a les **propriétés suivantes :**

1° Le rapprochement des individus en fonction de leurs profils sur les différents caractères et non en fonction des notes absolues sur ces caractères

La distance entre 2 points, appelée distance du χ^2, est

$$d^2(i,i') = \sum_{j \in J} \frac{1}{p(j)} \left[\frac{p(i,j)}{p(i)} - \frac{p(i',j)}{p(i')} \right]^2 .$$



La pondération par p(i) accentue le caractère original j pour l'individu i ; en effet si, p(i,j) est petit dans un ensemble où p(i) est grand, en pondérant par p(i) on accentue le fait que p(i,j) est petit.

La pondération par P(j) normalise, par l'étendue du caractère.

Dans la formule classique de la distance sans pondération par p(j), intervient la comparaison terme à terme des j éléments des profils de i et i' en donnant le même poids à chacun. Supposons que les effectifs mesurés par $p(j_o)$ de la colonne j_o soient considérables dans $d^2(i,i')$ le terme en j_o sera très grand par rapport aux autres et jouera un rôle excessif dans la détermination des proximités.

L'expression pondérée appelée distance du χ^2 a le mérite d'atténuer ces disparités.

La distance du χ^2 a également l'avantage de vérifier le principe d'"équilibre distributionnelle" si deux points P_J^i et $P_J^{i'}$ sont confondus et si on les considère comme un seul point affecté de la somme des masses i, et i', alors les distances ne sont pas modifiées entre les éléments de J. Cette propriété de la formule est fondamentale, elle explique la stabilité des résultats.

2° Identification par une application canonique des individus et des caractères.

On montre qu'il existe des relations simples entre les facteurs représentants I et ceux représentants J.
Les facteurs du nuage J sont proportionnels aux coordonnées des points représentatifs de I sur les axes factoriels du nuage I (et réciproquement).
Une représentation simultanée des deux nuages et donc possible (CORDIER 1968).

Exemple:R. RENE (1971)

Les différents clones de Panicum maximum d'Afrique de l'Est décrits sur des caractères morphologiques du type du tableau de la fig. 1 et 2 ont été étudiés par l'analyse factorielle des correspondances.

La fig. 18 montre sur le même graphique (Axes 1,2) la disposition des clones des uns par rapport aux autres les caractères qualitatifs qui accentuent les rapprochements.

Cette analyse a permis une étude de la variabilité naturelle des populations de Panicum d'Afrique de l'Est. Les fig. 19, 20 et 21 montrent les différences entre régions.

La région représentée par la fig. 20, (Meru Embu) présente l'originalité d'inclure des hybrides interspécifique avec Panicum maximum.

CONCLUSIONS

La très grande diversité des méthodes disponibles montre qu'aucune solution n'est, à elle seule, entièrement satisfaisante. Seule la confrontation des résultats obtenus par différentes démarches permet une appréhension convenable des structures des ensembles étudiés. L'utilisation convergente de plusieurs méthodes, leur apport propre et leurs significations biologiques sont illustrées et justifiées dans PERNES (thèse 1972). Une structure ne peut-être tenue pour assurée que si elle est révélée à partir d'algorithmes de classification différents, sinon on risque d'attribuer à un ensemble non organisé un système classificatoire qui n'est qu'un artéfact de la méthode employée. C'est par cette convergence de méthodes que l'utilisateur prudent peut palier à l'absence de tests difficiles à définir et acquérir une vision synthétique et progressive des structures qu'il révèle.

- BIBLIOGRAPHIE -

CORDIER B. Thèse de 3e cycle (1968) ISUP.

PERNES J. 1965 - Rapport ORSTOM multigraphié.

PERNES J. et COMBES D. 1968 - Rapport ORSTOM multigraphié.

PERNES J., COMBES D. 1970 - Cahiers de Biologie ORSTOM 14 : 13-34.

PERNES J., COMBES D., RENE-CHAUME R., 1970 - C.R. Acad. Sciences Paris 270 : 1992-1995.

PERNES J. 1972 - Thèse de Doctorat d'Etat (en préparation).

RAO R. - Advanced Statistical Methods in Biometric Research WILEY New-York 1957.

RENE-CHAUME R., PERNES J., COMBES D. 1969 - Rapport ORSTOM multigraphié.

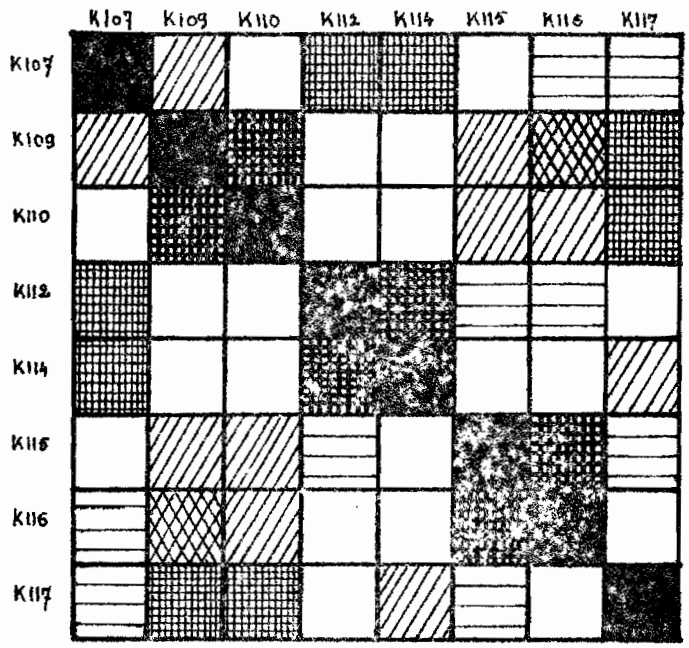
RENE-CHAUME R. (1971) - Rapport ORSTOM dactylographié.

SOKAL et SNEATH - Principal of Numerical Taxonomy Freeman 1963 San-Francisco.

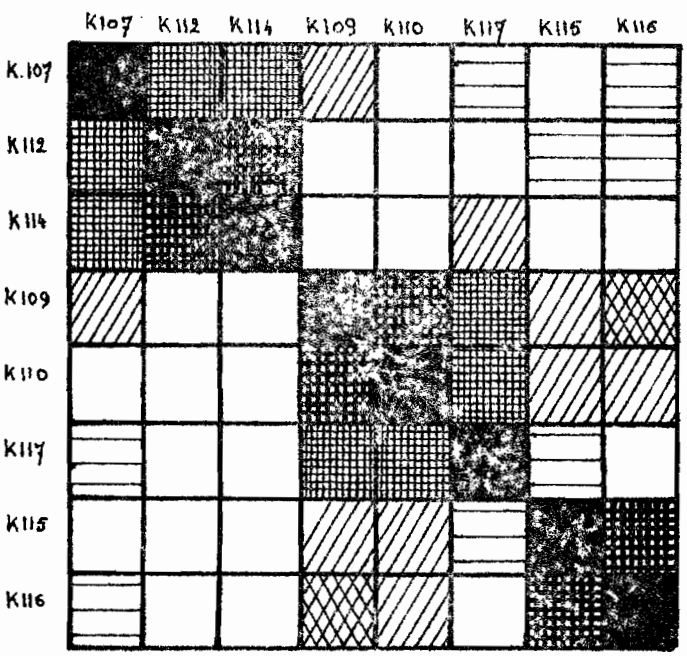
matrice des indices de Smirnov x100

	K107	K109	K110	K112	K114	K115	K116	K117
K107		-14	-33	+ 8	+9	-32	-21	-25
K109	-14		+27	-38	-39	-14	-9	+4
K110	-33	+27		-37	-32	-18	-15	+13
K112	+ 8	-38	-37		+60	-20	-26	-35
K114	+9	-39	-32	+60		-43	-48	-11
K115	-32	-14	-18	-20	-43		+65	-27
K116	-21	-9	-15	-26	-48	+65		-31
K117	-25	+4	+13	-35	-11	-27	-31	

- ombrages différentiels -



- avant permutations -



- après permutations -

. Dendrogramme (I.S)

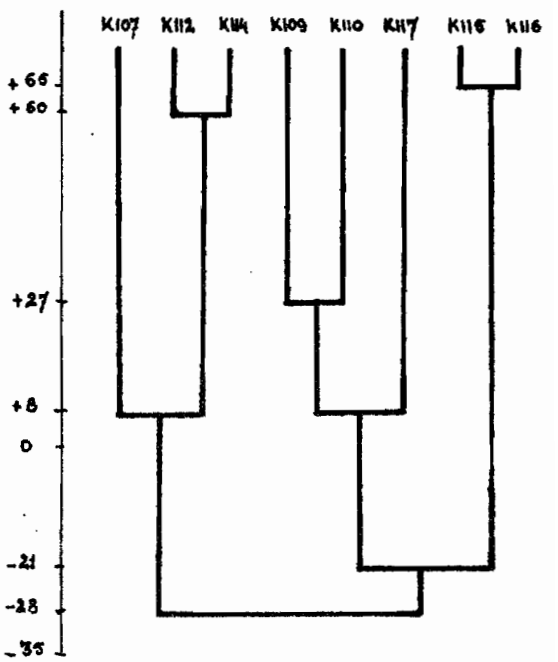
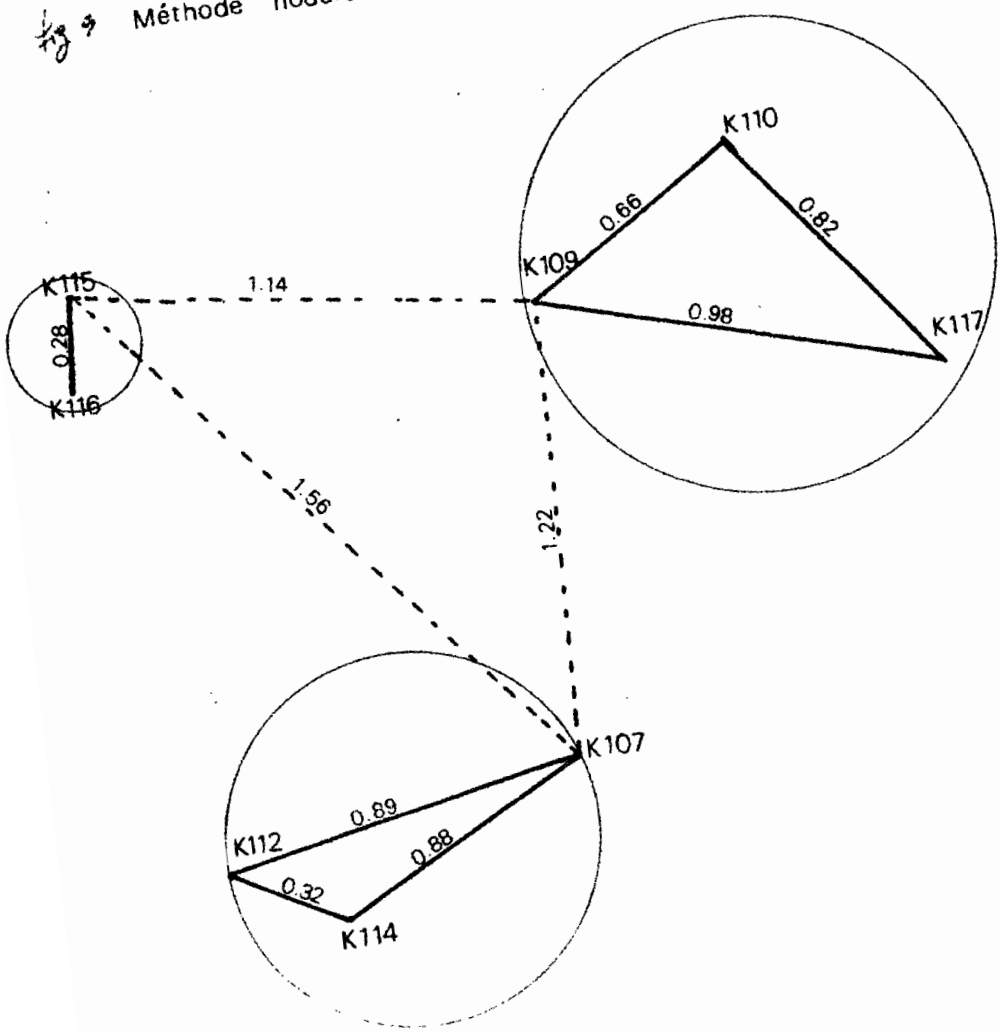
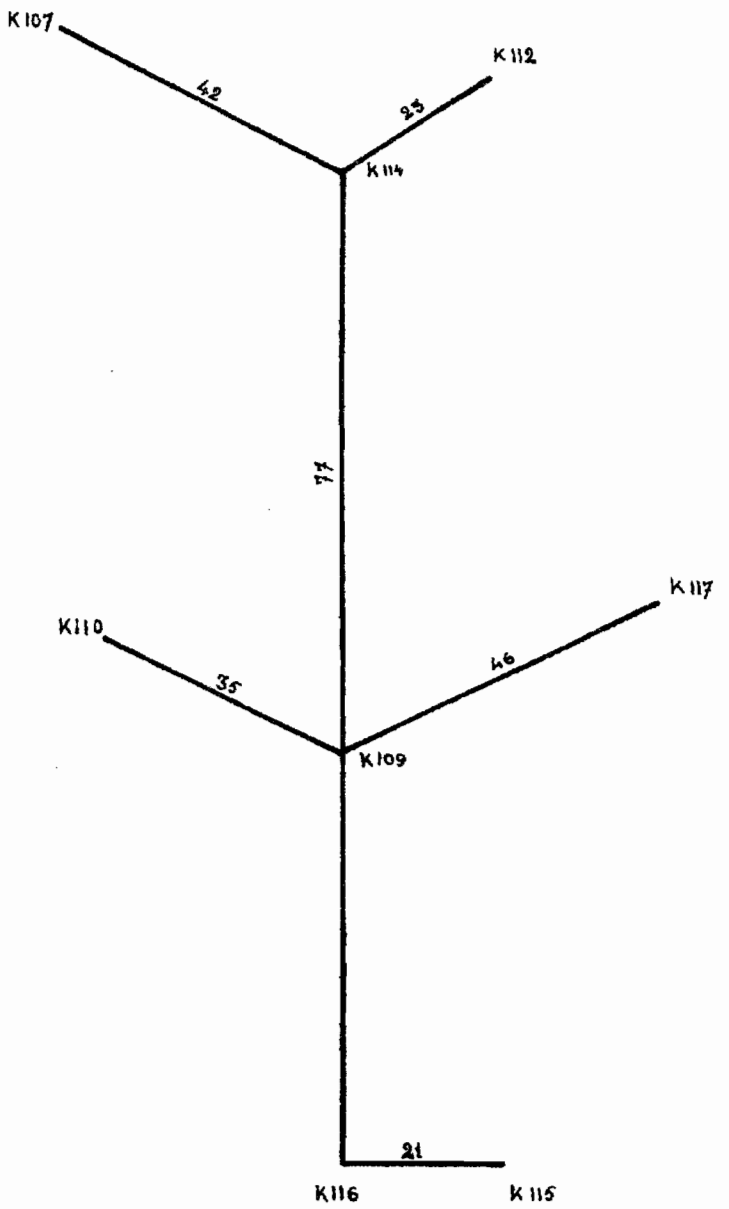


Fig 9 Méthode nodale



- Arbre de longueur minimum -



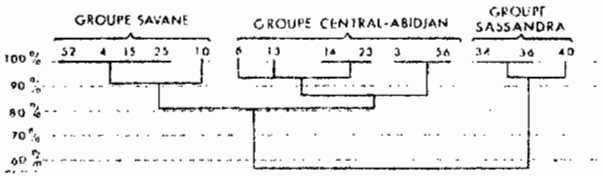


Fig. 11.— Dendrogramme après multiplication par graines

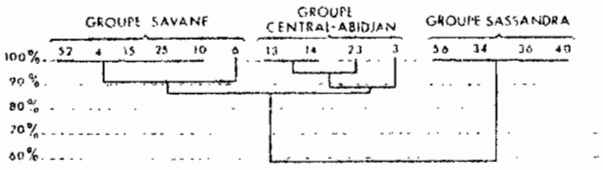


Fig.12. — Dendrogramme sans multiplication par graines

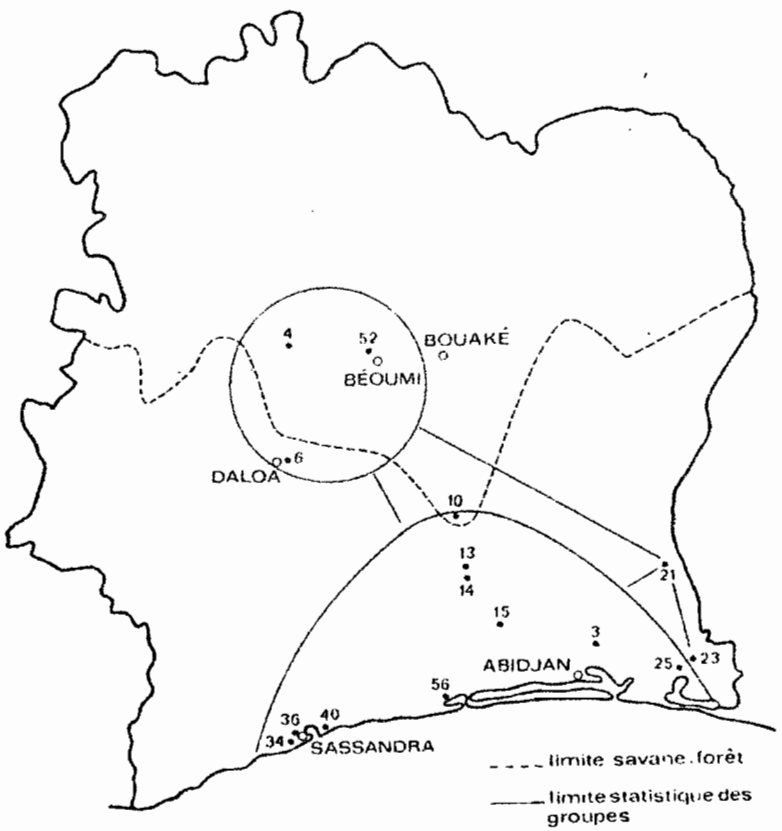


Fig.13. — Position géographique des 15 populations étudiées.

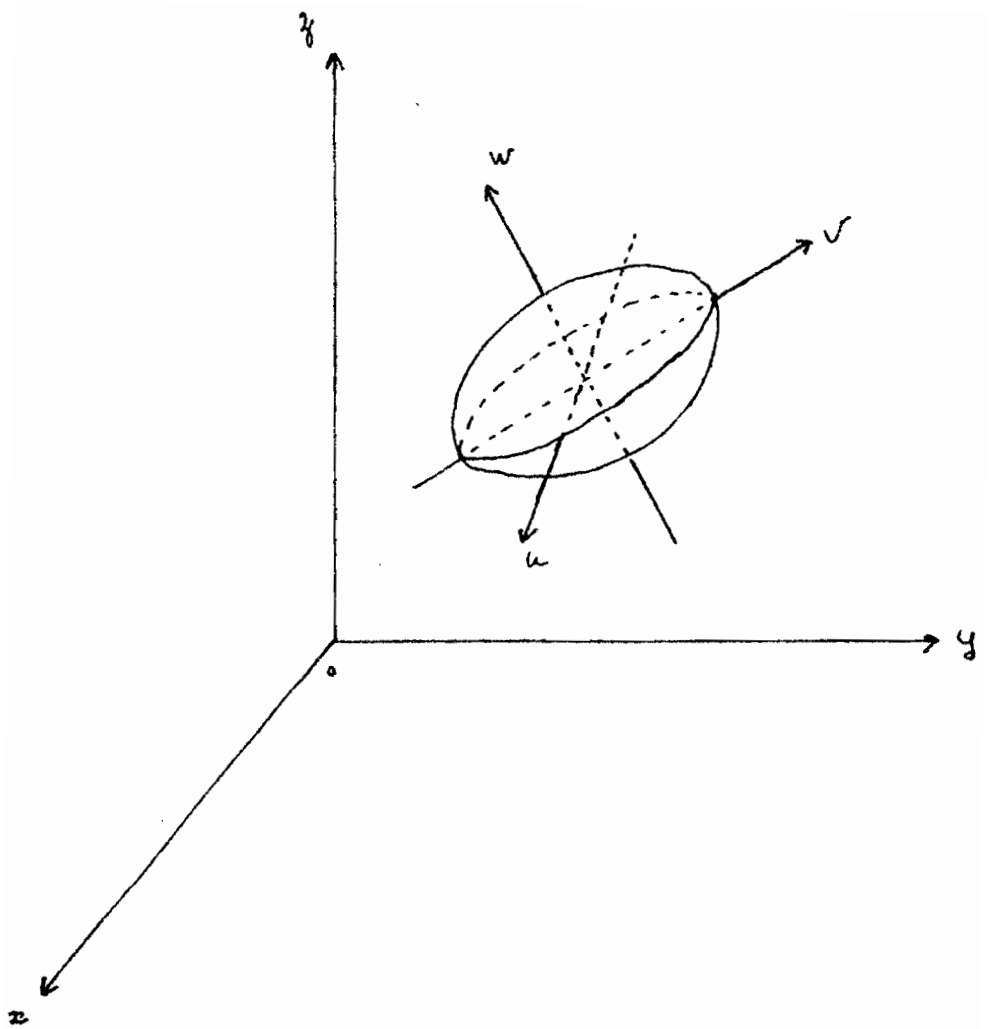


Fig. 15.

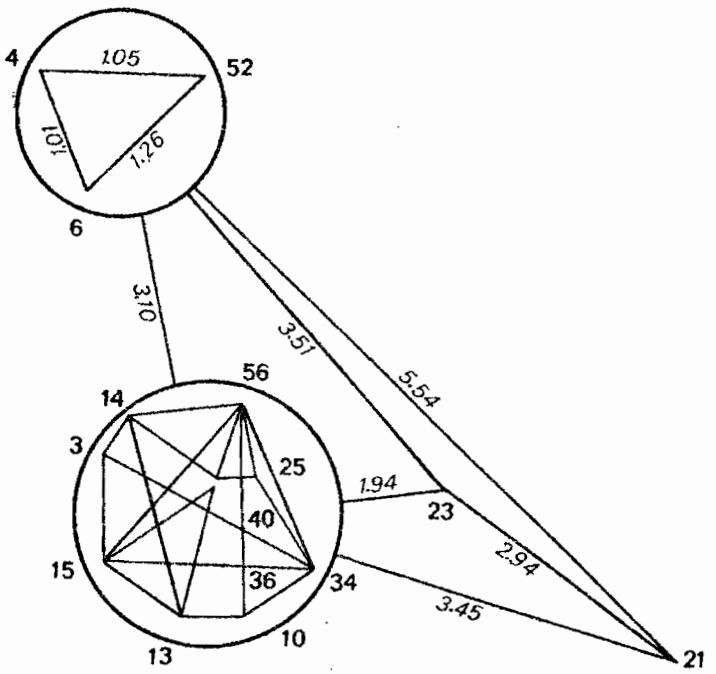


Fig. 14 — Positions relatives des diverses constellations obtenues par les distances D^2 de Mahalanobis.

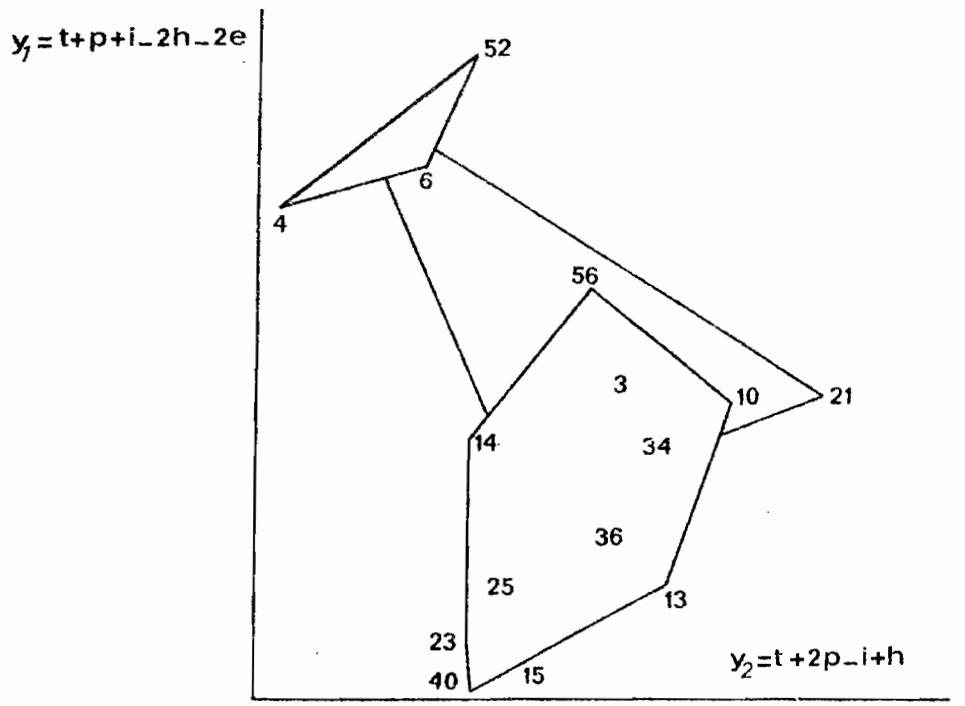


Fig. 16 — Représentation graphique des groupes obtenus par l'utilisation des deux composantes principales y_1 et y_2.