Boissiera 24: 369-381. 1975.

Les méthodes de taxonomie numérique

R. RÉNÉ-CHAUME

RÉSUMÉ

Quelques méthodes de taxonomie numérique sont décrites et appliquées à l'étude du polymorphisme des populations de *Panicum maximum.*

SUMMARY

Various methods of numerical taxonomy are described and applied to a study of polymorphism in populations of *Panicum maximum.*

Décrire un individu, le situer relativement à d'autres sont des démarches utilisées dans tous les domaines des sciences biologiques, physiques ou sociales. C'est la taxonomie (ou la taxinomie) numérique.

Le schéma théorique est simple: les différentes "unités" (individus, populations, etc.) sont observées et évaluées sur des caractères, puis leur "ressemblance" est calculée par des moyens mathématiques divers, enfin leur situation les unes par rapport aux autres est représentée graphiquement par différents procédés.

La détermination de la valeur propre d'une population et sa position par rapport à des populations voisines est un des problèmes essentiels en amélioration des plantes.

L'étude du polymorphisme des populations de *Panicum maximum* d'Afrique de l'Est nous a conduit à passer en revue différentes méthodes de taxonomie numérique. Ces méthodes sont nombreuses et particulièrement variées; il est apparu nécessaire d'en utiliser plusieurs, avec leurs avantages et inconvénients propres, seule la convergence des résultats permet de conclure sur la valeur des faits biologiques.

La nature des populations de *Panicum maximum* a pu ainsi être mise en évidence à l'aide de plusieurs méthodes d'analyse des données (cf. Réné-Chaume & al., 1969; Réné-Chaume, 1971; Pernès, 1975; Combes, 1975).

Nous allons passer en revue les différentes méthodes permettant la comparaison entre les individus étudiés et pour les caractères obtenus, qui aboutissent à une représentation visuelle d'un grand nombre de mesures (cf. Rao, 1957; Sokal & Sneath, 1963).

Tableau 1 A. — Aspect général de la plante.

N°	NOM	ETATS 0	1	2	3	4	5
1	Vigueur	280	type C	267	type II	K 187	
2	Remontaison	aucun pied n'est fleuri	moins de la moitié est fleurie	plus de la moitié est fleurie	tous les pieds sont fleuris		
3	Port du pied	étalé	demi étalé	dressé	rampant		
4	Port 3e feuille	dressée	cassée	retombante			
5	Densité, talles	faible	moyenne	forte			
6	Pilosité, tige	absence	vers le haut	uniforme			
7	Pilosité, nœud	absence	peu	beaucoup			
8	Anneau, rouge	absence	présence				
9	Pilosité, gaine, densité	absence	peu	beaucoup			
10	Pilosité, gaine, aspect	absence	durs et courts	durs et longs	duvets	mous et longs	

Tableau 1B. — Aspect des feuilles.

11	Pilosité, base du limbe	absence	ligule seule	peu		beaucoup	
12	Pilosité, angle gaine limbe	absence	peu	beaucoup			
13	Densité, pilosité limbe	absence	peu vers le haut			très dense	
14	Aspect, pilosité limbe	absence	duvet	mou court	mou long	dur court	dur long
15	Couleur limbe	vert jaune	vert	vert bleu			
16	Largeur limbe	< 1 cm	$1 << 2$ cm	$2 << 3$ cm	$3 << 4$ cm	> 4 cm	
17	Cercospora	absence	peu	beaucoup			
18	Colletotrichum	absence	peu	beaucoup			
19	Rouille	absence	peu	beaucoup			

Tableau 1C. — Aspect des inflorescences.

20	Forme	normale	racème				
21	Longueur ram. du verticille						
22	Position ram. secondaires	basses	incertaines	hautes			
23	Longueur ram. secondaires	courtes	longues				
24	Longueur des pédicelles	courts	longs	très longs			
25	Pilosité du verticille	absence	peu	beaucoup			
26	Pilosité de l'axe	absence	vers le haut sous le verticille	vers le bas sous le verticille	uniforme sous le verticille	au-dessus et en-dessous du vert	
27	Distribution des épillets	dispersés	uniforme	par groupes			
28	Pruine	absence	présence				

Tableau 1D. — Aspect des épillets.

N°	NOM	ETATS					
		0	1	2	3	4	5
29	Forme	tossé	normal	allongé			
30	Grosseur	petit	normal	gros			
31	Couleur	jaune foncé	vert pâle	vert			
32	Taches	absence	brunes	petites rouges	grandes rouges	petites violettes	grandes violettes
33	Pilosité	absence	présence				
34	Soies	absence	rares	sur presque tous	sur tous	plus de deux soies	soies longues
35	1re fleur	stérile	variable	toujours mâle			
36	Sorosporium	absence	peu	beaucoup			
37	Fusarium	absence	peu	beaucoup			
38	Cerebella	absence	peu	beaucoup			

Tableau 2. — Mesure de la ressemblance.

Caractères bivalents		*Caractères multivalents*	*Caractères continus*
SOKAL & MICHENER	$\frac{m}{n}$	ROGERS & TANIMOTO	DISTANCE MOYENNE
		SMIRNOV	DISTANCE EUCLIDIENNE
		GOODALL	DISTANCE GÉNÉRALISÉE DE MAHALANOBIS
JACCARD	$\frac{m_{JK}}{(n_{JK}+u)}$	DISTANCE DU χ_2	DISTANCE DU χ^2
			COEFFICIENTS DE CORRÉLATION
ROGERS & TANIMOTO	$m/(n+u)$		INDICE DE GOODALL

Les données

Choix des caractères

Le choix des caractères est au départ en lui-même un problème. Faut-il prendre tous les caractères qu'on a à sa disposition? Faut-il faire un tri, et là on risque d'orienter les résultats. Pour chaque matériel étudié les problèmes sont différents. Pour *P. maximum*, les caractères ont été recensés, un choix a été fait sur leur facilité d'observation et leur pouvoir discriminant. Les tableaux 1 A—D montrent quatre groupes de caractères: aspect général de la plante; aspect des feuilles; aspect des inflorescences; aspect des épillets. Ceux-ci pour les caractères discontinus. Pour les caractères continus les mesures de poids en matière verte et sèche de la plante, nombre de talles, longueur de gaine, et différentes mesures de l'inflorescence ont été utilisées.

Collecte des données

Les caractères qualitatifs sont observés en collection (sur une parcelle d'une quinzaine de plantes) dans un temps suffisamment limité pour éviter les variations de l'environnement (humidité, lumière, etc.). Les caractères quantitatifs sont mesurés dans des essais statistiques.

Codification des données

Un code simple est choisi pour faciliter les observations. Les tableaux 1 A—D montrent les différents états (0—5) ou codes pour chacun des caractères. Pour le calcul des indices de similarités (indices simples des caractères bivalents et multivalents, indice de Smirnov & Goodall) les données sont traitées directement. Pour certaines distances, pour les coefficients de corrélation, les données sont transformées en fonction des analyses statistiques dans lesquelles ils interviennent.

Le résultat de ces observations est transcrit sous forme de tableau où les individus figurent en colonnes et les caractères en lignes: c'est la "matrice de description" (Individus, *I*, × Caractères, *C*).

La mesure de la ressemblance

La matrice précédemment obtenue doit permettre la mise en correspondance des individus entre eux: c'est la "matrice de similarité" (Individus, *I*, × Individus, *I*).

La ressemblance entre deux individus pourra se calculer par différents moyens selon qu'on a affaire à des caractères bivalents (à deux états 0.1 ou +, — ou absence, présence), multivalents (à plusieurs états 0.1, ... 5, tableaux 1 A—D) ou continus (mesures quantitatives, longueurs, poids, etc.). On trouvera dans Sokal & Sneath (1963) la revue des ces différentes méthodes de calcul.

Pour Sokal & Sneath (op. cit.) la taxonomie numérique est strictement empirique, mais tous les auteurs ne sont pas d'accord sur ce point; Goodall (1966) par exemple raisonne en termes de probabilités, nous verrons plus loin comment il le fait.

Pour Sokal & Sneath (op. cit.) les caractères ont des poids égaux dans la création des classes; pour Smirnov et pour Goodall ou même dans l'analyse factorielle des correspondances, une pondération est indispensable.

Le tableau 2 montre trois groupes de mesures de la ressemblance pour les caractères bivalents, multivalents ou continus. Il est évident que les indices relatifs aux caractères multivalents peuvent être utilisés pour des caractères bivalents (mais non réciproquement). L'indice de Goodall peut être calculé pour toute sorte de caractères.

Pour les caractères bivalents le tableau 3 montre la signification des paramètres utilisés. Par exemple, l'indice le plus simple (Sokal & Michener) est le rapport entre le nombre de coïncidences (les individus sont au même état pour un caractère) et le nombre total de caractères.

Tableau 3. — Caractères bivalents.

Individu K \ *Individu 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
$n = m + u$	

Les indices simples ont le grand mérite de ne nécessiter aucun calcul mathématique autre que des dénombrements. Cependant ils ne reflètent pas toujours la réalité biologique. L'indice de Sokal & Michener tient compte dans la ressemblance des coïncidences positives ou négatives. Les coïncidences négatives n'ont pas toujours de sens (cf. Benzécri, 1973: "une baleine et une souris sont-elles semblables puisqu'elles n'ont ni l'une ni l'autre de plumes").

L'indice de Jaccard élimine complètement les coïncidences négatives du calcul.

Il est difficile de dire qu'un indice est meilleur qu'un autre. Tout dépend du matériel étudié.

Pour les caractères multivalents les indices de Rogers & Tanimoto et de Smirnov, dont le calcul bien que simple soit difficilement explicable par des formules, ont donné des résultats très comparables dans l'étude des populations de *P. maximum* (Réné-Chaume & al., 1969).

L'indice de Goodall est assez particulier. Dans l'étude des populations végétales, les variables décrivant les individus 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 (1966) a construit un indice de similarité basé directement sur la théorie des probabilités. Il définit une relation d'ordre entre couples d'individus pour la ressemblance. Deux couples sont également ressemblant (=) ou l'un est plus ou moins ressemblant que l'autre ($>$ ou $<$).

Soient un couple particulier et un caractère: Goodall définit la probabilité pour un couple pris au hasard d'être supérieur ou égal à ce couple particulier pour la ressemblance. Le complément à 1 de cette probabilité définit un indice de similarité pour ce caractère. En supposant l'indépendance des caractères il combine les probabilités. Goodall a défini aussi un indice de déviation qui ouvre la voie à un processus complet, contrôlé par des niveaux de signification.

L'indice de Goodall est, du point de vue théorique, très satisfaisant. Pratiquement on se heurte à des difficultés de calcul dès que les données sont nombreuses. De plus il n'est pas toujours en accord avec ce que les utilisateurs en attendent à la simple vue des données. L'utilisation de l'indice de Goodall dans le cas de comparaison de profils hydriques a donné des rapprochements sans réalités physiques.

La distance du x^2 sera détaillée dans l'analyse factorielle des correspondances.

Le tableau 4 montre les différentes formules de distance valables pour des variables continues.

Tableau 4. — Distances.

Moyenne	$d_{jk} = \frac{1}{n} \sum_{i=1}^{n} \left\| x_{ij} - x_{ik} \right\|$
Euclidienne	$\Delta_{jk} = \left(\sum_{i=1}^{n} (x_{ij} - x_{ik})^2 \right)^{1/2}$
Mahalanobis	$D^2_{jk} = \sum_{\alpha} \sum_{\beta} (\bar{x}_{\alpha j} - \bar{x}_{\alpha k})(\bar{x}_{\beta j} - \bar{x}_{\beta k}) \lambda^{\alpha\beta}$
x^2	$d^2(jk) = \sum_{l} \frac{1}{P(l)} \left[\frac{P(j,l)}{P(j)} - \frac{P(k,l)}{P(k)} \right]$

La représentation des données

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 ressemblance; l'ombrage le plus foncé correspondant aux valeurs de l'indice de similarité les plus grandes. On intervertit alors lignes et colonnes pour que les carreaux sombres soient le plus proche possible de la diagonale.

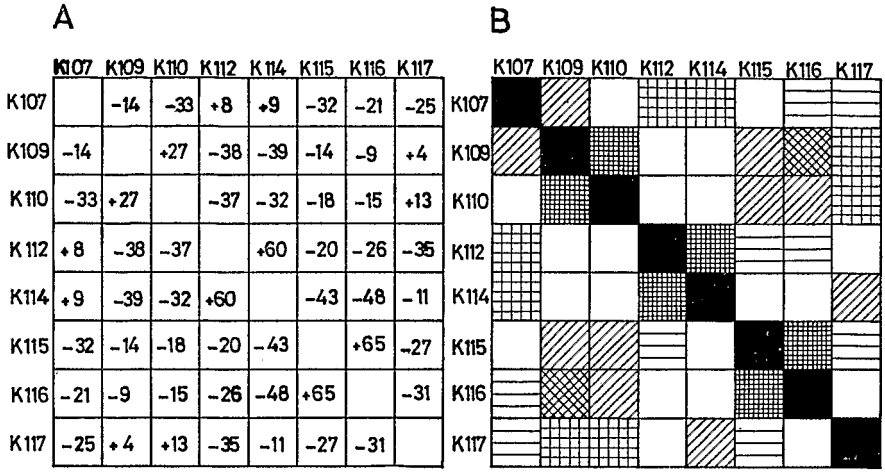
Cette méthode très simple donne de bons résultats tant à partir d'une matrice d'indice de similarité que d'une matrice de corrélation. Dans l'étude de *P. maximum* les ombrages reflétaient nettement la structure des populations. Selon que les populations sont en cours d'évolution ou non, la disposition des carreaux sombres autour de la diagonale est difficile ou non. La figure 1 A montre la matrice de similarité d'une population de *P. maximum* (région d'Amboselli, clones nos K107 à K117). La figure 1C (après permutation) montre trois phénotypes distincts.

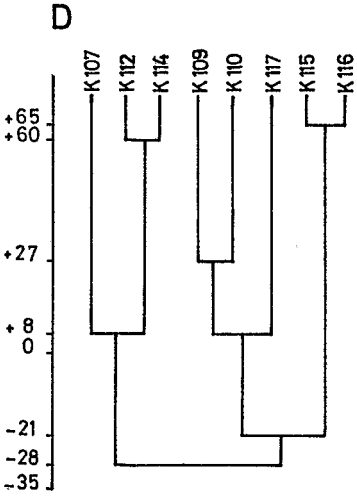
Dendrogrammes

La construction du dendrogramme ne conduit pas effectivement à la formation de 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.

A

	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	





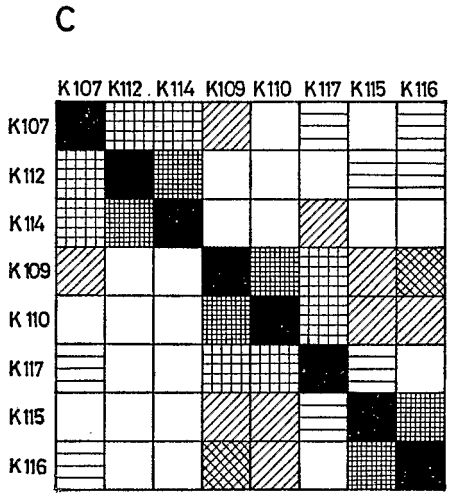


Fig. 1. — *Panicum maximum*, une population de la région d'Amboselli. *A*, matrice des indices de Smirnov x 100. *B-C*, ombrages différentiels, *B* avant permutation, *C* après permutation. *D*, Dendrogramme (cf. texte p. 374).

Une des méthodes consiste en tous les regroupements deux à deux des individus les plus ressemblants. On calcule ensuite une nouvelle matrice plus petite entre les groupes formés et les individus isolés en prenant les moyennes des indices. On recommence les regroupements deux à deux et ainsi de suite jusqu'à ce qu'il ne reste plus que deux groupes en présence.

L'inconvénient dans l'établissement d'un dendrogramme est que, dans le cas d'égalité des ressemblances, plusieurs solutions sont possibles. Ces solutions conduisent à des dendrogrammes légèrement différents.

Le dendrogramme (fig. 1 D) relatif à la région d'Amboselli à partir de la même matrice (fig. 1A) montre des associations analogues à celles des ombrages différentiels (fig. 1B-C).

Méthode nodale de Rogers & Tanimoto

La méthode nodale utilise des distances exprimées à partir des indices de ressemblance. Au lieu de partir comme dans la méthode des dendrogrammes d'éléments les plus ressemblants deux à deux, elle cherche à déterminer l'élément le plus représentatif de l'ensemble, globalement le moins dissemblable des autres et à constituer autour de lui un groupe central composé d'éléments qui lui sont semblables. La même opération sera faite sur tous les éléments restant après élimination de ce groupe.

Le premier groupe central est formé de façon:

— que l'adjonction d'un élément un peu plus éloigné que les autres n'augmente pas trop la distance intra-groupe;

— que la distance de tous les éléments du premier groupe au premier centroïde soit inférieure à la distance entre les deux premiers centroïdes.

La figure 2C montre une représentation par la méthode nodale des *Panicum* de la région d'Amboselli, toujours à partir de la matrice de la figure 1A. Les centroïdes sont dans l'ordre K109, K107 et K115. Les regroupements autour de ces centroïdes sont en accord avec les deux premières représentations.

Cette méthode est complexe et n'aboutit que s'il existe effectivement des groupes.

Arbre de longueur minimale

On suppose n points représentant les individus.

Un arbre raccordant ces points est un ensemble de segments joignant les points 2 à 2 tel que:

— il n'existe pas de boucle;

— chaque point est visité au moins une fois.

La représentation plane de l'arbre est approximative mais fournit une image des rapprochements. La détermination se fait par ordinateur.

La figure 2D montre la disposition des clones de *P. maximum* d'Amboselli par la méthode de l'arbre de longueur minimale. Ce graphique révèle les mêmes associations que les autres méthodes.

Analyse en composantes principales

On se trouve face à un grand nombre de données, par exemple 1000 individus décrits sur 60 caractères.

Si on voulait représenter exactement ces données il faudrait pouvoir construire un espace à 60 dimensions et disposer 1000 points, ce qui est impossible à notre œil.

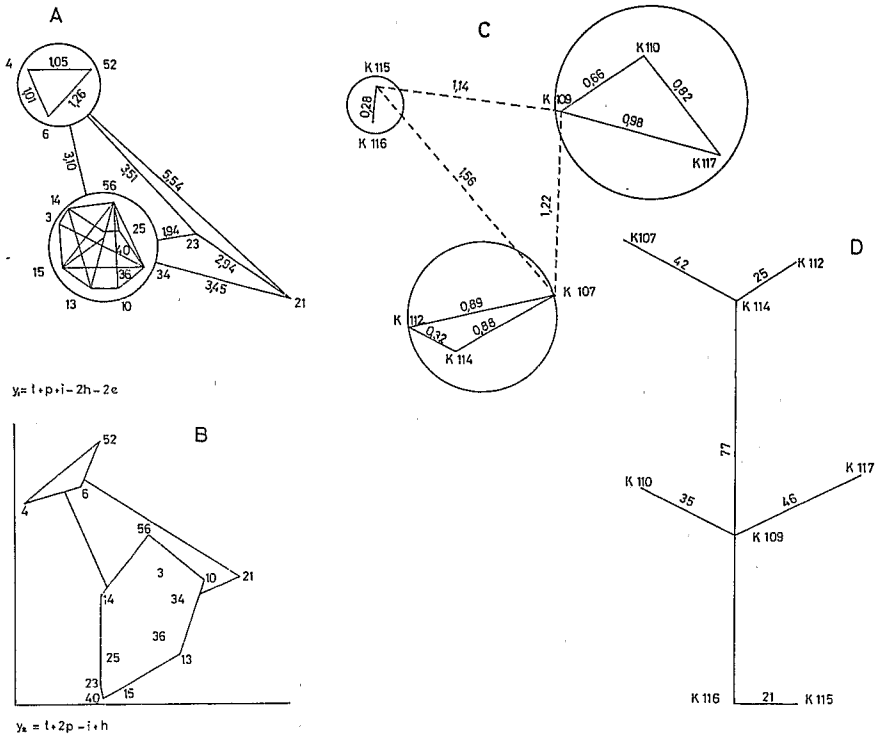


Fig. 2. — *Panicum maximum. A,* Positions relatives des diverses constellations obtenues par les distances D^2 de Mahalanobis. *B,* Représentation graphique des groupes obtenus par l'utilisation des deux composantes principales y_1 et y_2. *C,* Représentation par la méthode nodale de Rogers & Tanimoto des clones de la région d'Amboselli. *D,* La disposition des clones d'Amboselli par la méthode de l'arbre de longueur minimale.

On voudrait donc pouvoir représenter ces points dans un espace de plus faible dimension pour y voir plus clair. On pourrait, par exemple, trouver cinq nouvelles variables qui reconstitueraient approximativement les 60, mais un espace à 5 dimensions c'est encore beaucoup trop pour notre rétine.

Dans ce qui suit on ne fait aucune hypothèse particulière préalable — ceci pour insister sur le côté purement discriptif de l'opération.

Dans notre espace à 60 dimensions, chaque individu est un vecteur à 60 composantes (les mesures des 60 caractères). La ressemblance entre deux individus peut être mesurée par la distance euclidienne.

Soient 2 vecteurs x_1 et x_2: $d^2(x_1, x_2) = \sum_i (x_{1i} - x_{2i})^2$ (i allant de 1 à 60 et représentant les caractères).

On va chercher un sous-espace sur lequel on projettera tous les points en perdant le moins possible d'information (en déformant le moins possible les positions des points les uns par rapport aux autres).

On peut commencer par une droite; on cherchera la droite qui maximise l'ensemble des projections et si on suppose que le nuage de points est approximativement un ellipsoïde, la droite recherchée sera celle sur laquelle les projections seront bien étalées, c'est-à-dire le grand axe de l'ellipsoïde: ce sera la première composante principale. La seconde droite répondant aux conditions perdra un peu plus d'information que la première, etc.

On peut déterminer mathématiquement ces axes (vecteurs propres de la matrice définissant l'ellipsoïde) et leurs parts dans la quantité d'information globale (valeurs propres).

Dans le plan des deux premiers axes, on peut représenter les 1000 points de départ. On n'aura gardé que, par exemple, 75% de l'information globale mais on aura une vue plus claire de la disposition des points les uns par rapport aux autres; les axes suivants sont, selon leur importance, utilisés. On peut avoir ainsi plusieurs représentations planes de ces 1000 points.

La figure 2A montre une représentation graphique dans le plan des deux premières composantes principales. Les points représentent des clones de *Panicum maximum* de Côte-d'Ivoire.

La figure 2B montre les mêmes données traitées par la méthode des constellations (obtenues par les distances D^2 de Mahalanobis). Cette méthode d'un maniement moins simple ne sera pas détaillée. Les constellations formées sont en très bon accord avec l'analyse en composantes principales, mais apportent une information supplémentaire en isolant le clone 23.

Analyse factorielle des correspondances

On part ici encore de la matrice de description individu par caractère. Prenons un exemple à caractères continus (et sans le montrer pour les caractères qualitatifs la méthode reste valable).

Soit $k(i, j)$ le nombre correspondant au couple (i, j) [un tel tableau pourrait être étudié en analyse en composantes principales].

Il semble cependant plus naturel de chercher une méthode qui tienne compte du caractère probabiliste de ce type de données.

On pose $k = \sum k(i, j)$

$$p(i,j) = \frac{k(i,j)}{k} \qquad \left.\begin{array}{l} p(i) = \sum_j \frac{(k(i,j)}{k} \\ p(j) = \sum_i \frac{k(i,j)}{k} \end{array}\right\} \text{ lois marginales}$$

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.

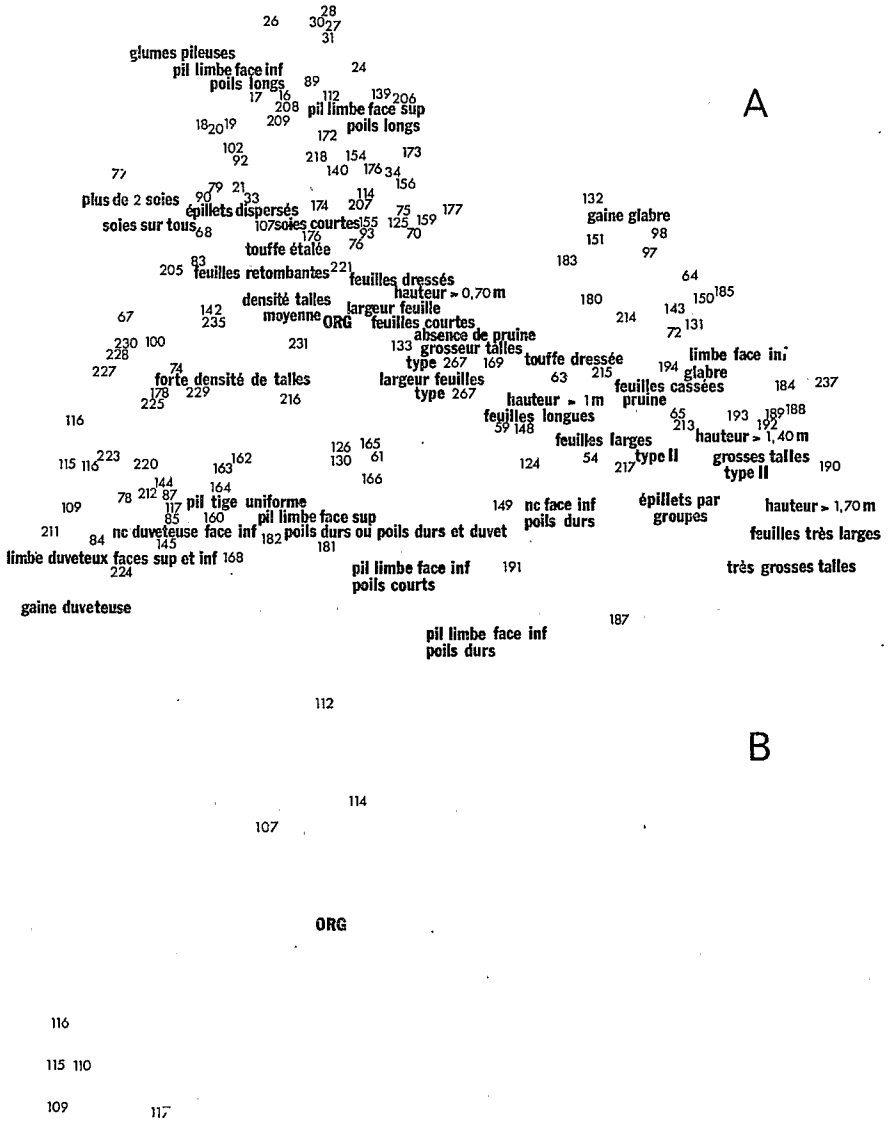


Fig. 3. — *Panicum maximum:* taille et pilosité. *A*, L'ensemble des clones prospectés. *B*, La population d'Amboselli. Voir texte p. 378.

Le rapprochement des individus est fait 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

$$d^2(i, i') = \sum_j \frac{1}{p(j)} \left(\frac{p(i,j)}{p(i)} - \frac{(p(i',j)}{p(i')} \right)^2$$

La pondération par $p(i)$ accentue le caractère original j pour l'individu i: en effet si $p(i, j)$ est petit pour un individu 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.

Comme en analyse en composantes principales on détermine les deux axes, le premier représentant le maximum d'information.

L'analyse peut être faite à la fois dans l'espace des individus et dans l'espace des caractères et on montre qu'il existe une application canonique de l'espace des individus sur l'espace des caractères.

Une représentation simultanée des deux nuages est possible.

La figure 3A représente l'ensemble des clones prospectés par Pernès & Combes en 1967 desquels on a extrait la population d'Amboselli précédemment étudiée. Le premier axe représente la taille des *Panicum*, à gauche les petits *Panicum*, à droite les plus grands. Le second axe différencie les clones sur les caractères de pilosité. Il en résulte trois grandes tendances: à droite les grands *Panicum* à feuilles et talle plutôt glabres, en bas à gauche les petits *Panicum* duveteux, en haut des *Panicum* petits à longs poils.

La figure 3B extraite de la précédente montre la population d'Amboselli. L'accord existe encore avec les autres représentations.

L'avantage de cette méthode est de traiter d'une façon satisfaisante et non empirique les caractères qualitatifs sur un grand nombre d'individus. D'autre part, la représentation simultanée de l'espace des individus et de l'espace des caractères permet une meilleure interprétation des faits biologiques.

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 compréhension convenable des structures des ensembles étudiés. L'utilisation convergente de plusieurs méthodes, leur apport propres et leurs significations biologiques sont illustrées dans Pernès (1975).

Une structure ne peut être tenue pour assurer que si elle est révélée à partir d'algorithmes de classification différents, sinon on risque d'attribuer à un ensemble non organisé un système de classification qui n'est qu'un artefact de la méthode employée. C'est par cette convergence de méthodes que l'utilisateur prudent peut pallier à l'absence de tests difficiles à définir et acquérir une vision synthétique et progressive des structures qu'il révèle.

RÉFÉRENCES BIBLIOGRAPHIQUES

Benzécri, J.-P. (1973) *L'analyse des données.* 2 vols. Dunod, Paris.

Combes, D. (1975) Polymorphisme et modes de reproduction dans la section des Maximae du genre Panicum (Graminées) en Afrique. *Mémoires ORSTOM 75* (Thèse, Univ. Paris-Sud, Centre Sci. Orsay, 1972).

Goodall, D. W. (1966) A new similarity index based on probability. *Biometrics* 22/4: 882-907.

Pernès, J. (1975) Organisation évolutive d'un groupe préférentiellement agamique: la section des Maximae du genre Panicum (Graminées). *Mémoires ORSTOM* 75 (Thèse, Univ. Paris-Sud, Centre Sci. Orsay, 1972).

Rao, R. (1957) *Advanced statistical methods in biometric research.* Willey, New York.

Réné-Chaume, R. (1971) Essai de description des populations de Panicum maximum Jacq. d'Afrique de l'Est par l'analyse factorielle des correspondances sur des caractères morphologiques qualitatifs. *Rapport ORSTOM,* 18 pp. Adiopodoumé.

— J. Pernès & D. Combes (1969) Essai de classification des populations de Panicum maximum Jacq. d'Afrique de l'Est sur des caractères morphologiques qualitatifs. *Rapport ORSTOM,* 24 pp. Adiopodoumé.

Sokal, R. R. & P. H. A. Sneath (1963) *Principals of numerical taxonomy.* Freeman & Co., San Francisco & London.