La classification hiérarchique automatique (clustering)

Axes AFC considérés	Pondération des coordonnées	Coefficient cophénétique
1 à 3	non	0.8419
1 à 7	oui	0.9036
1 à 30	oui	0.9047

Tableau 2. Coefficients cophénétiques des premiers axes de l'AFC
Matériel: 31 relevés du transect Jorette (chap. 7), donc 465 distances inter-relevés.
AFC par le programme CORRES. Calcul du coefficient cophénétique selon chapitre 3.2.2.

projection, comme on le fait couramment en phytosociologie.

Cette conclusion semble devoir être nuancée suivant le volume de données traitées, bien que nous ne l'ayons pas testé par le coefficient cophénétique. Lorsque l'on augmente le nombre de relevés soumis à l'analyse, la quantité d'information perçue par les premiers axes diminue. Alors que pour une trentaine de relevés les premiers axes reflètent une forte proportion de l'inertie (42% pour les 4 premiers axes de l'AFC Jorette, voir exemple sous 3.2.1), pour plus de cent relevés, cette proportion devient très réduite (par exemple 12,5 % de l'inertie pour les trois premiers axes de l'AFC, fig. 18 avec 131 unités).

Sa grande fidélité aux ressemblances floristiques, et le fait qu'elle ne propose pas de classification, font de l'AFC l'outil numérique primordial d'une syntaxonomie objective, devant les techniques de clustering.

4. La classification hiérarchique automatique (clustering)

4.1. *Introduction: les niveaux de choix*

Le détail des techniques est exposé par Sneath et Sokal (1973), Hartigan (1975), Lebart *et al.* (1982), Jambu (1978), Benzécri *et al.* (1980). Pratiquement, ce sont les programmes CLUSTAN (Wishart 1975) et CLTR (Wildi et Orloci 1983) qui ont été utilisés.

Remarques préliminaires:

–Un dendrogramme peut toujours être calculé, quelle que soit la structure taxonomique des données fournies: des classes sont mises en évidence de toute façon (Groenewoud 1983), même si par exemple une ordination selon un gradient était un bien meilleur reflet de la réalité.

–L'ordre des individus montré par un dendrogramme n'a aucun sens, car une rotation est possible autour de chaque liaison verticale (Jambu 1978, p. 71).

Plus que l'AFC, le clustering implique de nombreux choix à trois niveaux:

–préparation des données: normées ou non, soumises ou non au prétraitement par AFC (4.2);

–choix d'une mesure de ressemblance entre relevés (4.3);

–choix d'un algorithme de classification (4.4).

4.2. *Préparation des données*

L'idéal est de prendre les données brutes (matrice des relevés floristiques) comme base de la classification automatique; cependant, le grand nombre de variables dépasse souvent la capacité des ordinateurs disponibles: c'est le cas ici avec 600 espèces environ pour 273 relevés (tabl. 11), le double pour 131 groupements publiés (fig. 18). Par conséquent, la réduction des données s'impose. Deux possibilités existent:

–la troncature des listes d'espèces;

–la réduction par une AFC préliminaire: le relevé est repéré non plus par ses coordonnées sur tous les axes-espèces, mais par ses coordonnées sur les premiers axes factoriels (LACOSTE 1972, JAMBU 1978, p. 103, 122-3).

Nous avons adopté la seconde solution pour conserver les listes floristiques complètes, espèces accidentelles comprises. Nous n'avons pas testé la première possibilité, mais ce serait utile.

4.2.1. *Une technique de réduction des données par AFC*

4.2.1.1. *Choix du nombre d'axes factoriels*

De nombreux essais sur divers matériels (35 dendrogrammes non publiés, voir néanmoins tabl. 3) nous permettent de livrer les appréciations préliminaires suivantes (un test statistique suit sous 4.2.2.1):

–dans les essais effectués avec 2, 3 ou 4 axes factoriels, les groupements obtenus par clustering correspondent très bien aux images AFC sur ces premiers axes, mais la perte d'informations, donc la schématisation du résultat, est très grande;

–dans les essais effectués avec de nombreux axes (30 à 50), les groupes perdent toute cohérence avec les premières images factorielles et deviennent difficiles à comprendre. Nous expliquons et condamnons ce résultat ainsi: il n'est pas légitime de donner autant d'importance au 40ème axe factoriel (qui dans le contexte de l'analyse représente très peu d'informations) qu'au premier;

–dans les essais effectués avec les 5 à 7 premiers axes, la taxonomie intuitive des relevés originaux se trouve confirmée, de même que la synsystématique actuelle des relevés publiés; et en même temps la perte d'information est assez faible.

4.2.1.2. *Pondération des coordonnées*

L'utilisation des coordonnées sur un nombre restreint d'axes factoriels peut être soumise à la pondération des coordonnées proposée plus haut (3.2.1). Le nombre d'axes utilisés dans cette procédure variera suivant l'analyse, de manière à percevoir au moins la moitié (jusqu'à 75%) de l'inertie totale de l'AFC.

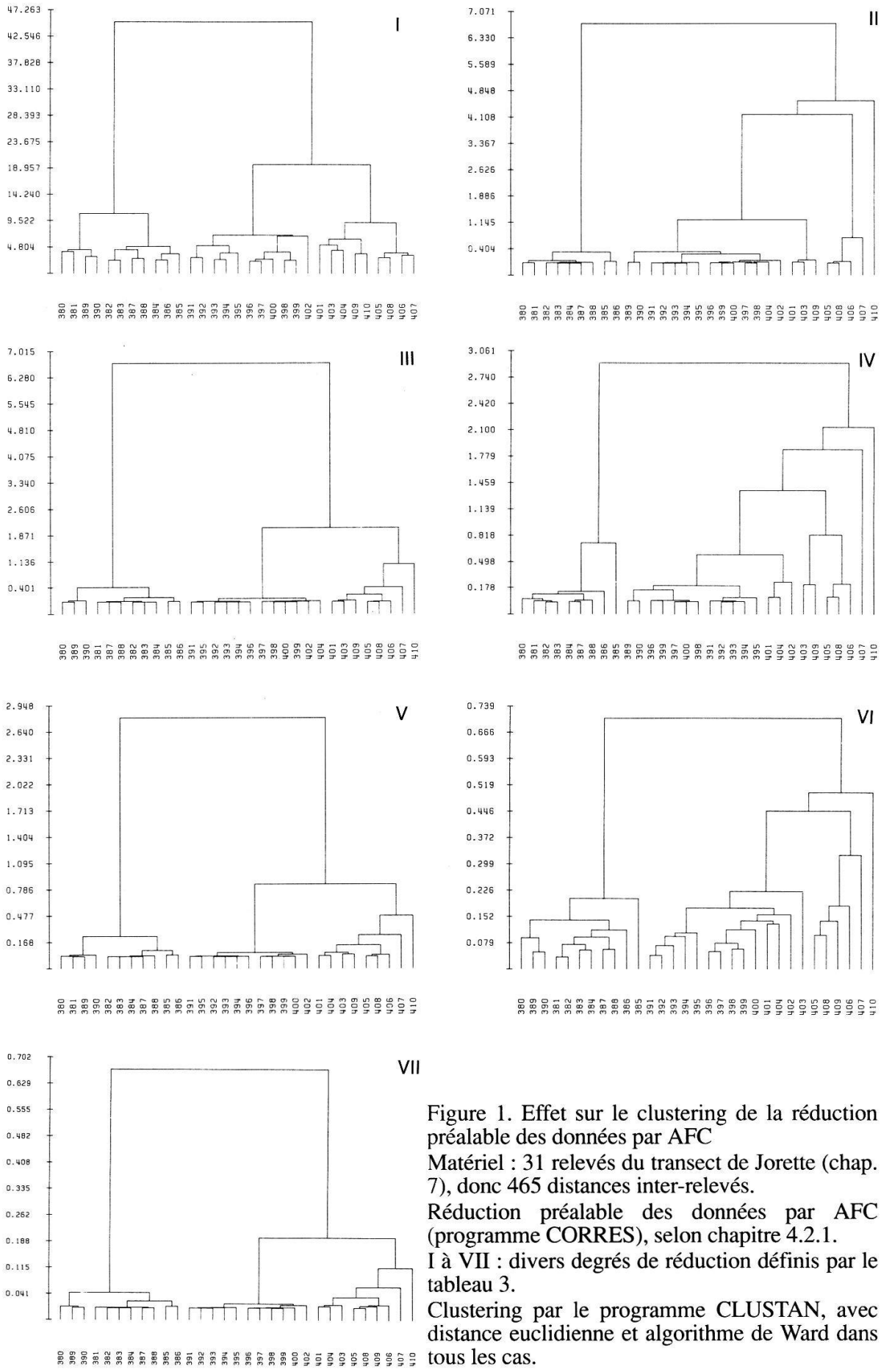


Figure 1. Effet sur le clustering de la réduction préalable des données par AFC

Matériel : 31 relevés du transect de Jorette (chap. 7), donc 465 distances inter-relevés.

Réduction préalable des données par AFC (programme CORRES), selon chapitre 4.2.1.

I à VII : divers degrés de réduction définis par le tableau 3.

Clustering par le programme CLUSTAN, avec distance euclidienne et algorithme de Ward dans tous les cas.

4.2.2. *Evaluation de la technique de réduction des données*

4.2.2.1. *Test statistique*

Ici encore nous utilisons le transect de Jorette pour comparer les résultats obtenus avec 3,7 et 30 axes, pondérés ou non. Ce matériel est présenté au chapitre 7 et les dendrogrammes à la figure 1. Le test utilise le coefficient cophénétique (voir 3.2.2) appliqué à la matrice des distances euclidiennes inter-relevés, et à la matrice des distances inter-relevés lues sur le dendrogramme (tabl. 3).

On s'attendrait à ce que le dendrogramme I, calculé à partir des relevés bruts, soit le plus fidèle aux ressemblances observables; or, ce n'est pas le cas dans ce test: plusieurs dendrogrammes établis sur des données réduites par une AFC préliminaire ont un coefficient cophénétique même supérieur (II, IV, V, VII)! De plus, la variation anarchique du coefficient cophénétique par rapport au nombre d'axes de l'AFC préliminaire et à l'option de pondération des coordonnées soulève la question de la validité de ces résultats: des dendrogrammes de 31 éléments comptent peut-être trop peu de noeuds, donc trop peu d'indices de noeuds, pour permettre une variation nuancée du coefficient cophénétique. Ce test apparaît dans notre cas comme un échec. Pour des raisons matérielles, il n'a pas pu être répété sur un plus grand échantillon. Nous en sommes donc réduits à une évaluation qualitative de ces dendrogrammes.

4.2.2.2. *Evaluation qualitative*

Dendrogramme I. Cette classification est théoriquement la meilleure puisqu'elle évite la perte d'information provoquée par l'AFC préliminaire. Elle servira de base pour la comparaison. Ce sont les groupes de relevés qu'elle propose qui sont le plus clairement séparés par des espèces différentielles. Une structure en trois groupes apparaît:

–A (380-390) les pâturages à sol profond les moins escarpés
–B (391-400) les pâturages plus rocailleux et plus raides
–C (401-410) les pelouses proches de la crête, moins pâturées

Dendrogramme II. 3 axes AFC non pondérés donnent une classification encore proche de l'image factorielle (1,2): les groupes A, B et C de la classification de référence (I) sont imparfaitement retrouvés (différences d'effectifs et de structure), et l'excentricité artificielle du relevé 410 (voir 5.1.1.) est très fortement marquée (isolé dès la 2ème ramification de l'arbre).

Dendrogramme III. L'effet de la pondération se fait fortement sentir sur les 3 premiers axes factoriels: le premier groupe (A) est retrouvé identique, les deux autres (B, C) sont semblables. Quant au relevé 410, quoique trop tôt isolé, il est classé dans le bon groupe (C). Remarquons qu'en raison du faible effectif de l'analyse (31 objets à classer), les trois premiers axes reflètent déjà 37% de l'inertie.

Dendrogramme IV. Avec 7 axes non pondérés, la classification est voisine de celle du dendrogramme II: structuration faible, regroupement différent de celui du dendrogramme de référence (I), effet d'enchaînement (chaining selon SNEATH et SOKAL 1973, p. 223) des relevés 410 et 407 dès les premières

Dendrogramme	Données	Nombre d'axes de l'AFC préalable	Pondération des coordonnées	Coefficient cophénétique
I	brutes	-	-	.7786
II	réduites	3	-	.7968
III	réduites	3	+	.7024
IV	réduites	7	-	.8017
V	réduites	7	+	.7803
VI	réduites	30	-	.7782
VII	réduites	30	+	.7807

Tableau 3. Coefficients cophénétiques de dendrogrammes obtenus après divers degrés de réduction des données
Matériel et dendrogrammes de la figure 1.
Calcul du coefficient cophénétique selon chapitre 3.2.2.

ramifications de l'arbre.

Dendrogramme V. La pondération sur 7 axes produit un arbre semblable au III (3 axes pondérés), mais meilleur car cette fois les 3 groupes (A, B, C) sont retrouvés identiques à ceux de l'arbre de référence (I). Le relevé 410 reste cependant un peu trop isolé.

Dendrogramme VI. Avec 30 axes, on a théoriquement toute l'information extraite de l'AFC. Sans pondération, l'effet d'enchaînement se fait sentir et, bien que le premier groupe (A) soit conforme à la référence (I), les deux derniers groupes (B, C) sont non conformes et peu structurés.

Dendrogramme VII. La classification partant de 30 axes pondérés est identique à la cinquième (V): on n'a rien gagné en passant de 56% à 100% d'inertie extraite. L'effet d'enchaînement qui induit l'isolement trop précoce du no 410 se manifeste comme dans le dendrogramme V.

4.2.2.3. Conclusion

Cette comparaison qualitative (4.2.2.2) autorise les conclusions suivantes:

–il est plus avantageux de prendre en compte un nombre d'axes factoriels moyen (par exemple 7), que de se limiter aux 3 premiers habituels;

–au-delà d'un certain nombre d'axes, la classification obtenue n'est plus améliorée. Dans le cas envisagé, l'optimum est réalisé avec 7 axes, qui portent 56% de l'information (dendrogramme V). Peut-être est-ce la proportion d'information portée par les axes qui est décisive, et non le nombre d'axes pris en considération;

–la pondération des coordonnées (4.2.1.2) améliore nettement le dendrogramme obtenu (par exemple dendrogramme III par rapport au dendrogramme II).

D'une manière générale, la réduction des données par l'AFC préliminaire avec pondération des données déforme le dendrogramme, même avec une technique optimale; mais la différence avec le dendrogramme établi sur les données brutes est faible. Cette méthode peut donc être recommandée pour traiter de grandes quantités de relevés.

4.3. *Choix d'un indice de similarité*

4.3.1. *Un exemple concret (figure 2)*

Un test utilisant 126 tableaux de relevés publiés, concernant des pelouses sèches montagnardes, a permis de comparer trois fonctions de ressemblance classiques:

–la distance euclidienne (SNEATH et SOKAL 1973, p. 124)

–le coefficient de VAN DER MAAREL (1969, p. 25)

–le coefficient de corrélation (WILDI et ORLOCI 1983, p. 31, opt. 3)

Pour ce test, les tableaux sont résumés par les constances de leurs espèces, codées selon 2.4.

A partir de ces trois indices de similarité, le même algorithme fournit trois dendrogrammes différents. Il n'existe pas de test permettant d'affirmer statistiquement quel est le meilleur de ces trois dendrogrammes. Nous sommes donc réduits à les comparer intuitivement. Pour ce faire, nous les simplifions au maximum en condensant en noeuds les groupes d'unités qui se retrouvent dans chacun des trois. La figure 2 présente les dendrogrammes condensés.

Une première constatation s'impose: la majorité des noeuds (A à U) se retrouvent dans les trois dendrogrammes. Si nous prenons pour référence les noeuds obtenus par le coefficient de VAN DER MAAREL, nous constatons qu'ils ne subissent que peu d'éclatements dans les deux autres dendrogrammes: éclatements minimes dans le cas de la distance euclidienne (noeuds N, P, Q), un peu plus importants avec le coefficient de corrélation (noeuds F, J, N, O, P, R, T). Cependant, dans ce dernier cas, plusieurs noeuds sont décomposés en parties (n él.) encore proches (noeuds F, J, O, T). En définitive, seuls les noeuds N et P (comprenant 10 unités sur les 126 au total) sont gravement inconstants et dépendent étroitement du choix de l'indice de similarité.

Une seconde constatation est que les divergences entre les classifications augmentent dans les niveaux supérieurs. Seul un groupe de trois noeuds (G, H, I) se retrouve dans les trois dendrogrammes; les 6 autres groupes (cadres traitillés) de 2 à 4 petits noeuds sont communs à deux dendrogrammes seulement. Le choix d'un indice de similarité a donc peu d'incidence aux niveaux taxonomiques inférieurs, mais il influence passablement les niveaux supérieurs de la classification.

4.3.2. *Discussion*

Un argument épistémologique nous fait abandonner le coefficient de corrélation: un coefficient de corrélation mesure l'interdépendance de deux variables, deux colonnes d'une matrice; dans cette matrice, chaque ligne est une paire de mesures des deux mêmes variables. Or, un tableau de deux relevés n'a rien à voir avec cette situation: les lignes ne sont pas des observations répétées de la même paire de variables, mais représentent des espèces indépendantes! Le coefficient de corrélation est donc a priori inadéquat pour comparer deux relevés. (Par contre, il prend tout son sens pour comparer la distribution de deux espèces dans un grand nombre de relevés).

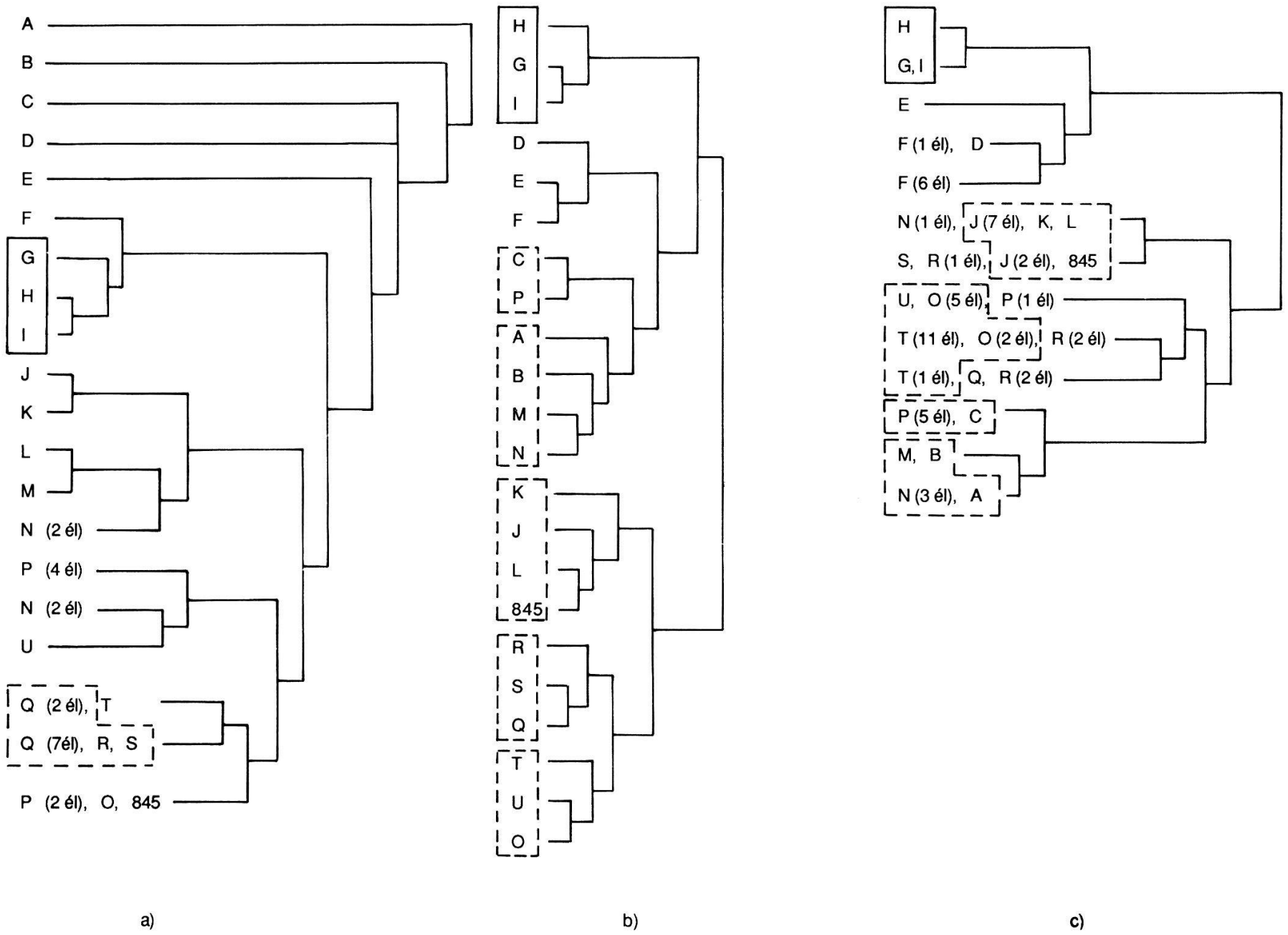
Trois arguments pratiques nous font adopter la distance euclidienne: c'est l'indice le plus compréhensible intuitivement, l'un des plus utilisés, et le seul compatible avec tous les algorithmes de clustering (WISHART 1975, p. 106). Mais cet indice suggestif a le défaut de dépendre du nombre d'espèces présentes dans les relevés, et de l'échelle de codification de l'abondance. Pour pallier à cet inconvénient, GOODALL (in WHITTAKER 1973, chap. 6.3.3.3), ORLOCI (ibidem, chap. 10.7.1.2) et WHITTAKER et GAUCH (ibidem, chap. 11.4.4) proposent de le diviser par la longueur du vecteur-relevé pour obtenir une "distance relative".

En conclusion, la profusion des indices de similarité disponibles (SNEATH et SOKAL 1973, chap. 4; HARTIGAN 1975, chap. 2; WISHART 1975, chap. 25; JAMBU 1978, chap. IV; WILDI et ORLOCI 1983, chap. 4.2) crée l'embarras du choix. Quelques discussions approfondies de leurs valeurs respectives

---

Figure 2. Condensé des dendrogrammes produits par 3 indices de similarité
Matériel : les 126 éléments sont des groupements végétaux résumés par leur colonne de constances, numérotés entre 801 et 950, selon la figure 18.
Programme CLTR: algorithme minimum variance (WILDI et ORLOCI 1983, p. 36, opt. 3)
Trois indices de similarité, selon WILDI et ORLOCI (1983, programme RESE, p. 31):
a) Distance euclidienne (option 4)
b) Coefficient de VAN DER MAAREL (option 7)
c) Coefficient de corrélation (option 3)
P : noeud d'éléments (JAMBU1978, p. 145) complet, se retrouvant dans au moins deux des dendrogrammes.
P (2 él) : une partie des éléments du noeud P (2 éléments)
Cadre plein: un groupe de noeuds qui se retrouve dans les 3 dendrogrammes
Cadre traitillé: un groupe de noeuds qui se retrouve dans 2 des 3 dendrogrammes

Noeuds :	éléments : nombre :	numéros :
A	4	802-805
B	4	806-809
C	3	880-882
D	7	815-819, 821, 837
E	3	822-824
F	7	820, 832-836, 917
G	7	815-868, 870-872
H	7	825-831
I	5	814, 840, 869, 873, 916
J	9	839, 902, 905, 907-908, 920, 924, 927-928
K	12	838, 849, 901, 906, 909-910, 912-914, 918, 929, 932
L	3	842, 847-848
M	3	810-811, 933
N	4	801, 812-813, 841
O	7	851, 859-860, 864, 891-892, 939
P	6	846, 877-879, 883, 890
Q	9	850, 852, 862, 921-922, 925-926, 930, 936
R	5	843-844, 855, 903, 923
S	4	904, 911, 935, 937
T	12	853-854, 856-858, 861, 893, 915, 919, 931, 934, 938
U	4	863, 874-876
(845)	1	845 toujours isolé


A
B
C
D
E
F
G
H
I
J
K
L
M
N (2 él)
P (4 él)
N (2 él)
U
Q (2 él), T
Q (7él), R, S
P (2 él), O, 845
H
G
I
D
E
F
C
P
A
B
M
N
K
J
L
845
R
S
Q
T
U
O
H
G, I
E
F (1 él), D
F (6 él)
N (1 él), J (7 él), K, L
S, R (1 él), J (2 él), 845
U, O (5 él), P (1 él)
T (11 él), O (2 él), R (2 él)
T (1 él), Q, R (2 él)
P (5 él), C
M, B
N (3 él), A

a) b) c)

(SNEATH et SOKAL 1973, chap. 4; LERMAN 1981, chap. 2) concluent qu'aucun indice ne s'impose de façon péremptoire et que le choix dépend du sens que l'on donne à la ressemblance dans le domaine étudié: «*even more important than recommendations on specific coefficients is the general recommendation that numerical taxonomists should consider carefully what it is they wish to measure*» (SNEATH et SOKAL 1973, p. 178).

4.4. Choix d'un algorithme

Il existe plusieurs algorithmes de clustering, qui produisent des classifications différentes. A notre connaissance, aucune étude méthodologique n'a tenté de justifier le choix de l'un d'entre eux.

Nous allons évaluer les cinq algorithmes suivants:

–*average linkage clustering* (SNEATH et SOKAL 1973, p. 228; WISHART 1975, chap. 10, option 3) = agrégation suivant la distance moyenne (BENZÉCRI *et al.* 1980, p. 180);

–*complete linkage clustering* (SNEATH et SOKAL 1973, p. 222; WISHART 1975, chap. 10, option 2) = agrégation suivant le saut maximum (BENZÉCRI et al. 1980, p. 191);

–*Ward's hierarchical grouping method* (SNEATH et SOKAL 1973, p. 241 et 283; WISHART 1975, chap. 10, option 6);

–*centroïd sorting* (WISHART 1975, chap. 10, option 4) = *unweighted pair-group centroid method* (SNEATH et SOKAL 1973, p. 234-235);

–*median* (*Gowen's method*) selon WISHART (1975, chap. 10, option 5) = *weighted pair-group centroid method* (SNEATH et SOKAL 1973, p. 235).

Nous utiliserons deux méthodes d'évaluation:

-un test statistique par le coefficient cophénétique sur deux échantillons de structure taxonomique différente (4.4.1. et 4.4.2);

-un test empirique fondé sur les décomptes d'espèces différentielles (4.4.3).

4.4.1. *Test statistique sur un domaine phytosociologique relativement homogène*

Le domaine phytosociologique choisi pour représenter une faible gradation floristique est l'ensemble des 31 relevés du transect de Jorette (voir chapitre 7). Cela correspond à 465 distances inter-relevés. Quatre des dendrogrammes sont présentés à la figure 3, un à la figure 1.

Le test du coefficient cophénétique (3.2.2) est appliqué (tabl. 4).

Dans cet essai, tous les algorithmes de clustering donnent des résultats fortement corrélés avec la matrice des distances brutes, à part l'algorithme «median». Les algorithmes «complete linkage» et «average linkage» donnent les meilleurs résultats.

4.4.2. *Test statistique sur un domaine phytosociologique relativement hétérogène*

Ce domaine phytosociologique hétérogène est constitué de 31 relevés de groupements végétaux publiés appartenant au *Caricion ferrugineae* ; relevés 1

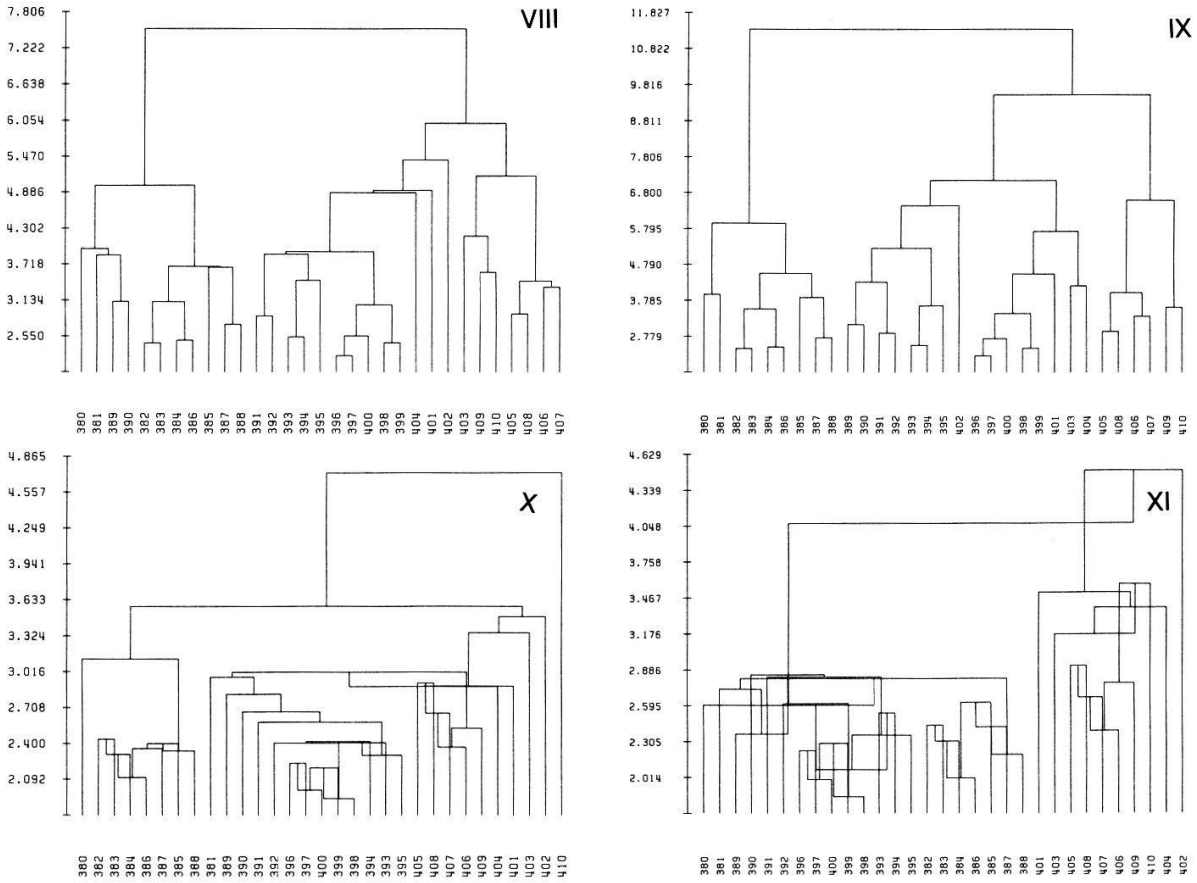


Figure 3. Effet du choix de l'algorithme de clustering
Matériel : 31 relevés du transect de Jorette (chap. 7).
Programme CLUSTAN, appliqué aux données brutes. Distance euclidienne.
VIII à XI : Algorithmes selon tableau 5 (voir WISHART 1975, chap. 10).

Dendrogramme	Algorithme	Option CLUSTAN	Coefficient cophénétique
IX	**complete**	**2**	**.7972**
VIII	**average**	**3**	**.7943**
I	**Ward's**	**6**	**.7786**
X	**centroïd**	**4**	**.7200**
XI	**median**	**5**	**.5604**

Tableau 4. Coefficients cophénétiques de divers algorithmes de clustering
Matériel et dendrogrammes de la figure 3, sauf I (*Ward's*): voir figure 1.
Calcul du coefficient cophénétique selon chapitre 3.2.2.

Dendrogramme	Algorithme	Coefficient cophénétique
XII	**average**	**0.8405**
XIV	**complete**	**0.8389**
XIII	**Ward's**	**0.8374**

Tableau 5. Coefficients cophénétiques de divers algorithmes de clustering
Matériel : 31 relevés pris dans des tableaux de différentes associations publiées, selon chapitre 4.4.2.
Clustering par le programme CLUSTAN (les dendrogrammes XII à XIV ne sont pas publiés), appliqué aux données brutes, avec la distance euclidienne.
Calcul du coefficient cophénétique selon chapitre 3.2.2.

à 10: *Serratulo-Semperviretum* (BERSET 1969), 11 à 26: *Pulsatillo-Anemonetum* (BÉGUIN 1972), 27 à 30 *Senecio-Semperviretum* (RICHARD 1977), 31 *Peucedano-Laserpitietum* (RICHARD 1977).

Les coefficients cophénétiques sont calculés pour 3 algorithmes (tabl. 5).

Le classement de performance des algorithmes est voisin de celui obtenu avec les 31 relevés du transect Jorette (voir tabl. 5). En outre, le nombre identique de relevés utilisés dans ces deux tests légitime la comparaison de ces coefficients. Les valeurs sont systématiquement supérieures avec le matériel des 31 relevés publiés; il est vraisemblable que des données plus différenciées induisent une moindre déformation de la réalité par le dendrogramme.

4.4.3. *Test empirique par les espèces différentielles*

C'est l'ensemble des 273 relevés originaux (chapitre 6.2) qui a servi à ce test. Trois algorithmes performants, *average linkage*, *complete linkage* et *Ward's* y sont appliqués, après une réduction des données par une AFC préliminaire (77 axes avec pondération des coordonnées). Les trois dendrogrammes produits sont condensés dans la figure 9. Pour notre test, nous ne considérons que les principaux noeuds de relevés, qui sont représentatifs de la variation de l'ensemble des pelouses à *Laserpitium siler* de la dition: les noeuds **1** à **5**, qui sont les plus grands noeuds communs aux 3 dendrogrammes, et le groupe **6-7**, qui est petit mais représente un extrême écologique. Selon l'algorithme utilisé, ces noeuds principaux sont regroupés en deux structures générales différentes, schématisées par les figures 4A et 4B respectivement.

Dans les tableaux de végétation ordonnés d'après cette disposition des noeuds (a ou b), on dénombre les espèces différentielles que chaque coupure met en évidence. Cette opération conduit au bilan suivant (tabl. 6): average (a) met en évidence plus d'espèces différentielles que les deux autres (503 contre 455). Cet algorithme a donc le meilleur pouvoir discriminant. Ce résultat corrobore ceux des tests par le coefficient cophénétique (tabl. 4 et 5).

	a)			b)		
	g	d	total	g	d	total
1ère coupure	42	97	138	46	58	104
2e coupure	36	41	77	34	69	103
3e coupure	60	44	104	41	47	88
4e coupure	30	74	104	37	41	78
5e coupure	34	46	80	63	19	82
			----			----
Total :			503 espèces			455 espèces

Tableau 6. Nombre d'espèces différentielles correspondant aux dendrogrammes produits par différents algorithmes
Matériel et dendrogrammes de la figure 4.
g, d : espèces différentielles à gauche, respectivement à droite de la coupure
Pour faciliter le traitement numérique d'un grand nombre de données, l'espèce est prise en compte comme différentielle lorsqu'elle a plus de 10% de constance d'un côté de la coupure et moins de 10% de l'autre.

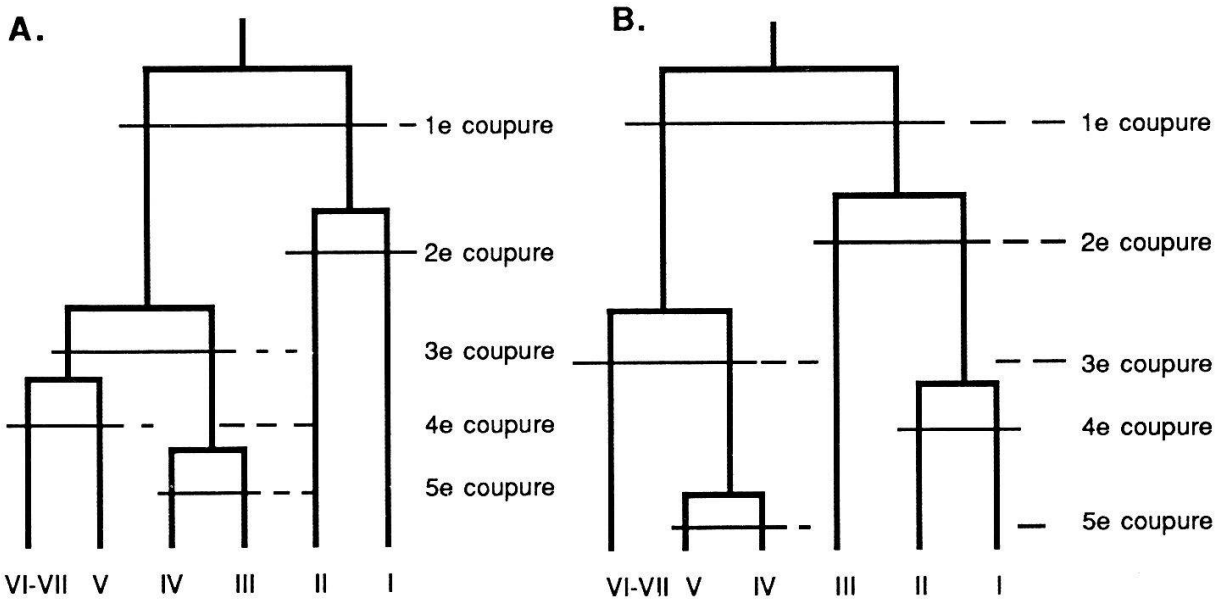


Figure 4. Structure générale des dendrogrammes produits par trois algorithmes
Matériel et dendrogrammes de la figure 9 (chap. 6.4.1.)
A: les liaisons entre les noeuds **1** à **7** de la figure 9 (chap. 6.4.1.) dans le dendrogramme «*average linkage*» figure 9c.
B: liaisons entre les noeuds **1** à **7** dans les dendrogrammes «*Ward's*» et «*complete linkage*» (fig. 9a et 9b), qui ne diffèrent que dans le détail.
1 à **7**: principaux noeuds de relevés communs aux trois dendrogrammes, identiques dans les deux cas.
Les coupures ne sont pas homologues pour A et B.

4.4.4. *Conclusions et commentaires sur les algorithmes*

Sur la base de ces essais, nous admettrons que l'algorithme *average* est le plus fidèle aux ressemblances observables entre les relevés. BENZÉCRI *et al.* (1980, p. 193) le considèrent comme l'un des meilleurs; ces auteurs semblent toutefois lui préférer la «maximisation de la variance interclasse» (p. 193), qui n'a pas été testée ici.

En outre, la pratique du clustering nous inspire les commentaires suivants:

–Si l'algorithme *Ward's* est légèrement moins fidèle aux ressemblances observées que average, il produit par contre des dendrogrammes plus structurés en groupes clairs, donc plus lisibles et plus faciles à condenser que average; il induit très peu d'enchaînement entre les unités classées.

–L'algorithme *complete linkage* présente souvent les mêmes avantages que *Ward's*.

–Par conséquent, nous disposons de 3 algorithmes au moins entre lesquels il n'est pas évident de choisir: *average*, *Ward's* et *complete linkage*.

–Les deux autres algorithmes testés, *centroïd* et *median*, sont moins

satisfaisants. Ils produisent des inversions de noeuds qui rendent difficile la lecture des dendrogrammes: un noeud est relié à un autre par une distance inférieure à celle qui a formé le premier noeud (voir fig. 3 X et XI). En outre, ils ont donné les moins bonnes réponses aux tests statistiques.

4.5. *Technique d'extraction de «noyaux stables»*

Les différences de qualité des divers dendrogrammes n'imposent pas clairement le choix d'une technique, et ceci à aucun des trois niveaux de choix (4.1). La meilleure manière d'obtenir une classification fiable est de confronter les résultats des meilleures techniques.

Nous proposons la procédure suivante, illustrée par la figure 5:

1) appliquer les meilleurs algorithmes de clustering (en particulier *average linkage*, *Ward's* et *complete linkage*) au même matériel;

2) chercher les noeuds homologues, c'est-à-dire les noeuds qui ont en commun une majorité d'éléments dans tous les dendrogrammes obtenus (fig. 5/1);

3) l'intersection de ces noeuds homologues constitue un noyau stable, un groupe d'unités reconnu par les diverses techniques de clustering utilisées (fig. 5/2).

Pour un exemple concret: les dendrogrammes bruts des figures 1 et 3 (4.2.2.1., 4.4.1) ont donné les dendrogrammes condensés de la figure 13 (7.3).

Cette procédure pourrait être automatisée. Elle résoud d'autre part deux problèmes méthodologiques:

–l'exigence d'un consensus entre diverses techniques élimine la question du choix de la «meilleure technique», toujours discutable;

–la confrontation de divers dendrogrammes impose une limite supérieure à l'envergure des noyaux stables, et partant, fournit une partie de la «stopping rule» nécessaire pour découper un dendrogramme en classes (GOODALL, in WHITTAKER 1973, chap. 19.5): dans l'exemple de la figure 9 (chap. 6.4.1), les noyaux stables 1 à 5 sont les plus grands que l'on puisse trouver: ils ne se réunissent pas en noeuds plus grands que l'on puisse retrouver dans les 3 dendrogrammes.

Par contre, la confrontation de divers dendrogrammes ne semble pas imposer de limite inférieure à la taille des noyaux stables: en analysant le noyau stable 5 (fig. 9) dans les dendrogrammes originaux, on constatera qu'il comporte 4 noyaux stables plus petits (fig. 6).

Figure 5. Extraction des noyaux stables et condensations des dendrogrammes

5/1 deux dendrogrammes (a et b) imaginaires indicés, portant sur 8 éléments (1 à 8). A à D: noeuds homologues dans les deux dendrogrammes.

5/2 Exemple d'intersection de deux noeuds homologues: Ca et Cb pour dégager le noyau stable **3**.

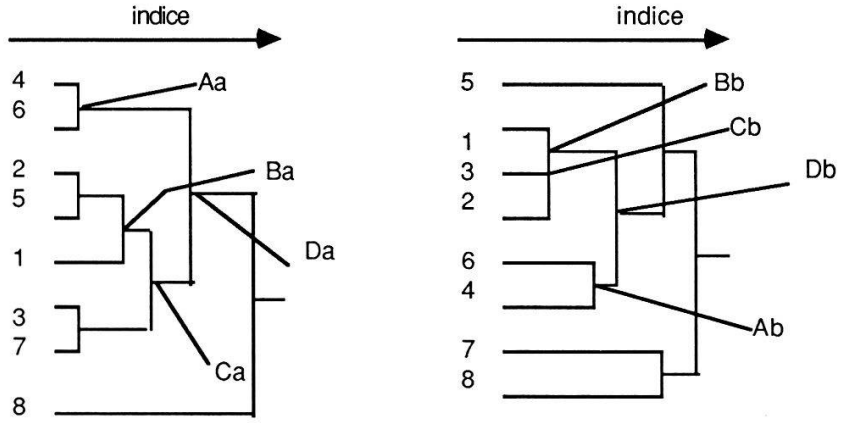
5/3 **1** à **4** : noyaux stables des noeuds homologues A à D.

5/4 Condensation minimale des deux dendrogramme.

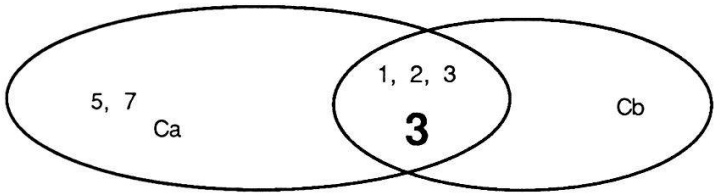
2 \
5 / signifie que les éléments du noyau **2** et l'élément 5 sont mélangés, c'est-à-dire que le noyau stable **2** ne forme pas un noeud à lui seul.

5/5 Condensation maximale des deux dendrogrammes. Dans les dendrogrammes condensés, les indices de similarité des noeuds sont négligés.

5/1



5/2



5/3

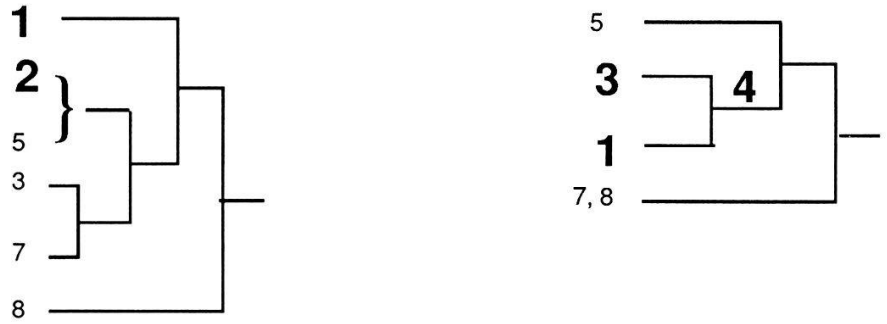
1 = 4 6

2 = 1 2

3 = 1 2 3 (= **2** et 3)

4 = 1 2 3 4 6 (= **1** et **3**)

5/4



5/5

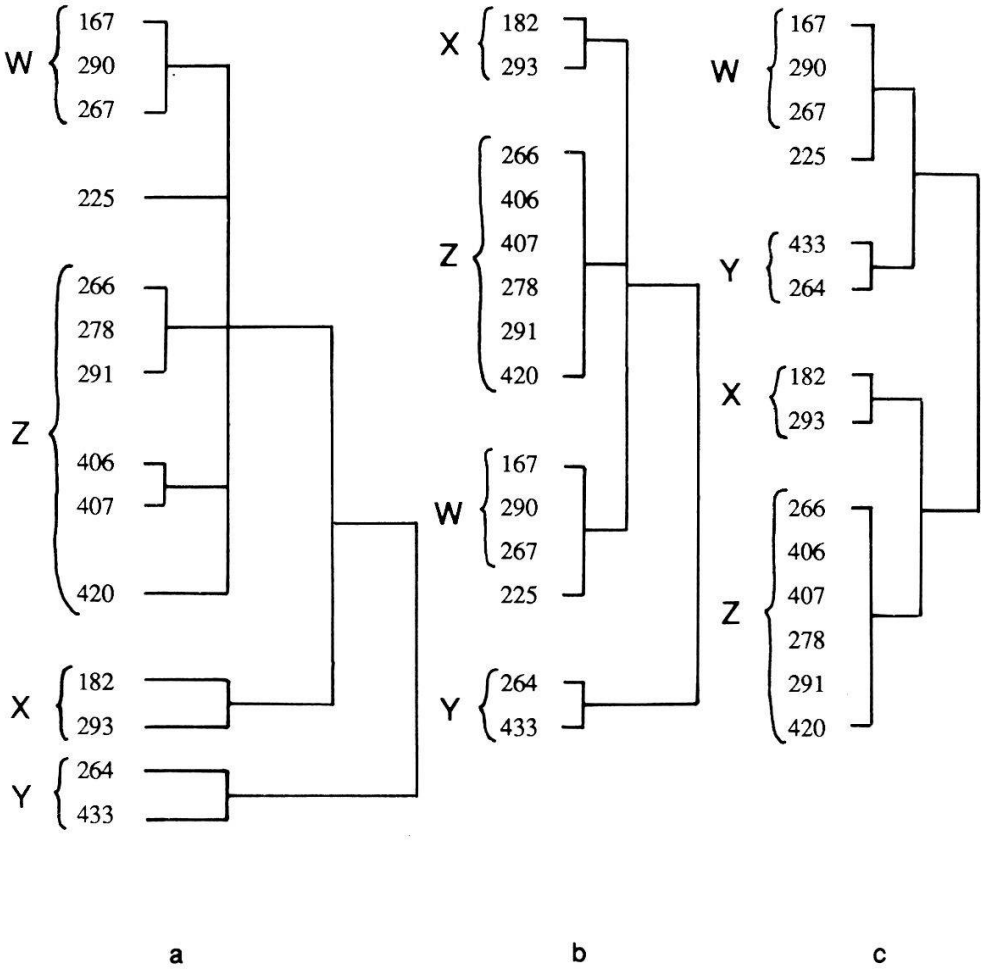


Figure 6. Petits noyaux dans un grand
Détail du noyau stable **5** dans les dendrogrammes de la figure 9 (chap. 6.4.1.):
a: *Average linkage*
b: *Ward's*
c: *Complete linkage*
Petits noyaux stables: abcd

4.6. *Bilan de la classification automatique (clustering)*

Dans le foisonnement des techniques proposées, nous avons tenté d'approcher un choix objectif, en particulier en recourant au critère du coefficient cophénétique. Nous sommes ainsi en mesure de recommander en phytosociologie la technique de réduction préalable des données par AFC, avec pondération des coordonnées (4.2), lorsque le nombre de variables dépasse la capacité de l'ordinateur disponible. Par contre, nous avons vu que le choix ne s'impose pas avec évidence entre les diverses techniques de clustering testées ici. Nous avons rejoint l'idée exprimée par Jambu (1978, p. 213) de confronter plusieurs dendrogrammes pour en extraire l'information commune, et l'avons concrétisée dans la technique d'extraction des «noyaux stables» (4.5).