NOTES DE STATISTIQUE
ET D'INFORMATIQUE

96/1

LA CLASSIFICATION NUMÉRIQUE: PRINCIPES ET APPLICATION

R. PALM

Faculté universitaire des
Sciences agronomiques

Centre de Recherches
agronomiques

GEMBLOUX
(Belgique)

LA CLASSIFICATION NUMÉRIQUE : PRINCIPES ET APPLICATION

R. PALM[1]

RÉSUMÉ

Les principes de quelques méthodes de classification numérique sont décrits et illustrés par un exemple relatif à la classification de 45 eaux minérales sur la base de leur composition chimique.

SUMMARY

The principles of some clustering techniques are described and illustrated by an example. The data consist of chemical contents for 45 types of mineral water.

1. INTRODUCTION

L'objectif de la *classification numérique* ou *automatique*[2] est de subdiviser un ensemble de n individus ou objets en un nombre k de classes ou groupes, cette subdivision se faisant à partir des observations relatives à p variables.

Les domaines d'application sont très variés et de nombreuses méthodes sont proposées dans la littérature. Notre but n'est pas de décrire toutes ces méthodes. Nous nous limiterons, au contraire, à la présentation de quelques techniques, en nous appuyant sur un exemple concret.

L'exemple retenu est dû à TOMASSONE *et al.* [1993]. Il concerne la classification d'une série d'eaux minérales sur la base des teneurs en six éléments chimiques, relevées sur les étiquettes. Aux 28 eaux considérées par TOMASSONE *et al.* [1993], nous avons ajouté 17 eaux supplémentaires. L'ensemble des données de

[1] Chargé de cours associé à la Faculté des Sciences agronomiques de Gembloux.

[2] En anglais : *classification, cluster analysis.*

base est repris en annexe. On notera que les compositions chimiques des eaux ont pu subir des modifications et que les teneurs relevées ne sont pas nécessairement les teneurs actuelles.

Nous présenterons d'abord quelques principes de base relatifs à la classification hiérarchique (paragraphe 2) puis nous passerons en revue les différentes étapes de la classification numérique (paragraphe 3). Ensuite, nous traiterons l'exemple présenté ci-dessus (paragraphe 4), avant de tirer les conclusions (paragraphe 5).

2. QUELQUES ALGORITHMES DE CLASSIFICATION HIÉRARCHIQUE

2.1. Généralités

Dans cette partie, nous nous proposons de décrire, de façon relativement détaillée, les principes à la base de quelques méthodes courantes de classification numérique. Toutefois, avant de présenter ces méthodes, nous consacrerons deux paragraphes au rappel de notions préliminaires, utiles pour la compréhension des algorithmes et des documents fournis par les logiciels statistiques et notamment par les logiciels Minitab et SAS.

Les principes qui seront présentés seront illustrés par un exemple numérique de taille réduite.

Dans ce but, nous avons sélectionné sept eaux minérales parmi les 45 eaux qui seront analysées plus en détail au paragraphe 4. Ces eaux ont été choisies de manière à ce que le sous-ensemble contienne à la fois des eaux semblables et des eaux très différentes. Elles sont identifiées par une lettre majuscule ou minuscule et la correspondance entre ce code et la dénomination de l'eau est donnée en annexe. Comme nous le justifierons au paragraphe 4.1, les variables ont été standardisées de manière à ce que les moyennes soient nulles et que les écarts-types estimés soient unitaires (tableau 1).

2.2. Notion de distance

La distance euclidienne entre deux objets, i et i', est définie par la formule suivante :

$$d_{ii'} = \sqrt{\sum_{j=1}^{p}(x_{ij} - x_{i'j})^2}\,,$$

x_{ij} étant l'observation relative à la variable j sur l'objet i. On peut encore éventuellement inclure le facteur $1/p$ sous le radical.

Ainsi, le carré de la distance entre les objets G et L est égal à :

$$d^2_{\mathrm{GL}} = [(-0{,}1527) - (-0{,}0750)]^2 + \ldots + [(-0{,}3811) - (-0{,}3622)]^2 = 4{,}6806$$

et la distance vaut donc :

$$d_{\mathrm{GL}} = 2{,}16\,.$$

Tableau 1. Composition chimique de sept eaux minérales : données centrées et réduites.

Code	HCO_3^-	SO_4^{--}	Cl^-	Ca^{++}	Mg^{++}	Na^+
G	– 0,1527	– 0,8480	– 0,5379	– 0,1490	– 0,2001	– 0,3811
L	– 0,0750	0,9677	– 0,2689	0,9493	0,1143	– 0,3622
O	– 0,8084	– 0,3679	– 0,4758	– 1,1777	– 0,8575	– 0,3937
S	– 0,0674	1,7358	– 0,2689	1,5749	0,1429	– 0,3937
T	– 0,7079	– 0,8742	– 0,4137	– 1,0943	– 0,7146	– 0,3622
c	2,1649	0,0424	2,2549	0,0040	2,0866	2,2676
o	– 0,3536	– 0,6559	– 0,2896	– 0,1072	– 0,5717	– 0,3748

Des calculs analogues peuvent être réalisés pour toutes les paires d'objets. On obtient alors la matrice des distances qui est symétrique et dont les éléments de la diagonale sont tous nuls (tableau 2).

La distance entre deux objets est une mesure de la proximité des objets dans l'espace des variables. On constate, par exemple, que les objets G et o sont les plus proches et que les objets O et c sont les plus éloignés.

Tableau 2. Tableau des distances euclidiennes entre les sept objets.

Distances	G	L	O	S	T	c	o
G	0	2,16	1,47	3,14	1,22	5,12	0,53
L	2,16	0	2,80	0,99	2,95	4,89	2,07
O	1,47	2,80	0	3,69	0,55	5,80	1,25
S	3,14	0,99	3,69	0	3,89	5,25	3,02
T	1,22	2,95	0,55	3,89	0	5,67	1,09
c	5,12	4,89	5,80	5,25	5,67	0	5,23
o	0,53	2,07	1,25	3,02	1,09	5,23	0

Après avoir défini la distance entre deux objets, voyons comment mesurer la distance entre deux groupes d'objets. Plusieurs définitions peuvent être données pour cette distance et nous en retiendrons trois, qui reposent sur l'examen des distances entre les paires d'objets, un objet appartenant à chacun des deux groupes.

La figure 1, qui schématise un exemple théorique de deux groupes de deux objets (A et B, d'une part, C et D, d'autre part) dans un problème à deux variables, permet de comprendre facilement ces trois définitions.

Soit d_{AC}, d_{AD}, d_{BC} et d_{BD}, les distances entre les paires d'objets. On peut définir la distance entre les deux groupes comme étant :

– la distance minimum, qui est la distance entre les deux objets les plus proches, c'est-à-dire d_{BC} ;

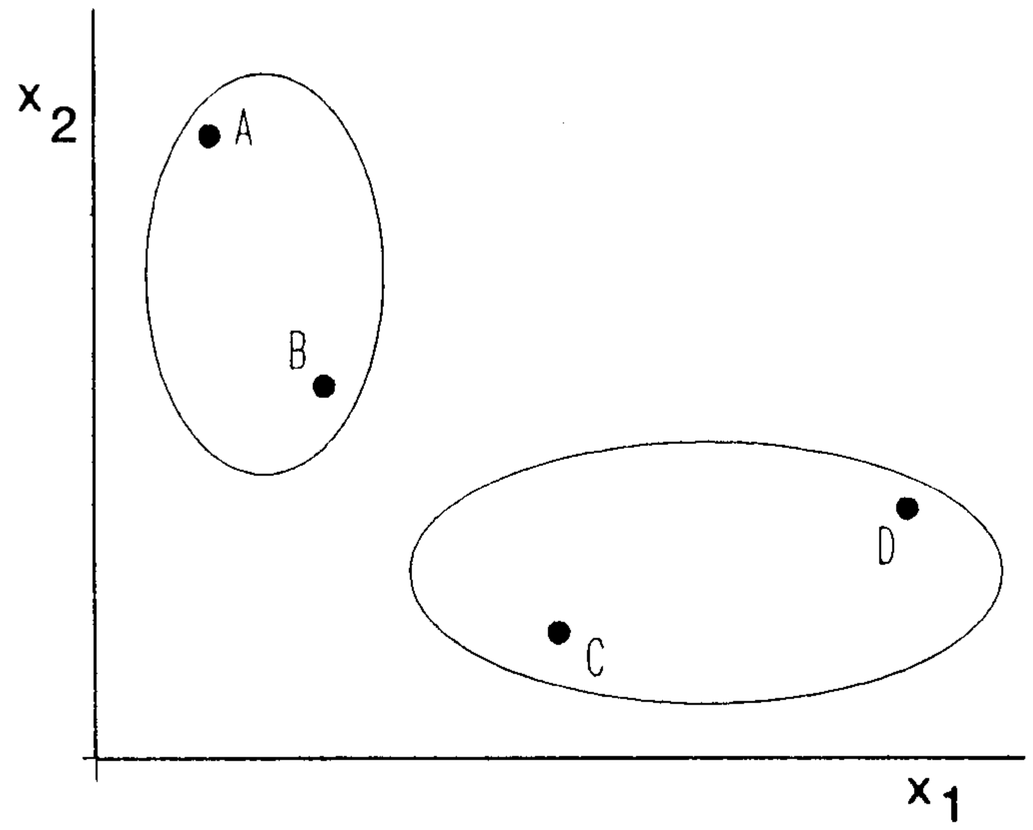

Figure 1. Représentation schématique de deux groupes de deux objets dans un espace à deux dimensions.

- la distance maximum, qui est la distance entre les deux objets les plus éloignés, c'est-à-dire d_{AD} ;
- la distance moyenne, qui est la moyenne quadratique des distances entre toutes les paires :

$$d_{\mathrm{moy}} = \sqrt{\left(d_{\mathrm{AC}}^2 + d_{\mathrm{AD}}^2 + d_{\mathrm{BC}}^2 + d_{\mathrm{BD}}^2\right)/4}\,.$$

L'utilisation de ces distances conduit aux méthodes de classification hiérarchique connues sous les noms de méthode du lien simple, méthode du lien complet et méthode du lien moyen, qui seront brièvement présentées au paragraphe 2.4.

A titre d'exemple, considérons le groupe formé des objets G, O, T et o et le groupe formé des objets L et S. Le tableau 3, construit à partir du tableau 2, donne les distances entre les paires d'objets. Pour ces deux groupes, la distance minimum est de 2,07 (distance entre o et L), la distance maximum est de 3,89 (distance entre T et S) et la distance moyenne est égale à 3,02.

Enfin, on peut encore définir une distance entre un objet donné et le centre de gravité du groupe auquel il appartient, appelé aussi *centroïde*[3]. Une telle distance se calcule comme la distance entre deux objets. Il suffit de remplacer les coordonnées du deuxième objet par les coordonnées du centre de gravité du groupe. De même, on peut encore définir la distance entre les centres de gravité de deux groupes.

---

[3] En anglais : *centroid*.

Tableau 3. Distances euclidiennes entre les objets de deux groupes.

Distances	L	S
G	2,16	3,14
O	2,80	3,69
T	2,95	3,89
o	2,07	3,02

2.3. Décomposition des sommes des carrés des écarts

Considérons la répartition suivante des sept objets en quatre groupes : le groupe 1 est constitué des objets G et o, le groupe 2 des objets O et T, le groupe 3 des objets L et S et, enfin, le groupe 4, du seul objet c. Nous verrons, au paragraphe 2.4, que cette partition est précisément une des partitions fournies par la classification hiérarchique. De façon plus condensée, on écrira qu'il s'agit de la partition :

$$\{(G,o),(O,T),(L,S),(c)\}\,.$$

Pour cette partition et pour chacune des variables, on peut effectuer la décomposition de la somme des carrés des écarts totale en deux parties : la somme des carrés des écarts entre les groupes et la somme des carrés des écarts dans les groupes. Les résultats sont donnés dans le tableau 4.

Tableau 4. Décomposition des sommes des carrés des écarts pour la partition $\{(G,o),(O,T),(L,S),(c)\}$.

Sources de variation	Degrés de liberté	HCO_3^-	SO_4^{--}	Cl^-	Ca^{++}	Mg^{++}	Na^+	Totaux
Entre les groupes	3	5,97	5,56	5,97	5,80	5,92	6,00	35,22
Dans les groupes	3	0,03	0,44	0,03	0,20	0,08	0,00	0,78
Totaux	6	6,00	6,00	6,00	6,00	6,00	6,00	36,00

En sommant les composantes des sommes des carrés des écarts pour les variables, on obtient une somme des carrés des écarts globale entre les groupes de 35,22, une somme des carrés des écarts globale dans les groupes de 0,78 et une somme des carrés des écarts globale totale de 36,00.

Le rapport entre la somme des carrés des écarts globale entre les groupes et la somme des carrés des écarts globale totale définit le coefficient R^2 relatif à la partition :

$$R^2 = 35{,}22/36{,}00 = 0{,}978\,.$$

Il s'agit d'une valeur globale ou moyenne pour les six variables et l'examen des valeurs pour chaque variable permet de constater que la variable Na^+ est la plus discriminante (R^2=1,000) alors que la variable SO_4^{--} est la moins discriminante (R^2=0,926).

Si on considère maintenant la partition obtenue en fusionnant les groupes 1 et 2 :

$$\{(G,o,O,T),(L,S),(c)\}$$

et si on calcule à nouveau les sommes des carrés des écarts globales, on obtient, respectivement, pour les sommes des carrés des écarts entre les groupes et dans les groupes les valeurs 33,77 et 2,23. Pour cette partition, le coefficient R^2 vaut 0,938. La diminution de R^2 liée à la fusion des groupes (G,o) et (O,T) est donc égale à 0,040.

La fusion des groupes (G,o) et (O,T) provoque donc une diminution de la somme des carrés des écarts entre groupes de 1,45 et une augmentation identique de la somme des carrés des écarts dans les groupes. Cette quantité est aussi la somme des carrés globale entre les deux groupes fusionnés.

On notera aussi que les notions de distances et de sommes des carrés des écarts sont liées. En effet, la somme des carrés des distances de tous les objets au centre de gravité de l'ensemble des objets est égale à la somme des carrés des écarts totale. De même, la somme des carrés des distances entre toutes les paires d'objets est égale à n fois la somme des carrés des écarts totale.

D'autre part, pour une partition donnée, la somme des carrés des distances entre les paires d'objets d'un groupe, divisée par l'effectif du groupe, est égale à la somme des carrés des écarts résiduelle dans ce groupe. La sommation sur tous les groupes donne la somme des carrés des écarts résiduelle globale.

Ainsi, pour l'exemple considéré, la somme des carrés des 21 distances entre objets est égale à 252 et, pour la partition en trois groupes examinée ci-dessus, la somme des carrés des écarts résiduelle est égale à :

$$\begin{aligned} SCE_r &= (d^2_{GO} + d^2_{GO} + d^2_{GT} + d^2_{OO} + d^2_{OT} + d^2_{OT})/4 + d^2_{LS}/2 \\ &= (0{,}53^2 + \ldots + 0{,}55^2)/4 + 0{,}99^2/2 = 2{,}23\,. \end{aligned}$$

2.4. Quelques stratégies d'agrégation

Nous avons signalé, dans l'introduction, qu'il existe de nombreuses méthodes de classification. Parmi celles-ci, nous allons nous intéresser plus particulièrement aux méthodes hiérarchiques agglomératives. Des informations relatives à d'autres méthodes seront données au paragraphe 3.3.

Le principe des méthodes hiérarchiques agglomératives est de prendre comme point de départ la partition des n objets en n classes d'un objet. A chaque étape ultérieure, on fusionne deux classes pour former une nouvelle classe. On passe donc, par fusions successives, de n groupes d'un objet à un seul groupe de n objets et, à l'issue de la classification, on dispose des partitions à n groupes, $n-1$ groupes, ... et 1 groupe.

NCL	-Clusters Joined-		FREQ	RMS STD	SPRSQ	RSQ	BSS
6	G	o	2	0.1524	0.003871	0.9961	0.13935
5	O	T	2	0.1578	0.004148	0.9920	0.14932
4	L	S	2	0.2863	0.013657	0.9783	0.49166
3	CL6	CL5	4	0.3107	0.040256	0.9381	1.44920
2	CL3	CL4	6	0.6713	0.313608	0.6245	11.28990
1	CL2	c	7	1.0000	0.624460	0.0000	22.48057

Figure 2. Classification des sept objets par l'algorithme de WARD : résultats obtenus par le logiciel SAS.

Les méthodes hiérarchiques agglomératives se distinguent par la stratégie d'agrégation qui est utilisée. Nous allons envisager d'abord l'algorithme de WARD ; ensuite nous décrirons les méthodes basées sur les mesures de distances entre groupes qui ont été décrites au paragraphe 2.2.

Le principe de la méthode de WARD est d'effectuer les regroupements de manière à ce que la différence de R^2 pour deux partitions successives soit aussi faible que possible.

Les figures 2 et 3 reprennent les résultats de la classification des sept objets par cet algorithme. La figure 2 a été obtenue par la procédure PROC CLUSTER du logiciel SAS [SAS, 1989] et la figure 3 résulte de la commande CLUOBS du logiciel Minitab [X, 1994]. Pour cette dernière, on a également fait appel à l'option donnant des informations relatives aux groupes formés, pour une partition donnée. La partition en trois groupes a été choisie à titre d'exemple.

On constate qu'il y a d'abord réunion des objets G et o (partition en six groupes), ensuite des objets O et T (partition en cinq groupes) et des objets L et S (partition en quatre groupes). Les trois dernières fusions concernent non plus des objets isolés mais des groupes préalablement formés : la partition en trois groupes s'obtient en fusionnant les deux groupes formés de deux objets, (G, o), d'une part, (O, T) d'autre part, et la partition en deux groupes s'obtient en fusionnant le groupe (G, o, O, T) et le groupe (L, S). Enfin, l'objet c est rattaché au groupe constitué des six objets.

Le logiciel SAS indique également diverses informations concernant chacune de ces fusions. Ainsi, il donne la taille du groupe nouvellement formé (colonne FREQ), la racine carrée de la moyenne des variances de l'ensemble des variables pour les objets constituant ce groupe (colonne RMS STD[4]), la diminution de R^2 (colonne SPRSQ[5]), la valeur de R^2 (colonne RSQ) et la diminution de la somme des carrés des écarts globale factorielle liée à la fusion qui vient d'être réalisée (colonne BSS[6]). Pour la partition en trois groupes, qui, comme nous

---

[4] En anglais : *root-mean-square standard deviation.*
[5] En anglais : *semi partial R-squared.*
[6] En anglais : *between cluster sum of squares.*

Hierarchical Cluster Analysis of Observations

Standardized Variables, Squared Euclidean Distance, Ward Linkage

Amalgamation Steps

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	6	99.17	0.279	1	7	1	2
2	5	99.11	0.299	3	5	3	2
3	4	97.07	0.983	2	4	2	2
4	3	91.38	2.898	1	3	1	4
5	2	32.82	22.580	1	2	1	6
6	1	-33.77	44.961	1	6	1	7

Final Partition

Number of clusters: 3

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	4	1.738	0.653	0.753
Cluster2	2	0.492	0.496	0.496
Cluster3	1	0.000	0.000	0.000

Cluster Centroids

Variable	Cluster1	Cluster2	Cluster3	Grand centrd
HCO3	-0.5056	-0.0712	2.1649	-0.0000
SO4	-0.6865	1.3517	0.0424	-0.0000
Cl	-0.4293	-0.2689	2.2549	0.0000
Ca	-0.6321	1.2621	0.0040	-0.0000
Mg	-0.5860	0.1286	2.0866	-0.0000
Na	-0.3779	-0.3779	2.2676	0.0000

Distances Between Cluster Centroids

	Cluster1	Cluster2	Cluster3
Cluster1	0.0000	2.9099	5.4234
Cluster2	2.9099	0.0000	5.0497
Cluster3	5.4234	5.0497	0.0000

Figure 3. Classification des sept objets par l'algorithme de WARD : résultats obtenus par le logiciel Minitab.

venons de le voir, s'obtient en fusionnant G et o, d'une part, avec O et T, d'autre part, on retrouve bien, aux arrondis près, les valeurs qui ont été calculées au paragraphe précédent.

Le logiciel Minitab indique le niveau d'agrégation, sous l'intitulé « *Distance level* ». Il s'agit en fait d'une quantité égale à deux fois la diminution de la somme des carrés des écarts globale factorielle liée à la fusion qui vient d'être réalisée (colonne BSS de SAS). L'introduction de ce facteur deux a comme effet de faire coïncider le paramètre avec le carré de la distance lorsque la fusion concerne deux objets isolés, comme on peut le constater pour les trois premières fusions. La colonne « *Similarity level* » s'obtient en prenant le complément à 100 du rapport de la quantité reprise dans la colonne « *Distance level* » au carré de la distance

maximum. Pour la première fusion, on a, par exemple :

$$100 - (0{,}279/5{,}802) = 99{,}17\,,$$

la distance maximale étant la distance d_{OC}.

Pour la partition en trois groupes, Minitab donne, pour chaque groupe, la somme des carrés des écarts dans les groupes, la distance moyenne et la distance maximum au centre de gravité du groupe. On retrouve bien les sommes des carrés des écarts résiduelles calculées au paragraphe 2.3, à partir des distances entre objets.

Le logiciel donne aussi les coordonnées des centres de gravité des groupes, les coordonnées du centre de gravité de l'ensemble des objets et les distances entre les centres de gravité des groupes. Ces éléments peuvent être utilisés lors de l'interprétation des résultats, comme nous le verrons au paragraphe 3.6. Les coordonnées du centre de gravité de l'ensemble des objets sont nulles car les variables ont été centrées.

On notera que pour l'algorithme de WARD, les valeurs de la colonne SPRSQ et BSS du logiciel SAS et « *Distance level* » de Minitab sont toujours non décroissantes et sont, à chaque étape, aussi faibles que possible.

Les méthodes du lien simple, du lien complet et du lien moyen se basent sur les mesures de distance entre groupes qui ont été définies au paragraphe 2.2. Dans le cas de la *méthode du lien simple*, appelée aussi *méthode du plus proche voisin*[7], le principe est de fusionner les groupes pour lesquels la distance minimum est la plus faible. Dans le cas de la *méthode du lien complet* ou *méthode du voisin le plus éloigné*[8], on fusionne les groupes pour lesquels la distance maximum est la plus faible et, dans le cas du *lien moyen*[9], on fusionne les groupes pour lesquels la distance moyenne est la plus faible.

Pour l'exemple considéré, les trois méthodes conduisent aux mêmes résultats que l'algorithme de WARD. Pour cette raison, nous ne présentons pas les détails de ces classifications, mais nous avons simplement repris, dans le tableau 5, les informations spécifiques à chaque algorithme qui sont fournies par le logiciel SAS. En effet, en plus des colonnes FREQ, RMS STD, SPRSQ et RSQ, définies à propos de l'algorithme de WARD, le logiciel SAS donne les distances minimales, les distances maximales ou les distances moyennes, selon l'algorithme retenu. On peut noter que lorsque la fusion concerne deux objets isolés, ces trois distances sont identiques. Par contre, lorsque les fusions impliquent un ou deux groupes préalablement formés, la distance maximum est évidement plus grande que la distance minimum et la distance moyenne a une valeur intermédiaire.

Le logiciel Minitab fournit également, sous l'intitulé « *Distance level* », les distances reprises dans le tableau 5. Toutefois, pour la distance moyenne, Minitab indique les carrés des distances.

---

(7) En anglais : *single linkage, nearest neighbour.*
(8) En anglais : *complete linkage, furthest neighbour.*
(9) En anglais : *average linkage.*

Tableau 5. Distances minimales, maximales et moyennes aux différentes étapes de la classification hiérarchique.

Nombres de groupes	Distances minimales	Distances maximales	Distances moyennes (quadratiques)
6	0,53	0,53	0,53
5	0,55	0,55	0,55
4	0,99	0,99	0,99
3	1,09	1,47	1,26
2	2,07	3,89	3,02
1	4,89	5,80	5,34

Pour la méthode du lien complet, choisie à titre d'exemple, le tableau 6 donne les distances maximales pour les différents niveaux de la hiérarchie. Le point de départ de la construction de ce tableau est le tableau des distances individuelles (tableau 2), qui montre que la distance maximum est minimale pour le couple (G, o). Ces objets sont donc fusionnés. Ensuite, en considérant les six groupes, la distance maximum est minimale pour le couple (O, T) et ces deux objets sont fusionnés. Pour cinq groupes, la distance maximum est minimale pour le couple (L, S). Après la fusion de ces deux objets, la distance maximum est minimale pour le couple de groupes (G, o) et (O, T), qui sont donc fusionnés. L'étape suivante conduit à la fusion du groupe (L, S) et du groupe (G, o, O, T). Enfin, la dernière fusion inclut l'objet c au groupe des six autres objets.

2.5. Les dendrogrammes

Les résultats d'une classification hiérarchique sont fréquemment représentés sous la forme d'un dendrogramme, qui schématise les fusions successives permettant de passer de n classes à une seule classe regroupant tous les objets. Un exemple de dendrogramme sera donné au paragraphe 4.1. Lorsque le nombre d'objets est grand et pour des raisons de clarté du graphique, on peut toutefois se limiter à la portion du dendrogramme la plus utile et correspondant, le plus souvent, à un nombre de groupes suffisamment réduit.

Un des problèmes pratiques qui se pose en relation avec ces dendrogrammes est le choix de l'échelle représentant les niveaux auxquels ont lieu les fusions car diverses options sont possibles.

D'une façon générale, on peut considérer que l'algorithme de classification utilisé détermine assez logiquement le choix de l'échelle. Si la classification est basée sur l'algorithme de WARD, on utilisera la valeur de R^2 ou, ce qui revient au même, la variance dans les groupes fusionnés. Si la classification est basée sur le lien moyen, on portera en ordonnée les valeurs du lien moyen et ainsi de suite. Toutefois, lorsqu'on souhaite comparer les résultats de classifications d'un même ensemble d'objets par différents algorithmes, il peut être utile d'adopter la

Tableau 6. Distances maximales entre groupes, pour les différents niveaux de la hiérarchie (les nombres en caractères gras correspondent aux valeurs minimales).

Distances maximales	L	O	S	T	c
(G,o)	2,16	1,47	3,14	1,22	5,23
L	0	2,80	0,99	2,95	4,89
O		0	3,69	**0,55**	5,80
S			0	3,89	5,25
T				0	5,67

Distances maximales	L	(O,T)	S	c
(G,o)	2,16	1,47	3,14	5,23
L	0	2,95	**0,99**	4,89
(O,T)		0	3,89	5,80
S			0	5,25

Distances maximales	(L,S)	(O,T)	c
(G,o)	3,14	**1,47**	5,23
(L,S)	0	3,89	5,25
(O,T)		0	5,80

Distances maximales	(L,S)	c
(G,o,O,T)	**3,89**	5,80
(L,S)	0	5,25

Distances maximales	c
G,o,O,T,L,S	**5,80**

même échelle pour tous les dendrogrammes, de manière à les rendre aussi comparables que possible. Une modification de la caractéristique servant à quantifier les niveaux auxquels se font les fusions peut en effet fortement modifier l'allure générale d'un dendrogramme.

En relation notamment avec le problème de comparaisons de classifications, on ne perdra pas de vue qu'un dendrogramme doit être vu comme un mobile, dans le sens où des rotations de 180° peuvent être appliquées à des portions d'un dendrogramme, ce qui modifie évidemment son aspect général.

3. LES ÉTAPES DE LA CLASSIFICATION NUMÉRIQUE

3.1. Généralités

Dans les paragraphes précédents, nous nous sommes limités à la présentation des notions de base. Ces notions ne couvrent pas tous les aspects de la classification numérique et il nous paraît utile de donner quelques informations complémentaires permettant d'élargir le champ d'application de la classification. Dans ce but, nous allons passer sommairement en revue les grandes étapes de la classification, qui sont la collecte des données, le choix d'un indice de similitude, le choix d'un algorithme de classification, la détermination du nombre de classes et l'interprétation des résultats.

3.2. Collecte des données

En ce qui concerne le choix des objets soumis à la classification, deux situations différentes peuvent se présenter. En effet, les objets étudiés peuvent constituer la population entière à laquelle on s'intéresse ou, au contraire, avoir été sélectionnés dans un ensemble plus vaste. Cette distinction est sans grande importance pour la classification proprement dite, mais entre en ligne de compte lors de l'interprétation des résultats.

Pour ce qui est des variables, une distinction doit être faite entre les variables quantitatives et les variables qualitatives, car la nature des variables influence le choix de l'indice de similarité (paragraphe 3.3). A ce sujet, on se rappellera que des variables qualitatives binaires peuvent être remplacées par des variables du type 0/1 et que des variables qualitatives à q modalités peuvent être remplacées par $q-1$ variables de type 0/1.

Le choix des variables à prendre en considération doit être pertinent compte tenu de l'objectif poursuivi et est laissé à l'appréciation du praticien. Ainsi, par exemple, dans le cas de la classification de relevés de végétation (objets) sur la base de la présence ou de l'absence ou bien de l'abondance-dominance de plantes (variables) il pourra être utile d'éliminer les plantes présentes dans un ou deux relevés seulement, à moins qu'elles ne revêtent une importance particulière aux yeux du phytosociologue. De même il pourra être utile éventuellement de supprimer les plantes présentes dans une majorité de relevés.

En relation avec le choix des variables, se pose le problème de la pondération de celles-ci. Lorsque les variables sont exprimées dans des unités différentes, il est le plus souvent utile de standardiser les données. Mais la question de la standardisation doit être posée même si les variables sont de nature identique, comme dans le cas évoqué précédemment de la composition chimique des eaux minérales.

Il faut noter cependant que la standardisation peut avoir comme effet la dilution des différences entre groupes pour les variables les plus discriminantes. Comme le signale EVERITT [1993], il serait plus efficace de standardiser les variables en utilisant les écarts-types dans les groupes mais, évidemment, ceux-ci sont *a priori* inconnus.

Tableau 7. Codification de la répartition de la coloration anthocyanique sous la forme de deux variables quantitatives.

Variable initiale	Variables codées	
	1	2
Absence	0	0
Mouchetée	1	0
Tachetée	0	1
Mouchetée et tachetée	0,5	0,5

Le poids d'une caractéristique est également lié au nombre de variables servant à décrire cette caractéristique. Pour illustrer ce phénomène, considérons le problème étudié par VANDERBORGHT [1986] relatif à la classification d'une centaine de variétés de haricots sur la base d'une trentaine de caractères morphologiques et agronomiques. Un de ces caractères concerne la répartition de la coloration anthocyanique de la gousse. Il s'agit d'une variable qualitative à quatre modalités, codée par l'auteur sous la forme de deux variables quantitatives de la manière décrite dans le tableau 7. Ces deux variables ont été ajoutées aux autres variables quantitatives. Toutes les variables ont ensuite été standardisées de manière à obtenir un écart-type unitaire. Ensuite, pour éviter que le caractère « répartition de la coloration anthocyanique » n'ait un poids plus important que les autres caractères, par le fait que deux variables servent à le décrire, l'auteur a divisé les valeurs standardisées de ces deux variables par $\sqrt{2}$. Une démarche analogue a été adoptée pour toutes les autres variables résultant d'un caractère qualitatif initial à plus de deux états. De cette manière, la variance totale de chacun des caractères initiaux est égale à l'unité, assurant ainsi une pondération identique.

Par ailleurs, la notion d'objets et de variables est une notion relative, fonction de l'objectif poursuivi. Ainsi, si on reprend le problème des relevés phytosociologiques, on peut s'intéresser soit à la classification des relevés, soit à la classification des plantes. Dans le premier cas, les relevés seront les objets et les plantes seront les variables, alors que dans le second cas les plantes constitueront les objets et les relevés les variables. De façon générale, une simple transposition de la matrice des données permet de résoudre les deux problèmes avec le même logiciel. Le logiciel Minitab propose cependant une commande spécifique permettant de classer les variables sans transposition préalable des données. Il s'agit de la commande CLUVARS. L'intérêt de cette commande est qu'elle permet la prise en considération de mesures de similitude particulièrement adaptées à la classification des variables (paragraphe 3.3).

3.3. Choix d'une mesure de similitude

Dans l'exemple des eaux minérales, nous avons utilisé la distance euclidienne comme mesure de la ressemblance existant entre deux objets. Il s'agit d'une solution couramment utilisée lorsqu'on dispose de variables quantitatives mais d'autres types de distances peuvent être envisagés. Ainsi, la distance du

χ^2 est particulièrement adaptée aux variables quantitatives résultant de comptages. Ces distances ont en commun d'être d'autant plus grandes que les objets sont considérés comme différents et d'autant plus petites que les objets sont ressemblants.

D'autres paramètres sont définis pour les variables alternatives. Ils expriment, sous diverses formes, la proportion de caractères communs à deux objets. Ces indices de similarité sont généralement compris entre 0 et 1 ou entre 0 et 100 et, contrairement aux distances, ils possèdent des valeurs d'autant plus grandes que les objets sont semblables. Ainsi, si on considère deux objets et si on désigne par P le nombre de caractères communs aux deux objets (co-présences), par A le nombre de caractères absents pour les deux objets (co-absences), par N le nombre de caractères présents sur un objet et absents sur l'autre (non-coïncidences) et par T le nombre total de caractères, on peut définir, par exemple, les rapports suivants :

P/T (indice de RUSSEL et RAO),
$(P + A)/T$ (indice de SOKAL et MICHENER),
$P/(P + N)$ (indice de JACCARQ),

mais bien d'autres indices sont encore proposés [CHANDON et PINSON, 1981 ; DAGNELIE, 1975 ; EVERITT, 1993, LEGENDRE et LEGENDRE, 1984]. Ces rapports se différencient essentiellement par le poids qui est donné aux co-absences, dans le numérateur et aux non-coïncidences dans le dénominateur. Une discussion de ces indices est donnée, notamment, par LEGENDRE et LEGENDRE [1984].

Le coefficient de corrélation entre deux objets peut également être utilisé pour mesurer la ressemblance entre ces objets. Ce coefficient est toujours compris entre -1 et 1. Il est d'autant plus grand que les objets se ressemblent et d'autant plus petit que les objets sont différents. Il peut être utilisé à la fois pour des résultats de mesures ou pour des données de type 0/1 [DAGNELIE, 1975]. Il convient bien lorsqu'il s'agit de classer des variables quantitatives. C'est d'ailleurs la solution retenue par défaut dans la commande CLUVARS de Minitab.

3.4. Choix d'un algorithme de classification

Quelques méthodes de classification hiérarchiques agglomératives ont été présentées au paragraphe 2. Les méthodes hiérarchiques conduisent toutes à un dendrogramme et les différences sont liées à la distance ou à l'indice de similarité utilisé et à la définition de la distance (ou de la similarité) entre un objet et un groupe d'objets ou entre deux groupes d'objets.

Nous avons déjà décrit les méthodes de WARD, du lien simple, du lien complet et du lien moyen. Trois autres méthodes sont encore disponibles à la fois dans les logiciels SAS et Minitab. Il s'agit de la méthode basée sur les distances entre centres de gravité, sur les distances médianes et de la méthode de MCQUITTY.

Dans le cas de la *méthode basée sur les distances entre centres de gravité*[10], la distance entre deux groupes est la distance entre les centres de

[10] En anglais : *centroid linkage*.

gravité des groupes. Dans le cas de la *méthode des médianes*[11], la distance entre deux groupes est la médiane des distances entre toutes les paires d'objets, un objet appartenant à chaque groupe. Enfin, dans la *méthode de* McQuitty[12], on fusionne les groupes pour lesquels la distance est la plus faible, la distance en question étant calculée de la façon suivante :

$$d_{mj} = (d_{kj} + d_{lj})/2\,,$$

d_{kj} et d_{lj} représentant les distances des groupes k et l par rapport au groupe j et d_{mj} représentant la distance par rapport au groupe j du groupe m, résultant de la fusion des groupes k et l.

D'autres méthodes de classification sont encore disponibles dans le logiciel SAS. Nous renvoyons le lecteur au guide de l'utilisateur pour plus d'information [SAS, 1989].

Il faut noter aussi que certains algorithmes ne sont pas valables pour certaines mesures de similitude. Ainsi, les méthodes du lien simple, du lien complet et du lien moyen peuvent être utilisées à la fois avec des mesures de distance ou de similarité. Quant à la méthode de Ward, elle n'est valable que si les caractéristiques mesurées sont quantitatives. Elle peut cependant s'utiliser aussi directement à partir de la matrice des distances euclidiennes entre toutes les paires d'objets.

Dans le cas de certains logiciels, il est nécessaire de transformer les indices de similarité de manière à avoir, comme pour les distances, des valeurs d'autant plus grandes que les objets sont différents. Cela peut se faire en prenant le complément à l'unité des indices de similarité définis au paragraphe 3.3. La commande CLUVARS de Minitab transforme, par exemple, automatiquement les coefficients de corrélation des variables i et j en distances :

$$d_{ij} = 1 - r_{ij}\,,$$

qui varient de zéro à un pour les corrélations positives et de 1 à 2 pour les corrélations négatives.

Il existe également des méthodes hiérarchiques divisives, dont le principe est de partir de l'ensemble des objets considérés comme appartenant à un seul groupe. A chaque étape, l'algorithme divise un groupe pour en former deux et le processus s'arrête lorsqu'on obtient la partition à n groupes. On notera qu'un algorithme agglomératif et un algorithme divisif basés sur un même critère et appliqués à un même ensemble d'objets ne donnent en général pas la même partition pour un nombre de groupes fixé, sauf dans le cas du lien simple.

Contrairement aux méthodes hiérarchiques, certaines méthodes produisent directement une partition en un nombre de classes fixé *a priori*. Ces méthodes sont dites non hiérarchiques. L'objectif poursuivi est de regrouper n objets en k classes de sorte que les objets d'une classe soient aussi semblables que possible et que les classes soient aussi différentes que possible.

---

[11] En anglais : *median linkage.*

[12] En anglais : McQuitty*'s similarity analysis.*

Les critères utilisés pour apprécier la qualité d'une partition sont souvent liés à l'équation fondamentale de l'analyse de la variance multivariée à un critère :

$$\mathbf{T} = \mathbf{H} + \mathbf{E}$$

où **T**, **H** et **E** représentent les matrices des sommes des carrés et des produits des écarts totale, factorielle (entre les groupes) et résiduelle (dans les groupes). On cherche, par exemple, à minimiser la trace ou le déterminant de **E**, ou à maximiser la trace de $\mathbf{E}^{-1}\,\mathbf{H}$.

La recherche de la solution optimale est théoriquement possible par l'étude de toutes les partitions répartissant les n objets en k groupes. Toutefois, le nombre de solutions à comparer est en général si grand que le problème devient insurmontable : pour 14 objets seulement, il y a plus de 10 millions de partitions possibles en 4 classes [BOUROCHE et SAPORTA, 1980]. En conséquence, on se contentera de solutions approchées.

Il existe différents algorithmes permettant de déterminer la partition en k groupes. Ils ont en commun de choisir ou de générer une partition initiale et de transférer ensuite les objets d'un groupe à un autre, de manière à optimiser le critère retenu. Pour cette raison, les méthodes sont souvent appelées méthodes de *transfert* ou de *réallocation*[13]. Elles se différencient par le choix de la configuration initiale, par le mode de calcul des nouveaux centres des groupes et par les critères d'arrêt du transfert des objets.

A titre d'illustration, on peut considérer la *méthode des centres mobiles*[14]. On choisit, dans un premier temps, k objets dont les coordonnées constituent les centres provisoires. Ces objets sont choisis, soit sur la base d'idées *a priori* concernant les groupes, soit de manière purement arbitraire. Dans ce dernier cas, on prend, par exemple, les k premiers objets ou k objets sélectionnés au hasard. On classe alors chacun des n objets dans le groupe dont il est le plus proche du centre. On recalcule ensuite les k nouveaux centres de la partition et on effectue une nouvelle partition en regroupant les objets dans les k groupes en fonction de leurs distances aux centres de ces groupes. Et ainsi de suite jusqu'à ce que le critère de classification ne s'améliore plus. Un premier inconvénient de la méthode est qu'on peut obtenir des classes ne contenant aucun objet, c'est-à-dire une partition en moins de k classes. Un second inconvénient est que la partition finale dépend de la partition de départ ; on peut donc n'atteindre qu'un optimum local. Il est, en pratique conseillé de recommencer plusieurs fois la classification à partir de centres de classes différents, afin de vérifier si les résultats sont stables.

La classification par la méthode des centres mobiles peut être réalisée par la commande KMEANS du logiciel Minitab ou par la procédure PROC FASTCLUS de SAS.

Dans le cas de classification d'unités spatiales, il peut se justifier de procéder à une classification en imposant une contrainte de contiguïté spatiale de manière

---

[13] En anglais : *relocation techniques.*

[14] En anglais : *k-means.*

à obtenir des groupes d'unités géographiquement contiguës. Des informations à ce sujet sont données par FOGUENNE [1994] et LANGE [1982], notamment.

Les méthodes que nous venons de présenter conduisent toutes à des partitions telles que chaque objet se trouve affecté à une et une seule classe. Mais il existe également des méthodes permettant un recouvrement. Dans ces méthodes un objet donné peut appartenir à deux ou plusieurs groupes [CHANDON et PINSON, 1981].

Le choix d'une méthode de classification conditionne, dans une certaine mesure, la nature des résultats. Ainsi, la méthode du lien moyen favorise *l'effet de chaînage*: à chaque fusion, les objets non encore regroupés, tendent à être incorporés aux groupes existants plutôt qu'à former de nouveaux groupes, ce qui peut empêcher la distinction de deux groupes, pourtant bien séparés, lorsque quelques points intermédiaires suffisamment rapprochés se situent entre les groupes. Au contraire, la méthode du lien complet provoque l'effet de dissection car elle peut affecter à des groupes différents des objets qui sont très proches. Enfin, la méthode de WARD favorise la fusion de groupes petits et a tendance à constituer des groupes de forme sphérique.

En pratique, si l'on n'a pas plus de raisons de choisir une méthode plutôt qu'une autre, il peut être utile d'examiner la stabilité des résultats obtenus avec des méthodes différentes. Ceux-ci seront d'autant plus stables qu'il existe réellement une structure dans les données.

Dans certains cas on peut aussi utiliser successivement des méthodes différentes. Ainsi, si le nombre d'objets à classer est grand, on peut utiliser d'abord un algorithme non hiérarchique, généralement plus rapide, pour constituer, dans un premier temps, un nombre élevé de groupes, une centaine par exemple. On soumet alors les centres de gravité de ces groupes à une classification hiérarchique. De même, lorsqu'on a choisi une partition suite à une classification hiérarchique basée sur le critère de WARD, on peut essayer de réduire la somme des carrés des écarts dans les groupes par la méthode de transfert, en prenant comme partition initiale le résultat de la classification hiérarchique.

3.5. Détermination du nombre de classes

Un problème commun à tous les algorithmes de classification est le choix du nombre de classes à retenir.

Pour les méthodes hiérarchiques agglomératives il existe dans la littérature un grand nombre de règles d'arrêt permettant de choisir, de façon automatique, le nombre de classes à retenir [BAAMAL, 1994]. Ces règles d'arrêt peuvent être regroupées en deux grandes familles :les règles non inférentielles et les règles inférentielles.

Les règles d'arrêt non inférentielles sont essentiellement basées sur l'évolution, au cours des regroupements successifs, des valeurs d'un critère mesurant la qualité de la classification.

Ainsi, on peut examiner des valeurs successives de R^2 lorsque le nombre de classes diminue. On repère le regroupement qui donne lieu à une diminution fort importante de R^2 et on arrête les fusions avant ce regroupement.

Par analogie avec la valeur F_{obs} de l'analyse de la variance univariée, on peut aussi déterminer pour chaque étape le rapport entre le carré moyen inter-classes et le carré moyen intra-classes. En désignant par $tr(\mathbf{H})$ et $tr(\mathbf{E})$, les traces des matrices des sommes de carrés et de produits des écarts factorielle et résiduelle et par g le nombre de groupes formés, on peut calculer la valeur :

$$\text{« } F_{obs} \text{ »} = \frac{tr(\mathbf{H})/(g-1)}{tr(\mathbf{E})/(n-g)},$$

et retenir le nombre de classes pour lequel cette valeur présente un maximum local ou global ou, à défaut, celui qui correspond à un accroissement maximum de cette valeur.

Le rapport défini ci-dessus est appelé « pseudo-F » car il présente une analogie avec la valeur F_{obs} calculée dans le cas de l'analyse de la variance univariée mais il ne possède pas une distribution F de SNEDECOR. Si les n objets étaient prélevés au hasard dans une distribution normale à p variables non correlées et de même variance et si la répartition des objets dans les g groupes était aléatoire, alors la valeur « F_{obs} » serait une valeur observée d'une variable F de SNEDECOR.

D'une manière similaire, on peut définir à chaque étape une valeur pseudo-T^2, en ne prenant en considération que les deux groupes faisant l'objet de la fusion. Cette valeur s'apparente à la valeur T^2 calculée dans le test de HOTELLING [DAGNELIE, 1975] en considérant toutefois une matrice de variances et covariances scalaire. Comme pour la valeur pseudo-F, on retient le nombre de groupes pour lequel la valeur est un maximum local ou global, ou, à défaut, présente un accroissement maximum.

Les règles d'arrêt inférentielles tiennent compte, d'une manière générale, uniquement des deux classes candidates à la fusion à chaque étape de la classification. Elles sont basées sur le calcul d'un critère mesurant le degré de séparation des deux groupes à fusionner, qui est comparé à une valeur critique déduite de la distribution d'échantillonnage du critère.

Le choix d'une règle d'arrêt reste un problème relativement complexe et il semble bien qu'aucune règle ne soit supérieure aux autres pour toutes les situations. Des informations à ce sujet sont données par BAAMAL [1994].

3.6. Interprétation des résultats

Lorsqu'une partition est retenue on s'efforce généralement de caractériser les groupes obtenus.

Pour des variables quantitatives, on peut tout d'abord calculer les moyennes et les écarts-types des différentes variables pour chacun des groupes. Ces paramètres permettent de localiser les groupes dans l'espace des variables et d'apprécier leur homogénéité. L'interprétation de ces valeurs peut cependant être assez lente si on dispose d'un grand nombre de variables et/ou de groupes.

Le calcul des centres de gravité des groupes et des distances entre les centres de gravité permet également de se faire une idée de la position relative des différents groupes dans l'espace des variables. Une autre solution consiste à

réaliser une analyse en composantes principales avec la représentation des objets dans les plans factoriels. Une telle solution est particulièrement avantageuse si les axes principaux sont facilement interprétables et si un nombre réduit d'axes prend en considération une part importante de la variabilité des données.

Enfin, d'autres techniques d'analyse multivariée peuvent également être utilisées.

4. APPLICATION

4.1. Classification des eaux minérales

Les principes décrits au paragraphe 2 ont été illustrés par un exemple ne comportant que sept objets. De plus, nous avons davantage insisté sur la signification des concepts et sur le mécanisme de la classification que sur l'interprétation concrète des résultats. Par contre, dans les paragraphes 4.1 et 4.2, nous allons examiner, essentiellement sous l'angle pratique, les résultats de la classification de l'ensemble des eaux minérales.

Le point de départ de l'exemple que nous nous proposons d'analyser est un tableau de données comportant 45 lignes et 6 colonnes (annexe). Chaque ligne correspond à un objet, c'est-à-dire à une eau minérale, et chaque colonne correspond à une variable, c'est-à-dire à la teneur en un élément chimique donné.

Les variables considérées sont quantitatives et toutes exprimées en mg par litre. Toutefois, l'examen des données montre que les ordres de grandeur sont fort différents :la teneur moyenne en HCO_3^- est environ 28 fois plus importante que la teneur moyenne en Mg^{++}, et, ce qui est plus fondamental encore, l'écart-type de la teneur en HCO_3^- est près de 30 fois plus grand que l'écart-type de la teneur en Mg^{++}.

Les différences importantes de variabilité nous semblent inhérentes aux caractéristiques étudiées et non refléter le caractère discriminant des variables. Pour cette raison, il nous paraît justifié de standardiser les variables, ce qui leur donne une importance comparable. L'absence de standardisation donnerait un poids nettement prépondérant à la teneur en HCO_3^- et en Na^+ et un poids très faible à la teneur en Mg^{++} et les résultats de la classification seraient assez différents.

Dans la mesure où nous avons affaire à des variables quantitatives uniquement, nous choisissons l'algorithme de WARD. Les résultats de cette classification sont donnés dans les figures 4 et 5. La première figure provient du logiciel SAS, qui a l'avantage de donner les valeurs de R^2 et la seconde figure provient de Minitab qui fournit, en option, des dendrogrammes plus lisibles que ceux de SAS. Dans ce dendrogramme, les objets sont cependant représentés par leur numéro d'ordre et non pas par le code. L'échelle utilisée pour la représentation des niveaux auxquels ont lieu les fusions correspond au double de l'augmentation de la variance résiduelle globale, comme nous l'avons détaillé au paragraphe 2.4.

On constate que le regroupement des eaux en une vingtaine de classes ne conduit qu'à une très faible perte d'information, puisque, pour 20 classes, R^2

NCL	-Clusters	Joined-	FREQ	RMS STD	SPRSQ	RSQ	BSS	Tie
44	B	l	2	0.0010	0.000000	1.0000	0.0000	
43	N	Q	2	0.0221	0.000011	1.0000	0.0029	
42	A	G	2	0.0383	0.000033	1.0000	0.0088	
41	E	J	2	0.0389	0.000034	0.9999	0.0091	
40	o	p	2	0.0466	0.000049	0.9999	0.0130	
39	R	T	2	0.0487	0.000054	0.9998	0.0142	
38	D	I	2	0.0514	0.000060	0.9998	0.0159	
37	CL39	j	3	0.0519	0.000068	0.9997	0.0181	
36	CL38	g	3	0.0538	0.000071	0.9996	0.0188	
35	CL41	F	3	0.0482	0.000071	0.9995	0.0188	
34	CL44	CL36	5	0.0503	0.000098	0.9994	0.0259	
33	C	e	2	0.0843	0.000162	0.9993	0.0427	
32	CL35	CL43	5	0.0554	0.000162	0.9991	0.0427	
31	CL42	H	3	0.0709	0.000195	0.9989	0.0516	
30	CL37	m	4	0.0706	0.000218	0.9987	0.0575	
29	CL34	CL40	7	0.0676	0.000344	0.9984	0.0908	
28	K	s	2	0.1294	0.000380	0.9980	0.1004	
27	CL31	P	4	0.0956	0.000395	0.9976	0.1042	
26	O	CL30	5	0.0917	0.000425	0.9972	0.1121	
25	i	r	2	0.1706	0.000661	0.9965	0.1746	
24	CL33	CL28	4	0.1430	0.000853	0.9957	0.2251	
23	CL27	q	5	0.1339	0.001006	0.9946	0.2656	
22	L	S	2	0.2203	0.001103	0.9935	0.2912	
21	CL29	CL32	12	0.0913	0.001182	0.9924	0.3121	
20	CL23	f	6	0.1695	0.001635	0.9907	0.4317	
19	M	h	2	0.2720	0.001682	0.9890	0.4439	
18	CL26	CL25	7	0.1585	0.001997	0.9870	0.5273	
17	CL21	CL19	14	0.1442	0.002375	0.9847	0.6270	
16	CL20	d	7	0.2040	0.002408	0.9823	0.6357	
15	CL24	k	5	0.2188	0.002956	0.9793	0.7803	
14	CL22	Y	3	0.3529	0.004557	0.9748	1.2031	
13	CL16	CL15	12	0.2439	0.004844	0.9699	1.2787	
12	X	c	2	0.5868	0.007826	0.9621	2.0660	
11	Z	a	2	0.6551	0.009753	0.9523	2.5747	
10	CL13	CL17	26	0.2387	0.011358	0.9410	2.9984	
9	CL11	n	3	0.7299	0.014460	0.9265	3.8175	
8	CL10	CL18	33	0.2714	0.017787	0.9087	4.6956	
7	CL14	U	4	0.6006	0.018931	0.8898	4.9979	
6	V	b	2	1.2520	0.035623	0.8542	9.4045	
5	CL8	CL9	36	0.3793	0.036650	0.8175	9.6757	
4	CL5	CL7	40	0.4942	0.077466	0.7401	20.4511	
3	W	CL12	3	1.7840	0.136843	0.6032	36.1265	
2	CL4	CL3	43	0.7985	0.247406	0.3558	65.3153	
1	CL2	CL6	45	1.0000	0.355806	0.0000	93.9328	

Figure 4. Résultats de la classification selon le critère de WARD.

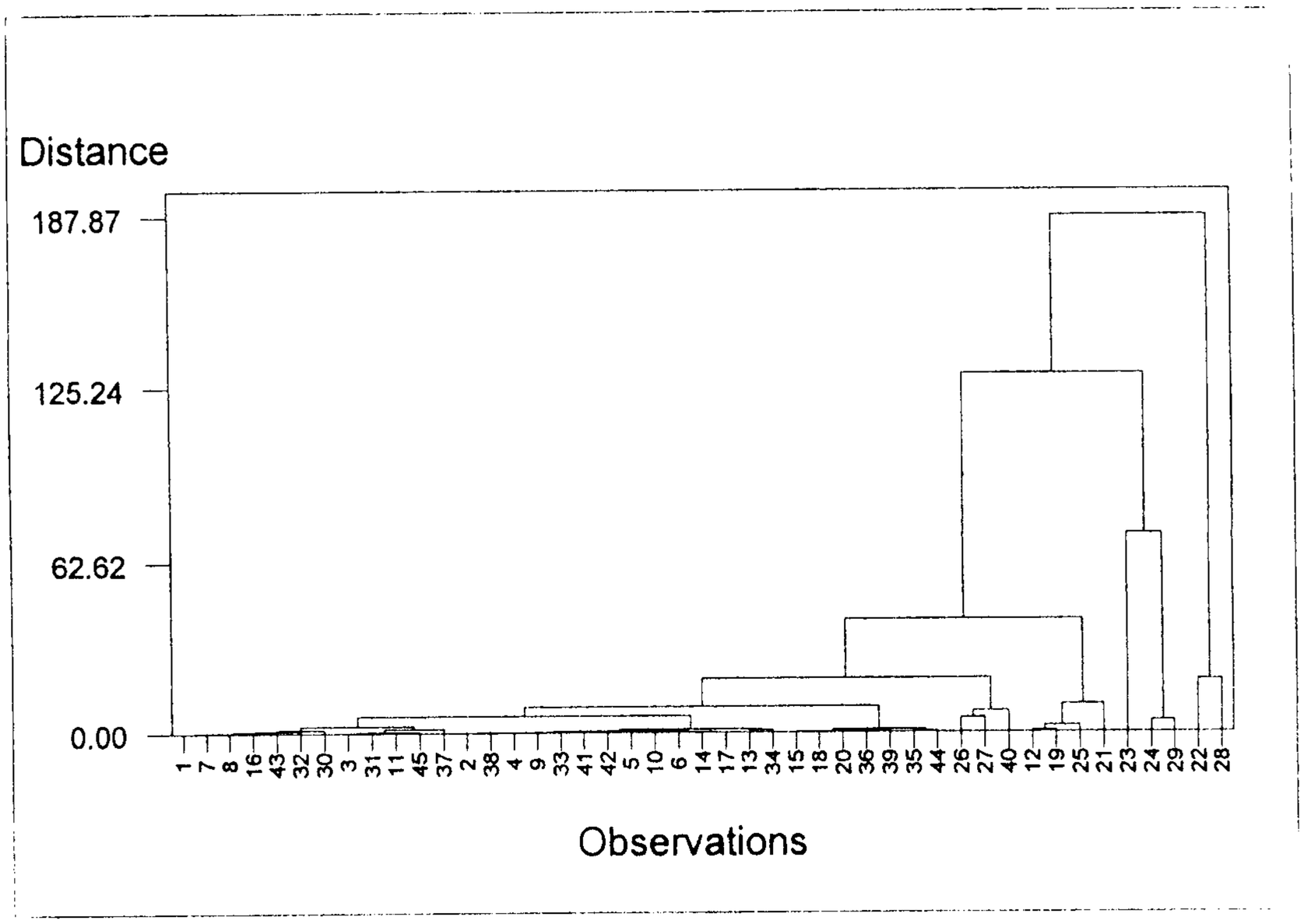

Figure 5. Dendrogramme relatif à la classification selon le critère de WARD.

vaut 0,99. Le regroupement en sept classes conduit à une valeur R^2 de 0,89 et les regroupements ultérieurs donnent lieu à des diminutions nettement plus importantes de R^2.

Le choix du nombre de classes à retenir dépend dans une large mesure des objectifs poursuivis par la classification. Cet objectif pourrait, par exemple, être la sélection d'une vingtaine d'eaux, aussi différentes que possible, qui seraient utilisées dans des études ultérieures. Dans ce cas, on pourrait retenir la partition en 20 classes et, pour les classes contenant plus d'une eau, sélectionner une eau au hasard dans chaque classe.

Si l'objectif est de classer les eaux en un nombre plus réduit de groupes dans le but de faire, du point de vue médical, des recommandations à des patients on pourrait retenir, par exemple, sept classes. Pour cette partition, le tableau 8 reprend la liste des eaux constituant chaque classe. Il donne également la moyenne et, pour les classes de plus d'une eau, l'écart-type de la teneur pour chaque élément chimique. Ces paramètres seront utiles pour la description des groupes (paragraphe 4.2).

On remarque immédiatement que le premier groupe reprend la majorité des eaux (33 eaux), trois groupes comportent de 2 à 4 eaux et trois eaux sont isolées.

Afin de vérifier dans quelle mesure les résultats dépendent de l'algorithme de classification utilisé, nous avons recommencé l'analyse en utilisant les au-

Tableau 8. Composition des sept classes d'eaux : moyennes et écarts-types (valeurs entre parenthèses) des teneurs.

Numéro des classes	Composition des classes	HCO_3^-	SO_4^-	Cl^-	Ca^{++}	Mg^{++}	Na^+
1	A,B,C,D,E,F, G,H,I,J,K,M,N, O,P,Q,R,T,d,e, f,g,h,i,j,k,l,m, o,p,q,r,s	235 (111)	25 (27)	20 (23)	61 (36)	11 (10)	22 (34)
2	Z,a,n	1.196 (778)	80 (87)	142 (102)	44 (27)	22 (18)	212 (211)
3	L,S,U,Y	625 (362)	216 (147)	45 (59)	212 (64)	34 (13)	47 (58)
4	W	386	1.058	6	451	66	8
5	X,c	1.865 (403)	161 (69)	94 (61)	129 (57)	121 (24)	415 (14)
6	V	2.147	48	610	50	8	1.136
7	b	4.263	182	329	78	9	1.744
Ensemble	–	542 (776)	79 (169)	53 (107)	85 (81)	20 (27)	117 (315)

tres méthodes de classification hiérarchiques décrites au paragraphe 3.4, en considérant toujours la partition en sept groupes.

La méthode du lien complet aboutit aux sept groupes décrits ci-dessus. Pour le lien simple, le lien moyen et la méthode basée sur la distance entre centres de gravité, les objets L, S, et Y d'une part et les objets n et a d'autre part, sont incorporés au groupe 1. Pour la méthode de McQUITTY, les objets L, S, Y et n sont incorporés au groupe 1. Enfin, la méthode basée sur les distances médianes fusionne les groupes 1 et 2 mais isole dans un groupe l'objet U.

On retrouve donc des similitudes entre ces différentes classifications. En particulier, toutes les méthodes écartent d'un très grand groupe, éventuellement subdivisible, cinq eaux très particulières (groupes 4, 5, 6 et 7).

D'autre part, nous avons vérifié s'il était possible d'améliorer les résultats de la classification de WARD en utilisant le méthode des centres mobiles. Lorsqu'on utilise comme point de départ les sept groupes formés par la méthode de WARD, aucun transfert d'objets n'a lieu.

A titre d'information, signalons aussi que nous avons utilisé la méthode des centres mobiles en considérant comme centres de gravité initiaux les coordonnées des sept premiers objets. Les résultats de cette classification sont assez différents et nettement moins bons que ceux présentés ci-dessus. En effet, pour la partition

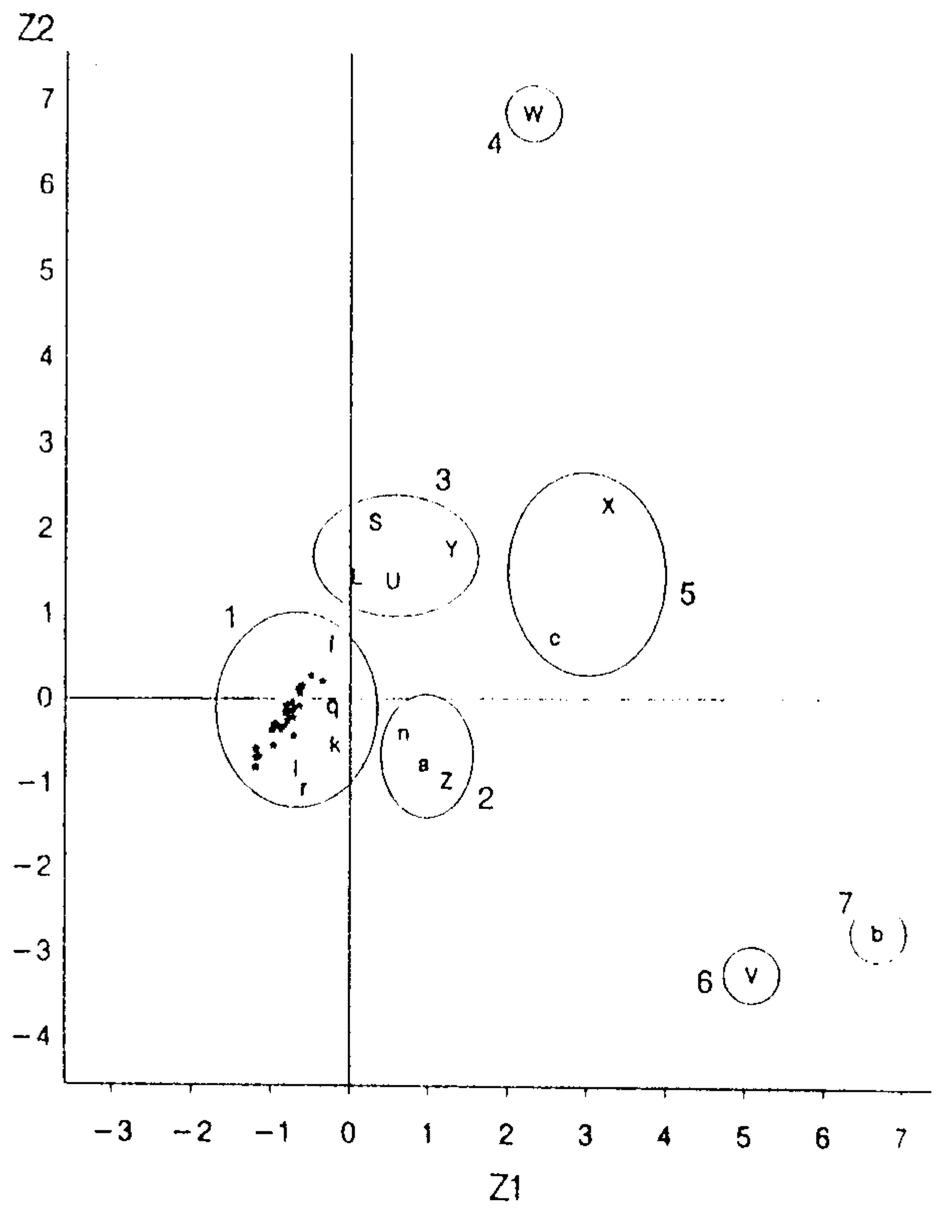

Figure 6. Graphique des objets dans le plan factoriel (z_1, z_2).

obtenue par la classification selon la méthode de WARD, la valeur de R^2 est égale à 0,89 (figure 4), alors que pour la partition obtenue à partir des centres de gravité correspondant aux sept premières observations, R^2 ne vaut que 0,78 : la variance résiduelle est donc deux fois plus importante dans le dernier cas que dans le premier. Ceci montre bien que la méthode des centres mobiles conduit à un minimum local de la variance résiduelle, ce minimum étant fonction des centres de gravité initiaux (paragraphe 3.4).

4.2. Description des groupes

Pour visualiser autant que possible les résultats de la classification, les données centrées réduites ont été soumises à une analyse en composantes principales. Les trois premières composantes expliquent respectivement 45, 34 et 11 pourcents de la variance. Le pourcentage cumulé, pour les trois composantes, est donc de 90 %, ce qui signifie qu'une bonne représentation des eaux peut être faite dans un espace à trois dimensions (z_1, z_2 et z_3), c'est-à-dire aussi qu'il suffit d'examiner les deux premiers plans factoriels (figures 6 et 7). Pour faciliter l'interprétation, nous avons reporté, sur ces plans factoriels les groupes décrits au tableau 8. On notera aussi que toutes les eaux du groupe 1 ne peuvent pas être identifiées par leur symbole du fait de la trop forte densité de points.

On remarque immédiatement la position très excentrique des quatre petits groupes (4, 5, 6 et 7) et la position nettement plus centrale des groupes 1, 2 et 3.

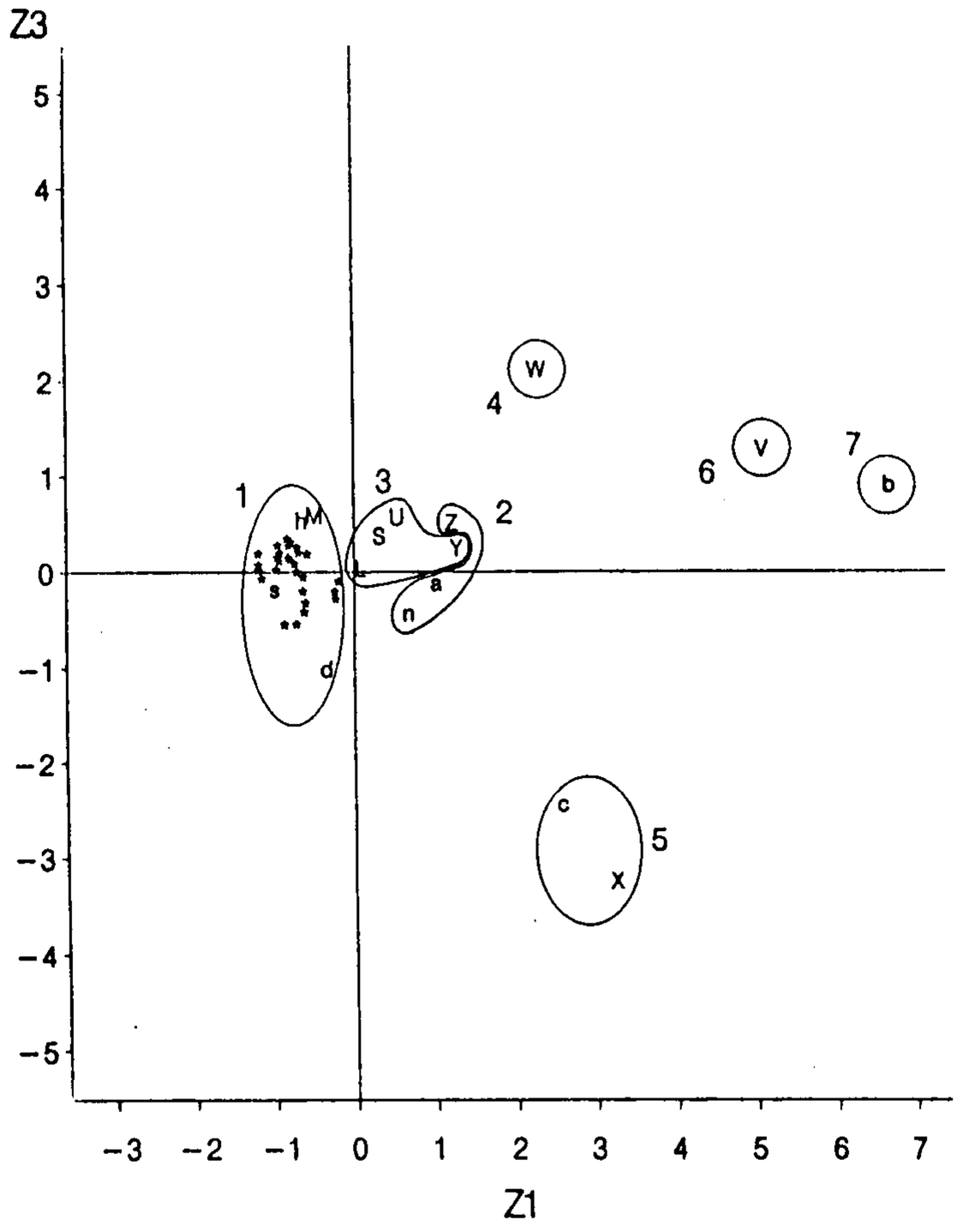

Figure 7. Graphique des objets dans le plan factoriel (z_1, z_3).

Tableau 9. Corrélation des variables initiales avec les trois premiers axes factoriels.

Variables	Axe 1	Axe 2	Axe 3
HCO_3^-	0,92	− 0,21	− 0,09
SO_4^{--}	0,42	0,79	0,31
Cl^-	0,78	− 0,42	0,17
Ca^{++}	0,31	0,85	0,24
Mg^{++}	0,46	0,57	− 0,66
Na^+	0,89	− 0,39	0,05

Nous n'avons pas repris le tableau des distances entre les centres de gravité des groupes, car les graphiques 6 et 7 mettent bien en évidence les distances relativement faibles entre les trois premiers groupes et entre les groupes 6 et 7, et les distances relativement grandes entre le groupe 4, d'une part, et les groupes 6 et 7, d'autre part.

Le tableau 9 donne les corrélations entre les premières composantes principales et les variables initiales. Le premier axe, corrélé positivement avec toutes

Eaux faiblement minéralisées	groupe 1
Eaux moyennement minéralisées et	
plutôt pauvres en SO_4^{--}, Ca^{++} et Mg^{++}	
et riches en Na^+ et Cl^-	groupe 2
plutôt riches en SO_4^{--}, Ca^{++} et Mg^{++}	
et pauvres en Na^+ et Cl^-	groupe 3
Eaux fortement minéralisées et	
très riches en SO_4^{--} et pauvres en Na^+ et Cl^-	groupe 4
très riches en Mg^{++}	groupe 5
très riches en Na^+ et Cl^- et	
riches en HCO_3^-	groupe 6
très riches en HCO_3^-	groupe 7

Figure 8. Schéma d'interprétation des groupes.

les variables peut être interprété comme un axe de richesse en ions et surtout en HCO_3^-, en Cl^- et en Na^+. Les eaux des groupes 4, 5, 6 et 7 sont des eaux assez fortement à très fortement minéralisées. Les eaux du groupe 1 sont, dans l'ensemble, des eaux faiblement minéralisées et les eaux des groupes 2 et 3 sont des eaux moyennement minéralisées. Le deuxième axe est corrélé positivement à la teneur en SO_4^{--}, en Ca^{++} et en Mg^{++} et négativement à la teneur en Na^+ et Cl^-. L'eau W est particulièrement riche en SO_4^{--} en Ca^{++} et en Mg^{++} et pauvre en Na^+ et en Cl^-, alors que pour V et b on observe la situation inverse. Le troisième axe est corrélé négativement à la teneur en Mg^{++}. Il met en évidence la richesse en Mg^{++} des eaux c et X. En fonction de ces constatations et en examinant le tableau 8, on peut arriver au schéma d'interprétation donné dans la figure 8.

Afin de vérifier que les quelques eaux très particulières n'influencent pas de façon trop importante les résultats de la classification, nous avons recommencé les calculs après élimination des cinq eaux appartenant aux groupes 4, 5, 6 et 7. Dans ces conditions, l'algorithme de WARD a conduit à une partition en trois classes qui correspondent exactement aux classes 1, 2 et 3 de la classification réalisée sur l'ensemble des objets. La valeur R^2 associée à cette partition est cependant nettement plus faible (0,50 au lieu de 0,89). De même la diminution des sommes des carrés des écarts liée à la fusion de deux groupes est modifiée. Ainsi par exemple, pour l'ensemble des eaux, la fusion de U avec le groupe (L, S, Y) donne lieu à une réduction de la somme des carrés des écarts entre les groupes

de 5,00 (figure 4) tandis que, après suppression des cinq eaux, la même fusion donne lieu à une réduction de la somme des carrés des écarts entre les groupes de 18,4. Cette discordance s'explique par la standardisation des données qui ramène à l'unité les variances des variables, quelle que soit la dispersion initiale.

De la comparaison des résultats, on peut donc conclure que la prise en considération des eaux de nature un peu particulière ne fait qu'augmenter le nombre de classes mais ne modifie pas fondamentalement les regroupements.

5. CONCLUSIONS

Dans cette note, nous avons présenté quelques méthodes de classification et nous avons, à plusieurs reprises, insisté sur la variété des techniques disponibles. L'utilisateur doit donc faire une série de choix qui, dans une certaine mesure, conditionnent les résultats. Les méthodes de classification numériques, appelées aussi méthodes de classification automatiques, ne sont en fait automatiques qu'à partir du moment où ces choix ont été réalisés.

Nous avons vu également que ces choix reposent souvent sur la connaissance du problème pratique à traiter et notamment sur la définition des objectifs poursuivis. Ces choix doivent, par conséquent, être réalisés par le praticien lui-même.

On notera aussi que, dans beaucoup de situations, les procédures de classification sont essentiellement des techniques descriptives pour des données multivariées et les solutions fournies par la classification numérique devraient conduire à un réexamen de la matrice des données plutôt qu'à une simple acceptation des groupes obtenus [EVERITT, 1993]. La classification peut, par exemple, faire apparaître des regroupements inattendus conduisant à de nouvelles hypothèses de travail. Elle peut aussi ne pas conduire à des regroupements attendus, faisant ainsi ressortir le faible pouvoir séparateur des paramètres utilisés. Dans l'optique de ce réexamen des données, le recours à d'autres méthodes d'analyse multivariée, comme par exemple l'analyse en composantes principales, peut être très utile.

Comme nous l'avons déjà signalé à plusieurs reprises, nous nous sommes limités à la présentation de quelques techniques de classification. Beaucoup d'autres méthodes ont été proposées. Des informations à leur sujet et des points de départ pour des études bibliographiques plus approfondies peuvent être trouvées dans de nombreux ouvrages d'analyse multivariée et dans les livres spécifiquement consacrés à la classification numérique, tels que, par exemple, ceux de CELEUX *et al.* [1989], CHANDON et PINSON [1981], EVERITT [1993], HARTIGAN [1975], JARDINE et SIBSON [1971].

Enfin, en ce qui concerne les deux logiciels utilisés, on retiendra surtout la facilité d'utilisation de Minitab et la possiblité qu'offre ce logiciel d'obtenir en option des dendrogrammes très lisibles. On regrettera cependant son manque de souplesse par rapport à SAS du point de vue des critères utilisés pour quantifier les niveaux auxquels s'effectuent les fusions.

BIBLIOGRAPHIE

BAAMAL L. [1994]. *Etude des règles d'arrêt en classification numérique* (Thèse de doctorat). Gembloux, Faculté des Sciences agronomiques, 257 p.

BOUROCHE J.M., SAPORTA G. [1980]. *L'analyse des données.* Paris, Presses universitaires de France, 127 p.

CELEUX G., DIDAY E., GOVAERT G., LECHEVALLIER Y., RALAMBODRAINY H. [1989]. *Classification automatique des données.* Paris, Dunod, 285 p.

CHANDON J.L., PINSON S. [1981]. *Analyse typologique. Théories et applications.* Paris, Masson, 254 p.

DAGNELIE P. [1975]. *Analyse statistique à plusieurs variables.* Gembloux, Presses agronomiques, 362 p.

EVERITT B. [1993]. *Cluster analysis.* New York, Wiley, 170 p.

FOGUENNE M.[1994]. *Classification des communes wallonnes selon leur ruralité* (Travail de fin d'études). Gembloux, Faculté des Sciences agronomiques, 93 p.

HARTIGAN J.A. [1975]. *Clustering algorithms.* New York, Wiley, 351 p.

JARDINE N., SIBSON R. [1971]. *Mathematical taxonomy.* New York, Wiley, 286 p.

LANGE B. [1982]. *Contribution à l'étude de la localisation des activités agricoles en Belgique* (Thèse de doctorat). Gembloux, Faculté des Sciences agronomiques, 316 p.

LEGENDRE L., LEGENDRE P. [1984]. *Ecologie numérique : la structure des données écologiques.* Paris, Masson, 335 p.

SAS INSTITUTE INC [1989]. *SAS/STAT. User's guide, version 6,* Fourth edition (2 volumes). Cary NC, SAS Institute Inc. 943 + 846 p.

TOMASSONE R., DERVIN C., MASSON J.P. [1993]. *Biométrie : modélisation de phénomènes biologiques.* Paris, Masson, 553 p.

VANDERBORGHT T. [1986]. *L'étude de la variabilité chez le haricot commun* (Phaseolus vulgaris *L.*) *par l'utilisation de méthodes statistiques multivariées appliquées à une banque de données* (Thèse de doctorat). Gembloux, Faculté des Sciences agronomiques, 229 p. + annexes.

X [1994]. *Minitab référence manual, release 10 for Windows* PA State College, Minitab, 984 p.

ANNEXE

Données relatives aux teneurs des eaux minérales en six éléments chimiques, en mg/l.

NUMERO ET NOM	CODE	HCO_3^-	SO_4^{--}	Cl^-	Ca^{++}	Mg^{++}	Na^+
1 AIX-LES-BAINS (F)	A	341	27	3	84	23	2
2 BECKERISH (F)	B	263	23	9	91	5	3
3 CAYRANNE (F)	C	287	3	5	44	24	23
4 CHAMBON (F)	D	298	9	23	96	6	11
5 CRISTAL-ROC (F)	E	200	15	8	70	2	4
6 ST.-CYR (F)	F	250	5	20	71	6	11
7 EVIAN (F)	G	357	10	2	78	24	5
8 FERITA (I)	H	311	14	18	73	18	13
9 ST. HYPPOLITE (F)	I	256	6	23	86	3	18
10 LAURIER (F)	J	186	10	16	64	4	9
11 OGEU (F)	K	183	16	44	48	11	31
12 ONDINE (F)	L	398	218	15	157	35	8
13 PERRIER (F)	M	348	51	31	140	4	14
14 RIBES (E)	N	168	24	8	55	5	9
15 SPA (B)	O	11	65	5	4	1	3
16 THONON (F)	P	332	14	8	103	16	5
17 VERI (E)	Q	196	18	6	58	6	13
18 VILADREAU (E)	R	59	7	6	16	2	9
19 VITTEL (F)	S	402	306	15	202	36	3
20 VOLVIC (F)	T	64	7	8	10	6	8
21 CASERTA (I)	U	1159	8	18	304	17	47
22 VICHY-CATALAN (E)	V	2147	48	610	50	8	1136
23 CONTREXEVILLE (F)	W	386	1058	6	451	66	8
24 FONYOD (H)	X	2150	210	51	169	138	405
25 HARMATVIZ (H)	Y	540	333	133	186	48	129
26 RIOM (F)	Z	1850	180	180	16	4	62
27 VALS (F)	a	1403	39	27	45	21	453
28 ST. YORRE (F)	b	4263	182	329	78	9	1744
29 APOLLINARIS (D)	c	1580	112	137	89	104	425
30 BATNA (ALGERIE)	d	373	29	18	58	43	13
31 BRU (B)	e	209	5	4	23	22	10
32 GRANDS BOIS (F)	f	399	80	22	124	29	6
33 LUCHEUX (F)	g	284	5	10	96	2	7
34 MONTIGNY (F)	h	105	126	33	78	5	28
35 NOPRI (B?)	i	300	63	20	2	1	140
36 NORM/RIEUDRIERE (F)	j	15	0	18	5	3	8
37 POSSOTOME (BENIN)	k	260	10	115	54	19	70
38 ROXANE (L)	l	262	23	9	91	5	4
39 SIDI ALI (MAROC)	m	98	32	19	22	7	27
40 SIDI HARAZEM (MAROC)	n	335	20	220	70	40	120
41 SPONTIN (B)	o	251	32	14	81	11	6
42 ST. JEAN BAPTISTE (F)	p	263	13	12	80	8	8
43 ST. LEGER (F)	q	476	55	36	60	23	92
44 VAL (B?)	r	220	18	75	3	2	125
45 VERA (I)	s	144	15	2	34	13	2

La collection

NOTES DE STATISTIQUE ET D'INFORMATIQUE

réunit divers travaux (documents didactiques, notes techniques, rapports de recherche, publications, etc.) émanant des services de statistique et d'informatique de la Faculté des Sciences agronomiques et du Centre de Recherches agronomiques de Gembloux (Belgique).

Quelques titres récents:

PALM R. [1989]. Quelques éléments de programmation linéaire. *Notes Stat. Inform.* (Gembloux) 89/1, 37 p.

DAGNELIE P. [1989]. Le choix d'une méthode d'analyse statistique et l'examen préliminaire des données. *Notes Stat. Inform.* (Gembloux) 89/2, 17 p.

PALM R. [1990]. La corrélation canonique: principes et application. *Notes Stat. Inform.* (Gembloux) 90/1, 28 p.

CARLETTI G. [1991]. Les micro-ordinateurs: présentation générale. *Notes Stat. Inform.* (Gembloux) 91/1, 27 p.

IEMMA A.F., PALM R. [1992]. Les matrices généralisées et leur utilisation dans le modèle linéaire. *Notes Stat. Inform.* (Gembloux) 92/1, 25 p.

PALM R. [1993]. Les méthodes d'analyse factorielle: principes et applications. *Notes Stat. Inform.* (Gembloux) 93/1, 38 p.

CLAUSTRIAUX J.J., DELVAUX A. [1994]. Traitement de données par le logiciel Minitab. *Notes Stat. Inform.* (Gembloux) 94/1, 20 p.

CLAUSTRIAUX J.J., DELTENRE A. [1994]. La gestion des données avec le logiciel V.M.S. *Notes Stat. Inform.* (Gembloux) 94/2, 15 p.

OGER R. [1994]. Les expériences de rotations: planification et analyse. *Notes Stat. Inform.* (Gembloux) 94/3, 34 p.

PALM R. [1994]. La régression linéaire pondérée: principes et application. *Notes Stat. Inform.* (Gembloux) 94/4, 20 p.

DELTENRE A. [1994]. La gestion de fichiers sur micro-ordinateur. *Notes Stat. Inform.* (Gembloux) 94/5, 23 p.

DELTENRE A., CLAUSTRIAUX J.J., AUSTRAET J. [1995]. La pratique du traitement de texte appliquée à la conception de documents scientifiques. *Notes Stat. Inform.* (Gembloux) 95/1, 27 p.

PRÉVOT H. [1995]. Introduction au logiciel SAS/GRAPH. *Notes Stat. Inform.* (Gembloux) 95/2, 32 p.

Faculté universitaire des Sciences agronomiques
Avenue de la Faculté d'Agronomie 8
5030 GEMBLOUX (Belgique)

D/1996/2371/1