Cah. ORSTOM, sér. Pédol., vol. VI, n° 1. 1968

LA RECHERCHE DES VARIABLES LES PLUS UTILES A LA MESURE DES DISTANCES D^2

par Raymond VAN DEN DRIESSCHE * et Marie-Claire MASBOU **

SUMMARY

The best subset out of v characters ($7 \leqslant v \leqslant 20$) for Mahalanobis' D^2 measurements between m groups ($3 \leqslant m \leqslant 30$) of observations can be located by a generalized use of Rao's F test for additional information D^2_v versus D^2_{v-1}.

The F tests are in fact applied to each of the $\binom{v}{v-1}\binom{m}{2}$ distances measured with all combinations of v characters taken $v-1$ at a time. A Cochran Q test for the distances × combinations two-way table of zeros (0 for a non significant F) and ones (1 for a significant F) provides a method for testing wether the v combinations differ significantly. Moreover, a distribution-free analysis of variance (Friedman's χ^2_r) of the groups × combinations two-way table formed by the group frequencies of non significant F tests determines wether the R rank totals per combination differ. As the number of combinations always exceeds 6, χ^2_r is distributed approximately as χ^2. When both tests are significant at the.05 level, the greatest R identifies the best combination of characters and, subsequently, a useless character.

The procedure can be repeated on the $\binom{v-1}{v-2}\binom{m}{2}$ distances computed with the $v-1$ best characters taken $v-2$ at a time. One makes all the F tests for additional information D^2_v versus D^2_{v-2} and the Q and χ^2_r tests, if significant, locate a second useless character.

At a given step, the highest R may be present under several combinations of characters. This case of tied highest R does not mean that all the 'useless' characters are to be deleted at the same time. The more conservative procedure of dropping a single character at a time, eventually at random, seems actually preferable to guard against possible interactions between characters.

Soil data from the Niger are treated to illustrate the procedure.

---

* Maître de recherche, Section pédologie, Services scientifiques centraux de l'ORSTOM, 93-Bondy, France.

** Analyste-programmeur, Service de calcul, Institut Blaise Pascal du CNRS, Paris.

RÉSUMÉ

La recherche des variables les plus utiles à la mesure des distances généralisées D^2 de Mahalanobis entre m groupes ($3 \leqslant m \leqslant 30$) repose sur la généralisation du test F de l'information additionnelle (Rao, 1965) des $\binom{m}{2} D_v^2$ par rapport aux $\binom{v}{v-1}\binom{m}{2} D_{v-1}^2$ calculés avec toutes les combinaisons des v variables $v-1$ à $v-1$ ($7 \leqslant v \leqslant 20$). Le tableau des 0 (F non significatifs) et des 1 (F significatifs) soumis au test Q de Cochran permet d'éprouver la différence entre les combinaisons de variables. Une analyse de la variance de Friedman appliquée au tableau des effectifs intra-groupe de tests F non significatifs permet, en outre, lorsque le χ_r^2 est significatif, de désigner la combinaison la plus utile sur la base de la somme des rangs R maximale.

La même méthode est suivie pour désigner la combinaison la plus utile des $v-1$ combinaisons des $v-1$ variables $v-2$ à $v-2$ sur la base de $\binom{v-1}{v-2}\binom{m}{2}$ tests F, d'un test Q, et d'un χ_r^2 significatifs.

Le rejet des variables se poursuit ainsi jusqu'à l'obtention d'un Q ou d'un χ_r^2 non significatif. Le nombre minimal de variables est fixé à 6 pour que la distribution de χ_r^2 ne s'écarte pas trop de celle de χ^2.

En présence de R maximaux ex aequo, il semble prudent de ne pas écarter, d'emblée, toutes les variables « inutiles » et de suivre le processus habituel qui consiste à rejeter une seule variable à la fois.

Des données d'argile, de sable grossier, de carbone, d'azote, de fer total, de calcium et de magnésium échangeables, de capacité d'échange, de pH, transformées en logarithmes décimaux, provenant de sols ferrugineux tropicaux du Niger, servent d'exemple numérique simple. Trois profondeurs de prélèvement et deux secteurs de prélèvement, séparés par une centaine de kilomètres, constituent les 6 groupes. Les effectifs intra-groupe sont de 33 et de 37 échantillons. Trois combinaisons ex aequo se dégagent après la première série de tests ; combinaisons sans sable, sans calcium, sans magnésium. Les 8 variables de la combinaison sans sable donnent, ensuite, sous forme de 8 combinaisons de 7 variables, de nouvelles distances intergroupes qui sont éprouvées par rapport aux distances d'origine à 9 variables. Le test de Cochran et celui de Friedman sont significatifs et le R maximal est donné par trois combinaisons ex aequo : sans calcium, sans magnésium, sans capacité d'échange. L'élimination du calcium, après celle du sable, entraîne le calcul des distances avec 7 combinaisons de 6 variables. Les tests Q et χ_r^2 sont significatifs et le R maximal provient de la combinaison sans magnésium. Cet élément peut donc aussi être négligé. Les variables qui semblent les plus intéressantes des 9 variables étudiées pour dissocier les échantillons superficiels, moyens et profonds des sols ferrugineux tropicaux lessivés du centre-ouest et du centre-est du Niger sont donc l'argile, le carbone, l'azote, le fer total, la capacité d'échange et le pH KCl.

INTRODUCTION

La méthode des distances généralisées D^2 de MAHALANOBIS (1925 paru en 1927, 1930, 1936, 1948), utilisée pour dissocier des groupes d'observations et dont il a déjà été question dans les Cahiers de Pédologie (VAN DEN DRIESSCHE et MAIGNIEN, 1965), est d'une grande rigueur, car elle fait appel aux corrélations totales d'ensemble et aux moyennes centrées réduites intra-groupe ; car elle transforme, en outre, les variables corrélées en variables indépendantes. L'obtention des D^2 est toutefois rendue laborieuse par les

nombreuses étapes de calcul et la vérification préalable des conditions de validité : normalité des variables et homogénéité des variances intra-groupe.

Les variables obtenues dans les laboratoires sont coûteuses ; celles qui sont notées sur le terrain manquent de précision. Aussi est-il nécessaire de mettre au point des méthodes pour dépister les variables les moins utiles, les variables qui séparent et identifient mal les groupes ; étant entendu que la notion d'utilité couvre les $\binom{m}{2}$ distances entre les m groupes.

MÉTHODE

Lorsque v variables ($7 \leqslant v \leqslant 20$) sont disponibles pour mesurer les distances D^2 entre m groupes ($3 \leqslant m \leqslant 30$), une méthode peut être proposée. Elle consiste à dépister la moins utile des v variables et à répéter l'opération, successivement sur $v-1$, $v-2$, ..., w variables. L'abandon de l'une des variables entraîne le calcul des D^2 sur les $v-1$ variables restantes. Cette nouvelle matrice des D^2_{v-1} est comparable à toutes les matrices de D^2_{v-1} obtenues sur $v-1$ variables. Or, il y a v combinaisons sans répétition de v variables $v-1$ à $v-1$. Avec ces v combinaisons on calcule v matrices de D^2_{v-1}.

Un test de l'hypothèse de nullité selon laquelle les $v-w$ variables abandonnées d'un calcul de distance unique D^2, entre deux groupes d'effectifs n_1 et n_2 ($n_1+n_2>v+1$), n'apportent aucune séparation ou discrimination supplémentaire est donné par RAO (1965, p. 482). Il s'agit du test de l'information additionnelle des D^2_v par rapport aux D^2_w :

$$F = \frac{n_1+n_2-v-1}{v-w} \cdot \frac{n_1 n_2 (D^2_v - D^2_w)}{(n_1+n_2)(n_1+n_2-2)+n_1 n_2 D^2_w}$$

Le test est significatif lorsque F dépasse la valeur critique donnée par les tables à $v-w$ et n_1+n_2-v-1 degrés de liberté au risque 5 %. Il intéresse la seule distance mesurée entre deux groupes et est étendu, ici, à toutes les distances mesurées entre m groupes. Du fait que $\binom{v}{v-1}$ matrices de D^2_{v-1} sont calculées, il y a $\binom{v}{v-1}\binom{m}{2}$ tests F. Si k et l désignent deux groupes quelconques d'effectifs n_k et n_l ($n_k+n_l>v+1$), chaque test

$$F = (n_k+n_l-v-1) \cdot \frac{n_k n_l (D^2_v - D^2_{v-1})}{(n_k+n_l)(n_k+n_l-2)+n_k n_l D^2_{v-1}}$$

est fait avec 1 et n_k+n_l-v-1 degrés de liberté. Dans un tableau, dont les lignes sont les distances G1G2, G1G3, ... GkGl, ..., G$m-1$Gm, et les colonnes les combinaisons de variables 1, 2, ..., v, on introduit le résultat de chaque test : le chiffre 0 pour un test non significatif, le chiffre 1 pour un test significatif. Non significatif, le test de RAO traduit, pour la distance en cause, l'inutilité de la variable exclue de la combinaison. Un résultat significatif indique que l'abandon de la variable donne une distance différente de celle qui est mesurée avec toutes les variables ; la variable est utile.

Une analyse de ce tableau de 0 et 1 est tentée. L'hypothèse de nullité : « il n'y a pas de différence entre les combinaisons pour l'ensemble des distances » est éprouvée par le test Q de COCHRAN :

$$Q = \frac{(v-1)[v\Sigma M^2 - (\Sigma M)^2]}{v\Sigma L - \Sigma L^2}$$

dans lequel M est le nombre de tests non significatifs par combinaison, L le nombre de tests non significatifs par distance et v le nombre de combinaisons. Le test Q (in SIEGEL, 1956, p. 162) se fait comme χ^2 au risque 5 % avec $v-1$ degrés de liberté. Non significatif, Q ne permet pas de rejeter l'hypothèse d'une absence de différence entre combinaisons de variables ; il constitue une indication de l'absence d'influence de la variable abandonnée de chaque combinaison sur l'ensemble des distances. Un test significatif entraîne le rejet de l'hypothèse de nullité et permet d'identifier la combinaison la plus utile (ou son complément, la variable la moins utile) — pour l'ensemble des distances — en se basant sur le M maximal. En fait, ce test ne donne pas entière satisfaction car la combinaison qui mesure le plus de D^2_{v-1} non significativement différents des D^2_v correspondants donne, parfois, pour un seul des groupes, des D^2_{v-1} significativement plus faibles que les D^2_v.

Il semble par conséquent nécessaire de compter, par groupe, les tests F non significatifs. On se trouve ainsi en présence d'un tableau dont les lignes sont les groupes G1, G2, ... Gk, ... Gl, ..., Gm et les colonnes les combinaisons de variables 1, 2, ... v. Ce tableau d'effectifs 0 à $m-1$ est soumis à l'analyse de la variance de FRIEDMAN ; analyse qui est libérée de la fonction de répartition. L'hypothèse de nullité : « pas d'influence dissemblable des combinaisons de variables sur la mesure des distances à partir de chaque groupe » est faite. Le tableau est remplacé, à cet effet, par un tableau de rangs intra-ligne. Sur chaque ligne, l'effectif le plus faible a le rang 1, le plus élevé le rang v et les effectifs ex aequo la moyenne des rangs correspondants. La somme des rangs par colonne, R, entre dans le test :

$$\chi^2_r = \frac{12\Sigma R^2}{mv(v+1)} - 3m(v+1)$$

et la consultation de la table de χ^2 au risque 5 % se fait avec $v-1$ degrés de liberté (in SIEGEL, op. cit., p. 168). Si χ^2_r n'est pas significatif, toutes les combinaisons se valent et l'identification d'une variable moins utile que les autres est impossible. S'il est significatif, c'est la combinaison au R maximal qui est choisie.

Lorsque le R maximal est donné, dans la même analyse de FRIEDMAN, par plusieurs combinaisons se présentant ex aequo, il est tentant de conserver un ensemble de variables d'où sont exclues toutes les variables ex aequo. Mais la prudence s'impose car ces variables sont reconnues inutiles individuellement. L'inutilité d'une variable n'est peut-être que la conséquence de la présence d'autres variables au sein de la combinaison choisie. Il vaut mieux, par conséquent, n'éliminer qu'une seule variable et attendre les étapes suivantes pour s'assurer du comportement des autres variables ex aequo. En effet, après le choix d'une combinaison et l'abandon d'une première variable on recommence le processus entier de recherche. Les distances D^2_v servent de référence. Il y a $\binom{v-1}{v-2}$ combinaisons de variables, avec lesquelles on calcule $\binom{v-1}{v-2}\binom{m}{2}$ distances D^2_{v-2} qui sont, toutes, soumises au test de RAO :

$$F = \frac{n_k + n_l - v - 1}{v-(v-2)} \cdot \frac{n_k n_l (D^2_v - D^2_{v-2})}{(n_k+n_l)(n_k+n_l-2) + n_k n_l D^2_{v-2}}$$

avec 2 et n_k+n_l-v-1 degrés de liberté. Un test de COCHRAN est appliqué au tableau distances × combinaisons des 0 et 1. Non significatif, Q montre l'impossibilité de choisir une combinaison à $v-2$ variables ; significatif, il fait place à une analyse de la variance de FRIEDMAN du tableau groupes × combinaisons des effectifs de F non significatifs. Un χ_r^2 non significatif arrête la recherche de la plus utile des $v-1$ combinaisons. Un χ_r^2 significatif désigne la combinaison la plus utile.

On en arrive, ainsi, en plusieurs étapes, à extraire du lot de v variables celles dont la disparition n'affecte pas significativement les distances mesurées entre les m groupes.

MÉCANISATION

Le programme est écrit en langage Fortran IV pour ordinateur CDC* 3600. Il couvre, à partir des corrélations totales d'ensemble et des moyennes centrées réduites intra-groupe, le calcul des D^2 et les tests nécessaires au dépistage des variables les moins utiles. L'ordinateur CDC 3600 du CNRS a pour mémoire centrale 2 bancs de 32 768 mots (32 K) chacun (mots de 48 bits) et son cycle mémoire est de 1,4 μs. Le programme et les données de 20 variables dans 30 groupes occupent 14 417+18 900 mots de cette mémoire, en plus des sous-programmes d'usage général. L'utilisation du programme est commandée par quelques cartes contrôle : une carte pour l'identification du travail ; une carte avec le nombre de variables et le nombre de groupes ; une carte renfermant les sigles des variables (dans le format 20 A3) ; une ou deux cartes avec les effectifs intra-groupe (24 I3/6 I3). Les données sont présentées sous forme de cartes renfermant la moitié inférieure de la matrice des corrélations totales d'ensemble (9 F8.5) et de cartes contenant le tableau groupes × variables des moyennes centrées réduites intra-groupe (dans le format 8 F9.4). Il a suffi de 21 secondes, dont 13 secondes de lecture du programme en binaire, pour effectuer l'application qui suit, à 9 variables et 6 groupes. L'exécution proprement dite d'un autre travail, à 10 variables et 18 groupes, a duré 32 secondes.

UNE APPLICATION

Il convient d'essayer la méthode dans une région prospectée par les pédologues de l'ORSTOM. D'anciens bordereaux de laboratoire, renfermant des données d'analyse des profils examinés au Niger par BOULET et GAVAUD (1963) entre le 13e et le 15e degré de latitude N. et entre le 4e et le 8e degré de longitude E., sont utilisés à cet effet**.

Seuls des profils de sols ferrugineux tropicaux lessivés, renfermant à chacune des profondeurs moyennes 5 cm, 75 cm et 190 cm des données d'argile granulométrique (ARG), de sable grossier (SBG), de carbone (C), d'azote (N), de fer total (FE), de calcium échangeable (CAE), de magnésium échangeable (MGE), de capacité d'échange (T) et de pH au chlorure de potassium (PHK), sont transcrits. Un dépouillement préliminaire montre qu'après transformation logarithmique décimale, les données de ces variables

* Control Data Corporation, Minneapolis.

** La diffusion des bordereaux normalisés remonte à décembre 1965.

se distribuent, aux trois profondeurs, selon des courbes unimodales, légèrement dissymétriques, dont les variances sont homogènes. Ces profils, au nombre de 70, se répartissent en deux secteurs, séparés par un degré de longitude (107 km). Les secteurs sont d'effectif 33 profils à l'ouest et 37 profils à l'est. Six groupes sont ainsi considérés :

G1 33 échantillons de surface à l'ouest
G2 37 échantillons de surface à l'est
G3 33 échantillons mi-profonds à l'ouest
G4 37 échantillons mi-profonds à l'est
G5 33 échantillons profonds à l'ouest
G6 37 échantillons profonds à l'est

Les 9 variables, disponibles sous forme de corrélations totales d'ensemble (tableau I) et de moyennes centrées réduites intra-groupe (tableau II), servent au calcul des 15 distances D_9^2 entre les 6 groupes.

TABLEAU I

Corrélations totales d'ensemble

Pooled estimates of correlation

−0.54295							
0.58296	−0.41209						
0.53294	−0.36480	0.74752					
0.62900	−0.38453	0.51490	0.50015				
0.36863	−0.10368	0.34731	0.34794	0.56345			
0.40073	−0.23807	0.34895	0.33709	0.46563	0.49454		
0.61922	−0.35353	0.50836	0.51375	0.63241	0.67939	0.62920	
0.03458	0.05160	0.04664	0.02774	0.27064	0.52976	0.23305	0.12970

TABLEAU II

Moyennes centrées réduites intra-groupe

Normalized mean values of characters

−0.6426	0.0166	1.6553	0.6567	0.0716	0.2185	−0.4597	−0.4321
0.6096							
−1.0938	0.2236	1.2334	0.7367	−1.1015	0.4652	0.0473	−0.1833
1.3630							
0.8238	−0.1715	0.0033	−0.1430	0.8223	−0.5365	−0.2654	0.0510
−1.0182							
0.0196	0.2586	−0.5919	−0.2103	−0.4139	0.0541	0.2860	0.3231
−0.3960							
0.7039	−0.3650	−0.7870	−0.5563	0.8824	−0.3037	−0.2052	−0.2214
−0.2752							
0.1891	0.0376	−1.5131	−0.4839	−0.2608	0.1023	0.5970	0.4627
−0.2832							

Ces distances sont présentées dans l'ordre croissant et par groupe (tableau III). Mesurées, à la même profondeur, entre les profils du secteur ouest et ceux du secteur est, les distances sont faibles, mais augmentent avec la profondeur (G1G2 = 5,25 ; G3G4 = 5,69 et G5G6 = 6,48). Les distances sont grandes entre les échantillons de surface et les échantillons extraits à mi-profondeur, tant à l'ouest qu'à l'est (G1G3 = 15,97 et G2G4 = 15,71). Elles sont, au contraire, très faibles, à l'ouest comme à l'est, entre les échantillons mi-profonds et profonds (G3G5 = 1,84, et G4G6 = 2,16).

TABLEAU III

Distances généralisées D_9^2 entre les 6 groupes avec les 9 variables

Values of D_9^2 based on 9 characters

clay and coarse sand

carbon and nitrogen

total iron

exchangeable calcium and magnesium

cation exchange capacity

pH

Distances measured between 6 groups (3 depths × 2 sectors) of data from leached ferruginous tropical soils in the Niger

G1		G2		G3		G4		G5		G6	
G2	5,25	G1	5,25	G5	1,84	G6	2,16	G3	1,84	G4	2,16
G4	15,10	G4	15,71	G4	5,69	G3	5,69	G4	6,19	G5	6,48
G3	15,97	G6	26,13	G6	8,95	G5	6,19	G6	6,48	G3	8,95
G5	20,08	G3	26,85	G1	15,97	G1	15,10	G1	20,08	G2	26,13
G6	26,71	G5	29,65	G2	26,85	G2	15,71	G2	29,65	G1	26,71

Les 9 variables sont introduites dans les $\binom{9}{8}$ combinaisons sans répétition des 9 variables 8 à 8. Les 15 distances D_8^2 sont calculées avec chacune des 9 combinaisons. Les 9 séries de D_8^2 sont comparables, le nombre de variables étant le même. Elles figurent, arrondies à la première décimale, dans le tableau IV. Les remarques faites à l'examen du tableau III restent valables pour chacune des séries, tant à propos de l'augmentation avec la profondeur des distances faibles entre échantillons prélevés à l'ouest et à l'est (G1G2, G3G4, G5G6), qu'à propos des grandes distances entre échantillons de surface et échantillons mi-profonds (G1G3 et G2G4) ou, encore, des distances très faibles entre échantillons mi-profonds et profonds (G3G5 et G4G6 du tableau IV). Il s'agit toutefois de voir si les distances D_8^2 sont significativement différentes des distances D_9^2 correspondantes. Ainsi, l'omission de FE, dans la cinquième combinaison de 8 variables, donne une distance G5G6 de 3,5 et la question se pose de savoir si la distance 3,5 est significativement différente de la distance 6,5 mesurée avec les 9 variables. Le test F de l'information additionnelle

TABLEAU IV

Distances D_8^2 avec les 9 combinaisons à 8 variables

Values of D_8^2 based on the 9 subsets of 8 characters

	1	2	3	4	5	6	7	8	9
G1G2	5,3	5,2	5,0	4,7	1,9	5,2	5,1	4,7	4,2
G1G3	11,9	15,6	10,1	16,0	14,7	15,7	15,9	15,9	14,4
G1G4	12,4	15,0	6,3	14,7	14,7	15,1	14,5	14,5	14,5
G1G5	15,4	20,1	9,1	20,1	18,6	20,0	19,9	20,1	19,4
G1G6	22,8	26,7	8,2	25,6	26,5	26,7	25,6	26,0	26,2
G2G3	22,8	26,2	23,2	26,3	18,3	26,7	26,8	26,6	21,5
G2G4	13,0	15,4	9,7	15,7	14,4	15,7	15,6	15,7	12,5
G2G5	24,9	29,6	21,8	29,3	20,4	29,6	29,6	28,9	26,2
G2G6	22,1	26,1	11,8	26,0	24,4	26,1	25,8	26,1	23,1
G3G4	5,5	5,6	5,4	5,2	2,5	5,5	5,4	5,4	5,4
G3G5	1,8	1,6	1,1	1,8	1,8	1,8	1,8	1,7	1,6
G3G6	9,0	8,6	5,4	7,8	6,4	8,8	8,3	8,6	8,6
G4G5	5,9	6,1	6,1	6,0	2,6	6,1	6,1	5,4	6,2
G4G6	2,0	2,1	0,4	2,0	2,1	2,2	2,1	2,2	2,2
G5G6	6,4	6,5	5,5	5,7	3,5	6,5	6,0	5,6	6,5

TABLEAU V

Tests F de l'information additionnelle des D_9^2 par rapport aux D_8^2

Results of the $\binom{9}{8}\binom{6}{2}$ variance ratio F tests for additional information D_9^2 versus D_8^2

	ARG 1	SBG 2	C 3	N 4	FE 5	CAE 6	MGE 7	T 8	PHK 9	TESTS NON SIGNIFICATIFS
1 G1G2	0	0	0	0	1	0	0	0	1	7
2 G1G3	1	0	1	0	0	0	0	0	1	6
3 G1G4	1	0	1	0	0	0	0	0	0	7
4 G1G5	1	0	1	0	0	0	0	0	0	7
5 G1G6	1	0	1	0	0	0	0	0	0	7
6 G2G3	1	0	1	0	1	0	0	0	1	5
7 G2G4	1	0	1	0	1	0	0	0	1	5
8 G2G5	1	0	1	0	1	0	0	0	1	5
9 G2G6	1	0	1	0	0	0	0	0	1	6
10 G3G4	0	0	0	0	1	0	0	0	0	8
11 G3G5	0	0	1	0	0	0	0	0	0	8
12 G3G6	0	0	1	1	1	0	0	0	0	6
13 G4G5	0	0	0	0	1	0	0	1	0	7
14 G4G6	0	0	1	0	0	0	0	0	0	8
15 G5G6	0	0	1	1	1	0	0	1	0	5
TESTS NON SIGNIFICATIFS	7	15	3	13	7	15	15	13	9	97

de $D_9^2 = 6{,}5$ par rapport à $D_8^2 = 3{,}5$ pour les deux groupes G5 et G6 d'effectifs 33 et 37 s'écrit :

$$F = (33+37-9-1) \cdot \frac{33 \times 37(6{,}5-3{,}5)}{(33+37)(33+37-2)+33 \times 37 \times 3{,}5} = 24{,}3$$

24,3 dépasse la valeur critique 4,0 donnée par les tables de F au risque 5 % avec 1 et 60 degrés de liberté. F est donc significatif et l'absence de FE donne une distance significativement plus courte entre les deux lots d'échantillons profonds. Le chiffre 1, par opposition à 0, est utilisé pour signaler que le test est significatif. Ce 1 figure à l'intersection de la 15e ligne et de la 5e colonne du tableau V, qui renferme $\binom{9}{8}\binom{6}{2}$ résultats de tests D_9^2 versus D_8^2. Le test de COCHRAN, appliqué au tableau V, a pour but de déceler les différences significatives entre les 9 combinaisons de 8 variables ; il a pour expression :

$$Q = \frac{8[9(7^2+15^2+\ldots+9^2)-97^2]}{(9 \times 97)-(7^2+6^2+\ldots+5^2)} = 49{,}1$$

49,1 dépasse la valeur critique 15,5 donnée par les tables de χ^2 au risque 5 % avec 8 degrés de liberté. Q est donc significatif et les combinaisons de 8 variables diffèrent significativement. On remarque que trois combinaisons n'entraînent aucune modification dans la mesure des distances ; il s'agit des combinaisons dont sont respectivement absentes les variables SBG, CAE, MGE. On peut penser qu'il n'en est pas toujours ainsi et que l'homogénéité des données utilisées y est pour beaucoup. Aussi est-il bon de passer à l'analyse de la variance des effectifs intra-groupe de tests non significatifs, tels qu'ils apparaissent dans le tableau VI aux 6 lignes (les 6 groupes) et 9 colonnes (les 9 combinaisons de variables). On sait qu'il faut,

TABLEAU VI

Effectifs intra-groupe de tests F non significatifs D_9^2 par rapport aux D_8^2

Group frequencies of non significant F tests D_9^2 versus D_8^2

	ARG 1	SBG 2	C 3	N 4	FE 5	CAE 6	MGE 7	T 8	PHK 9
G1	1	5	1	5	4	5	5	5	3
G2	1	5	1	5	1	5	5	5	0
G3	3	5	1	4	2	5	5	5	3
G4	3	5	2	5	2	5	5	4	4
G5	3	5	1	4	2	5	5	3	4
G6	3	5	0	3	3	5	5	4	4

sur chaque ligne, attribuer un rang aux effectifs pris dans l'ordre croissant et que les effectifs ex aequo reçoivent des rangs moyens. Ainsi, pour le sixième groupe, le rang 1 est attribué à l'effectif 0 de la combinaison sans C (dont le rôle semble important) ; les rangs 2,3, 4, sous forme de moyenne 3, aux combinaisons ex aequo sans ARG, sans N et sans FE ; la moyenne des rangs 5 et 6 (soit 5,5) aux combinaisons ex aequo sans T et sans PHK ; enfin, la moyenne des rangs 7, 8, 9 (soit 8) aux combinaisons ex æquo sans SBG,

TABLEAU VII

Rangs intra-ligne des effectifs de F non significatifs D_9^2 par rapport aux D_8^2
Within-line ranking for the frequencies of non significant F tests D_9^2 versus D_8^2

	ARG 1	SBG 2	C 3	N 4	FE 5	CAE 6	MGE 7	T 8	PHK 9
G1	1,5	7,0	1,5	7,0	4,0	7,0	7,0	7,0	3,0
G2	3,0	7,0	3,0	7,0	3,0	7,0	7,0	7,0	1,0
G3	3,5	7,5	1,0	5,0	2,0	7,5	7,5	7,5	3,5
G4	3,0	7,5	1,5	7,5	1,5	7,5	7,5	4,5	4,5
G5	3,5	8,0	1,0	5,5	2,0	8,0	8,0	3,5	5,5
G6	3,0	8,0	1,0	3,0	3,0	8,0	8,0	5,5	5,5
R	17,5	45,0	9,0	35,0	15,5	45,0	45,0	35,0	23,0

sans CAE, sans MGE. Du tableau des rangs intra-ligne (tableau VII), les totaux par colonne, R, entrent dans le calcul du χ_r^2 de FRIEDMAN :

$$\chi_r^2 = \frac{12(17{,}5^2 + 45{,}0^2 + \ldots + 23{,}0^2)}{6 \times 9 \times 10} - 18 \times 10 = 35{,}1$$

35,1 dépasse la valeur critique 15,5 donnée par les tables de χ^2 au risque 5 % avec 8 degrés de liberté. χ_r^2 est significatif et les combinaisons de variables diffèrent donc, à la fois, dans les distances individuelles D_8^2 qu'elles mesurent, et dans les effectifs intra-groupe de tests F non significatifs des D_9^2 par rapport aux D_8^2. Le choix de la combinaison la plus utile se fait sur la base du R maximal 45,0 qui caractérise, toutefois, trois combinaisons ex aequo : la deuxième sans sable, la sixième sans calcium, la septième sans magnésium.

TABLEAU VIII

Distances D_7^2 avec les 8 combinaisons à 7 variables, après abandon de SBG
Values of D_7^2 based on the 8 subsets of 7 characters, coarse sand excluded

	ARG 1	C 2	N 3	FE 4	CAE 5	MGE 6	T 7	PHK 8
G1G2	5,2	5,0	4,7	1,9	5,2	5,0	4,6	4,1
G1G3	11,9	9,4	15,6	14,5	15,4	15,6	15,6	14,0
G1G4	12,4	5,9	14,6	14,5	15,0	14,4	14,4	14,4
G1G5	15,0	9,0	20,0	18,6	20,0	19,9	20,1	19,4
G1G6	22,4	8,0	25,6	26,4	26,7	25,6	26,0	26,2
G2G3	22,7	22,2	25,6	18,0	26,1	26,1	25,9	20,8
G2G4	13,0	9,0	15,4	14,2	15,4	15,3	15,4	12,1
G2G5	24,7	21,5	29,2	20,4	29,5	29,6	28,8	26,1
G2G6	22,0	11,4	26,0	24,4	26,1	25,7	26,1	23,0
G3G4	5,5	5,3	5,2	2,5	5,4	5,3	5,3	5,4
G3G5	1,5	0,9	1,5	1,5	1,6	1,5	1,5	1,3
G3G6	8,6	5,3	7,5	6,2	8,6	7,9	8,2	8,3
G4G5	5,7	6,0	5,9	2,4	6,0	6,0	5,3	6,1
G4G6	1,9	0,4	1,9	2,0	2,1	2,0	2,1	2,1
G5G6	6,4	5,5	5,7	3,5	6,5	6,0	5,6	6,5

La deuxième combinaison est retenue, c'est elle qui sert de point de départ à la deuxième étape du processus de recherche.

Les 8 combinaisons des 8 variables ARG, C, N, FE, CAE, MGE, T, PHK, prises 7 à 7, donnent les $\binom{8}{7}\binom{6}{2}$ distances D_7^2 qui entrent dans le tableau VIII.

Les tests F effectués pour déterminer si les D_9^2 apportent une information additionnelle par rapport aux D_7^2 figurent dans le tableau IX. Le test $Q = 46{,}5$ est significatif avec 7 degrés de liberté et l'analyse de la variance de FRIEDMAN appliquée au tableau X des effectifs intra-groupe de F non significatifs donne

TABLEAU IX

Tests F de l'information additionnelle des D_9^2 par rapport aux D_7^2

Results of the $\binom{8}{7}\binom{6}{2}$ *variance ratio F tests for additional information* D_9^2 *vs* D_7^2

	ARG 1	C 2	N 3	FE 4	CAE 5	MGE 6	T 7	PHK 8
G1G2	0	0	0	1	0	0	0	1
G1G3	1	1	0	0	0	0	0	0
G1G4	1	1	0	0	0	0	0	0
G1G5	1	1	0	0	0	0	0	0
G1G6	1	1	0	0	0	0	0	0
G2G3	1	1	0	1	0	0	0	1
G2G4	1	1	0	0	0	0	0	1
G2G5	1	1	0	1	0	0	0	1
G2G6	1	1	0	0	0	0	0	1
G3G4	0	0	0	1	0	0	0	0
G3G5	0	1	0	0	0	0	0	0
G3G6	0	1	1	1	0	0	0	0
G4G5	0	0	0	1	0	0	0	0
G4G6	0	1	0	0	0	0	0	0
G5G6	0	1	0	1	0	0	0	0

TABLEAU X

Effectifs intra-groupe de tests F non significatifs D_9^2 par rapport aux D_7^2

Group frequencies of non significant F tests D_9^2 *versus* D_7^2

	ARG 1	C 2	N 3	FE 4	CAE 5	MGE 6	T 7	PHK 8
G1	1	1	5	4	5	5	5	4
G2	1	1	5	2	5	5	5	0
G3	3	1	4	2	5	5	5	4
G4	3	2	5	3	5	5	5	4
G5	3	1	5	2	5	5	5	4
G6	3	0	4	3	5	5	5	4

TABLEAU XI

Rangs intra-ligne des effectifs de F non significatifs D_9^2 par rapport aux D_7^2

Within-line ranking for the frequencies of non significant F tests D_9^2 versus D_7^2

	ARG 1	C 2	N 3	FE 4	CAE 5	MGE 6	T 7	PHK 8
G1	1,5	1,5	6,5	3,5	6,5	6,5	6,5	3,5
G2	2,5	2,5	6,5	4,0	6,5	6,5	6,5	1,0
G3	3,0	1,0	4,5	2,0	7,0	7,0	7,0	4,5
G4	2,5	1,0	6,5	2,5	6,5	6,5	6,5	4,0
G5	3,0	1,0	6,5	2,0	6,5	6,5	6,5	4,0
G6	2,5	1,0	4,5	2,5	7,0	7,0	7,0	4,5
R	15,0	8,0	35,0	16,5	40,0	40,0	40,0	21,5

TABLEAU XII

Distances D_6^2 avec les 7 combinaisons à 6 variables, après abandon de SBG et CAE

Values of D_6^2 based on the 7 subsets of 6 characters, coarse sand and calcium excluded

	ARG 1	C 2	N 3	FE 4	MGE 5	T 6	PHK 7
G1G2	5,2	4,9	4,6	1,7	5,0	4,6	3,9
G1G3	11,4	9,1	15,4	14,4	15,3	15,4	12,0
G1G4	12,3	5,9	14,6	14,5	14,4	14,2	14,1
G1G5	14,6	8,8	19,9	18,6	19,8	19,8	18,6
G1G6	22,2	7,9	25,6	26,4	25,5	25,9	25,8
G2G3	22,5	22,1	25,5	17,9	26,1	25,4	17,2
G2G4	13,0	9,0	15,4	14,0	15,3	15,4	11,1
G2G5	24,5	21,5	29,2	20,2	29,5	28,3	24,2
G2G6	22,0	11,4	26,0	24,3	25,7	26,0	21,8
G3G4	5,3	5,2	5,0	2,5	5,2	4,5	4,6
G3G5	1,4	0,9	1,5	1,5	1,5	1,5	1,1
G3G6	8,5	5,2	7,3	6,2	7,8	7,6	7,7
G4G5	5,6	5,9	5,8	2,4	5,9	4,5	6,0
G4G6	1,8	0,3	1,9	2,0	2,0	2,1	2,1
G5G6	6,4	5,5	5,7	3,5	6,0	4,9	6,4

$\chi_r^2 = 33{,}8$ également significatif. La somme maximale des rangs 40,0 caractérise trois combinaisons de variables (tableau XI) : la combinaison sans calcium, celle sans magnésium et celle sans capacité d'échange.

La combinaison sans calcium est retenue pour la troisième étape. Sept combinaisons (tableau XII) de 7 variables ARG, C, N, FE, MGE, T, PHK, prises 6 à 6, donnent $\binom{7}{6}\binom{6}{2}$ distances D_6^2 qui sont comparées aux distances D_9^2 de référence (tableaux XIII, XIV et XV). Le caractère nettement significatif du test Q

TABLEAU XIII

Tests F de l'information additionnelle des D_9^2 par rapport aux D_6^2

Results of the $\binom{7}{6}\binom{6}{2}$ variance ratio F tests for additional information D_9^2 vs D_6^2

	ARG 1	C 2	N 3	FE 4	MGE 5	T 6	PHK 7
G1G2	0	0	0	1	0	0	1
G1G3	1	1	0	0	0	0	1
G1G4	1	1	0	0	0	0	0
G1G5	1	1	0	0	0	0	0
G1G6	1	1	0	0	0	0	0
G2G3	1	1	0	1	0	0	1
G2G4	1	1	0	0	0	0	1
G2G5	1	1	0	1	0	0	1
G2G6	1	1	0	0	0	0	1
G3G4	0	0	0	1	0	0	0
G3G5	0	1	0	0	0	0	0
G3G6	0	1	1	1	0	0	0
G4G5	0	0	0	1	0	1	0
G4G6	0	1	0	0	0	0	0
G5G6	0	0	0	1	0	1	0

TABLEAU XIV

Effectifs intra-groupe de tests F non significatifs D_9^2 par rapport aux D_6^2

Group frequencies of non significant F tests D_9^2 versus D_6^2

	ARG 1	C 2	N 3	FE 4	MGE 5	T 6	PHK 7
G1	1	1	5	4	5	5	3
G2	1	1	5	2	5	5	0
G3	3	1	4	2	5	5	3
G4	3	2	5	3	5	4	4
G5	3	2	5	2	5	3	4
G6	3	1	4	3	5	4	4

TABLEAU XV

Rangs intra-ligne des effectifs de F non significatifs D_9^2 par rapport aux D_6^2

Within-line ranking for the frequencies of non significant F tests D_9^2 versus D_6^2

	ARG 1	C 2	N 3	FE 4	MGE 5	T 6	PHK 7
G1	1,5	1,5	6,0	4,0	6,0	6,0	3,0
G2	2,5	2,5	6,0	4,0	6,0	6,0	1,0
G3	3,5	1,0	5,0	2,0	6,5	6,5	3,5
G4	2,5	1,0	6,5	2,5	6,5	4,5	4,5
G5	3,5	1,5	6,5	1,5	6,5	3,5	5,0
G6	2,5	1,0	5,0	2,5	7,0	5,0	5,0
R	16,0	8,5	35,0	16,5	38,5	31,5	22,0

(28,0 avec 6 degrés de liberté) et du test χ_r^2 (26,9 avec 6 degrés de liberté) autorisent le choix de l'unique combinaison qui présente la plus grande somme de rangs (38,5) : la combinaison sans magnésium. Les variables argile, carbone, azote, fer, capacité d'échange et pH sont retenues.

Comme la distribution de χ_r^2 suit approximativement celle de χ^2 pour plus de 5 degrés de liberté, il est préférable d'arrêter, ici, l'exclusion des variables. En pédologie, cela ne constitue nullement un obstacle.

Cette application numérique est choisie en raison de sa simplicité et de l'homogénéité de ses données : mêmes pédologues, même laboratoire, mêmes méthodes physico-chimiques, même groupe de sol et même région. L'importance des effectifs et la transformation des données assurent, d'autre part, un maximum de validité à l'application des distances généralisées D^2 de MAHALANOBIS.

CONCLUSIONS

La méthode décrite plus haut implique des calculs d'une telle complexité qu'un ordinateur est indispensable. Le temps machine est heureusement minime et facilite l'utilisation du programme.

BIBLIOGRAPHIE

MAHALANOBIS (P.C.), 1927. — Analysis of race-mixture in Bengal (Presidential Address, Anthropological Section, Indian Science Congress, Banaras, 1925). *J. Proc. Asiat. Soc. Beng.*, vol. XXIII, pp. 301-333.

MAHALANOBIS (P.C.), 1930. — On tests and measures of group divergence. Part 1 : Theoretical formulae. *J. Proc. Asiat. Soc. Beng., new ser.*, vol. XXVI, pp. 541-588.

MAHALANOBIS (P.C.), 1936. — On the generalized distance in statistics. *Proc. Nat. Inst. Sci. India*, vol. II, n° 1, pp. 49-55.

MAHALANOBIS (P.C.), 1948. — Historical note on the D^2-statistic. *Sankhya*, vol. 9, n° 2 et 3, pp. 237-240.

RAO (C.R.), 1965. — *Linear statistical inference and its applications.* Wiley, New York, London, Sydney, XVIII, 522 p.

SIEGEL (S.), 1956. — *Nonparametric statistics : for the behavioral sciences.* Mc Graw-Hill, New York, Toronto, London, XVII, 312 p.

VAN DEN DRIESSCHE (R.), MAIGNIEN (R.), 1965. — Application d'une méthode de la statistique approfondie à la pédologie. *Cah. ORSTOM, sér. Pédol.*, vol. 3, n° 1, pp. 79-88.