UN PROBLÈME DE CLASSIFICATION NUMÉRIQUE

par

Raymond VAN DEN DRIESSCHE*

SUMMARY

Grouping of individual soil profiles has been achieved on a regional basis, by a clustering process satisfying following conditions

average inter-clusters distance > average intra-cluster distance

intra-extra-cluster distance > average intra-cluster distance.

Unit sample size and a spurious correlation between classification variates imposed a distribution-free measure of distance. Clustering was repeated twice on the cluster means of the previous step.

RÉSUMÉ

Un regroupement régional de sols ferrallitiques à partir des distances généralisées Δ et des constellations qu'elles engendrent, en trois phases successives, apporte une solution à un problème caractérisé essentiellement par une corrélation illusoire entre les variables de classification et par un effectif unitaire.

La classification des sols ferrallitiques au niveau de la sous-classe est définie et l'utilisation simultanée de trois variables de l'horizon B2 préconisée (AUBERT et SEGALEN 1966). Les limites intra-sous-classe choisies pour ces variables (pH, S/T et S) restent à confronter avec les données d'un certain nombre de profils.

C'est par une analyse multivariable de 295 profils, décrits dans vingt rapports de prospection en provenance de São Tomé, du Gabon, de la Guadeloupe, de la Côte d'Ivoire, du Congo Brazzaville, du Congo Kinshasa, de la Guyane française, du Cameroun, de la Guinée, du Dahomey,

* Chargé de recherche . Services scientifiques centraux, Bondy.

Communication à la Réunion de Pédologie, 7 et 8 oct. 1966, Paris.

des districts de Huíla et Huambo en Angola et de Madagascar, que le problème est abordé. La nomenclature des horizons fait défaut dans les descriptions de profils de la Côte d'Ivoire, du Congo Brazzaville, du Cameroun, de la Guinée, du Dahomey, de l'Angola, de Madagascar, et l'identification, *a posteriori*, de l'horizon B_2 repose notamment sur la teneur en carbone des différents horizons. Dans tous les cas, une vérification des valeurs numériques que prennent les variables dérivées S/T et S a été faite à partir des variables d'origine Ca, Mg, K, Na, T. Ces rapports sont présentés par des équipes différentes ; ils couvrent des régions écologiques variées ; certains négligent Mg dans le calcul de S ; leur examen simultané se heurte à l'absence de standardisation. Un dépouillement des données dans un cadre régional semble donc justifié.

L'effectif est unitaire - un seul horizon d'un seul profil - et la corrélation entre les variables S/T et S est illusoire. Il est, par conséquent, exclu, même en recourant aux variables S et T, rendues gaussiennes et homoscédastiques par transformation, de mesurer des distances généralisées D^2 entre les profils. D'autres méthodes, comme les composantes principales, les corrélations multiples maximales, les partitions basées sur l'analyse de la variance ne conviennent guère mieux.

Une mesure de distance robuste, c'est-à-dire libérée de la fonction de répartition, s'impose* , la nouvelle mesure de distance (HIERNAUX 1965)

$$\Delta_g^{P1\,P2} = 10\,000 \sum_1^n \left\{ \left[g_i^{P1} - g_i^{P2} \right]^2 / a_i \right\} /n$$

ne fait appel qu'à l'étendue mondiale a_i de chacune des n variables et aux observations g_i de ces variables dans chaque profil p_1, p_2,... Elle convient, dès lors que les trois variables sus-mentionnées ont pour étendue mondiale :

$$a_{pH} = 3,5$$
$$a_{S/T} = 100\,\%$$
$$a_S = 10 \text{ mé}$$

dans la classe des sols ferrallitiques ; que les variables sont présentes dans la description des 295 profils et que les effectifs sont compris entre 17 et 30 profils par région. Le nombre de distances, égal au nombre de combinaisons sans répétition des profils pris deux à deux, est ainsi compris entre $\binom{17}{2} = 136$ et $\binom{30}{2} = 435$ par région. Les Δ ne servent qu'à rechercher des constellations qui répondent à des conditions bien définies (VAN DEN DRIESSCHE 1965). La constellation englobe les profils les plus ressemblants quant à l'ensemble des trois variables. Les conditions imposées aux distances moyennes $\overline{\Delta}$ et individuelles Δ sont :

$$\overline{\Delta} \text{ interconstellations} > \overline{\Delta} \text{ intra-constellation ;}$$
$$\text{et} \quad \Delta \text{ intra-extra-constellation} > \overline{\Delta} \text{ intra-constellation.}$$

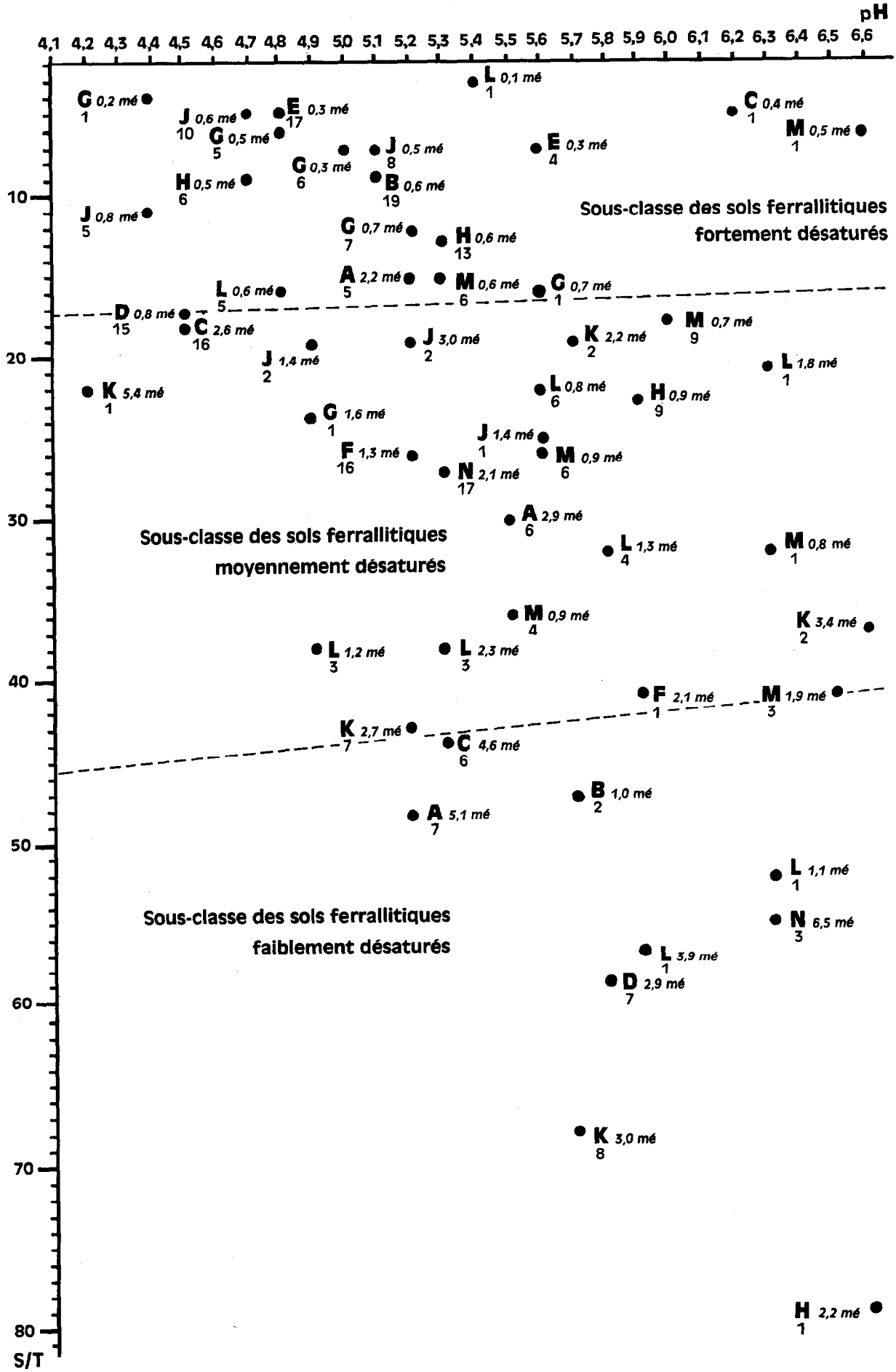
Distances et constellations sont toutefois calculées en trois phases : une première sur les profils individuels : une deuxième sur les moyennes intra-constellation et les données des profils isolés par la première phase ; une troisième, enfin, sur les moyennes intra-constellation et profils non encore regroupés de la deuxième phase. Cette procédure est inédite, qu'il s'agisse de l'utilisation de Δ en pédologie ; du choix d'effectifs unitaires pour la mesure de distance ; de la comparaison de $\overline{\Delta}$ et non de $\overline{D}^2$ dans la recherche des constellations ou du regroupement des données en trois phases successives. Il convient, par conséquent, de l'envisager comme UNE solution possible d'un problème particulier de classification et non comme une méthode à préférer, dans les cas de normalité et d'homoscédasticité, aux distances généralisées D^2 entre des ensembles de nombreux profils.

Le tableau qui suit donne, par région, les constellations ou regroupements de profils obtenus et les moyennes de pH, de S/T et de S qui les caractérisent.

* Non-paramétrique désigne un problème et non une méthode (KENDALL and STUART 1961).

Régions	Effectifs	Moyennes		
		pH	S/T %	S mé
A São Tomé	18 profils en 3 constellations de 7 profils	5,2	48	5,1
	6	5,5	30	2,9
	5	5,2	15	2,2
B Gabon	21 profils en 2 constellations de 19 profils	5,1	9	0,6
	2	5,7	47	1,0
C Guadeloupe	23 profils en 3 constellations de 16 profils	4,5	18	2,6
	6	5,3	44	4,6
	1	6,2	5	0,4
D Côte d'Ivoire	22 profils en 2 constellations de 15 profils	4,5	17	0,8
	7	5,8	59	2,9
E Congo Brazzaville	21 profils en 2 constellations de 17 profils	4,8	5	0,3
	4	5,6	7	0,3
F Congo Kinshasa, Maniema	17 profils en 2 constellations de 16 profils	5,2	26	1,3
	1	5,9	41	2,1
G Guyane française	21 profils en 6 constellations de 7 profils	5,2	12	0,7
	6	5,0	7	0,3
	5	4,8	6	0,5
	1	4,4	4	0,2
	1	4,9	24	1,6
	1	5,6	16	0,7
H Cameroun	29 profils en 4 constellations de 13 profils	5,3	13	0,6
	9	5,9	23	0,9
	6	4,7	9	0,5
	1	6,6	79	2,2
J Guinée	28 profils en 6 constellations de 10 profils	4,7	5	0,6
	8	5,1	7	0,5
	5	4,4	11	0,8
	2	4,9	19	1,4
	2	5,2	19	3,0
	1	5,6	25	1,4
K Dahomey	20 profils en 5 constellations de 8 profils	5,7	68	3,0
	7	5,2	43	2,7
	2	5,7	19	2,2
	2	6,6	37	3,4
	1	4,2	22	5,4
L Angola, Huíla	25 profils en 9 constellations de 6 profils	5,6	22	0,8
	5	4,8	16	0,6
	4	5,8	32	1,3
	3	4,9	38	1,2
	3	5,3	38	2,3
	1	6,3	52	1,1
	1	6,3	21	1,8
	1	5,4	3	0,1
	1	5,9	57	3,9
M Angola, Huambo	30 profils en 7 constellations de 9 profils	6,0	18	0,7
	6	5,6	26	0,9
	6	5,3	15	0,6
	4	5,5	36	0,9
	3	6,5	41	1,9
	1	6,3	32	0,8
	1	6,6	6	0,5
N Madagascar	20 profils en 2 constellations de 17 profils	5,3	27	2,1
	3	6,3	55	6,5

Le graphique S/T pH reprend les moyennes du tableau. La lettre désignant la région a pour coordonnées pH et S/T. La valeur de S, en mé, et l'effectif, en nombre de profils, sont indiqués. Les limites intra-sous-classes proposées par AUBERT et SEGALEN (*op. cit.*) sont tracées. Les regroupements de profils semblent satisfaisants.


pH
4,1 4,2 4,3 4,4 4,5 4,6 4,7 4,8 4,9 5,0 5,1 5,2 5,3 5,4 5,5 5,6 5,7 5,8 5,9 6,0 6,1 6,2 6,3 6,4 6,5 6,6
S/T
10
20
30
40
50
60
70
80
Sous-classe des sols ferrallitiques fortement désaturés
Sous-classe des sols ferrallitiques moyennement désaturés
Sous-classe des sols ferrallitiques faiblement désaturés
L 0,1 mé 1
G 0,2 mé 1
J 0,6 mé 10
E 0,3 mé 17
G 0,5 mé 5
C 0,4 mé 1
M 0,5 mé 1
J 0,5 mé 8
E 0,3 mé 4
G 0,3 mé 6
H 0,5 mé 6
B 0,6 mé 19
J 0,8 mé 5
G 0,7 mé 7
H 0,6 mé 13
A 2,2 mé 5
M 0,6 mé 6
G 0,7 mé 1
L 0,6 mé 5
D 0,8 mé 15
C 2,6 mé 16
M 0,7 mé 9
J 3,0 mé 2
K 2,2 mé 2
J 1,4 mé 2
L 1,8 mé 1
K 5,4 mé 1
L 0,8 mé 6
H 0,9 mé 9
G 1,6 mé 1
J 1,4 mé 1
F 1,3 mé 16
M 0,9 mé 6
N 2,1 mé 17
A 2,9 mé 6
L 1,3 mé 4
M 0,8 mé 1
M 0,9 mé 4
K 3,4 mé 2
L 1,2 mé 3
L 2,3 mé 3
F 2,1 mé 1
M 1,9 mé 3
K 2,7 mé 7
C 4,6 mé 6
B 1,0 mé 2
A 5,1 mé 7
L 1,1 mé 1
N 6,5 mé 3
L 3,9 mé 1
D 2,9 mé 7
K 3,0 mé 8
H 2,2 mé 1

Comme on le voit, la solution ne porte pas sur des données de masse mais n'en prend pas moins l'aspect d'une vaste entreprise de calculs, qui couvre plus de 5 000 Δ et 39 matrices de $\overline{\Delta}$. Grâce au programme STAR écrit par Mme MASBOU, à l'Institut Blaise Pascal du C.N.R.S., les calculs prirent moins de trois minutes sur l'ordinateur CDC 3 600 de cet Institut. Compiler les rapports et publications, vérifier et retranscrire les données, décoder les résultats sont autant d'opérations manuelles qui peuvent aussi être mécanisées par l'adoption de la FICHE ANALYTIQUE et l'utilisation du programme SOLS (pour CDC 3 600) que nous devons à la collaboration de Mme POTIN de l'I.B.P.

Source des données

BOURGEAT (F.), HERVIEU (J.), RIQUIER (J.) - 1964 - Présentation de quelques profils de sols ferrallitiques. Etude du milieu pédogénétique dans les environs de Tananarive. Centre O.R.S.T.O.M. de Tananarive, 87 p., multigr.

CARDOSO (J. Carvalho), GARCIA (J. Sacadura) - 1962 - Carta dos solos de São Tomé e Príncipe. *Mem. Junta Invest. Ultram.*, 2a sér., n° 39, Lisbonne, 306 p.

CHATELIN (Y.) - 1964 - Les sols dès massifs cristallins et cristallophylliens des Monts de Cristal, des Monts de N'Djolé et du chaînon de Lambaréné-Chinchoua. Centre O.R.S.T.O.M. de Libreville, 21 p., multigr.

CHATELIN (Y.) - 1964 - Les sols du bassin sédimentaire côtier entre Libreville et Lambaréné. Centre O.R.S.T.O.M. de Libreville, 61 p., multigr.

COLMET-DAAGE (F.) - 1965 - Fiches analytiques de sols ferrallitiques de la Guadeloupe. Inédit.

DABIN (B.) - 1963 - Etude pour la reconversion des cultures de caféier dans la République de Côte d'Ivoire. B.D.P.A. et O.R.S.T.O.M., Paris, 332 p., multigr.

DELHUMEAU (M.) - 1964 - La route de N'Djolé à La Lara. Centre O.R.S.T.O.M. de Libreville, 17 p., multigr.

DELHUMEAU (M.) - 1964 - La vallée du moyen Ogooué de Booué à Junkville. Centre O.R.S.T.O.M. de Libreville, 27 p., multigr.

GRAS (F.) - 1965 - Etude pédologique d'une zone témoin dans la région de Tsiaki (avec carte au 1/50 000). Centre O.R.S.T.O.M. de Brazzaville, 74 p., multigr.

HERVIEU (J.) - 1961 - Profils-types de sols malgaches. Annexe : résultats analytiques. Centre O.R.S.T.O.M. de Tananarive, non pag., multigr.

JAMAGNE (M.) - 1963 - Contribution à l'étude des sols au Congo oriental (Maniema). *Pédologie*, XIII, 2, p.271-414.

LEVEQUE (A.) - 1962 - Mémoire explicatif de la carte des sols des Terres Basses de Guyane française. *Mém. O.R.S.T.O.M.*, 3, Paris, 86 p.

LEVEQUE (A.) - 1963 - Les sols développés sur le bouclier antécambrien guyanais. Centre O.R.S.T.O.M. de Cayenne, 244 p., multigr.

MARIUS (C.) - 1965 - Etude pédologique de la feuille au 1/50 000 de Cayenne. Centre O.R.S.T.O.M. de Cayenne, 56 p., multigr., annexes.

MARTIN (D.) - 1965 - Etudes pédologiques dans le Centre Cameroun (Nanga-Eboko à Bertoua). 2. Résultats analytiques. Centre O.R.S.T.O.M. de Yaoundé, 47 p., multigr.

MISSAO DE PEDOLOGIA DE ANGOLA - 1959 - Carta geral dos solos de Angola. 1. Distrito da Hufla. *Mem. Junta Invest. Ultram.*, 2a sér., n°9, Lisbonne, VIII-482 p.

MISSAO DE PEDOLOGIA DE ANGOLA - 1961 - Carta geral dos solos de Angola. 2. Distrito da Huambo. *Mem. Junta Invest. Ultram.*, 2a sér., n° 27, Lisbonne, VIII-275 p.

PEREIRA-BARRETO (S.) - 1963 - Etude pédologique du secteur sud de la surface d'études pédo-agronomiques dans les hauts-plateaux du Fouta-Djalon. Ministère de l'Economie rurale, Inspection des Eaux et Forêts, Service de la conservation des sols, Conakry, 88p., multigr., annexes.

VOLKOFF (B.) - 1964 - Etude des sols. Région des Dongas, nord Dahomey. Centre O.R.S.T.O.M. de Cotonou, 89 p., multigr.

WILLAIME (P.) - 1964 - Contribution à l'étude des sols de la basse vallée du Mono. Centre O.R.S.T.O.M. de Cotonou, 74 p., multigr., annexes.

Bibliographie

AUBERT (G.), SEGALEN (P.) - 1966 - Projet de classification des sols ferrallitiques. Première version. Communication à la Réunion de Pédologie, 7 et 8 oct., O.R.S.T.O.M., Paris, 18 p., multigr.

HIERNAUX (J.) - 1965 - Une nouvelle mesure de distance anthropologique entre populations utilisant simultanément des fréquences géniques, des pourcentages de traits descriptifs et des moyennes métriques. *C.R. Acad. Sci.*, Paris, 260, p. 1748-1750.

KENDALL (M.G.), STUART (A.) - 1961 - *The advanced theory of statistics. vol. 2. Inference and relationship.* Griffin, Londres, 676 p.

VAN DEN DRIESSCHE (R.) - 1965 - La recherche des constellations de groupes à partir des distances généralisées D^2 de Mahalanobis. *Biom.-Prax.*, VI, p.36-47.