SCIENCE DU SOL 1988 - Vol. 26/2 - pp. 81-102 - PLAISIR-FRANCE

UN ALGORITHME INTERACTIF POUR LA CLASSIFICATION DES HORIZONS DE LA COUVERTURE PEDOLOGIQUE : DIMITRI

M.-C. GIRARD[1] et D. KING[2]

RESUME

L'analyse de la couverture pédologique peut se faire à partir de structures homogènes tridimensionnelles telles que les horizons par exemple. Il est intéressant de définir une typologie de ces volumes pour introduire ensuite, une analyse structurale de la couverture pédologique. A partir de cette typologie, il est aussi possible de comparer tous les horizons réels à des horizons de référence : ceci peut donc déboucher sur une typologie des comportements des sols, sur une thématisation des données sols ou sur une classification des sols.

On peut caractériser ces volumes par des variables qualitatives ordonnées. Un algorithme de classification DIMITRI est proposé : il peut fonctionner à partir de différentes mesures de distance mathématique. Il permet d'obtenir un ensemble de références caractérisées par des valeurs modales et marginales. L'algorithme, programmé sur micro-ordinateur, est inter-actif et propose au pédologue des interventions précises dans le choix de certains paramètres du traitement. Il contribue à l'élaboration d'un système de Pédologie Assistée par Ordinateur.

INTRODUCTION

On dispose actuellement en France de banques de données (BERTRAND et al, 1979 - DUVAL et KING, 1982) qui permettent d'obtenir rapidement, par divers traitements, de l'information pédologique sous des formes différentes. Depuis quelques dizaines d'années, on a testé, en pédologie, diverses méthodes statistiques pour évaluer leur intérêt pour la pédologie : DAGNELIE, 1985 - HOLE et HIRONOKA, 1960 - MOORE et RUSSEL, 1966 - CAMPBELL et al., 1970 - WEBSTER, 1971, 1977 - GRIGAL et al., 1969 - BOTTNER et al., 1975 - de GRUIJTER, 1977. Ces méthodes se sont révélées fructueuses et on a montré quelles en étaient les contraintes (MAUCORPS et GIRARD, 1976). Elles ont aussi permis une réflexion sur les concepts pédologiques ainsi que sur les raisonnements utilisés par les pédologues (GIRARD, 1983 - KING, 1986 - LEGROS, 1982).

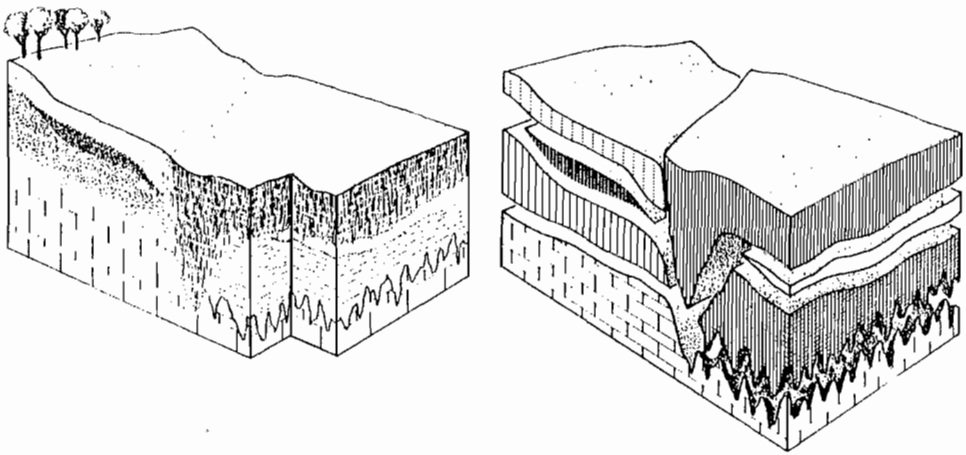
La méthode présentée ici est issue d'une réflexion sur la façon de faire du pédologue lorsqu'il classe des entités pédologiques en vue de définir une typologie. On a donc cherché à suivre le mieux possible la démarche du pédologue et tenté de lui procurer une méthode statistique adaptée aux contraintes d'étude de la couverture pédologique.

MOTS CLES : horizon, taxonomie, cartographie, algorithme, statistique.
KEY WORDS : horizon, taxonomy, mapping, algorithm, statistics.

(1) Institut National Agronomique Paris-Grignon.
Science des sols et hydrologie, 78850 Thiverval-Grignon.
(2) Institut National de la Recherche Agronomique, SESCPF Ardon, 45160 Olivet.

I. MODELISATION DE LA COUVERTURE PEDOLOGIQUE

Pour en faciliter l'analyse, on peut concevoir la couverture pédologique comme un ensemble d'horizons (GIRARD, 1983) dont l'organisation dans les trois dimensions constitue une réalité simplifiée à laquelle on se réfère. On fait donc l'hypothèse de l'existence d'une partition de la couverture pédologique en volumes homogènes : les horizons (fig. 1). Cette hypothèse, reprise par G. PEDRO (1986) et par D. BAIZE (1987) avait déjà été suggérée en 1969 par M.-C. GIRARD et avancée en 1978 par R. BOULET, F.X. HUMBEL, et Y. LUCAS.



A) Couverture pédologique.

B) Décomposée en compartiments constitués de volumes homogènes : les horizons.

Figure 1 : Partitionnement de la couverture pédologique en horizons.
Figure 1 : Partitioning of soil mantle in horizons.

A) Caractérisation des volumes

On fait aussi l'hypothèse qu'il est raisonnable de caractériser ces horizons à partir d'un ensemble de variables de description et d'analyse de laboratoire. Cette hypothèse est plus difficile à justifier si l'on veut définir des « horizons d'interprétation » (GIRARD M.-C. et BAIZE D., 1987). En effet, le plus souvent ceux-ci nécessitent de connaître les relations entre horizons (transitions, contrastes, position dans la couverture pédologique...), et entre le sol et son environnement.

Il est cependant possible de définir des entités correspondant à des volumes pédologiques, à partir des seules variables de description et d'analyse (GIRARD, 1977 et KING, 1986) ; on les appellera « Unités d'information » (UI).

B) Les unités d'informations

Celles-ci sont définies par un ensemble de variables intrinsèques que l'on peut considérer comme constantes dans tout le volume correspondant à l'unité d'informa-

tion (variables que l'on pourrait nommer variables « horizoniques »). Les variables qui nécessitent l'observation de l'ensemble du pédon, (et que l'on pourrait nommer variables « pédoniques »), comme par exemple les fentes, ne sont pas prises en compte pour définir les unités d'information.

Les variables utilisées pour la caractérisation des unités d'information sont le plus souvent qualitatives ordonnées ou quantitatives. Mais il faut être conscient de « l'inégalité qualitative des données numériques » (GEORGES, 1970). De plus, les variables quantitatives sont fréquemment interprétées sur la base de classes : classes de texture, classes de pH, etc. On retombe sur des données qualitatives ordonnées. On ne perd donc guère de précision en caractérisant directement une unité d'information par un ensemble de variables qualitatives ordonnées.

Pour caractériser la couverture pédologique à partir de ce modèle, il est nécessaire de placer chaque volume auquel correspond chaque unité d'information dans l'espace tridimensionnel. Pour cela, il faut connaître les coordonnées géographiques en x et y, la profondeur et l'épaisseur des volumes. Cela permet de reconstituer la couverture pédologique en déterminant la position relative des différents volumes constitutifs.

C) Un algorithme typologique

Les unités d'informations identifiées dans l'étude d'une portion de la couverture pédologique sont très nombreuses. Il est le plus souvent nécessaire de les regrouper.

Pour ce faire, on propose un algorithme : DIMITRI, qui réunit les unités d'informations en ensembles. Dans chaque ensemble, les valeurs modales pour chaque variable permet de définir un individu statistique : l'Unité de référence.

II. UNE METHODE DE CLASSIFICATION NON HIERARCHIQUE : DIMITRI

DIMITRI, basé sur la DIstance MInimum de TRI, est une méthode de classification non hiérarchique fondée sur le même principe que la méthode des « nuées dynamiques » (DIDAY, 1972).

On choisit un certain nombre d'individus-type : les noyaux. Chaque individu est regroupé avec le noyau dont il est le plus proche (la distance noyau-individu est donc minimum). Si les individus-types correspondent à des références non remises en question on obtient un classement des individus. Par contre, si l'on se pose la question de la pertinence des noyaux, on ne peut pas s'arrêter là. On continue la démarche en redéfinissant chaque noyau à partir de la population statistique définie dans l'étape précédente. En continuant ainsi, jusqu'à ce que l'on obtienne une stabilité pour chaque noyau, on obtient une classification.

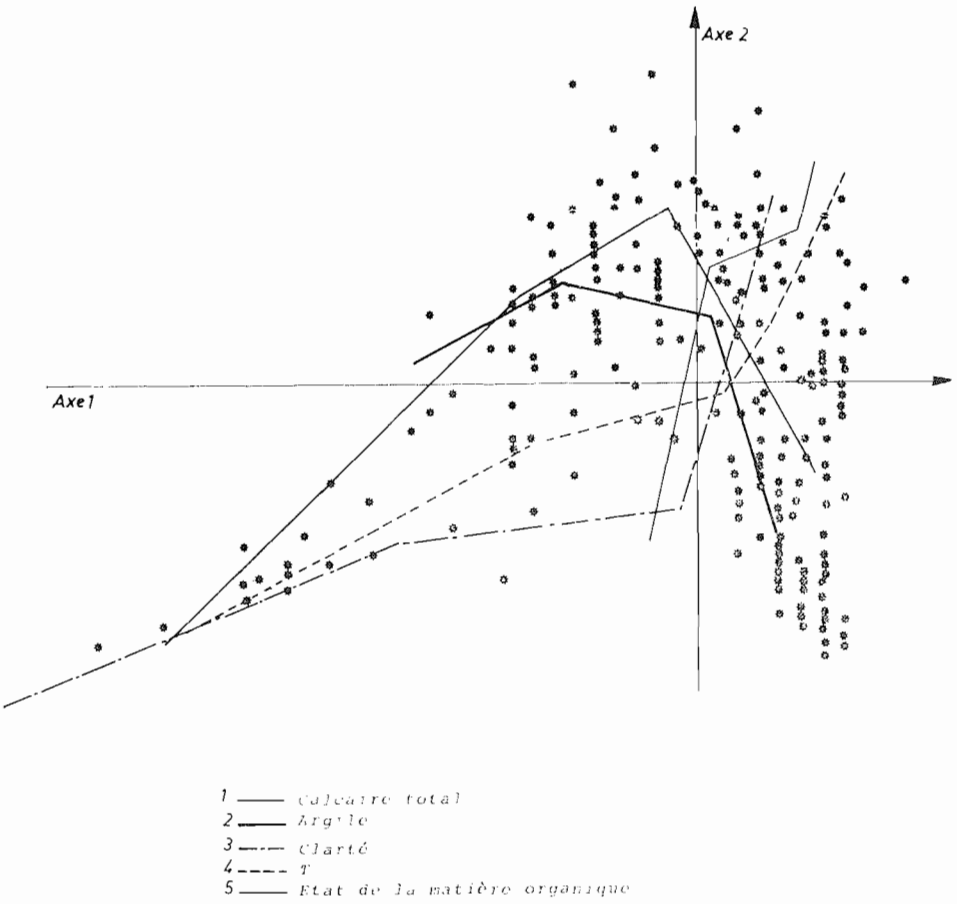
A) Le codage des variables

Les variables qualitatives ordonnées caractérisant les unités d'information comportent à peu près le même nombre de modalités (le plus souvent autour de cinq) afin qu'ultérieurement l'influence du nombre des modalités ne se fasse pas trop sentir dans le calcul des distances.

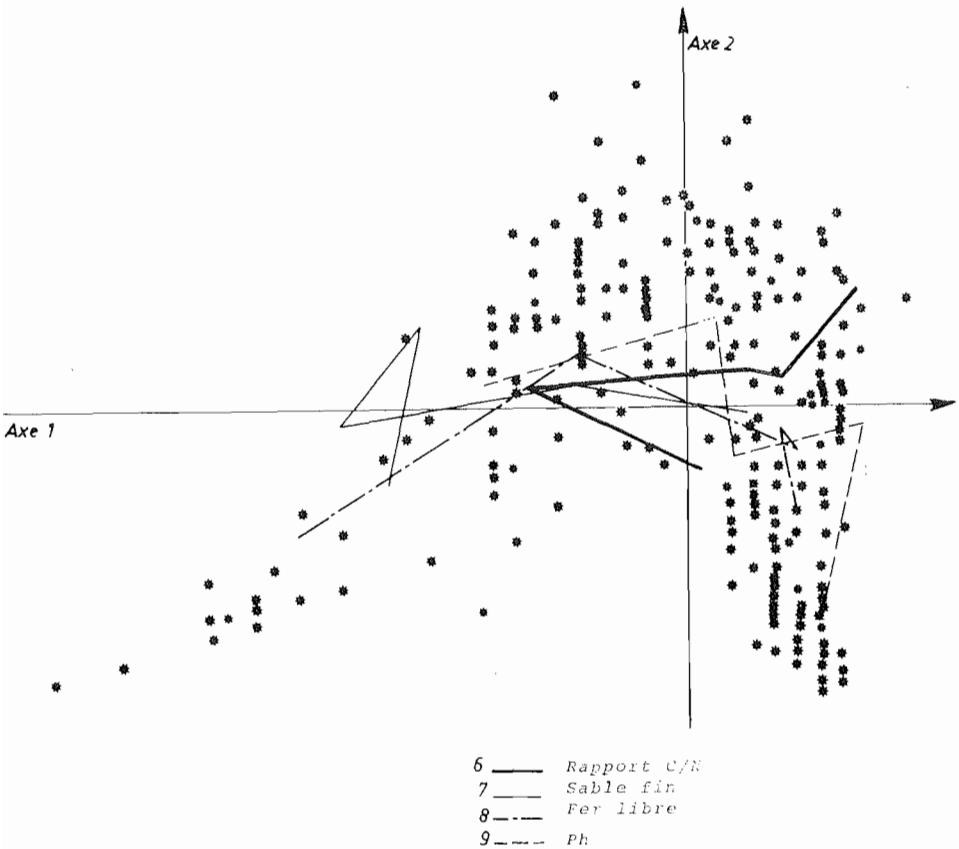
On peut vérifier la qualité du codage des variables quantitatives par une analyse factorielle des correspondances (fig. 2). Les modalités définies lors du codage doivent s'organiser dans les plans factoriels suivant des cheminements en cohérence avec le nuage de points-individus. C'est le cas des variables : état de la matière organique (1), clarté (2), argile (3), capacité totale d'échange (4), et calcaire total (5) de la figure 2a. Par contre les variables : rapport C/N (6), fer libre (7), sable fin (8), et pH (9) n'ont pas un codage satisfaisant (fig. 2b) car leur cheminement ne correspond pas à l'extension générale du nuage de points.

Figure 2 : Projection sur les axes principaux 1 et 2 de 216 horizons décrits par 30 variables.

Figure 2 : Projection on main axis (1 and 2) of 216 horizons described by 30 variables.



A) Les variables : état de la matière organique (1), clarté (2), argile (3), capacité totale d'échange (4), calcaire total (5) ont un codage satisfaisant.



B) Les variables : rapport C/N (6), sables fins (7), fer libre (8), pH (9) n'ont pas un codage satisfaisant.

B) Le choix des noyaux

Chaque noyau est défini par le même ensemble de variables que celui utilisé pour les individus. Si l'on ne sait rien sur l'organisation des individus, on peut tirer au hasard un certain nombre d'unités d'information pour constituer les noyaux. On peut aussi utiliser un sous-ensemble des individus à classifier et effectuer une classification ascendante (JAMBU, 1978). Une partition de l'arbre hiérarchique obtenu (fig. 3) peut servir au choix des noyaux.

Enfin, les noyaux peuvent être choisis par le pédologue :

— en fonction de ce qu'il estime être les horizons de référence d'après l'étude de terrain (fig. 4) ;

— en fonction de références déjà déterminées régionalement, nationalement ou internationalement (GIRARD et BAIZE, 1987) ;

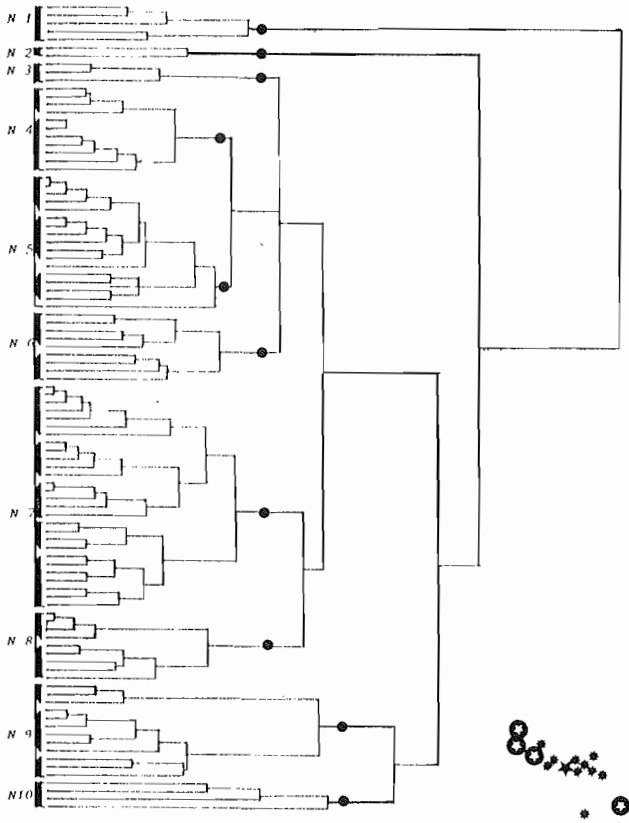


Figure 3 :

Choix des noyaux à l'aide d'une partition de l'arbre hiérarchique issu de la classification ascendante de Jambu.

Figure 3 :

Choice of nuclei, using a partioning of hierarchical tree issued from the ascending Jambu classification.

Les points noirs représentent le choix des noyaux retenus.

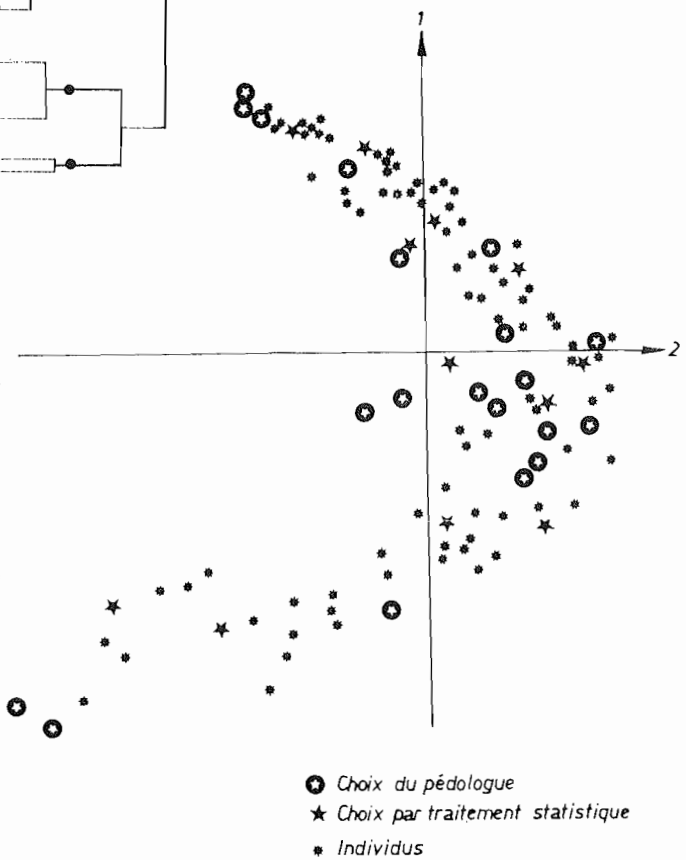
Figure 4 :

Exemples de choix des noyaux par le pédologue et par traitement des données.

Figure 4 :

Examples of nuclei choice by the soil scientist using data numerical handling.

Les deux types de choix et l'ensemble des unités d'information sont représentées sur le plan factoriel 1-2 d'une Analyse en Composante Principale.



— en créant des noyaux sur la base : d'une analyse de l'histogramme de chaque variable, d'une AFC, ou d'un modèle pédogénétique.

On verra plus loin qu'il est possible de réintroduire des noyaux au cours des diverses itérations. En conséquence le nombre de noyaux choisi au début du processus n'est pas absolument fondamental. Si le nombre de noyaux est trop important, certains constitueront des ensembles vides. Si le nombre est trop petit, alors la variabilité à l'intérieur d'un noyau indiquera qu'il est possible de le partitionner en plusieurs (cf. fig. 6).

C) Notion de distance

La distance d (A, B) entre deux individus A et B, caractérisés chacun par V variables est égale à la somme des différences entre A et B, pour chaque variable. Ces différences sont prises en valeur absolue ou élevées au carré.

Deux variables qui n'ont pas la même étendue maximum, n'auront pas le même poids dans le calcul de la distance. Il y aurait alors une pondération implicite. En conséquence, on utilise un facteur correctif : fc, pour rendre comparable ces étendues maximum (par exemple : valeur maximale - valeur minimale).

Afin que le nombre de variables pris en compte pour calculer la distance ne joue pas, on divise la somme des différences par le nombre de variables.

Une distance est donc de la forme :

$$d\ (A, B) = \frac{1}{v} \sum_{1}^{v} C_v \qquad C_v = \frac{1}{fc} (v_A - v_B)^2$$

C_v étant la contribution de chaque variable v à la distance.

De nombreuses distances existent, plusieurs ont été employées par les pédologues.

1. La distance euclidienne classique

Elle a pour avantage d'être indépendante du nombre total d'individus comparés. Mais l'élévation au carré de la différence a pour effet de donner un poids plus grand à de grandes différences entre quelques variables qu'à de petites différences entre un très grand nombre de variables. CIPRA et al. (1970), CUANALO et WEBSTER (1970), KING (1986) ont employé cette distance.

2. La distance du χ^2

Elle permet de comparer des variables qualitatives. Elle a été encore peu utilisée en pédologie, peut-être parce qu'on ne dispose pas de descriptions entièrement qualitatives.

3. La distance généralisée de MAHALANOBIS (1936)

Elle permet de comparer des groupes d'individus. Elle a été utilisée par HUGHES et LINDLEY (1955), VAN DEN DRIESSCHE (1965), VAN DEN DRIESSCHE et MAIGNIEN (1965), PRUSINKIEWICZ (1969).

4. La mesure des distances de HIERNAUX (1965)

Elle utilise un facteur de pondération dénommé « étendue mondiale » qui est défini par le spécialiste qui traite ses données. Elle permet d'utiliser des variables de natures différentes : moyennes, pourcentages, fréquences. Pour cette raison, elle semble attractive pour l'étude de la comparaison des sols. Elle a été utilisée en pédologie par VAN DEN DRIESSCHE (1966) et par GIRARD (1968, 1969).

5. La métrique L 1

Elle se base sur la valeur absolue de la différence entre variables. Il est nécessaire que toutes les variables soient du même type. Elle a été utilisée par MOORE et RUSSEL (1967).

6. L'indice de similarité de GOWER (1971)

L'indice de similarité proposé par J. GOWER a été utilisé par RAYNER (1966), BONNERIC (1978), SIMMONNEAUX (1987), OLIVER et WEBSTER (1987). Il permet de prendre en compte en même temps des variables : à deux états, qualitatives, et quantitatives. Il faut définir un poids, qui peut être le même, pour chaque variable.

7. La relation de dissemblance de RANG

Diverses autres distances ou coefficients de dissemblance ont été utilisés en pédologie (GIRARD, 1983). Nous retiendrons la relation de dissemblance de rang proposée par VAN DEN DRIESSCHE et GARCIA-GOMEZ (1972) (voir annexe) et programmée par AUBRY (1975). Elle est proche de celle proposée par KENDALL et STUART (1966). Elle est indépendante de la fonction de répartition.

Il est possible de calculer une telle « distance » sur un ensemble d'individus, même s'il y a quelques données manquantes. Deux individus dont les valeurs ne diffèrent que par une valeur estimée de l'ordre de l'erreur, peuvent être considérés comme ex-æquo : on les code avec les mêmes valeurs et leur distance sera donc nulle. Les ex-æquo qui peuvent alors être très nombreux, sont pris en compte dans le calcul de la distance. Cette « distance de rang » présente surtout un intérêt lorsqu'on la calcule entre de nombreux individus.

D) Organisation de chaque itération au sein de DIMITRI

Une itération comporte les 7 phases suivantes (fig. 5) :

1) On calcule la distance entre chaque horizon et chaque noyau. schématisé par un point sur la figure 5 (point 1).

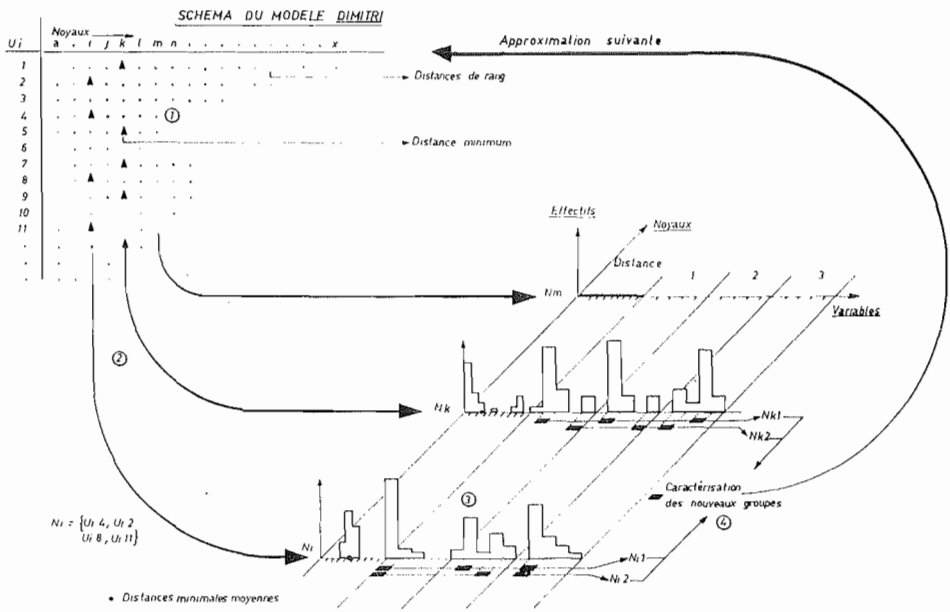


Figure 5 : Schéma de fonctionnement de l'algorithme DIMITRI.
Figure 5 : Scheme of DIMITRI algorithm working.

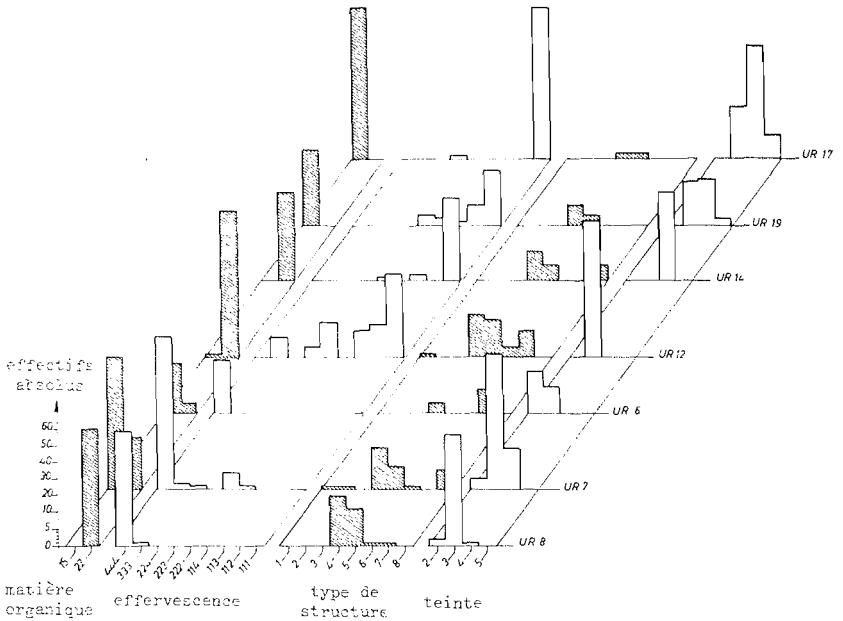
2) On affecte chaque unité d'information, au noyau qui est à la distance minimum, schématisée par un triangle noir sur la figure 5 (point 1).

3) L'ensemble des unités d'information affectés (selon 2) à un même noyau constitue la population statistique qui définit un groupe (fig. 5, point 2). Comme toutes les unités d'information sont rattachées à l'un des noyaux de départ, chaque unité d'information se trouve dans le groupe dont elle est le plus proche, ou le moins loin !

4) Pour chaque groupe, on dresse un histogramme pour chacune des variables (fig. 6) ainsi que pour les valeurs des distances minimum. On analyse ces histogrammes pour en définir le mode ou la médiane qui seront pris comme nouvelle définition du noyau de l'itération suivante (fig. 7). Les noyaux seront constitués à partir d'un ensemble réel et non à partir de concepts. Lorsque les histogrammes montrent une bimodalité, il est possible alors de prendre deux sous-populations, et de caractériser deux noyaux, là où il n'y en avait qu'un seul (fig. 5, point 3 : noyau i ou k). Il est aussi possible qu'il existe un noyau pour lequel aucun horizon ne soit le plus proche (l'effectif des histogrammes est nul). On ne retiendra pas ce noyau pour une autre itération (noyau m, fig. 5, point 4).

5) L'analyse des groupes par le pédologue est facilitée par un programme conversationnel. Il permet de visualiser tel ou tel histogramme, et de prendre des décisions pour déterminer la caractéristique d'un groupe.

6) On construit alors de nouveaux noyaux (fig. 5, point 4) à partir des groupes en utilisant la médiane ou le mode, selon les places respectives de ces deux critères dans la population, mais aussi entre les divers groupes (fig. 7). On peut :



On remarque la bimodalité du noyau 12 pour l'effervescence et le type de structure. On peut le décomposer en deux noyaux : l'un, effervescent (modalités comprises entre 222 et 333 et à structure grenue ou grumeleuse (modalités 6 et 7) ; l'autre peu ou pas effervescent (111 à 114) et à structure polyédrique anguleuse ou non (4 ou 5.

Figure 6 : Histogramme de 4 variables pour 7 noyaux.
Figure 6 : Histogram of 4 variables for 7 nuclei.

— reprendre les anciens noyaux avec une nouvelle caractérisation,
— décomposer les anciens noyaux bimodaux,
— supprimer les anciens noyaux quand aucun horizon n'a constitué de groupe avec ce noyau (fig. 5, noyau m),
— rajouter un noyau que le pédologue estime pouvoir exister, ou qu'il pense intéressant d'adjoindre à l'ensemble des noyaux définis,
— fusionner deux groupes très proches.

7) On dispose alors de nouveaux noyaux que l'on peut comparer, dans une seconde itération, à la population des horizons réels.

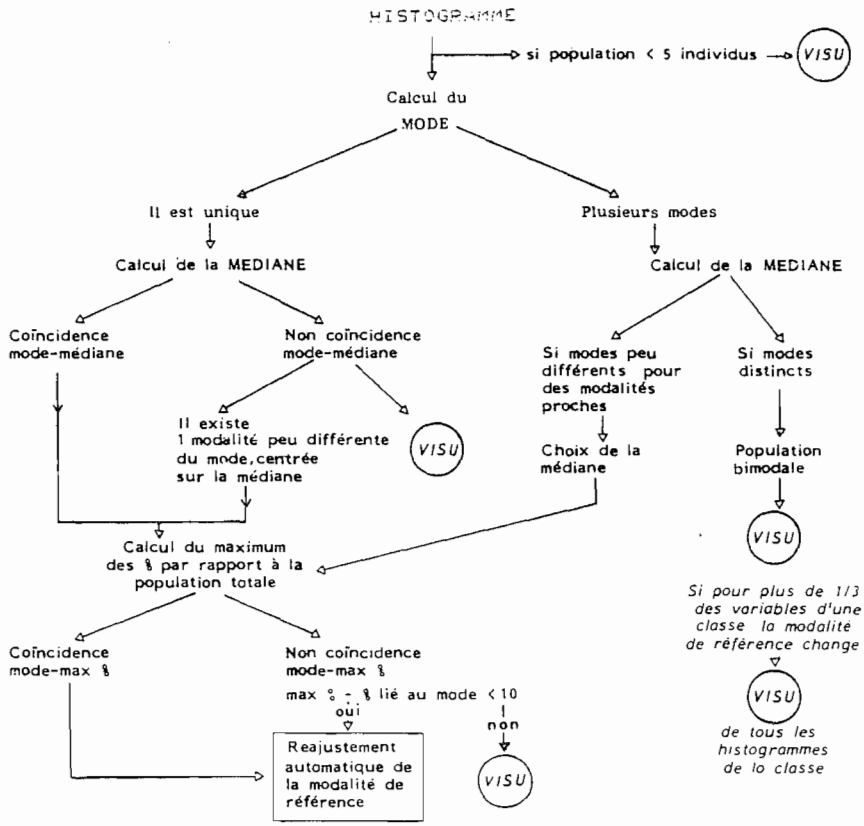


Figure 7 : Organigramme pour le choix semi-automatique des modalités caractérisant un noyau.

Figure 7 : Organigram for semi-automated choice of modalities describing a nucleus.

E) Les itérations

A chaque itération, on calcule la somme des distances minimum. Plus cette somme est faible, plus on estime que l'ensemble des noyaux est satisfaisant pour caractériser l'ensemble des unités d'information étudiées. On réitère les approximations tant que cette somme diminue. Si elle augmente entre l'approximation n et n + 1, on retient l'approximation n. On voit sur le tableau I que la quatrième approximation est meilleure que la cinquième.

Tableau I : Valeur des distances minimum moyennes de 15 noyaux pour 8 approximations
Tableau I : Mean minimum distance values for 15 nuclei for 8 approximations.

	1ère App. 30 v		2ème App. 30 v		3ème App. 18 v		4ème App. 18 v		5ème App. 10 v		6ème App. 18 v		7ème App. 18 v		8ème App. 18 v	
	F		F		F		F		F		F/S		F/S		F/S	
UR	$\bar{d}$	N	$\bar{d}$	N	$\bar{d}$	N	$\bar{d}$	N	$\bar{d}$	N	$\bar{d}$	N	$\bar{d}$	N	$\bar{d}$	N
4	21,3	13	22,5	14	21,0	12	23,4	11	28,6	16	24,5	19	22,9	17	25,5	18
6	25,9	13	25,3	12	27,1	15	20,8	10	21,2	11	32,5	10	31,1	8	16,1	8
8	22,7	21	26,0	21	22,6	23	19,1	18	21,5	26	24,2	25	19,7	21	15,1	19
15			23,0	8	20,5	10	20,5	6	18,7	11	15,8	13	17,1	15	15,1	14
12	28,5	26	23,1	21	20,0	19	20	12	16,8	18	30,1	7	18,8	13	16,9	10
13	25,6	19	25,4	19	26,7	18	17,1	10	16,3	14	18,1	18	14,2	13	15,6	13
14	19,7	16	15,1	13	14,7	14	12,3	10	14,6	17	11,9	14	2,8	8	6,1	11
9	30,8	17	31,0	21	26,2	18	21	6	20,6	10	24,2	29	24,8	35	22,8	31
11	25,0	11	23,5	13	24,9	13	24,5	5	23,6	14	33	12	27,3	8	20	10
10	27,3	16	28,4	15	24,1	13	20,6	5	21,5	17	23,9	7	19,4	7	20,8	9
16	19,3	12	17,0	11	17,3	12	14,6	8	13,9	9	23	4	26,3	4	19,8	7
17	14,8	21	13,5	21	11,7	19	7,7	15	7,4	25	11,5	34	9,0	28	8	27
19	29,7	22	26,8	19	29,9	21	18,5	13	18,9	19	25,6	17	17,8	18	16,7	22
18	14,4	9	11,8	8	9,1	9	1,7	6	5,1	7	6,1	7	9,7	9	1,5	6
7				0		0			36,5	2			31,0	12	22,5	11
S	5198		4969		4635		3793		3919		4529		4117		3643	
$\bar{d}$	24,1		23,0		21,5		17,6		18,1		21		19,0		16,8	

$\bar{d}$: distance moyenne des unités d'information à un noyau pour une approximation.
N : nombre d'unités d'information appartenant au noyau.
S : somme des distances minimum pour une approximation.
$\bar{\bar{d}}$ = S/N : moyenne des distances minimum pour toutes les unités d'information lors d'une approximation.

A chaque approximation, on peut juger des divers noyaux les uns par rapport autres autres en comparant la distance moyenne $\bar{d}$ de chaque noyau. Plus cette distance est faible, meilleure est la caractérisation du noyau : c'est le cas du noyau 18 pour la première approximation (tabl. I).

Entre deux approximations, il est intéressant d'étudier les unités d'information qui ont changé d'affectation entre les noyaux. On dresse un tableau où on met en ligne les anciens noyaux et en colonne les nouveaux noyaux (Tab. II). A l'intersection, on indique le nombre d'unités d'information. On trouve ainsi celles qui n'ont pas changé d'affectation, et celles qui, à la suite du changement des noyaux entre les deux approximations, ont changé d'affectation. Il est facile d'étudier quels sont les transferts qui ont eu lieu lorsqu'on a fait disparaître un noyau, quand on a fait apparaître un nouveau noyau (noyaux 15, tab. II), ou qu'on en a subdivisé un ancien en deux nouveaux. Il est possible d'évaluer le mouvement des unités d'information en calculant le nombre de celles qui ont changé d'affectation, soit pour un noyau, soit pour l'ensemble des noyaux. Un noyau sera d'autant mieux défini qu'il est plus stable au cours des approximations et donc que le mouvement est faible. C'est le cas des noyaux 4, 6, 16 et 18 (tab. II).

Dans un ensemble d'horizons à traiter, il existe toujours des horizons de transition entre horizons principaux, ou entre unités d'organisation de la couverture pédologique. Il n'y a pas lieu de créer des noyaux pour ces horizons. Mais ceux-ci, par leur caractère de transition se distinguent très bien par l'analyse du mouvement dans le modèle DIMITRI. En effet, ces horizons correspondent aux unités d'information qui sont constamment en mouvement entre les noyaux. Aussi est-il fréquent de trouver un noyau dans lequel se retrouvent tous ces horizons. C'est le cas du noyau 12 (tab. II).

En conséquence, il y a toujours du mouvement entre deux approximations. On peut estimer que la meilleure image des horiozns est obtenue quand les noyaux sont les plus stables possible, et quand la somme des distances minimum est la plus faible possible.

Tableau II : Etude des transferts des Unités d'information entre deux approximations.
Tableau II : Study of information units transfert between 2 approximations.

		2e APPROXIMATION (Ap. 2)																	|L|+|C| = M	L+C = D
	UR	4	6	8	15	12	13	14	9	11	10	16	17	19	18	7	L	C		
	4																0	+ 1	1	1
1re	6	1															— 1	+ 0	1	— 1
A	8								1								— 1	+ 1	2	0
P	15																0	+ 8	8	8
P	12			1			2		3		1						— 7	+ 2	9	— 5
R	13				1	1											— 2	+ 2	4	0
O	14				3												— 3	+ 0	3	— 3
X	9																0	+ 4	4	4
I	11																0	+ 2	2	2
M	10					1				1							— 2	+ 1	3	— 1
A	16													1			— 1	+ 0	1	— 1
T	17				1												— 1	+ 1	2	0
I	19				3								1				— 4	+ 1	5	— 3
O	18									1							— 1	+ 0	1	— 1
N	7																0	+ 0	0	— 0
	C	1	0	1	8	2	2	0	4	2	1	0	1	1	0	0			46	0

Tableau II : Lu ligne par ligne, ce tableau indique pour chaque noyau de la première approximation, combien il y a d'unités d'information qui lui ont été otées lors de la deuxième approximation(d'où le signe négatif en bout de ligne) et où elles se sont réparties.

Lu colonne par colonne, ce tableau indique combien d'unités d'information sont venues s'adjoindre au noyau de la seconde approximation (d'où le signe positif en bas de colonne) et de quel noyau de la première approximation il provient. Ainsi le noyau 15, nouveau dans la deuxième approximation, s'est constitué à partir de 1 unité d'information de l'ancien noyau 13, de 3 de l'ancien 14, de 1 de l'ancien 17 et de 3 de l'ancien 19.

D = L + C : la somme algébrique des lignes et colonnes indique la variation du nombre d'unités d'information pour chaque noyau entre les deux approximations.

M = |L| + |C| : le mouvement M indique le nombre d'unités d'information qui a changé d'affectation (de noyau) entre les deux approximations. Dans le cas présenté, le mouvement concerne 46 Unités d'information sur les 216 traitées.

Quand on est arrivé à cette phase, on dénomme les noyaux des REFERENCES.

III. UTILISATIONS ET APPLICATIONS PEDOLOGIQUES DE DIMITRI

DIMITRI essaie de formaliser mathématiquement l'approche utilisée par le pédologue lorsqu'il saisit de l'information, qu'il élabore une classification, ou qu'il cartographie.

Cet algorithme a déjà été testé dans diverses régions : en Bourgogne par M.-C. GIRARD (1983), dans le Marais de l'Ouest par D. KING (1986), dans le Massif Central par Mme O. DUVAL, dans la plaine de Kairouan en Tunisie par C. DEROUICH (1983).

A) Le choix des distances et codage

Le regroupement des unités d'information dans une même référence se fait individu par individu sur la base d'une distance minimum. Les études menées ont permis de constater que 60 à 70 pour cent des unités d'information se regroupent de la même façon autour des mêmes références, même si l'on change de distance. Les 30 ou 40 % restant correspondent à des intergrades d'un point de vue typologique ou à des horizons de transition d'un point de vue cartographique.

Les choix les plus importants concernent donc les variables et leur codage.

B) Les données manquantes

Les fichiers de données pédologiques complets sont très rares et les données manquantes posent toujours problème au statisticien. Les calculs de distances d'une unité d'information aux différents noyaux sont effectués en ne tenant compte que des seules données présentes. Ceci n'affecte pas les calculs de distances des autres unités d'information puisqu'il n'y a, à aucun moment, création d'une matrice générale des distances.

C) Les apports et les perspectives de cette méthode

DIMITRI est programmé sur micro-ordinateur (KING et DUVAL, 1988). Le programme est inter-actif. L'intervention du pédologue existe à des endroits précis qui sont toujours les mêmes :

— choix des variables dans le calcul de la distance,

— choix de la distance : distance euclidienne, métrique L 1, distance de HIERNAUX, distance de rang,

— choix des noyaux et de leur caractérisation au début et au cours de l'algorithme.

En effet, il est possible à tout moment de créer de nouveaux noyaux. On juge de l'intérêt de cet ajout lors de l'itération suivante : il y a, ou non, des unités d'information qui viennent alimenter le noyau (Tab. II).

On ajoute des noyaux quand on détecte des unités de transition, on en supprime quand on désire se situer à un niveau d'organisation différent. Pour ce faire, la simple valeur de la distance n'est pas toujours assez précise. On pallie cet inconvénient par l'examen des contributions de chaque variable.

Ces possibilités de choix ont été retenues en tenant compte d'une part, de la grande variété des milieux pédologiques et d'autre part, de la diversité non moins grande des objectifs : agronomique, taxonomique, pédogénétique...

Le déroulement du reste du programme se fait indépendamment de toute intervention, ce qui assure une objectivité certaine.

Cependant, le problème de la convergence n'est pas assuré. Il existe vraisemblablement des minimums locaux.

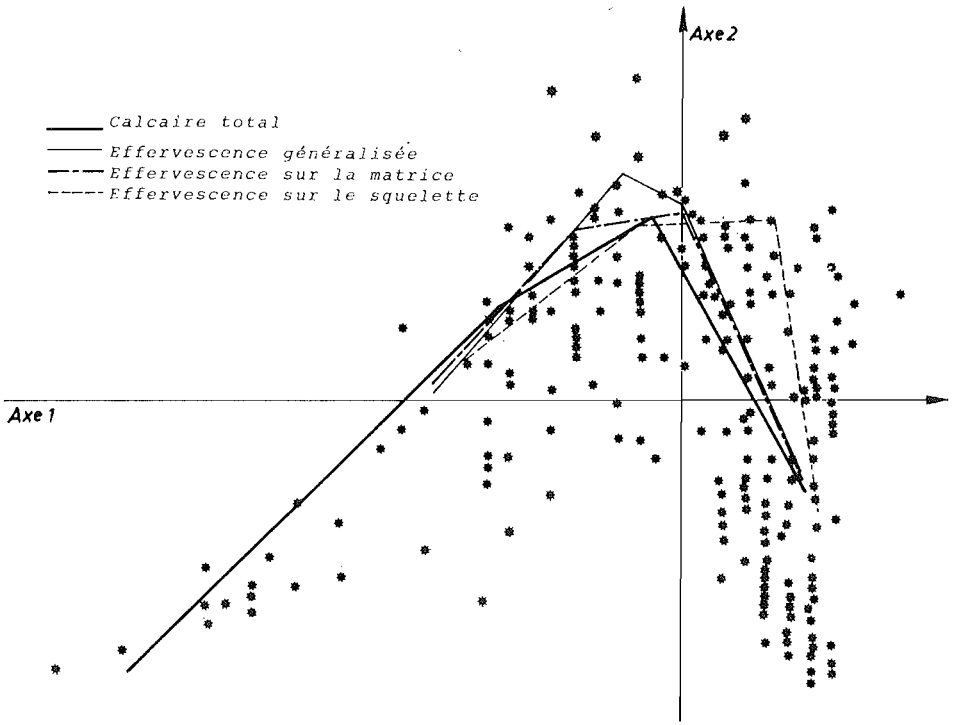
Sur le tableau I, on arrive à une convergence à la quatrième approximation pour les 216 unités d'informations. On a repris ces mêmes unités d'information dans un fichier plus large de 956 unités d'information. On s'aperçoit alors que la convergence a lieu lors de la huitième approximation. Il y avait donc un minimum local à la quatrième approximation.

Il est aussi possible d'utiliser DIMITRI sans intervention du pédologue si l'on fixe une règle de choix des caractéristiques des noyaux lors des itérations.

Ainsi l'algorithme DIMITRI permet de suivre pas à pas les classements fournis par le calcul des distances. Une série de méthodes sont proposées pour vérifier à tout moment les conséquences du choix de tel ou tel paramètre du modèle.

D) La recherche des variables utiles

La caractérisation des unités d'information se fait par des variables plus ou moins redondantes. De telles variables suivent le même cheminement sur les plans d'une AFC. On estime ainsi que le calcaire total et les trois effervescences mesurées sur le terrain (fig. 8) ont le même rôle dans la description des Unités d'information. On peut donc supprimer ces variables redondantes. On diminue ainsi le nombre de variables et on refait une approximation (tableau I : approximations 2 à 5). Tant que la distance minimum moyenne diminue, on peut estimer que le nouvel ensemble de variables est satisfaisant pour décrire l'ensemble des unités d'information étudiées.



Variables redondantes ayant le même « cheminement » sur le plan principal 1-2 d'une analyse factorielle des correspondances.

Figure 8 : Le calcaire total (1) et les trois effervescences généralisées (2), sur la matrice (3), sur le squelette (4), ont le même cheminement.

Figure 8 : Redundants variables with the same « trip » on the main plane (1-2) of factorial analysis.

E) Les références

La notion de référence fournit une description précise et chiffrée du milieu pédologique.

Si l'on a besoin de comparer un horizon à ces références, il suffit alors de calculer les distances de cet horizon à toutes les références : la distance minimum

indique la référence la plus proche de l'horizon. La première phase du modèle DIMITRI exécute ce calcul. Il est possible d'utiliser DIMITRI comme instrument de classement d'horizons dans un système typologique.

De cette façon, on dispose d'une méthode objective pour le choix des regroupements des unités d'information en unité de référence, ce qui est important si celles-ci servent de base à une modélisation de la couverture pédologique.

F) L'organisation des références

Il est possible de définir l'organisation des références à partir des distances calculées par DIMITRI. Pour chaque noyau, la somme des distances minimales représente en fait la distance intra-constallation (VAN DEN DRIESSCHE, 1965 - GIRARD, 1969).

On peut aussi supprimer dans un noyau 15 % des unités d'information (les plus éloignés) et calculer la somme des distances minimum des horizons restants. La comparaison des divers noyaux sur la base de cette distance a pour avantage d'éviter les queues de distribution. On retrouve le noyau 12 qui a une queue de distribution constituée d'unités d'information qui sont à grande distances : 10,2, pour une moyenne de 4,8 (Tab. III).

Tableau III : Valeurs des distances minimum moyennes des noyaux lorsque l'on enlève ou pas, la queue de distribution, et valeur des distances minimum moyennes de la queue de distribution.

Tableau III : Values of mean minimum distance of nuclei when end of distribution is withdrawn or not, and mean minimum distance values of end of distribution

			Suppression des 15%		Les 15 %	
UR	$\bar{d}$	N	$\bar{d}$	N	$\bar{d}$	N
4	5,3	51	4,4	43	9,8	8
6	3,4	29	2,6	25	8,8	4
8	6,0	116	5,4	98	9,4	18
15	3,8	77	3,2	66	7,4	11
12	4,8	80	4,1	69	10,2	11
13	4,7	68	4,0	60	10,0	8
14	3,6	59	3,0	50	7,1	9
9	6,1	49	5,4	42	9,8	7
11	5,3	75	4,7	65	9,1	10
10	6,0	64	5,5	55	9,0	9
16	5,2	19	4,5	16	9,0	3
17	3,0	104	2,5	86	5,3	18
19	5,3	58	4,5	49	9,7	9
18	0,9	8	0	?	7,0	1
7	5,6	99	4,7	84	10,8	15
S	4689		3447		1242	
$\bar{\bar{d}}$	4,9		4,2		8,8	

$\bar{d}$: distance moyenne pour une UR d'une approximation

N : nombre d'Ui appartenant à l'UR

S : Somme des distances minimum pour une approximation

$\bar{\bar{d}}$: moyenne des distances minimum pour toutes les Ui d'une approximation.

On peut ensuite calculer les distances des noyaux entre eux, ce qui correspond à des distances inter-constellations.

Pour définir plus finement une référence, il est nécessaire de définir ses parties modale, périmodale, et marginale (GIRARD, 1981). Cela se fait sur la population statistique qui sert à caractériser la référence. Ces diverses parties peuvent se définir par approximation entre deux références proches l'une de l'autre (fig. 9).

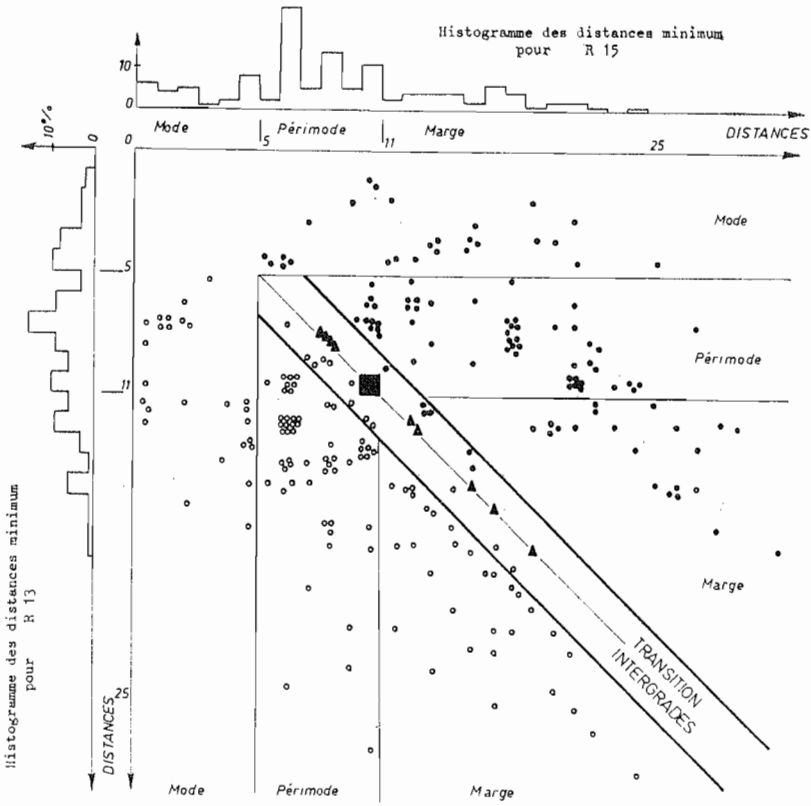


Figure 9 : Représentation des unités d'information par rapport à deux horizons de référence (n° 13 et n° 15) proches l'un de l'autre. Définition des modes, périmode, et marge.

Figure 9 : Information units representation towards 2 neighbouring reference horizons (n° 13 and 15). Définition of mode, perimode, and margin.

Chaque unité d'information est représentée par un point dans un graphe ayant pour coordonnée la valeur des distances à chacune des références. On constate qu'il n'existe aucune unité d'information qui soit à la fois à une distance inférieure à 0,005 de chacune des deux références. Les unités d'information qui ont une distance inférieure à cette valeur vis-à-vis d'une seule des deux références constituent l'ensemble qui permet de définir le mode de cette référence. La distance inter-référence (indiquée par un carré sur la figure 9) permet de situer le seuil entre le périmode et la marge. Les unités d'information qui sont à égale distance des deux références correspondent à des intergrades en matière taxonomique et à des transitions en matière cartographique. Il est possible dans un système de dénomination de les appeler par les deux noms correspondant respectivement aux deux références (GIRARD, BAIZE, 1987).

Le modèle DIMITRI permet ainsi de définir trois parties dans une population caractérisant une référence : le mode, le périmode, la marge. Le tableau IV donne l'exemple de l'évolution du taux d'argile pour ces trois parties dans le cas d'une référence :

— 64 pour cent des individus de la partie modale ont un taux d'argile compris entre 35 et 45 pour cent ;

— 53 pour cent des individus de la partie périmodale ont un taux d'argile compris entre 35 et 45 pour cent ;

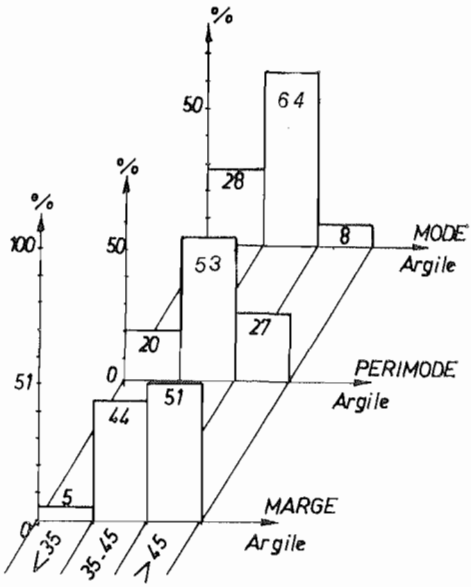
— 44 pour cent des individus de la partie marginale ont un taux d'argile compris entre 35 et 45 pour cent, alors qu'on trouve 51 pour cent des individus qui ont plus de 45 pour cent d'argile.

Ainsi, lorsqu'on passe de la partie modale, vers la partie marginale, le taux d'argile est de plus en plus fort.

La méthode proposée facilite la description des transitions plus ou moins nettes d'un milieu pédologique connu puisqu'elle est basée sur une discrétisation de la couverture pédologique. Il faut noter que les références qui sont caractérisées par le calcul comme des intermédiaires, correspondent soit à des horizons intergrades dans l'espace typologique, soit à des transitions repérables sur le terrain dans l'espace géographique.

Tableau IV : Répartition de l'argile selon trois classes d'abondance (moins de 35 %, 35-45 %, 45 % et plus) en fonction du mode, du périmode et de la marge d'une référence.

Tableau IV : Clay distribution into three classes of abundancy (less than 35 %, 35-45 %, more than 45 %) according mode, perimode and margin of a reference.



Y \ X	< 35	35-45	> 45
mode	28	64	8
périmode	20	53	27
marge	5	44	51

Les valeurs dans le tableau correspondent au pourcentage d'unités d'information ayant un taux X d'argile dans la partie Y de la distribution.

CONCLUSIONS

La plupart des pédologues adoptent une démarche logique lorsqu'ils recherchent une organisation au sein de la couverture pédologique : milieu vaste et continu.

Le travail présenté ici a consisté à analyser cette démarche et à en programmer les différentes phases :

— saisie de l'information de manière plus ou moins systématique (variables manquantes) et avec des variables de diverses natures (qualitatives et quantitatives),
— recherche des horizons caractéristiques,
— classification à des fins typologiques ou cartographiques,
— comparaison d'un individu réel à un concept.

Le modèle DIMITRI a pour principal résultat d'établir, à partir de variables qualitatives ordonnées ou quantitatives :
— des horizons de référence qui peuvent constituer la base d'une typologie,
— des volumes de référence qui peuvent constituer la base d'une analyse spatiale de la couverture pédologique,
— des couches de référence (par exemple un labour) qui peuvent constituer la base d'une étude des comportements agronomiques.

La façon d'aborder un milieu est certainement dictée par l'organisation de celui-ci. Testé sur des milieux très différents, le modèle DIMITRI a fourni des résultats satisfaisants. La multiplication des expériences sur d'autres sites permettrait de vérifier les possibilités d'extension spatiale de ce modèle à différents milieux pédologiques.

Le modèle basé sur DIMITRI est surtout une tentative de modélisation de la pensée du pédologue qui analyse la couverture pédologique.

reçu pour publication : novembre 1987
accepté pour publication : juin 1988

REMERCIEMENTS :

Nous remercions tout particulièrement MM. D. BAIZE, J.-P. LEGROS et PAGES qui nous ont donné leur conseil lors de la relecture de ce texte.

ALGORITHM FOR SOIL MANTLE HORIZONS CLASSIFICATION : DIMITRI

Analysis of soil mantle for mapping and taxonomy may be performed by studying organisation of heterogeneous volumes horizons (fig. 1). These horizons are defined by a set of qualitative ordinale variables.

An algorithme for realizing a non hierarchical typology : DIMITRI is proposed. The first step consist in coding variables (fig. 2). Then « nucleus-horizons » are chosen (fig. 3 and 4) as well as a distance : for example the rank distance (annex.). In a second step horizons nearer to each nucleus are gathered. A population constitued solely of real horizons is such defined for each nucleus. A first approximation is obtained (fig. 5). In a third step, the soil scientist with the help of histograms (fig. 6 and 7) defines the nucleus. A new approximation is then realized and iterated in order to obtain a minimum distance between real horizons and nucleus (tab. I). Moves of real horizons changing nucleus gathering for two successive approxamations may be studied (tab. II). For the last approximation (tab. III) nucleus are named : Reference. A typology of horizons is thus obtained.

DIMITRI enables to take in account horizons even if they present missing data. New nuclei may be added during computation while the number of variables may decrease (fig. 8). The Reference may be described by their modal, perimodal and marginal parts (fig. 9 and tab. IV). The results show all the interest of this algorithm for mapping.

Bibliographie

AUBRY A.-M., 1975. — Programmes Fortran pour distances de Rang, Constellations et correlations. Init. Doc. Techn. ORSTOM, 43 p.

BAIZE D., 1986. — Couvertures pédologiques, cartographie et taxonomie, Science du Sol, v. 24, n° 3, pp. 227-243.

BERTRAND R., FALIPOU P., LEGROS J.-P., 1979. — STIPA. Notice pour l'entrée des descriptions et analyses de sols en banque de données. Doc. SES Montpellier. INRA - IRAT, n° 487, 119 p.

BONNERIC Ph., 1978. — Conception et réalisation d'un système cartographique appliqué à la pédologie. Mémoire CNAM, 108 p., Montpellier.

BOTTNER P., GRANDJOUAN G. et NEDELKA E., 1975. — Classification des sols par une méthode multivariable. Application à une séquence bioclimatique méditerranéo-alpine sur roches mères calcaires. Géoderma, 14, pp. 15-45.

BOULET R., HUMBEL F.X., LUCAS Y., 1978. — Analyse structurale et cartographie en pédologie. Cah. ORSTOM, Pédologie, 4, pp. 309-351.

CAMPBELL N.A., MULCAHY M.J., Mc ARTHUR W.M., 1970. — Numerical classification of soil profiles or the basis of field morphological properties. Austr. J. Soil RES. 8, (1), pp. 43-58.

CIPRA J.E., BIDWEL O.W., ROHLF F.J., 1970. — Numerical taxonomy of soils from nine orders by cluster and centroid-component analysis. Soil Sc. Soc. Am. Proc., vol. 34, n° 2, pp. 281-287.

CUANALO C. de la H.E., WEBSTER R., 1970. — A comparative study of numerical classification and ordination of soil profiles in a locality near Oxford. Part. I : analysis of 85 sites. The journal of Soil Sc. Vol. 21, n° 2, pp. 340-352.

DAGNELIE P., 1955. — Les sols forestiers de l'Ardenne : analyse statistique complémentaire. Application de tests non paramétriques. Bull. Inst. Agro. de Gembloux, t. XXIII, n° 2, pp. 107-150.

DEROUICH C., 1983. — Essai de formalisation des limites entre taxons. Mémoire de spécialisation de l'Ins. Nat. Agro. de Tunis. Doc. multig. 111 p., Tunis.

DIDAY E., 1971. — Une nouvelle méthode en classification automatique et reconnaissance des formes : la méthode des nuées dynamiques. Rev. Stat. Appli., XIX, 2, pp. 283-300.

DUVAL O. et KING D., 1982. — Notice pour l'entrée des descriptions et analyses de sols en banque de données : STIPA, INRA-IRAT, Orléans, 2e édition.

GEORGES P., 1970. — Les méthodes de géographie. Que sais-je ? n° 1398, P.U.F., 128 p.

GIRARD M.-C., 1968. — Approche statistique de la notion de Série. Thèse 3e cycle, 200 p., INA P-G Grignon.

GIRARD M.-C., 1969. — Statistique et pédologie détaillée introduction de la mesure des distances δ de Hiernaux. Sc. du Sol, n° 1, pp. 37-62.

GIRARD M.-C., 1977. — L'horizon mis à l'épreuve statistique. Sc. du Sol, pp. 219-230.

GIRARD M.-C., 1981. — Qu'est-ce que les pédologues demandent aux traitements informatiques des données de sol ? SOLS, n° 6, pp. 17-35.

GIRARD M.-C., 1983. — Recherche d'une modélisation en vue d'une représentation spatiale de la couverture pédologique. Thèse Etat, 430 p., INA-PG, SOLS, n° 12, Grignon.

GIRARD M.-C. et BAIZE D., 1987. — Référentiel Pédologique Français. — Première proposition. Doc. multig. AFES-INRA. Juillet, 145 p.

GRIGAL D.F., and ARNEMAN H.F., 1969. — Numerical classification of some forest Minnesota soils. Proc. Soil Soc. Am. 33, (3), pp. 433-438.

GRUIJTER J.J. de, 1977. — Numerical classification of soil and its application in survey. Agri. research reports 855. Pudoc. Wageningen, 117 p.

GOWER J.C., 1971. — A general coefficient of similarity and some of its properties. Biometrics, 27, pp. 857-874.

HIERNAUX J., 1965. — Une nouvelle mesure de distance anthropologique entre populations utilisant simultanément des fréquences géniques, des pourcentages de trait descriptifs et des moyennes métriques. C.R. Acad. Sc. 260, pp. 1748-1750, Paris.

HOLE F.D., HIRONOKA M., 1960. — An experiment in ordination of some profiles. Proc. Soil. Sc. Soc. Am. 24, pp. 309-312.

HUGHES R.E. and LINDLEY D.V., 1955. — Application of biometric methods to problems of soil classification in ecology. Nature, 175, pp. 806-807, London.

JAMBU M., 1978. — Classification automatique pour l'analyse des données. 310 p., Dunod Edit., Paris.

KENDALL M.G. and STUART A., 1966. — The advanced theory of statistics. Vol. 3, Design and analysis, and time series. Griffin, London, 552 p.

KING D., 1986. — Modélisation cartographique du comportement des sols. Thèse doc. ing. 243 p., INRA, Versailles.

KING D. et DUVAL O., 1988. — Notice pour l'utilisation du programme DIMITRI, INRA, Orléans.

LEGROS J.-P., 1982. — L'évolution granulométrique au cours de la pédogénèse. Approche par simulation par ordinateur. Thèse d'état. INRA, Montpellier, 429 p.

MAHALANOBIS P.C., 1936. — On the generalized distance in statistics. Proc. Nat. Inst. Sci. India. (2), 1, 49-55.

MAUCORPS J. et GIRARD M.-C., 1976. — Essai de classification des sols calcaires par traitement statistique. Comparaison avec la classification française. Pédologie XXVI, 3, pp. 225-254. Gand.

MOORE A.W. and RUSSEL J.S., 1966. — Potential use of numerical analysis and Adansonian concepts in soil science. Australien Journal of Science, 29, pp. 141-143.

MOORE A.W. and RUSSEL J.S., 1967. — Comparison of coefficients and grouping procedures in unumerical analysis of soil trace element data. Geoderma, 1, pp. 139-156.

PEDRO G. et KILIAN J., 1986. — Les travaux pédologiques et les études des milieux physiques réalisés par les organismes français de recherche pour le développement dans les régions chaudes. In Sols et Eaux, ORSTOM, pp. 5-65.

OLIVER M.A. and WEBSTER R., 1987. — The elucidation of soil pattern in the Wyre Forest of the West Midlands. England. I. Multivariate distribution. Journ. of Soil Science, vol. 38. 2, 279-291.

OLIVER M.A. and WEBSTER A., 1987. — The elucidation of soil pattern in the Wyre Forest of the West Milands. England. II. Spatial distribution. Journ. of Soil Science, vol. 38. 2, 293-307.

PRUSINKIEWICZ Z., 1969. — Application of multivariate statistical analysis and computers in investigations of the genetic homogeneity of glacial deposits. Zesz. Nauk. UAM geographia 8. Nadbitka poznan, pp. 149-165.

RAYNER J.H., 1969.— Classification of soils by numerical methods. Journ. of Soil Science, vol. 17, n° 1, pp. 79-92.

SIMONNEAUX V., 1987. — Mesure de la ressemblance entre des groupes de sondages à la tarrière et des profils de référence. DAA INA-P-G-INRA 61 p., Grignon, Montpellier.

VAN DEN DRIESSCHE R., 1965. — La recherche de constellations de groupe à partir des distances généralisées de MAHALANNOBIS. Biométrie, 41, 1, pp. 36-47.

VAN DEN DRIESSCHE R., 1966. — Un problème de classification numérique. Cah. ORSTOM, sér. Pédo. vol. IV., n° 1, pp. 125-132.

VAN DEN DRIESSCHE R. et MAIGNIEN R., 1965. — Application d'une méthode de la statistique approfondie à la pédologie. Cah. ORSTOM, sér. pédo., vol. III, fas. 1, pp. 79-88.

VAN DEN DRIESSCHE R. et GARCIA-GOMEZ A., 1972. — Distances non paramétriques entre profils. Rev. Ecol. Biol. Sol., t. IV, n° 4, pp. 257-264.

WEBSTER R., 1971. — Wilks's criterion : a measure for comparing the value of general purpose soil classifications. Journ. of Soil Sc., vol. 22, n° 2, pp. 254-260.

WEBSTER R., 1977. — Quantitative and numerical methods in soil classification and survey. Monographs on soil survey. Clarendon Press, Oxford.

Annexe

Distance de rang

Il ne s'agit pas à proprement parler d'une distance, mais d'un indice de dissimilartié. Il est indépendant de la fonction de répartition des variables, ce qui est très utile en pédologie puisque la plupart du temps on ne connaît pas les fonctions de répartition des variables. On le dénommera « distance de rang ».

Elle se calcule de la façon suivante :

$$D\ (A,\ B) = \frac{1}{V} \sum_{1}^{V} \frac{1}{Y_i} (R_{Ai} - R_{Bi})^2$$

où :

D (A, B) : est la distance de rang entre les individus A et B

R_{Ai} : est le rang de l'individu A pour le variable i

R_{Bi} : est le rang de l'individu B pour le variable i

V : est le nombre de variables présentes pour les individus A et B

i : est une des V variables

Y_i : est un facteur de correction qui tient compte de l'effectif de chaque varible, du nombre d'ex-æquo.

$$Y_1 = \frac{1}{12} (m_i^3 - m_i - \sum_{i}^{ei} (t^3_{qi} - t_{qi}))$$

avec :

m_i : effectif de la variable i pour l'ensemble des individus étudiés

ei : nombre de lots (q) identiques (= ayant le même rang) pour la variable i

t_{qi} : nombre d'individus identiques dans le lot q, pour la variable i.