SCIENCE DU SOL 1988 - Vol. 26 - N° 4 - pp. 239-254 - PLAISIR-FRANCE

REFLEXIONS SUR LA CLASSIFICATION DES PROFILS DE LA COUVERTURE PEDOLOGIQUE PROPOSITION D'UN ALGORITHME : VLADIMIR

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

RESUME

L'algorithme proposé : VLADIMIR, classifie les profils. On choisit des « profils noyaux » qui initialisent des groupes, puis à l'aide d'une distance mathématique, on réunit les profils étudiés au « profil noyau » le plus proche.

Pour ce faire, deux méthodes principales sont proposées :

1) sur toute l'épaisseur du profil, ou à certaines profondeurs, on choisit des niveaux de comparaison entre le profil étudié et les « profils noyaux » ;

2) on se met à certains niveaux, on compare le profil étudié aux « profils noyaux » sur une certaine épaisseur.

Le choix des variables servant à caractériser les profils est important et conduit à des classements qui diffèrent selon les objectifs. La pondération des variables a peu d'effet sur le résultat de la classification. Le choix des niveaux de comparaison des profils influe davantage. Mais les effets sur des profils peu différenciés, donnés en exemple, est faible. L'effet du choix initial des noyaux n'est pas facile à mettre en valeur car les réajustements entre approximations sont faits par le pédologue.

VLADIMIR peut être utilisé aussi bien à des fins cartographiques qu'à des fins taxonomiques ou agronomiques.

MOTS-CLES : Profil - Taxonomie - Algorithme - Statistique - Agronomie.
KEY WORDS : Profile - Taxonomy - Algorithm - Statistics - Agronomy.

INTRODUCTION

Nous avons montré les avantages d'une approche tridimensionnelle de la couverture pédologique (GIRARD, KING, 1988). On peut privilégier l'horizon en tant qu'unité volumique d'échantillonnage (chemin 1, figure 1). Ce choix s'appuie sur des considérations pédologiques puisque l'horizon constitue une entité de base de l'organisation et du fonctionnement de la couverture pédologique.

Ce choix est fondé également sur une analyse statistique simple puisque l'horizon constitue un élément de base rarement subdivisé en sous-unités, permettant ainsi la construction du tableau statistique nécessaire à la branche des mathématiques appelé e « analyse des données » (BENZECRI et al., 1973).

(1) Institut National de la Recherche Agonomique, SESCPF, Ardon, 45160 Olivet.

(2) Institut National Agronomique Paris-Grignon, Science des sols et hydrologie, 78850 Thiverval-Grignon.

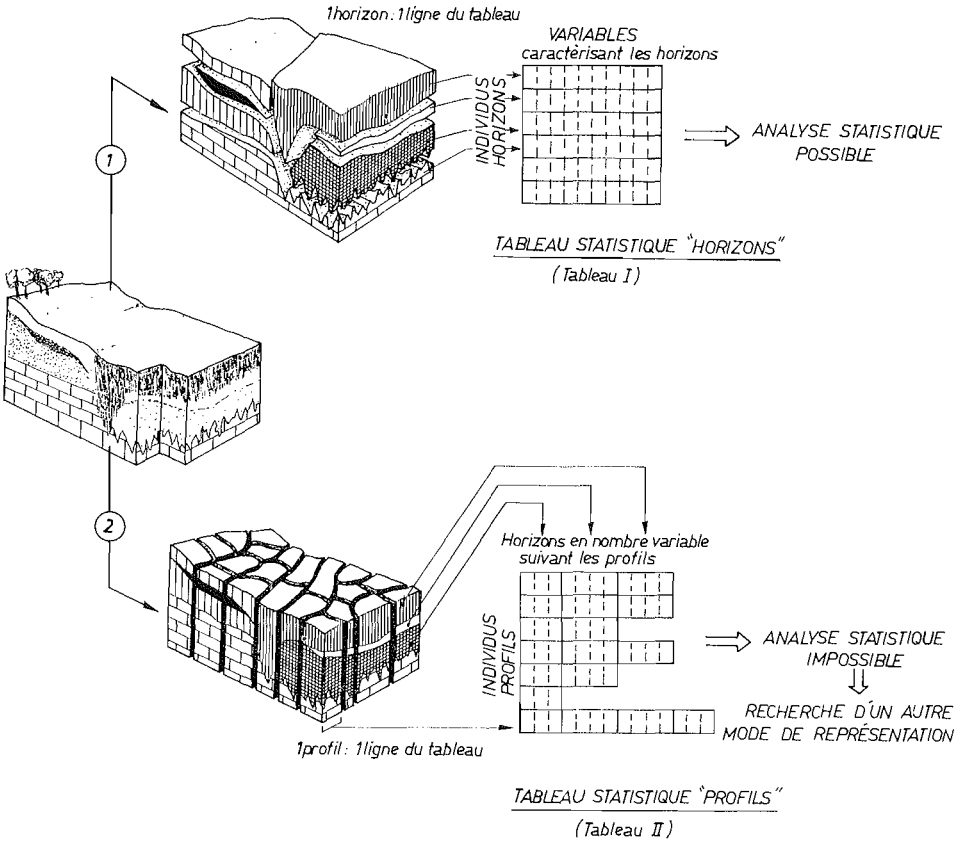


Figure 1 : Schéma des deux principales méthodes d'organisation des données pédologiques :
1) Découpage par volumes homogènes (horizons)
2) Découpage par volumes hétérogènes, organisés (profils)
Sketch of the two main methods for organizing soil data :
1) Cutting into homogeneous volumes (horizons)
2) Cutting into heterogeneous organized volumes (profiles).

Ainsi, l'hypothèse de l'existence d'une partition de la couverture pédologique en volumes homogènes que sont les horizons semble être la meilleure manière d'étudier et de comprendre l'organisation spatiale de la couverture pédologique. Il n'en reste pas moins qu'il est nécessaire d'étudier globalement la couverture pédologique en considérant la position relative de ces volumes que sont les horizons les uns par rapport aux autres sur une même verticale : le profil (chemin 2, figure 1).

C'est principalement dans le cadre d'applications thématiques (aptitudes culturales, drainage...) qu'il faut prendre des décisions en proposant un découpage de la couverture pédologique qui tienne compte globalement du sol (superposition des horizons) et de son environnement : topographie, substrat,... (BOULAINE, 1980).

Pour ce faire, un algorithme nommé « VLADIMIR » a été élaboré pour la classification des profils pédologiques. Il est basé sur les mêmes grands principes que DIMITRI, et pour marquer sa filiation, un nom de même consonance a été choisi. Dans les deux cas, il s'agit de proposer une méthode de classification semi-automatique simulant les travaux des pédologues et agronomes sur le terrain : intégration des variables de nature diverse, prise en compte des connaissances antérieures, introduction de données au fur et à mesure de l'avancement des travaux, classement des individus ayant des données manquantes, etc...

La mise en forme d'un tableau statistique pour des entités « profils » pose un problème d'organisation logique des données : on étudie dans la première partie les différentes propositions bibliographiques. Un choix parmi ces méthodes est justifié et permet, dans une seconde partie, d'exposer le détail de l'algorithme VLADIMIR. Enfin dans une dernière partie, on discute, à travers divers tests de sensibilité, des possibilités de cet algorithme.

I - PRESENTATION DE DIVERS MODES D'ORGANISATION STATISTIQUE DES DONNEES POUR LA REPRESENTATION DES PROFILS

Dans toute analyse numérique, chaque individu est caractérisé par une série de variables retenues pour le décrire. On peut ainsi rapporter l'information à une matrice dite « tableau statistique » dont les lignes sont les individus et dont les colonnes sont les variables.

Dans de nombreuses études pédologiques, l'objectif est de classer des profils alors que les variables portent sur les horizons. Cela pose le problème de la construction du tableau statistique. Une analyse des différentes approches a été faite par M.-C. GIRARD (1983). On peut en retenir les principaux cas suivants.

A) Les individus statistiques sont les horizons

Dans ce cas, (chemin 1, figure 1) l'horizon est considéré comme unité d'organisation et unité d'échantillonnage. Si chaque horizon est décrit par un même ensemble de variables descriptives et analytiques, on peut construire aisément le tableau statistique et lui appliquer différentes méthodes mathématiques (Tab. I).

	V_1	V_2	V_v
H_1			
⋮			
H_h			

Tableau I : Tableau statistique où les lignes sont des horizons et les colonnes sont les variables qui les caractérisent.

Matrix where lines represent horizons and columns the characterizing variables.

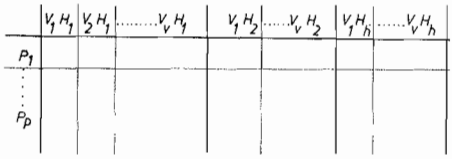
H : horizons variant de 1 à h
V : variables variant de 1 à v

L'étude statistique des sols a été aborbée de cette manière par de nombreux auteurs ; RAYNER, 1966 - MUIR et al., 1970 - GRIGAL et ARNEMAN, 1969 - MARTIN et AUBRY, 1975 - GIRARD, 1976 - FOURNIER et KING, 1982 - GIRARD, 1983. C'est sur ce tableau qu'est fondé l'algorithme DIMITRI. Si l'on veut classer des profils, il faut pouvoir construire un tableau statistique où les lignes du tableau sont des profils. Sinon, et c'est le cas de l'intéressante méthode de RAYNER (1966), en se basant sur une matrice d'horizons, on doit pouvoir comparer de l'ordre de 10 000 horizons, ce qui dépasse les capacités des micro-ordinateurs courants.

B) Les individus statistiques sont des profils

Dans tous les exemples qui suivent (chemin 2, figure 1), les lignes du tableau statistique sont toujours des profils. Les différents cas varient selon la façon dont on caractérise les colonnes.

1. Le cas le plus simple consiste à utiliser un nombre de colonnes égal au nombre de variables décrivant ces horizons (Tab. II). Dans un tel tableau, le nombre de colonnes complètes est variable d'une ligne à l'autre.



P : profil variant de 1 à p
H : horizons variant de 1 à h (h varie selon p)
V : variables variant de 1 à v (v varie selon h)

Tableau II :

Tableau statistique où les lignes sont des profils. Le nombre de colonnes varie selon le nombre d'horizons décrivant chaque profil.

Matrix where lines are profiles. The number of columns varies according to the number of horizons describing each profile.

Peu d'auteurs ont étudié de telles matrices (CHABANEL et al., 1975) étant donné la quasi impossibilité d'analyser mathématiquement un ensemble de données non structurées de façon homogène et pouvant présenter un grand nombre de données manquantes.

2. Afin d'éviter les problèmes précédents, de nombreux auteurs ont proposé de ne choisir que quelques variables particulières caractérisant les horizons reconnus comme typiques (CIPRA et al., 1970 - ARKLEY, 1971 - MAUCORPS et GIRARD, 1976). On peut également retenir des variables « indicatrices de pédogénèse » en fournissant des valeurs de relation entre horizons (MOORE et al., 1972 - BOTTNER et al., 1975 - MARTIN et AUBRY, 1975) (Tab. III).

P : profil variant de 1 à p
A, A/B, ...C : horizons choisis pour les p profils.

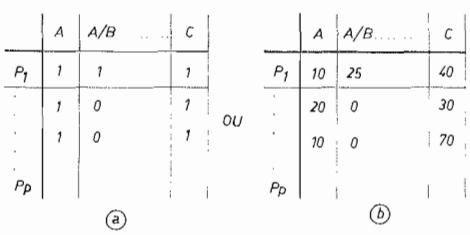
Tableau III :

Tableau statistique où les lignes sont des profils. Le nombre de colonnes varie selon le nombre d'horizons choisis et de variables choisies.

Matrix where lines are profiles. The number of columns varies according to the selected number of horizons and variables.

L'inconvénient majeur de cette méthode réside dans le choix a priori des variables et des types d'horizons. Cela demande une bonne connaissance du terrain étudié et une référence à une taxonomie déjà construite. Un second inconvénient provient de la dénomination a priori d'un horizon pour indiquer telle variable dans telle colonne. Ces choix, sources d'erreur, sont souvent discutables (GRIGAL et ARNEMAN, 1969 - SCHELLING, 1970).

3. On peut aller plus loin dans ce sens, en reportant dans chaque colonne l'indication de la présence ou de l'absence d'un horizon dit « diagnostic », ou l'épaisseur de l'horizon (Tab. IV).



P : profil variant de 1 à p
A, A/B, ...C : horizons choisis pour les p profils.

Tableau IV :

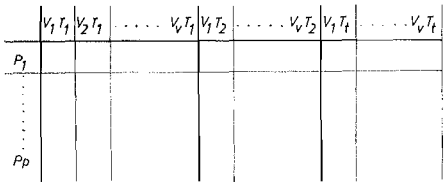
Tableau statistique où les lignes sont des profils. Le nombre de colonnes varie selon le nombre d'horizons choisis. On indique :
a) la présence(1) ou l'absence (0), ou
b) l'épaisseur de l'horizon (en cm)

Matrix where lines are profiles. The number of columns varies according to the selected number of horizons.
a) gives presence (1) or absence (o),
b) horizon thickness (cm).

Pour appliquer une telle méthode, il faut disposer d'une typologie d'horizons (GIRARD, 1977). L'attribution d'un nom à un horizon est souvent discutable, aussi est-il souhaitable que cette typologie provienne d'une classification appliquée à un tableau d'horizons (Tab. I).

Le tableau de type IV s'applique très bien aux analyses multidimensionnelles telles que les Analyses Factorielles des Correspondances, ou à des distances de chi² (KING, 1976 - GIRARD, 1983).

4. Une autre solution consiste à ne plus tenir compte de l'organisation des horizons et à découper chaque profil en un nombre fixe de tranches. On attribue à chaque tranche les valeurs des variables de l'horizon qu'elle recoupe (MOORE et RUSSEL, 1966 - GRIGAL et ARNEMAN, 1969 - CAMPBELL et al., 1970 - DE GRUITJTER, 1977 - PAVAT, 1986) (Tab. V).



P : profil variant de 1 à p
T : tranches variant de 1 à t
V : variables variant de 1 à v

Tableau V :

Tableau statistique où les lignes sont des profils. Le nombre de colonnes est égal au produit du nombre de variables par le nombre de tranches.

Matrix where lines are profiles. The number of columns is equal to the number of variables multiplied by the number of layers.

Le nombre de tranches étant fixe, et le même pour chaque profil, le tableau statistique se présente sous une forme analysable. La multiplication du nombre de tranches rapproche cette méthode de celle qui consiste à caractériser chaque profil par une série de profils d'éléments (RUSSEL et MOORE, 1968 - DE GRUIJTER, 1977) ou par une succession de tranches caractérisées par des horizons de référence (GIRARD, 1983).

L'inconvénient majeur de cette méthode est qu'une tranche fixée à une profondeur donnée peut recouper des horizons très différents. L'épaisseur et la profondeur des horizons sont donc deux variables qui jouent indirectement un rôle de tout premier plan dans une telle analyse.

C) Le choix de la méthode

Au travers de cette étude bibliographique et d'essais réalisés sur différents terrains (KING, 1976 - FOURNIER et KING, 1980 - MAUCORPS et GIRARD, 1976 - SIMMONNEAUX, 1987), il apparaît qu'aucune méthode ne permet de décrire d'une façon complète l'organisation pédologique. La méthode proposée par RAYNER (1966) est intéressante car elle compare les profils sur la base des horizons. Mais elle ne permet pas de tenir compte des épaisseurs des horizons au moment de leur comparaison. Ainsi, il n'y aura pas de différence entre deux profils qui auraient la même succession d'horizons, alors que les épaisseurs respectives des horizons seraient très différentes. En conséquence, on propose de réaliser une analyse en plusieurs étapes, chacune d'elles s'appuyant sur un type de tableau statistique différent.

1. Première étape

On construit le tableau statistique des horizons (Tab. I).

On lui applique l'algorithme DIMITRI. On dispose ainsi d'une typologie d'horizons de référence.

2. Deuxième étape

On construit un tableau statistique de profils (Tab. V), où chaque profil est divisé en un nombre fixe de tranches. Ce tableau sert de base pour le calcul d'un indice de similarité entre profils.

3. Troisième étape

On élabore un dernier tableau (Tab. IV a) où les lignes sont des profils et où les colonnes marquent la présence ou l'absence des horizons de référence déterminés par la première étape. Ce tableau permet de construire et de réajuster les « profils noyaux » au cours des approximations successives de VLADIMIR.

La nécessité de disposer de l'information recueillie sous plusieurs formes de tableaux n'est que le reflet de la complexité de l'organisation pédologique qu'il faut traduire sous une forme standard « simple ». Il faut remarquer que le tableau de type IV : présence/absence ou épaisseur d'horizons de référence au sein des profils, correspond à une méthode couramment utilisée par les pédologues. C'est le cas des horizons diagnostics employés par de nombreux systèmes typologiques : par exemple la Soil Taxonomy, le système FAO ou plus récemment le Référentiel Pédologique Français (GIRARD et BAIZE, 1987). Il en est de même pour le tableau V : découpage du profil en niveaux de comparaison. Ce tableau correspond à l'importance relative des variables « profondeur » et « épaisseur » des horizons de référence.

II - PRESENTATION DE L'ALGORITHME VLADIMIR

VLADIMIR est construit en reprenant des étapes similaires à celles de DIMITRI : choix a priori de « profils noyaux », définition d'une distance, approche par itérations successives en contrôlant l'évolution des groupes formés.

Après un bref rappel de l'algorithme DIMITRI, on détaillera les compléments proposés dans l'algorithme VLADIMIR pour prendre en compte les tableaux statistiques de profils pédologiques.

A) Rappel sur l'algorithme DIMITRI

Basé sur le principe de la méthode des « nuées dynamiques » de DIDAY (1971), DIMITRI classe les horizons étudiés à partir d'horizons types préalablement choisis : les « horizons noyaux ». Le tableau étudié est alors de type I.

Pour chaque horizon étudié, DIMITRI calcule la distance à chaque noyau et rattache cet horizon au noyau le plus proche (DIstance MInimale de TRI). Un retour aux données de base après classement de tous les horizons est possible grâce à la production d'histogrammes des variables pour chaque groupe formé autour des noyaux. Simples à interpréter, ces histogrammes permettent le contrôle de la classification et d'éventuels réajustements des noyaux. La construction de nouveaux noyaux permet de réitérer le classement et de procéder par approximations successives. A la fin du traitement, chaque horizon étudié est rattaché à un noyau nommé alors : « horizon de référence ».

B) Choix des « profils noyaux »

Comme pour DIMITRI, les noyaux peuvent être choisis de façons différentes : d'une façon aléatoire, par un raisonnement pédologique, par un raisonnement statistique.

Il est toujours possible de tirer au hasard des profils au sein d'une population étudiée. Les approximations successives permettent alors de réajuster ce choix.

On peut considérer certains profils comme caractéristiques :
— par rapport à un raisonnement pédogénétique,
— en tenant compte d'études antérieures,

— en utilisant des classifications régionales ou internationales, et donc les prendre comme noyaux.

Si l'on ne veut pas introduire d'hypothèses préalables, on peut construire les « profils noyaux » par raisonnement statistique. Pour cela, on établit, à partir d'une classification de type DIMITRI, un tableau de données où les lignes sont des profils et les colonnes sont les horizons de référence (Tableau de type IV). A partir de ce tableau, une analyse factorielle des correspondances fournit une image des profils. On choisit alors les profils les mieux centrés (KING, 1986).

Une telle méthode ne tient pas compte de la profondeur et de la position relative des horizons les uns par rapport aux autres. Pour cette raison, on propose une approche complémentaire qui fournit la position fréquentielle des horizons au sein des profils (GIRARD, 1983 - Programme SUSOU de la bibliothèque LOGOS, KING et DUVAL, 1988). D'une part on édite les histogrammes des profondeurs du haut et du bas de tous les horizons regroupés autour d'un horizon noyau donné ; d'autre part, on construit des tableaux croisés permettant de connaître le nombre de fois où deux horizons sont adjacents dans le même profil et le nombre de fois où deux horizons sont simplement présents simultanément dans le même profil (fig. 2). Ces différents tableaux et graphiques permettent de reconstituer manuellement les profils dont la succession d'horizons est la plus fréquente, ainsi que les profondeurs où il y a le plus fréquemment des limites entre horizons. Il serait possible d'automatiser cette étape.

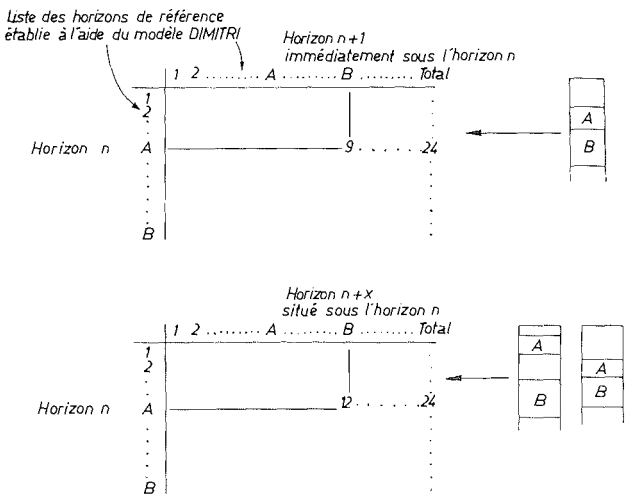


Figure 2 :

Construction des « profils noyaux » à l'aide du tableau de fréquence de la présence simultanée dans un même profil d'horizons de référence obtenus par DIMITRI.

Building « nucleus profiles » with frequency table of simultaneous occurence in a same reference horizons profile as obtained by DIMITRI.

C) Choix d'une distance

La discussion sur les différents types de distances est la même pour DIMITRI (GIRARD et KING, 1988) que VLADIMIR. Les options proposées dans le programme informatique sont aussi les mêmes. En particulier, les algorithmes permettent de classer tous les individus, même s'ils comportent des données manquantes (KING et DUVAL, 1988).

Le problème se situe plus particulièrement au niveau du choix du type de tableau statistique sur lequel on applique de tels calculs de distance. On a choisi de diviser chaque profil en un nombre de tranches défini, fixant ainsi un niveau de comparaison (Tableau de type V). Pour un profil réel P_i donné, on peut caractériser sa distance

à un profil de référence P_R par la somme des distances entre les horizons rencontrés à chaque niveau de comparaison (fig. 3).

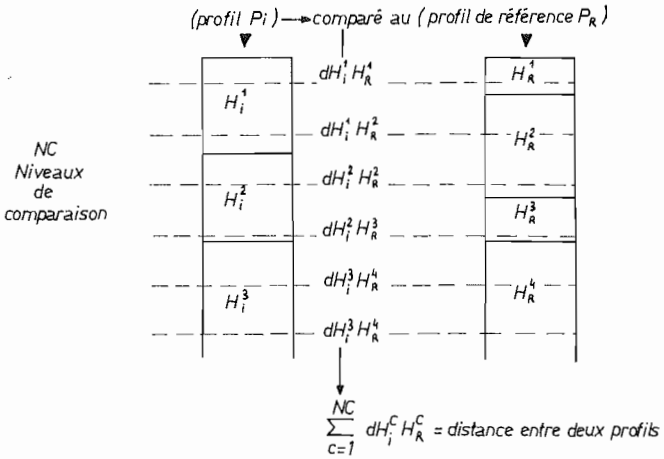


Figure 3 :

Méthode pour le calcul de la distance mathématique entre deux profils par niveaux fixes de comparaison.

Method for computing mathematical distance between two profiles using fixed comparison levels.

Le choix initial du nombre et de la profondeur des niveaux de comparaison est libre : on peut travailler d'une manière systématique (tous les x cm, par exemple), ou d'une façon raisonnée en tenant compte de l'organisation du milieu. Enfin on peut également établir un choix arbitraire en donnant plus d'importance à certaines profondeurs en fonction de problèmes agronomiques.

Le choix du nombre et de la profondeur des niveaux permet d'ajuster les comparaisons en fonction des objectifs. L'inconvénient d'une telle méthode est d'opérer sans tenir compte des fluctuations possibles des profondeurs des horizons d'un profil à l'autre. Dans un cas extrême, des profils ayant la même succession d'horizons, mais à des profondeurs légèrement différentes peuvent être considérés comme très distants (SIMONNEAUX, 1987).

On propose donc de ne plus se situer sur un niveau de comparaison, mais sur une tranche dont l'épaisseur est choisie (figure 4). A un niveau de comparaison Nt donné, correspond un horizon du profil P_i à classer. On recherche tous les horizons

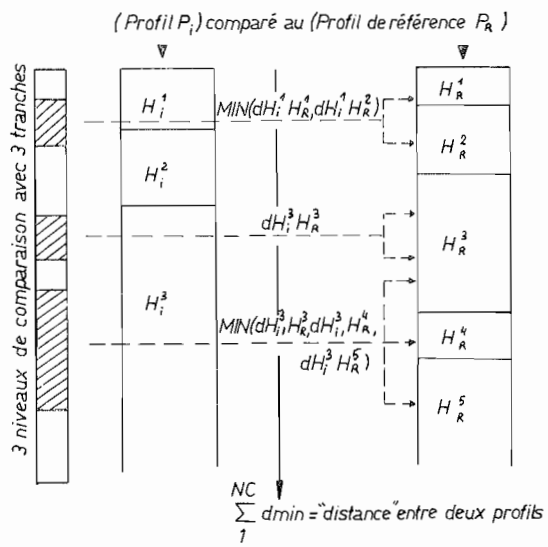


Figure 4 :

Méthode pour le calcul de la distance mathématique entre deux profils par niveaux de comparaison sur des tranches d'épaisseur variable.

Method for computing mathematical distance between two profiles using comparison levels on layers with varying thickness.

du profil de référence P_R recoupé par la tranche enveloppant le niveau Nt. On calcule l'ensemble des distances entre l'horizon du niveau Nt du profil étudié P_I, et tous les horizons trouvés dans P_R. On ne conserve que la distance minimum pour ce niveau. On fait la sommation globale sur l'ensemble des niveaux pour définir la distance entre profils (Annexe 1).

Par cette méthode de « tranches de comparaison ajustables » l'algorithme VLADIMIR tente de se rapprocher d'une vision globale du profil. Le profil est conçu, non pas comme une sommation de tranches, mais comme un ensemble d'entités que sont les horizons.

D) Réajustement des noyaux

C'est une phase importante puisqu'elle permet de contrôler si chaque noyau est représentatif des individus qu'il a regroupés autour de lui lors d'une approximation. Avec DIMITRI, ce travail est réalisé par l'observation approfondie des histogrammes des variables intervenant dans le calcul de la distance. Avec VLADIMIR, il serait trop long d'examiner l'ensemble des histogrammes dont le nombre peut être très grand. Pour 20 noyaux décrits avec 30 variables sur 10 niveaux de comparaison, il existe 6 000 histogrammes.

En fait, la donnée qui synthétise le mieux l'information pour un horizon est le type d'horizon de référence DIMITRI auquel il se rattache. Ainsi, pour chaque groupe et pour chaque niveau de comparaison, on édite seulement l'histogramme des types d'horizons de références rencontrés. L'analyse des histogrammes permet de redéfinir les « profils noyaux » soit en précisant la profondeur de ces horizons, soit en ajoutant ou en retranchant des horizons complets, soit enfin en définissant de nouveaux « profils noyaux ».

E) Itérations et approximations

Les différentes phases d'une itération pour VLADIMIR sont analogues à celles pour DIMITRI. Elles se résument en quatre étapes :

1) établissement des « profils noyaux » initiaux ;
2) classement de chaque profil auprès du noyau mathématiquement le plus proche ;
3) analyses des histogrammes ;
4) réajustement des profils noyaux et possibilité d'un nouveau classement.

Entre deux itérations, on peut étudier les profils qui ont changé d'affectation. On observe ainsi les noyaux stables et le nombre de mouvements d'une itération à l'autre, c'est-à-dire le nombre d'individus ayant changé de groupe (programme CROISA, in LOGOS, KING et DUVAL, 1988). On cesse les itérations lorsque la moyenne des distances minimum ne diminue plus. Les profils noyaux sont alors appelés : Profils de Référence.

III - DISCUSSION : TEST DE SENSIBILITE

Pour mieux guider l'utilisateur devant les choix multiples, on a testé la sensibilité des méthodes sur trois points :

— le choix des variables,
— les pondérations possibles, avec en particulier le choix du nombre de niveaux de comparaison,
— le choix des « profils-noyaux ».

A) Le choix des variables

Le choix des variables et de leur codage est un point essentiel pour la constructions d'un tableau statistique de bonne qualité (BENZECRI, 1973 - LEBART et al., 1982). DIMITRI et VLADIMIR n'échappent pas à la règle puisque des modifications apportées à ces choix préliminaires provoquent entre 10 % et 40 % de mouvement. Il est donc essentiel de raisonner et de justifier le choix des variables.

Le terrain qui a servi pour ces essais est situé en Charente-Maritime, sur des sols argileux sodiques de dépôts alluviaux de bordure de mer. Ces sols présentent des problèmes cartographiques (peu de différenciation et paysage monotone) et des problèmes de comportement (dégradation de la structure entraînant un mauvais drainage). Il est essentiel dans un tel cas de savoir si des critères de perception immédiate utilisés pour la cartographie suffisent à discriminer les principaux comportements reconnus à partir d'analyses plus fines de laboratoire.

A partir de 33 fosses pédologiques, on a adopté 3 attitudes correspondant à 3 objectifs différents (KING, 1987) :

— point de vue taxonomique (TAXO), le plus global possible,

— point de vue agronomique (AGRO), en tentant de discriminer des comportements,

— point de vue cartographique (CARTO), en tenant compte des contraintes dues au nombre restreint de variables disponibles.

La comparaison des résultats des différentes classifications montre une assez bonne adéquation entre les groupes formés autour des références correspondant aux trois attitudes (Tab. VI).

Tableau VI : Comparaison entre des classifications de profils pédologiques obtenues à partir de choix de différentes variables.

TAXO : choix du maximum de caractères descriptifs et de variables d'analyses de laboratoire.

AGRO : choix de critères agronomiques liés aux problèmes de comportements régionaux.

CARTO : choix de critères cartographiques pertinents et fiables.

A : comparaison entre les références AGRO et TAXO

B : comparaison entre les références AGRO et CARTO

(les lettres affectées aux différentes références n'ont pas de rapport entre elles, les références ont été ordonnées sur les tableaux selon une progression évolutive des processus de pédogénèse).

Comparison between soil profiles classification obtained from the choice of variables.

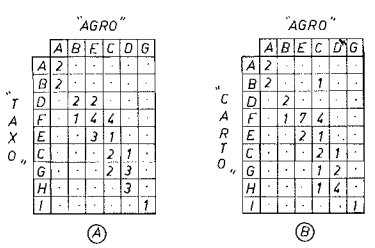
TAXO : choice of maximum descriptive characters and laboratory analysis variables.

AGRO : choice of agronomical characters linked to problems of regional behaviours.

CARTO : choice of pertinent and reliable mapping criteria.

A : comparison between AGRO and TAXO

B : comparison between AGRO and CARTO



Le mouvement est de 11 sur 33 pour la comparaison entre les groupes AGRO et TAXO, et de 10 entre les groupes AGRO et CARTO. Lorsqu'un groupe pour l'une des classifications éclate en plusieurs groupes pour une autre classification, les individus se répartissent autour de références mathématiquement proches. Par ailleurs on constate qu'il existe une bonne correspondance entre des références extrêmes (en haut à gauche, ou en bas à droite des tableaux VIa et VIb). Cela s'explique par le caractère particulièrement typé de ces sols : sols argileux de dépôts récents, très sodiques ; sols humifères en surface et tourbes. A l'inverse, pour une référence centrale comme par exemple « F-CARTO » qui est un sol argilo-sodique dessalé à amas gypseux en profondeur, on n'a pas pu distinguer à l'aide des variables cartographiques, des profils différents ayant pourtant des comportements différents. Il s'agit de sols dont les caractères morphologiques actuels ne permettent pas de discerner une évolution de l'équilibre chimique des ions impliquant certainement une modification des comportements (DAMOUR et al., 1984).

En conclusion, une forte inadéquation entre plusieurs classifications ne doit pas entraîner un rejet des algorithmes. Au contraire, différents choix raisonnés de variables mettent en évidence les individus pour lesquels des travaux plus fins sont nécessaires. De plus, ces matrices de comparaison nuancent d'une façon chiffrée les possibilités de thématisation agronomique à partir des cartes pédologiques.

B) La pondération des variables, le choix des niveaux de comparaison

Les programmes DIMITRI et VLADIMIR de la bibliothèque LOGOS offrent un choix d'options possibles pour le type de distance. On n'analysera pas ces options mais on retiendra le problème de la pondération.

Pour accentuer le choix d'une variable, il est possible d'affecter un coefficient multiplicatif à celle-ci. Plusieurs essais montrent que les groupes se modifient très peu en fonction de ces pondérations différentes : moins de 10 % du mouvement. Les individus à l'origine de ce faible mouvement sont à des distances quasi-égales de plusieurs noyaux. Il s'agit donc d'individus difficiles à classer et l'affectation de poids est une méthode rapide pour les repérer.

Le choix des niveaux de comparaison dans VLADIMIR est une autre façon d'affecter des pondérations. En privilégiant certaines profondeurs, on donne plus de poids à telle ou telle tranche reconnue comme importante pour l'objectif choisi. Il est possible de tenir compte des épaisseurs des différents horizons dans la comparaison des profils, ce qui est très important dans la pédogénèse et les applications de la pédologie. Un des inconvénients de la méthode proposée par RAYNER (1966) est justement de ne pas en tenir compte.

Pour tester le choix des niveaux de comparaison, on a analysé trois classements VLADIMIR, à partir d'un tableau de type « profil », avec les mêmes noyaux initiaux. Les profondeurs ont été choisies (figure 5) :

— soit par un raisonnement pédologique : nombreux niveaux en surface du fait de l'épaisseur plus faible des horizons supérieurs (17 niveaux),
— soit d'une façon systématique avec un pas de 5 cm (50 niveaux),
— soit sur 10 niveaux seulement, mais en regardant sur une tranche.

Les groupes obtenus varient peu d'une option à l'autre. En comparant les niveaux choisis pédologiquement à ceux choisis systématiquement, il n'y a que 8 profils sur 33 qui changent d'affectation, et ce entre des noyaux intermédiaires, pédologiquement peu différenciés.

Le choix des 50 niveaux de comparaison donne plus de poids aux horizons de profondeur. Les 8 profils concernés par le mouvement ressemblent plus à tel noyau pour la surface, et à tel autre pour la profondeur. La majorité des profils n'a pas changé de groupe, prouvant ainsi la relation existant entre les phénomènes de surface et de profondeur.

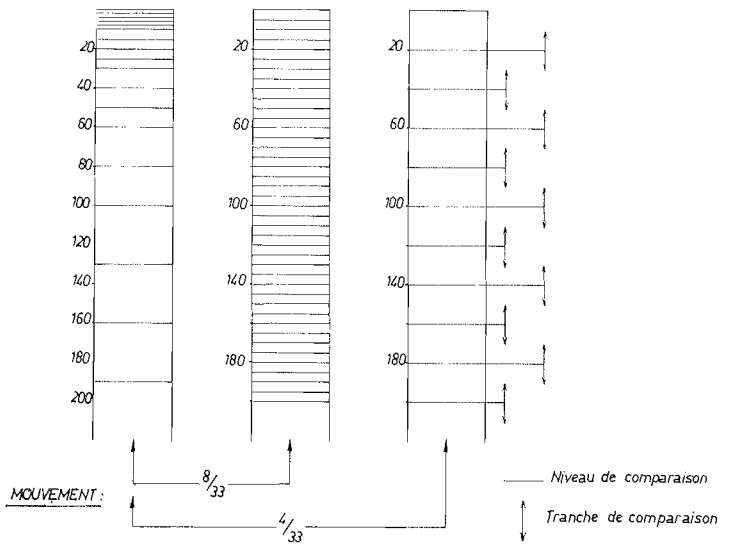


Figure 5 : Choix des différents niveaux de comparaison et indication des mouvements.
Choice of comparison levels and changes illustration.

Avec seulement 10 niveaux, espacés régulièrement, le mouvement est encore plus faibles : 4 profils sur 33. Ces 4 profils font partie des 8 précédemment trouvés.

Finalement, la multiplication du nombre de niveaux de comparaison apporte peu d'éléments supplémentaires. Pour le milieu étudié, VLADIMIR est peu sensible aux niveaux de comparaison. La méthode met surtout en évidence les profils « mal classés », ceux qui souvent sont intermédiaires entre plusieurs références.

D'autres essais ont été réalisés en faisant varier l'épaisseur de la tranche d'observation autour de chaque niveau de comparaison. Les mouvements obtenus sont quasi nuls. Cela s'explique par la faible différenciation des profils. Les horizons adjacents sont souvent très proches mathématiquement et cette option ne change pas le choix des distances minimum. Des tests de sensibilité concernant l'option « épaisseur de la tranche d'observation » devront être pratiqués sur des milieux plus différenciés où par exemple, un horizon peu épais est déterminant pour diagnostiquer un type de sol ou un problème agronomique.

C) Le choix des noyaux et leur ajustement

Le choix des noyaux initiaux peut être fait de manière a priori, statistique ou aléatoire. La première méthode permet à l'utilisateur connaissant le milieu de choisir des noyaux. Ce choix, dans la plupart des cas, diminue le nombre d'itérations. Cependant, on risque de favoriser des minimums locaux entraînant une non convergence du système. Le choix peut aussi être fait à partir de la projection des profils sur le plan 1-2 d'une analyse factorielle des correspondances. La comparaison de ces différentes approches montre que les profils choisis a priori sont des individus bien typés, qui privilégient des situations extrêmes au détriment de leur représentativité statistique (GIRARD et KING, 1988).

La comparaison entre plusieurs approches est difficile à réaliser d'une façon parfaitement objective, dans la mesure où les réajustements des noyaux se font manuellement. Mais ceci constitue l'un des fondements de la méthode : on laisse l'utilisateur maître des réajustements, même si ceux-ci ne respectent pas parfaitement la logique statistique. Par exemple, on peut préférer garder une référence bien typée, même si elle ne correspond pas au centre du nuage de points qui se

sont regroupés avec elle. Cela aura pour effet de ne pas accueillir dans ce nuage d'individus de transition.

L'inconvénient d'une telle méthode de réajustement manuel est d'être très longue lorsque les variables et les noyaux sont très nombreux. C'est pourquoi une méthode de réajustement semi-automatique des histogrammes (KING, 1986) est proposée. Celle-ci est basée sur la recherche du mode et de la médiane pour chaque variable de chaque noyau avec les histogrammes des variables sur la population totale, et sur la population du nuage. Seuls les cas litigieux sont présentés à l'opérateur pour qu'il prenne la décision finale.

Une telle proposition reste à affiner. Dans ce contexte, il sera possible de mettre en place des tests rigoureux de comparaison entre des classifications initialisées avec des choix différents de noyaux.

CONCLUSION

Face aux contraintes d'aménagement du milieu où l'on doit gérer des sols dans leur globalité, on propose, à la suite de DIMITRI, l'algorithme VLADIMIR pour la classification des profils. Celui-ci est bâti sur un modèle semblable à DIMITRI, simulant de façon statistique, les approches empiriques de classification des sols, dans le cadre de travaux thématiques de cartographie. On a tenu compte au maximum des contraintes inhérentes à des travaux réalisés sur de grandes étendues : bonne connaissance d'un nombre limité de sites (ceux-ci pouvant servir de noyaux), information plus éparse sur des sites très nombreux (l'algorithme permet de travailler même avec des données manquantes), introduction de nouvelles données au cours de l'étude (on peut à tout moment réaliser une nouvelle approximation)...

La méthode est conçue pour être « assistée par ordinateur » laissant le pédologue ou l'agronome responsable de ses choix. Avec ces algorithmes il est ainsi possible de comparer de façon rigoureuse différentes approches. Pour mieux cerner l'importance des différents choix possibles, au préalable ou au cours des approximations, on a testé la sensibilité de ces algorithmes à partir de profils pédologiques appartenant au « marais de Rochefort ».

Le choix des variables est l'élément premier dans le calcul de la distance. Si ce choix affecte peu les individus bien typés, il met en évidence les individus « intergrades ». Ceux-ci ne respectent pas une liaison générale « caractères d'observation - critères de comportemnt » et nécessitent un complément d'études.

Le choix de pondérations différentes pour les variables retenues affecte peu les groupes formés autour des noyaux. Il en est de même du choix des niveaux de comparaison. Ceci montre la cohérence du milieu et les relations existant par exemple, entre les horizons supérieurs et inférieurs. Des tests complémentaires, dans des milieux plus contrastés, sont nécessaires pour confirmer ce résultat.

Le choix des noyaux et leur réajustement correspondent à la phase de réflexion du pédologue au cours des itérations. Il est donc difficile de faire des comparaisons rigoureuses. Une proposition faite antérieurement pour un ajustement automatique des noyaux devrait permettre de tester différents choix de noyaux initiaux.

Les tests de sensibilité montrent la primauté du choix des variables, permettant ainsi d'envisager une comparaison entre des approches correspondant à différents objectifs tels que cartographiques, taxonomiques, ou agronomiques.

Reçu pour publication : Juin 1988
Accepté pour publication : Novembre 1988

AN ALGORITHM FOR CLASSIFYING THE SOIL MANTLE ACCORDING TO ITS PROFILE : VLADIMIR

While statistical analysis of the soil mantle into horizons is fairly easy (fig. 1), classifying profiles is not.

Vladimir is an algorithm designed to classify soil profiles. « Nucleus profiles » are chosen to initialize groups, then a mathematical distance is used to gather the studied profiles to the nearer « nucleus profile ». For that two main methods are proposed :

1) On the whole profile depth, or in given areas, comparison levels between the studied profile and « the nucleus profiles » are chosen (fig. 3).

2) Choosing some levels, the studied profile is compared to corresponding « nucleus profiles » examined over a given depth (fig. 4).

The choice of variables used to characterize profiles is important and leads to classifications varying according to different aims (table VI). Different weighting of the variable has little influence on the classification's result. The choice of levels for profiles has more influence (fig. 5), but the effects on few differentiated profiles, given in example, are small. The effect of the choice of nucleus is difficult to show as adjustments between approximations are made by the soil scientist.

Vladimir may be used for mapping as well as for taxonomy or agronomy.

BIBLIOGRAPHIE

ARKLEY R.J., 1971. — Factor analysis and numerical taxonomy of soils. Soil Sci. Soc. Am. Proc. 35(2), 312-315.

BENZECRI J.-P., 1973. — L'analyse des données. Vol. 1, 619 p. et 2, 615 p., Dunod, Paris.

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, 15-46.

BOULAINE J., 1980. — Pédologie appliquée. Collection Sciences agronomiques, 220 p., Paris.

CAMPBELL N.A., MULCAHY M.J., Mc ARTHUR W.M., 1970. — Numerical classification of soil profile on the basis of field morphological propertise. Aust. J. Soil Res. 8, 1, 43-58.

CHABANEL D., AFFAGARD A., JACQUIER C., DOSSO M., RUELLAN A., 1975. — Approche statistique de la pédologie : un exemple d'application de méthodes d'analyse multidimensionnelle à des données pédologiques de terrain. Sc. Agro. de Rennes, 61-68.

CIPRA J.E., BIDWELL O.W., ROHLF F.J., 1970. — Numerical taxonomy of soils from nine orders by cluster and centroid. Soil Sci. Soc. Am. Proc., 34, 2, 281-287.

DAMOUR L., CAMUS P., LAFON E., 1984. — Régime de drainage dans les sols argileux salés sodiques des Marais de l'Ouest. In : « Fonctionnement hydrique et comportement des sols ». AFES, p. 283-293.

DIDAY E., 1971. — La méthode des nuées dynamiques. Rev. Stat. Appl. vol. 19(2), 19-34.

FOURNIER B., KING D., 1982. — Apport complémentaire de deux méthodes en cartographie pédologique à grande échelle (les problèmes d'une carte automatique du sol). Sols, 6, 67-86. Grignon

GIRARD M.-C., 1976. — Recherche d'une méthodologie pédologique en matière de traitement statistique des données de sol. Application à la taxonomie et à la cartographie. Science du sol, 3, 177-204.

GIRARD M.-C., 1977. — L'horizon mis à l'épreuve statistique. Sc. du Sol, 4, 219-230.

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. 1re proposition, 145 p., Juillet, Grignon - Orléans.

GIRARD M.-C. et KING D., 1988. — Un algorithme pour la classification des horizons de la couverture pédologique : DIMITRI. Science du Sol, 26, 2, 81-101.

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

GRUIJTER J.J., 1977. — Numerical classification of soils and its application in surveys Agri. Research report 855. Center of Agricultural. Publishing and Doc. Pudoc. Wageningen, 117 p.

KING D., 1976. — Modélisation pédologique et cartographique automatique (à l'échelle du 1/5 000). Mémoire DAA INA PG, Grignon, 82 p.

KING D., 1986. — Modélisation cartographique du comportement des sols. Thèse doc. ing., INA PG, Grignon, 243 p.

KING D., 1987. — Modélisation de l'approche cartographique du comportement des sols. Science du sol, 26, 2, 27-41.

KING D. et DUVAL O., 1988. — Logiciels pour l'étude de la géographie des sols. Notice de présentation des programmes de la bibliothèque LOGOS. Version 3.1 INRA-SESCPF, Orléans, 112 p.

LEBART L., MORINEAU A., FENELON J.-P., 1982. — Traitement des données statistiques. Méthodes et programmes. Dunod, 500 p.

MARTIN D., AUBRY A.M., 1975. — Comparaison de profils du Congo par des distances Cah. ORSTOM, sér. Pédol., XII-2, 175-190.

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

MOORE A.W., RUSSEL J.S., 1966. — Potential use of numerical analysis and Adansonian concepts in Soil Sciences. Aust. J. of Sci. 29, 141-143.

MOORE A.W., RUSSEL J.S., WARD W.T., 1972. — Numerical analysis of soils : a comparison of three soil profile model with field classification. J. Soil Sci., 23, 2, 193-209.

MUIR J.W., HARDIE H.G.M., INKSON R.H.E., ANDERSON A.J.B., 1970. — The classification of soil profiles by traditional and numerical methods. Géoderma, 4, 81-90.

PAVAT J.L., 1986. — Contribution à l'étude de la ressemblance entre types de sols. Application aux secteurs de référence. DEA. Grignon, Montpellier, 77 p.

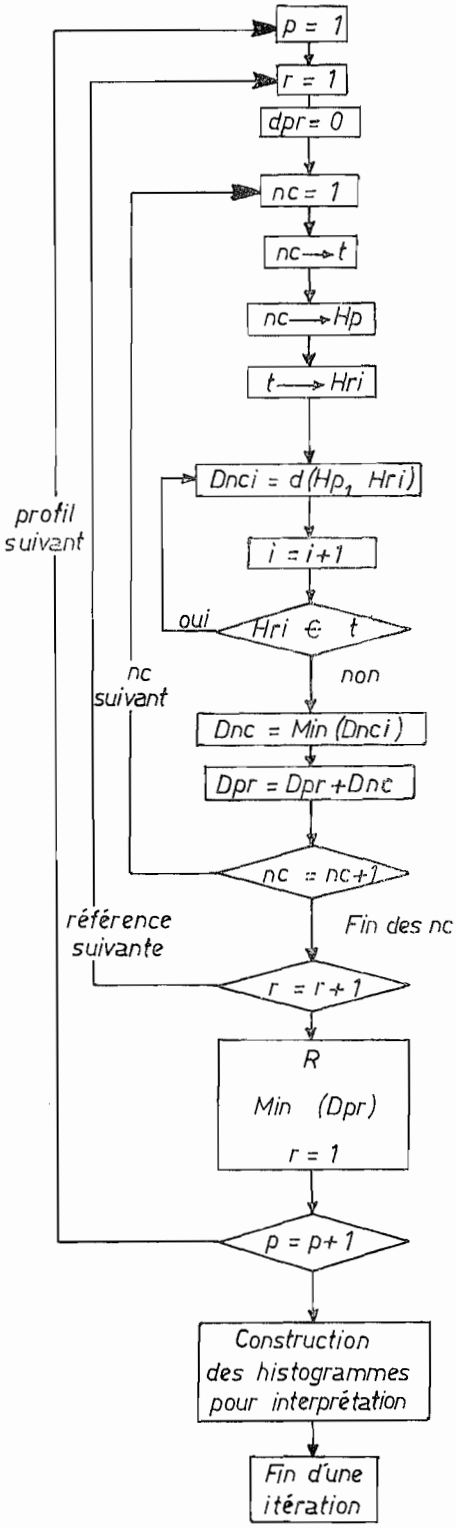
RAYNER J.H., 1966. — Classification of soils by numerical methods. J. Soil Sc., 17, 79-82.

RUSSEL J.S., MOORE A.W., 1968. — Comparison of different depth weighting in the numerical analysis of anisotropic soil profile data. Comm. 9e Congrès Inter. Sci. Sol, vol. 4, 205-213. Adelaïde.

SCHELLING J., 1970. — Soil genesis, soil classification, and soil survey. Géoderma,4, 165-193.

SIMONNEAUX V., 1987. — Mesure de la ressemblance entre des groupes de sondages à la tarière et des profils de référence. Application au classement des sols Mémoire DAA INA PG., Grignon, 61 p.

Annexe 1 : Algorithme d'une itération pour VLADIMIR. *VLADIMIR'S algorithm.*



P : numéro du profil à classer

r : numéro du profil de référence en cours (la distance entre les profils p et r est initialisée à 0).

nc : numéro du niveau de comparaison situé à une profondeur choisie préalablement.

A chaque niveau nc correspond une tranche t d'une épaisseur préalablement choisie.

Pour un niveau nc, on recherche le premier horizon Hp recoupé au sein du profil p.

Pour une tranche t, on recherche le premier horizon recoupé à partir du haut pour le profil de référence r.

Calcul de la distance entre les deux horizons sélectionnés. Celle-ci comporte différentes options non détaillées ici (KING, DUVAL, 1988).

On prend l'horizon situé en dessous de Hri (il y a autant de valeur pour i que d'horizons recoupés par la tranche t).

Cet horizon est-il recoupé par la tranche ? (profondeur du haut de l'horizon compris entre le haut et le bas de la tranche).

Calcul de la plus petite des distances.

Sommation des distances à chaque niveau de comparaison.

On passe au nc suivant. Puis quand on a exploré tous les niveaux de comparaison : fin des nc.

On passe à la référence suivante.

Le profil p est affecté au profil de référence le plus proche.

On passe au profil suivant.