SCIENCE DU SOL - 1992 - Vol. 30, 2, 117-131

Vérification de la qualité de l'information dans les bases de données de sol

J.-P. LEGROS (1)
P. FALIPOU (1)
F. DUNAND-DIVOL (2)

RÉSUMÉ

Des banques de données pédologiques ont été constituées un peu partout dans le monde et sont alimentées chaque année par de nombreuses descriptions et analyses de sols. On passe d'abord en revue les quelques méthodes permettant de vérifier la qualité des informations entrant dans ces banques. Ensuite, on propose une méthode complémentaire de détection des erreurs, basée sur un contrôle de cohérence entre les caractères morphologiques et analytiques du sol pris deux par deux et parfois trois par trois. La figure n°1 indique quelles informations sont recoupées pour rechercher, sur cette base, la présence éventuelle d'une erreur. Mais, pour détecter une incohérence, il faut bien connaître la gamme des valeurs que peut prendre une variable, quand la valeur d'une autre est fixée. Cela peut être obtenu graphiquement (ex. : figures 2 et 3) en exploitant l'ensemble des données préalablement enregistrées dans les banques. Une méthode est alors proposée pour essayer de retrouver, en cas d'incohérence constatée, quel est l'élément du couple dont la valeur est entachée d'erreur. A ce niveau on exploite, de manière groupée, tous les résultats des tests de cohérence réalisés en même temps pour le même horizon (tableau I).

Le programme de contrôle, appelé CHECKUP, est alors utilisé pour examiner la qualité globale des données entrant en banque ou la qualité de données déjà enregistrées (mais non employées pour bâtir le système de contrôle). On s'aperçoit que les erreurs sont assez nombreuses et ont toutes sortes d'origines. La plus spectaculaire touche l'estimation de la texture, sur le terrain (tableau III). Il faut donc faire un double effort : d'une part éliminer dans les banques déjà constituées les informations peu fiables, d'autre part tenter d'améliorer la situation future et prendre des dispositions pour alimenter les banques avec des données de qualité irréprochable. C'est cependant difficile. Le système CHECKUP peut laisser passer des erreurs ou bien alarmer inutilement l'utilisateur (risques de première et de seconde espèce en statistique). Quels que soient les efforts faits au niveau de l'automatisation, le dernier contrôle appartiendra au spécialiste du sol.

MOTS CLÉS : banque de données - contrôle de qualité - erreur - description de sol - analyse de sol.

ACCURACY OF SOIL DATA BASE INFORMATION

Pedological data bases have been set up throughout the world and each year numerous soil descriptions and analyses are added. Some of the methods which may be

(1) Institut National de la Recherche Agronomique, place Viala - 34060 Montpellier Cedex 1
(2) Institut des Sciences de l'Ingénieur, Université des Sciences et Techniques du Languedoc. Filière Informatique et Gestion, place E.-Bataillon, 34095 Montpellier Cedex 5

used to verify the quality of the information fed into these data banks will be reviewed. Then a complementary method of error detection, based on the verification of the coherence between morphological and analytical characteristics of the soil considered two by two or occasionally three by three, will be proposed. Figure 1 indicates which data are crossmatched in this data base to find possible errors. However, in order to detect an inconsistency, the range of values of a variable for a given value of another variable must be known. This may be obtained graphically (see for example figures 2 and 3) by considering all the data previously entered into the data base. A method is proposed to find which variable is erroneous when a pair of values is incoherent : all the results of coherence tests for a given horizon are grouped and compared simultaneously (table I).

The quality control program, called CHECKUP, is then used to study the overall quality of the data entered into the base or the quality of data already in the base (but not used to set up the checking). There are numerous errors with many sources. The most obvious arises from the estimation of texture in the field (table III). Two lines of action are thus required : firstly to remove unsound data from existing data bases and secondly to try to improve the situation in the future by ensuring that the data used to build up data banks are absolutely reliable. This is, however, difficult. CHECKUP may not pick up some errors or else unnecessarily worry the user. Whatever the improvements made in automatic quality control, the ultimate check must be that of the soil science specialist.

KEY-WORDS : *data base - accuracy - incoherency - soil description - soil analysis*

INTRODUCTION

Le développement de l'informatique a accru de manière considérable les possibilités de traitement des données. Cela a donc conduit à l'idée que les informations les plus diverses ont une valeur et qu'elles doivent être sauvegardées dans des banques. Mais, les traitements les plus sophistiqués n'ont aucun intérêt si les informations de départ ne sont pas fiables. Parallèlement au développement des Systèmes d'Information Géographique et des Systèmes de Gestion de Bases de Données, il faut donc concevoir des systèmes dont l'objet est de contrôler la qualité des données enregistrées.

Nous nous sommes attelés à cette tâche dans le cadre de la validation des descriptions et analyses de sols stockées dans les banques de données du type "STIPA" (BERTRAND *et al.,* 1984 ; BONNERIC *et al.,* 1985). Nous allons rendre compte des efforts réalisés de la façon suivante : dans une première partie nous passerons en revue les méthodes permettant de trouver les erreurs de différentes catégories affectant éventuellement les banques de données de sols, dans une seconde partie nous présenterons le système de diagnostic mis au point pour aider à la détection de ces erreurs, et, dans une troisième partie, nous indiquerons comment nous avons testé le système élaboré.

Les références bibliographiques sont nombreuses en ce qui concerne les problèmes de précision dans les bases de données graphiques (erreur sur le contenu d'une plage cartographique, erreur de positionnement, erreur liée à la superposition de deux cartes...). Au contraire, il n'existe apparemment rien en Science du Sol concernant le sujet que nous avons résolu d'aborder.

I. MÉTHODES PERMETTANT LA DÉTECTION DES ERREURS

Lorsqu'une information parvient à une banque de données, la validation peut faire appel à trois types de méthodes. Celles-ci sont d'ailleurs souvent utilisées les unes à la suite des autres car elles se complètent.

La première méthode correspond aux *contrôles de validité* appelés encore contrôles directs. Elle consiste à vérifier que l'information satisfait un certain nombre de règles connues et bien définies. Par exemple, certaines références sont obligatoires, en particulier le numéro permettant d'identifier le profil considéré. Par ailleurs, certaines données doivent être numériques, d'autres ne peuvent dépasser telle valeur ou telle longueur, etc...

La vérification de ces règles d'intégrité intervient dès que possible, par exemple lorsque l'information est saisie au clavier de l'ordinateur. Toute tentative d'enfreindre les règles établies entraîne un message d'erreur ou une alarme sonore : le logiciel refuse les données proposées ou refuse de continuer sans obtenir une information définie comme obligatoire. En d'autres termes, on élimine dès le départ toutes les causes susceptibles d'entraîner un blocage ultérieur à l'occasion de tel ou tel type de manipulation des données. A ce stade, les données sont "propres", au moins du point de vue informatique.

La seconde méthode correspond aux *contrôles de vraisemblance* appelés encore contrôles indirects. Elle consiste à évaluer si les données saisies ont des valeurs acceptables compte tenu d'informations connues des spécialistes mais souvent inconnues du système informatique chargé du contrôle. Ce type de vérification n'est que très partiellement confié à la machine. Par exemple, on constatera l'impossibilité de décrire un sol en France à 5 900 m d'altitude alors que le point le plus haut du pays est le Mont-Blanc (4 807 m) ! Mais il est presque impossible, à l'heure actuelle, de détecter automatiquement ce genre de bévue car les erreurs concevables et contraires au bon sens sont en nombre infini. Quel système expert pourrait savoir mieux qu'un spécialiste si les baobabs, les basaltes, les chotts, les vertisols, etc... sont des termes acceptables pour décrire des choses observées en des endroits déterminés du globe.

A ce stade, la vérification de l'information implique que le spécialiste puisse visualiser les données enregistrées antérieurement par lui ou par d'autres. L'expérience montre que les erreurs sont plus facilement détectées si les "sorties" proposées sont d'une grande lisibilité. Dans STIPA, on utilise deux procédés : d'une part l'édition en langage clair, d'autre part le dessin automatique des profils enregistrés. Le programme de dessin, initialement conçu pour illustrer les rapports pédologiques, se révèle un bon moyen pour faire apparaître les erreurs. Par exemple, un horizon est dessiné avec beaucoup de cailloux alors que l'observateur ayant vu le profil réel se souvient d'un sol pratiquement dépourvu d'éléments grossiers ! L'erreur lui apparaît alors avec une très grande évidence.

La troisième méthode, complémentaire des deux premières, représente les *contrôles de cohérence.* Elle consiste à examiner si les données saisies concernant un même sol sont cohérentes et compatibles entre elles. Cette méthode est employée par tous les spécialistes du sol. Par exemple, ceux-ci s'inquiètent d'un pH trop bas dans un sol calcaire. Cette recherche de cohérence est d'autant plus facile que les informations saisies sont

redondantes. La matière organique est concernée. Sa teneur, d'abord estimée sur le terrain, est ensuite mesurée au laboratoire. A la limite, on peut se demander s'il est naturel de dupliquer ainsi les efforts de caractérisation du sol. A notre avis, la réponse est oui, pour plusieurs raisons.

– D'abord, cette redondance offre, nous venons de le voir, des regroupements très utiles pour détecter certaines erreurs (erreurs d'appréciation sur le terrain, interversion éventuelle de deux échantillons de terre, erreur d'analyse, erreur de transcription des données lors de la saisie informatique).

– Ensuite, cela sert à définir dans chaque cas, quelle méthode d'analyse sera la plus adaptée. Ainsi, on ne mesure pas le carbone de la même façon si celui-ci est très abondant (plus de 50 % : perte au feu) ou très peu représenté (moins de 0,5 % : méthode ANNE).

– Enfin, cela permet d'éviter, un certain nombre d'analyses car celles-ci sont coûteuses. Ainsi, lorsque des profils semblables, dans la même unité de sol forestier, ont la même couleur, il paraît possible de supposer qu'ils contiennent la même quantité de matière organique et il est raisonnable de mesurer la teneur en carbone sur quelques - uns d'entre eux seulement. Pour les autres, l'appréciation de terrain sera la seule information disponible.

C'est la redondance fréquente des informations que nous allons exploiter en mettant au point des tests de cohérence. Pour cela, l'utilisation de l'ordinateur est indispensable, sinon en théorie, du moins en pratique. En effet, lorsque les lots de données à contrôler sont très importants, il est presque impossible à un lecteur de détecter toutes les incohérences sans le secours d'une machine par suite d'un effet de fatigue entraînant une perte rapide de vigilance.

Une précision doit être apportée concernant les données dites "saisies" ou "enregistrées". Dans le système STIPA, ces données en cours de vérification n'appartiennent pas encore complètement à la banque de données. Elles sont en transit sur un fichier d'attente. C'est seulement lorsqu'elles ont été vérifiées de toutes les façons possibles qu'elles sont injectées dans le système de stockage. Cela permet de confier l'enregistrement à des non spécialistes (étudiants, stagiaires) sans trop redouter les dégâts qui pourraient être faits à une banque représentant plus de dix ans d'efforts constants.

II. MISE AU POINT DU CONTRÔLE DE COHÉRENCE

Le contrôle de cohérence qui est effectué ici, se situe *exclusivement au sein de l'horizon*. On ne teste pas, par exemple, la cohérence entre la teneur en matière organique de l'horizon et des caractères de l'environnement tels que type de couvert végétal, altitude, etc... En d'autres termes, l'ambition de notre travail est limitée, des extentions du contrôle pourraient être prévues à l'avenir.

1. Principe

L'idée centrale du travail est de bâtir des tests de cohérence basés sur l'expérience. Pour cela, on utilise les données déjà stockées en banque et on examine les relations apparaissant entre les variables caractéristiques des horizons. Prenons le cas de la relation entre fer libre et fer total. Il s'agit d'abord de prouver que cette relation existe, ensuite d'en donner une évaluation (ex. : corrélation), enfin de définir pour le futur

quelles teneurs en fer libre sont compatibles avec chaque valeur possible du fer total. Hors de la gamme ainsi définie pour le fer libre, la probabilité d'une erreur est grande.

Cette démarche se heurte à plusieurs difficultés. D'abord, il n'est pas complètement pertinent de ramener les contrôles de cohérence à des vérifications ne faisant intervenir que deux valeurs à la fois. Dans certains cas, nous devrons d'ailleurs considérer non pas des couples mais des groupes de trois caractères. Nous ne sommes pas allés au-delà pour des raisons pratiques, mais c'est certainement une lacune.

La seconde difficulté tient au volume limité des données disponibles pour établir ces tests. Nous disposons en théorie de 5 000 profils dans la banque STIPA de Montpellier. A quatre horizons par profil, en moyenne, cela devrait correspondre à 20 000 horizons permettant de définir un échantillon caractérisé par autant de couples de valeurs tels que taux de saturation/pH ou couleur/teneur en carbone. En réalité, les profils de la banque ne sont pas tous convenablement décrits et analysés. En conséquence, les relations intéressant des couples d'informations sont bâties sur des échantillons d'horizons dont l'effectif est inférieur à 5 500. Le nombre exact est précisé, au fur et à mesure, pour toutes les statistiques présentées dans cet article.

La troisième difficulté rencontrée dans la démarche est liée aux choix à effectuer. La découverte d'une relation entre deux variables, x et y, ne suffit pas pour fournir un élément de contrôle de cohérence. Pour que la relation soit intéressante, il faut qu'elle mette en cause des caractères du sol très souvent décrits et analysés. Sinon, on établirait des tests dont l'intérêt pratique serait limité à quelques cas exceptionnels. Par ailleurs, on sait que les corrélations peuvent avoir une origine indirecte et n'impliquent pas des relations de cause à effet entre les variables. Pour être retenues dans notre système de contrôle, les relations devaient avoir une explication pédologique claire.

Au total, la méthode permet d'établir les relations schématisées sur la figure n° 1. Elles sont au nombre de quinze et font intervenir les variables figurées dans les rectangles. Pour chaque relation, un test de cohérence est effectué si l'information nécessaire est disponible, c'est-à-dire si les horizons ont été convenablement décrits et analysés.

Les relations n° 5 et 6 sont les seules à intéresser simultanément trois variables :

– CEC avec granulométrie et carbone,

– C/N avec pH.

Dans le détail, chaque relation présente des aspects spécifiques. Il faut distinguer, en particulier, le cas des variables quantitatives et celui des variables qualitatives.

En général, la valeur prise par une variable quantitative x est compatible avec une large gamme de valeurs d'une seconde variable quantitative y. On est donc en présence d'un nuage de points. On le construit pour les deux variables x et y à partir des données existantes et déjà stockées (figure 2). Par la suite, lorsque de nouvelles données sont introduites dans la banque, deux cas principaux sont distingués concernant un nouveau couple d'informations (xi, yi) :

– si le point correspondant s'inscrit au coeur du nuage préalablement défini, aucun message n'est déclenché,

– si le point correspondant est situé hors du nuage ou près de sa limite, une erreur est suspectée et donne lieu à l'émission d'un signal d'alarme.

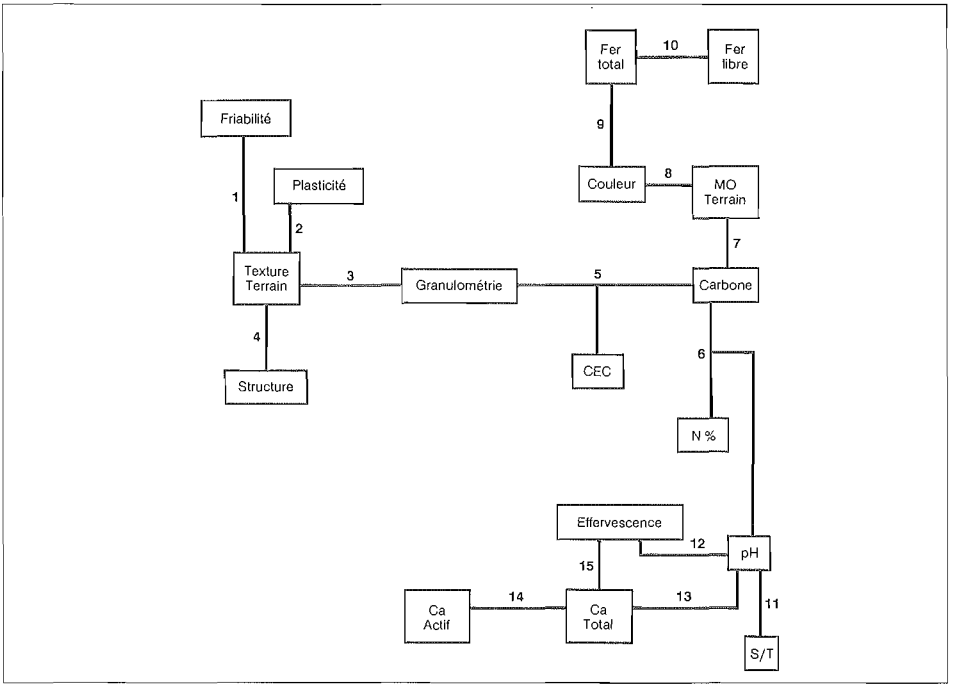


Figure 1 : Schéma général des variables et des relations
(Les abréviations sont les suivantes : MO pour matière organique, CEC pour capacité d'échange, S/T pour taux de saturation). 1, 2, 3, etc. : n° de la relation.

Schematic representation of the variables and the relations between them.
The following abbreviations are used : M.O. for organic matter, CEC for cation exchange capacity, S/T for base saturation. Numbers indicate the number assigned to the relationship.

En d'autres termes, on est amené à supposer que l'enveloppe du nuage préalablement dessiné représente le domaine de tous les cas possibles. Ce serait exact si le nuage représentait tous les sols du monde. Nous en sommes loin. Les tests que nous bâtissons sont donc applicables seulement à des sols semblables à ceux déjà stockés. On peut admettre que la banque STIPA de Montpellier renferme des profils caractérisant bien la moitié sud de la France (le démontrer exigerait de longs développements). L'utilisation, hors de cette zone, des tests proposés ne pourrait donc se faire sans adaptation. Il faudrait caler de nouveau certains paramètres.

Pour automatiser les contrôles, il faut décrire le nuage de points à l'intérieur du système informatique. Pour cela, il y a plusieurs solutions. La méthode généralement retenue consiste à découper ce nuage et le reste du plan en petits compartiments carrés ou rectangulaires en menant des parallèles aux axes x et y. Chaque compartiment correspond à une zone autorisée ou interdite pour le point correspondant à un nouveau couple d'informations (xi, yi).

Cette division des nuages en compartiments correspond de fait à la transformation des variables quantitatives en variables qualitatives ordonnées. Cela permet d'étendre

ce système de contrôle de cohérence à des informations faisant intervenir à la fois une variable quantitative et une autre qui est qualitative (ex. : pH mesuré/effervescence appréciée en 4 classes). Le cas de deux variables qualitatives peut être traité de la même façon (ex. : couleur/teneur en matière organique décrite sur le terrain sur la base d'un découpage en 7 classes). On se retrouve alors avec un tableau à double entrée indiquant, par oui ou non, si telle classe pour la variable x est compatible avec telle classe pour la variable y.

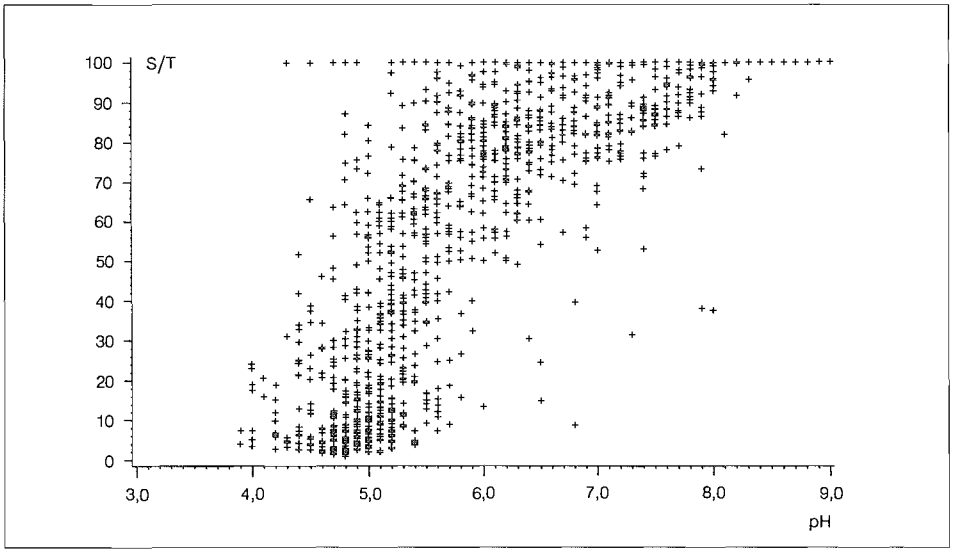


Figure 2 : Exemple de relation entre deux variables quantitatives (pH et taux de saturation)
Pour 1104 horizons et couples. Les taux de saturation supérieurs à 100 (sols calcaires) ont été arbitrairement ramenés à 100. De nombreux points, sur la figure, sont exactement superposés.
Origine : données STIPA - Montpellier

An example of the relationship between two variables (pH and base saturation).
For 1104 horizons and pairs of variables. Values of base saturation greater than 100 (calcareous soils) have arbitrarily been set to 100. Many points on the figure are exactly superposed.

Cependant, quand le nuage a une forme simple, on peut le décrire plus facilement, par exemple en le limitant par des droites dont le tracé est plus ou moins logique ou arbitraire. C'est ainsi que nous avons procédé pour la relation fer libre/fer total (figure 3). D'une part la quantité de fer libre ne peut excéder la quantité de fer total (sauf en cas d'erreur). D'autre part, une très grande majorité des points est située au-dessus de la droite :

Fer libre ‰ = 1/2 Fer total ‰ – 2

On retient donc en définitive comme tests :

1/2 Fer total ‰ – 2 < Fer libre < Fer total

Il reste à déterminer sur quelle base statistique on doit éliminer les points les plus extérieurs au nuage. Ce sera l'objet du paragraphe qui suit.

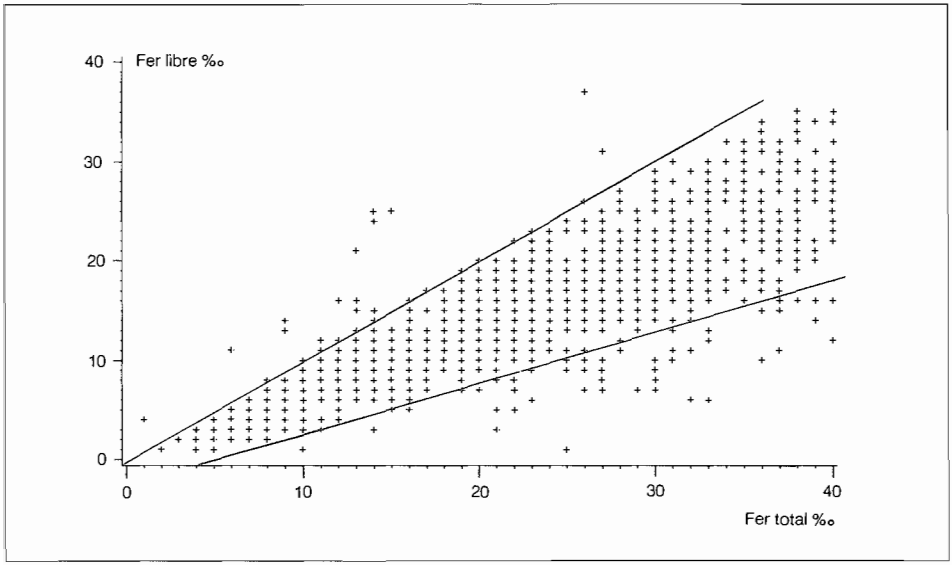


Figure 3 : Relation fer libre - fer total
Pour 1506 horizons et couples. Les valeurs supérieures à 40 ‰ ont été éliminées. De très nombreux points sont exactement superposés.
Origine : données STIPA - Montpellier

Free iron - total iron relationship
For 1506 horizons and pairs of variables. Values greater than 40 ‰ have been excluded. Many points are exactly superposed.

2. Problème de la qualité des données servant à bâtir le test

Nous ne sommes pas complètement sûrs de la qualité des données déjà enregistrées en banque et servant à bâtir les limites d'acceptation. Bien au contraire, des erreurs sont évidentes (voir figure n° 3, les horizons pour lesquels on a plus de fer libre que de fer total !). Les causes peuvent être diverses : erreur d'échantillonnage, erreur de copie de chiffres, erreur d'analyse etc... Quoiqu'il en soit, on ne doit pas définir les limites d'acceptation comme correspondant exactement à l'enveloppe externe des nuages de points servant à bâtir les tests. Nous avons adopté une position plus restrictive tendant à laisser à l'extérieur du nuage les points aberrants ou, lorsqu'on ne sait pas reconnaître ceux-là, les 5% de points les plus excentrés. Cela veut dire qu'on accepte la non-efficacité des tests dans, *a priori*, 5% des cas (si, par hasard, ces points étaient bons). Cette marge de 5% est habituelle dans les tests de statistique, mais elle n'est pas toujours adaptée. Nous le verrons plus loin.

3. Détection de la variable en cause

Détecter une incohérence ne permet pas de trouver laquelle des deux valeurs du couple est erronée. Par exemple, il est impossible qu'un horizon soit à la fois très acide (pH 5) et calcaire (présence de CO_3Ca). Dans un tel cas, il y a donc une erreur au niveau de la relation 13. Mais cette erreur porte-t-elle sur la teneur indiquée en CO_3Ca ou sur la valeur du pH ?

On peut répondre à ce type de question et trouver le responsable au sein des couples de données posant problème, ceci dans les cas simples. En restant dans le cadre de l'exemple pris, retournons au schéma de la figure 1 et supposons que les relations 6, 11, 12 et 13 ne soient pas vérifiées (valeurs incompatibles) au contraire des relations 15, 5 et 7. Le pH est probablement en cause puisque d'une part sa valeur est incompatible avec les valeurs des variables normalement corrélées avec lui et puisque, d'autre part, ces variables ont des valeurs paraissant convenables compte tenu du reste du contexte. On a donc sept indices permettant de suspecter la valeur du pH. Mais cela reste dans le domaine de l'hypothèse probable. On n'acquiert pas de certitude.

Cela nous a conduit à déterminer l'état que devraient prendre certaines relations (V = vérifiées, F = fausses ou non vérifiées, c'est-à-dire valeurs incompatibles) pour que l'on puisse détecter les variables dont la valeur a le plus de chances d'être en cause (tableau I).

Tableau I : Etat des relations et interprétation correspondante (V = vérifiée et F = non vérifiée)
Relationship and corresponding interpretation. (V = verified, F = non verified)

NUMÉROS ET ÉTATS DES RELATIONS (se référer à la figure 1)														DIAGNOSTIC DÉDUIT
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
F	F	F	F	V											Erreur probable sur la Texture terrain
V	V	F	V	F											Erreur probable sur la Granulométrie
		V		F		V									Résultat CEC douteux
		V		F	F	F	V								Erreur probable sur la teneur en Carbone
				V	V	F	F	V							Appréciation douteuse de la teneur MO terrain
						V	F	F	V						Appréciation douteuse de la Couleur
							V	F	F						Teneur en Fer total douteuse
				V	F	V				F	F	F		V	Valeur du pH suspecte
					V					V	F	V	V	F	Appréciation douteuse de l'Effervescence
					V					V	V	F	F	F	Erreur probable sur la teneur en Calcaire total

Exemple d'utilisation (ligne 3) : si la granulométrie est celle que permettait de prévoir la texture (relation n°3 = V) ; si la teneur en Carbone correspond à la quantité de matière organique prévue sur le terrain (relation n°7 = V), alors le non accord entre granulométrie, teneur en Carbone et CEC (relation n°5 = F) a probablement pour origine une erreur sur la CEC.

Ce genre de démarche se heurte cependant à trois difficultés. D'abord, elle n'est applicable qu'à des variables situées au milieu du schéma relationnel. Ainsi, il sera dif-

ficile de conclure avec certitude si c'est la friabilité ou la teneur en fer libre qui pose problème. Ensuite, on ne sait pas très bien où s'arrêter dans la liste des conditions qu'il faut poser et vérifier pour conclure valablement. Ainsi, dans l'exemple traité plus haut, on peut se demander s'il est réellement utile d'exiger que la relation 7 soit vérifiée pour détecter une erreur sur le pH... Enfin, et surtout, tout cela revient à dire que pour fournir un diagnostic complet et définitif, il faut disposer simultanément de tous les résultats d'analyse nécessaires de façon à définir une valeur bien précise (vérifiée ou non) pour chacune des relations en cause. Cela se produit rarement. Dans le cas général, des observations ou des analyses font défaut, si bien qu'on ne peut pas aller aussi loin et que l'algorithme se contente de détecter les incohérences, sans déterminer où se situe exactement l'erreur. Pour être plus précis, dans le cadre d'une banque STIPA, on dispose de toute l'information nécessaire au mieux une fois sur cinq. Au pire, c'est-à-dire lorsque de nombreuses conditions sont à tester simultanément (voir dans le tableau n° I le cas du pH, de l'effervescence ou de la teneur en calcaire total), la responsabilité de l'erreur est établie moins d'une fois sur vingt.

4. Mise en oeuvre technique

L'algorithme de contrôle appelé CHECKUP a fait l'objet d'un mémoire de l'Institut des Sciences de l'Ingénieur de Montpellier (ISIM) soutenu antérieurement (DUNAND-DIVOL, 1988).

Une programmation classique en langage PASCAL a paru être adaptée. On a testé aussi un système expert (HAMEX). Mais le nombre de règles à écrire pour résoudre notre problème était très important (description des compartiments sur les schémas de relation entre variables), ce qui nous a conduit à abandonner cette seconde approche. Il faudra peut-être, dans le futur, reprendre la question et tester d'autres générateurs de systèmes experts, particulièrement ceux permettant la manipulation de tableaux.

III. CALAGE, VALIDATION ET EMPLOI EN ROUTINE

1. Calage et validation sur un jeu d'essai

Le calage définitif peut se faire suivant deux principes :

– soit régler les tests pour des sols bien déterminés (ex. vertisols) et changer certains paramètres lorsqu'on change de type pédologique,

– soit définir des limites de confiance plus larges rendant les tests moins sensibles mais valables, en principe pour tous les sols d'une région.

C'est cette seconde façon que nous avons utilisée dans la mesure où les nuages de points ont été dessinés au départ pour tous les sols de la banque STIPA, sans considération d'origine précise au sein de la partie sud de la France. Nous ne prétendons pas que cette option soit la meilleure. Elle correspond plutôt à une sorte de première approximation.

Dans ces conditions, le jeu d'essai sur lequel nous avons testé CHECKUP était constitué pour représenter, lui aussi, un vaste éventail d'horizons appartenant à des sols divers, dispersés sous climat alpin (sols humifères), sous climat océanique (sols limoneux acides et lessivés) et sous climat méditerranéen (sols calcaires) (voir tableau n° II). Il est évident que ce jeu d'essai n'a pas pris en compte des données figurant dans le lot servant initialement à définir la forme des nuages de points.

Tableau II : Environnement des horizons ayant servi de test pour CHECKUP.
Origin of the horizons used to test CHECKUP.

TYPE DE CLIMAT	ZONE GÉOGRAPHIQUE	EFFECTIF
Alpin	Chamonix	200
Océanique	Ouest Massif Central	170
Méditerranéen	Languedoc	910

Avant de procéder aux essais, il faut introduire quelques réflexions d'ordre statistique. En procédant à un contrôle, on court les risques dits de première et de seconde espèce, c'est-à-dire qu'on peut signaler comme correspondant à une erreur une information valable et accepter comme valable une information fausse.

Il est difficile de mesurer avec précision le volume des erreurs des deux types. En pratique, on s'assure que les contrôles automatiques réalisés par CHECKUP, et dont le résultat est présenté à l'opérateur, ne sont ni trop sévères ni trop souples. Un contrôle trop sévère conduit à déclencher trop de signaux d'alarme sous lesquels l'utilisateur est submergé si bien qu'il n'y accorde plus d'attention. Un contrôle trop souple laisse passer les erreurs et ne sert plus à rien. Nous avons considéré comme raisonnable une proportion de 5 signaux d'alarme émis pour 100 contrôles effectués. Dans ce cas, en effet, puisqu'on teste 15 relations, cela correspond à 15 x 5/100 = 0,75 soit environ un message d'alarme par horizon contrôlé, soit encore 3 ou 4 messages pour un profil complet. Cela paraît suffisant, au moins dans un premier temps. En fait, le nombre véritable de messages reçus dépend pour chaque profil : du nombre d'horizons, de la qualité de la caractérisation (description, analyses) et de la sévérité plus ou moins grande des tests. Il convient maintenant d'examiner ce dernier point.

La réalisation des essais permet d'affiner les tests, par approximations successives. Par exemple, le pourcentage d'erreur signalé au niveau de la relation 4 (texture de terrain, structure) semblait trop élevé (plus de 15%). Nous nous sommes aperçus que cela était dû à l'existence de structures polyédriques ou polyédriques subanguleuses dans des sols assez largement sableux, alors que nous avions considéré cela comme anormal et devant donner lieu à un message d'alarme. Nous avons donc été conduits à une plus grande tolérance à ce sujet. Par exemple, aussi, nous avions cru pouvoir fixer à pH 7,5 la limite en-dessous de laquelle un sol ne peut pas contenir du carbonate de calcium (à 5% des cas près). En fait, il vaut mieux fixer cette limite à pH 7,3. Fournir ici le détail de tous les tests est impossible. Pour 15 tests, il faudrait évidemment 15 tableaux ou formules. Les tableaux inclus dans le programme CHECKUP peuvent être consultés sur écran et éventuellement modifiés.

Lorsque CHECKUP est utilisé en routine, en cas de message d'alarme, le dernier mot reste à l'utilisateur. Celui-ci a son attention attirée sur une situation méritant enquête, mais il reste maître du rejet ou de l'acceptation.

On aimerait pouvoir quantifier exactement le nombre, le niveau et le type des erreurs résiduelles après contrôle. Mais cela est impossible car nous ne pouvons préjuger de la

réaction, bonne ou mauvaise, de l'utilisateur qui reçoit un message d'alarme. En définitive, notre approche reste donc empirique, même si elle constitue un léger progrès par rapport à la situation antérieure.

2. Exemples de résultats concrets liés à l'emploi en routine de CHECKUP

Il ne serait pas intéressant de reprendre ici le menu détail de tous les résultats obtenus concernant la fiabilité des informations arrivant dans la banque STIPA de Montpellier. Quelques cas précis méritent cependant l'attention.

Texture de terrain/granulométrie (relation 3) :

La texture appréciée sur le terrain ne concorde pas assez bien et pas assez souvent avec la granulométrie mesurée au laboratoire. Le tableau n° III définit ce que nous appelons la concordance entre texture et granulométrie. Elle est établie sur 5426 essais texturaux réalisés par de nombreux pédologues entre 1979 et 1990. Chacun de ces essais a été suivi d'une analyse granulométrique.

Tableau III : Concordance texture/granulométrie sur 5426 couples-horizons dans la banque STIPA Montpellier
Agreement between field texture and particle size distribution for 5426 pair - horizons in the data bank STIPA - Montpellier

DIAGNOSTIC	EFFECTIF	POURCENTAGE
Concordance parfaite	2137	39 %
Concordance approximative (ex : LSA pris sur le terrain pour un LS)	2622	48 %
Incohérence (ex : Argile sèche prise pour un sable)	667	13 %

La concordance est jugée "parfaite" lorsque, dans le triangle de texture GEPPA, la granulométrie mesurée tombe dans la case texturale prédite. La concordance est jugée "approximative" lorsque la texture et la granulométrie correspondent à deux cases adjacentes. On estime qu'il y a présomption d'incohérence lorsque la texture prédite et la granulométrie mesurée sont séparées par au moins une case libre dans le triangle de texture. Un message d'alarme est alors émis. Cette situation a été rencontrée 667 fois sur l'échantillon de 5426 tests (soit 13% des cas). Elle correspond à un décalage considérable représentant, par exemple, une différence de teneur en argile de 30 points.

On sait qu'il n'y a pas équivalence stricte entre texture et granulométrie. La destruction des ciments, préalablement à l'analyse granulométrique, explique quelques discordances. Cependant, une analyse fine, au cas par cas (ROQUES, 1990), nous a montré que l'appréciation texturale était souvent en cause, c'est-à-dire généralement imprécise et parfois complètement fausse.

Les décalages suivants sont trop fréquents pour être considérés comme aléatoires :

– dans 12% des cas, les LAS (déterminés par analyse) sont pris sur le terrain pour des LS. Cela correspond à 86 discordances dans le lot de 5426 essais pour 725 LAS vrais ;

– dans 10% des cas, les ALS sont prises sur le terrain pour des LSA. Cela correspond à 50 discordances ;

– dans 8 à 9% des cas, les argiles sont prises sur le terrain pour des LAS. Cela représente 40 discordances.

D'une manière plus générale, les observateurs montpelliérains sous-estiment les proportions de particules fines et sur-estiment les proportions de sable, car celui-ci est plus sensible au toucher.

Dans ces conditions, nous estimons qu'il n'y a pas lieu de diminuer encore la sévérité du contrôle exercé par CHECKUP. En effet, il apparaît à l'évidence que de nombreux observateurs ne sont pas capables, sur le terrain, de définir, même approximativement, la texture. Le système de contrôle, en les mettant en face du problème, va les aider progressivement à affiner leur diagnostic.

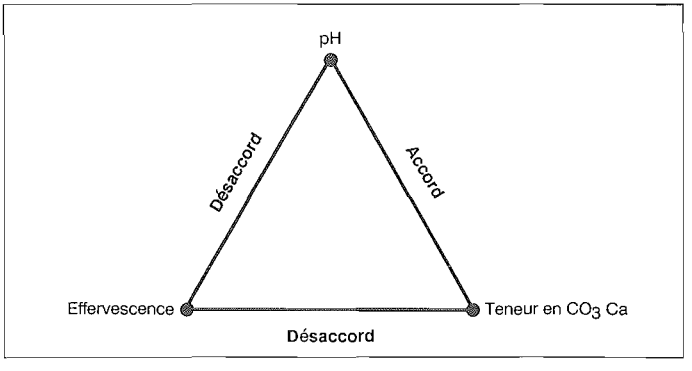
Relation pH/effervescence et calcaire total/effervescence (n° 12 et 15) :

Nous n'avons pas rencontré d'erreur systématique dans la banque STIPA, sauf en ce qui concerne les sols du Languedoc-Roussillon (2700 profils stockés). Pour ceux-ci, on constate d'une part que le pH est élevé par rapport à l'effervescence constatée (10% des cas) et d'autre part que la teneur en CO_3Ca est élevée par rapport à l'effervescence constatée (25% des cas). Notons tout de suite qu'il n'y a pas correspondance entre les deux chiffres fournis (10 et 25) car on ne dispose pas régulièrement des trois valeurs pour chacun des horizons étudiés. Dans ces conditions, la relation 12 est testée sur nombre de couples différents de la relation 15.

Si on applique la méthode exposée plus haut pour détecter la responsabilité de l'erreur (conditions vrai-faux), on est conduit à suspecter l'effervescence (figure 4).

Figure 4 : Concordance entre pH, CO_3Ca et effervescence.

Cohorence between pH, $CaCO_3$ content and effervescence.



L'hypothèse la plus probable est la présence de dolomite dans les sols. Ce minéral est abondant dans les calcaires de l'arrière-pays et dans les calcaires des Causses. Sous

l'effet de l'acide chlorhydrique, il donne une effervescence faible. Mais il faut penser aussi aux calcaires marneux. Ils sont abondants dans le Midi. Ils demeurent dans les échantillons de terre sous forme d'éléments grossiers qui sont effrités par un broyage excessif et provoquent donc la recarbonatation de la terre fine avec accroissement du pH correspondant. Ce cas, s'il se produit, est à méditer. Il révèle les risques et les limites de la méthode proposée dans la figure 4 et dans le tableau I. En effet, l'opinion minoritaire n'est pas forcément la plus mauvaise ! Par ailleurs, cet exemple montre aussi que les analyses de laboratoire ne sont pas toujours assez fiables pour conduire à négliger les observations qualitatives faciles à faire sur le terrain.

Relations 3, 5, 9, 10, 7...

Ces relations sont l'occasion, avant même d'être testées, de vérifier l'ordre de grandeur des chiffres qu'elles mettent en cause. Or, on s'aperçoit que les erreurs sont relativement nombreuses à ce niveau. Dans STIPA, pour éviter la manipulation des virgules, nous avons exprimé les résultats d'analyse avec des unités variables : par exemple, le pH est en dixièmes, l'argile en pour mille, le carbone en pour dix mille, le phosphore en ppm, etc...

A l'enregistrement, cela provoque des confusions et des décalages de colonnes se traduisant par la division ou la multiplication par dix des valeurs stockées. Cela n'est pas toujours détecté par la lecture des fiches d'analyses, soit par étourderie, soit parce que l'esprit humain corrige automatiquement. Ainsi, une teneur en argile de 234 % n'étonne presque personne ! On comprend qu'il s'agit de ‰ et on ne cherche pas plus loin. Mais l'erreur existe alors dans la banque de données et va se propager quand on fera des calculs statistiques du type : relation entre capacité d'échange de cations et teneur en argile. Les quelques points aberrants seront situés loin du groupe des autres. Si on oublie de les éliminer, ils tireront fortement la droite de corrélation et introduiront une erreur majeure !

CONCLUSION

L'exploitation statistique de milliers de données de sols déjà stockées dans les banques STIPA permet de bâtir des tests de validation pouvant servir à contrôler les informations qui seront saisies dans l'avenir.

Nous avons indiqué, en première partie, qu'il n'était pas possible de détecter toutes les erreurs concevables. Mais le programme CHECKUP, capable de réaliser 75 tests de contrôle dans un profil de 5 horizons, prend à sa charge une partie importante du travail de vérification. Un laboratoire, une équipe de cartographie pourront détecter facilement (et avant leurs clients) des incohérences entre le pH, la teneur en CO_3Ca, l'effervescence, etc...

Dans l'état actuel de fonctionnement, CHECKUP a déjà montré quelles erreurs étaient les plus fréquentes, sinon les plus graves, dans les lots de sols décrits et analysés.

On constate aussi que la responsabilité des erreurs est très partagée. Elle incombe à l'observateur travaillant sur le terrain (texture), au laborantin chargé de préparer les échantillons (broyage), à la personne assurant la saisie informatique (mauvaise recopie des chiffres), etc... Un contrôle d'ensemble s'impose donc.

A l'avenir, des améliorations devraient être recherchées concernant le programme CHECKUP. Ainsi, le contrôle pourrait être affiné en tenant compte du type de sol en cause. La classification du sol serait d'abord lue automatiquement dans les données d'entrée correspondant à un profil saisi au clavier puis le contrôle aurait lieu avec des tests et des limites de confiance adaptés. On pourrait aussi étendre les tests à des variables ne figurant pas uniquement au niveau de la caractérisation des horizons. Par ailleurs, une réflexion devrait être menée concernant l'attitude à tenir, dans différents cas, lorsque des erreurs ont été détectées. En effet, il est presque toujours impossible de retourner sur le terrain et parfois difficile de recommencer une analyse. Une solution convenable est de rejeter purement et simplement l'information douteuse. Une autre consiste à la conserver en l'état. Cela peut être intéressant dans le cas de l'appréciation texturale. Le décalage avec la granulométrie sert à estimer la précision du test de terrain et a parfois sa signification propre (problème des ciments détruits à l'analyse granulométrique). Une troisième voie est d'améliorer les chiffres ou les observations pour que tout coïncide. Dans le cas présent, il s'agit de corriger après coup, sur les fiches de profils, les estimations de texture. Certains cartographes agissent de la sorte.

Il ne nous appartient pas de faire des recommandations. CHECKUP est un système d'alarme, pas une méthode de correction des erreurs...

Enfin, il faut tirer les conséquences des erreurs faites et essayer d'en réduire certaines en modifiant de manière adéquate les systèmes de saisie de l'information. Mais cette transformation nécessite aussi une réflexion d'ensemble sur les problèmes de structuration des données. C'est ce qui est fait par ailleurs (LEGROS et NORTCLIFF, 1990 ; GAULTIER, 1990).

Reçu pour publication : Février 1992
Accepté pour publication : Juillet 1992

BIBLIOGRAPHIE

BERTRAND R., FALIPOU P., LEGROS J.-P., 1984. – Notice pour l'entrée des descriptions et analyses de sols en banques de données. STIPA-RITDS. Agence de Coopération Culturelle et Technique, Paris. 136 p.

BONNERIC P., NAVARRO R., FALIPOU P., 1985. – Notice pour la gestion informatique de la banque de données. STIPA-RITDS. Agence de Coopération Culturelle et Technique, Paris. 83 p. + annexes.

DUNAND-DIVOL F., 1988. – *Interface de vérification de la qualité des descriptions et analyses de sols (STIPA).* Rapport de stage ISIM-INRA, Montpellier. 78 p.

GAULTIER J.-P., 1990. – Projet DONESOL. Etude détaillée. Doc. INRA Science du Sol. 4 tomes, 200 p.

LEGROS J.-P. et NORTCLIFF S., 1990. – Conception d'un vocabulaire pour la description du milieu naturel et des sols. *Pédologie,* XL, 195-213.

ROQUES Th. , 1990 - *Estimation de la granulométrie d'un échantillon à partir de l'appréciation de la texture et de la cartographie.* DEA USTL-INRA-ENSA, Montpellier. 26 p. + annexes.

Achevé d'imprimer
sur les presses de l'Imprimerie MAUGEIN-LACHAISE
R.N. 89 - 19360 Malemort
en Octobre 1992