ACADÉMIE DE MONTPELLIER

UNIVERSITÉ MONTPELLIER II
--- SCIENCES ET TECHNIQUES DU LANGUEDOC ---

THÈSE

présentée à l'Université des Sciences et Techniques du Languedoc
pour obtenir le diplôme de DOCTORAT
Spécialité : Physiologie Biologie des Organismes et Populations
Science du Sol

FORMALISATION DES LOIS DE DISTRIBUTION DES SOLS POUR AUTOMATISER LA CARTOGRAPHIE PÉDOLOGIQUE A PARTIR D'UN SECTEUR PRIS COMME RÉFÉRENCE

Cas de la petite région naturelle
Moyenne Vallée de l'Hérault

par

Philippe LAGACHERIE

Soutenue le 4 décembre 1992 devant le jury composé de :

J.C.	REMY	ENSAM	Président
P.	BRABANT	ORSTOM	Rapporteur
P.	BURROUGH	Université d'Utrecht	Rapporteur
J.P.	CHEYLAN	GIP-RECLUS	
Y.	ESCOUFIER	ENSAM-INRA-USTL	
J.C.	FAVROT	INRA	
M.C.	GIRARD	INAPG	
J.P.	LEGROS	INRA	Directeur de Thèse

Institut National de la Recherche Agronomique
Laboratoire de science du sol
Montpellier

AVANT PROPOS

Avant d'exposer les résultats et les conclusions de ce travail, je tiens à remercier tous ceux qui ont contribué, par leurs conseils ou leur collaboration, à la réalisation de cette thèse.

Je tiens tout d'abord à exprimer ma gratitude et ma profonde estime à Jean Claude FAVROT, directeur du laboratoire de science du sol de l'INRA de Montpellier. J'ai bénéficié depuis le début de ma carrière de son soutien et de ses conseils. C'est peu de dire que ce travail lui doit beaucoup .

Jean Paul LEGROS, directeur de recherche INRA, a assumé la direction de ce travail. J'ai apprécié sa grande disponibilité et ses conseils avisés. Je lui exprime ici toute ma reconnaissance.

Jean Claude REMY, professeur à l'ENSA Montpellier, a bien voulu présider le jury de cette thèse. Il témoigne ainsi de l'intérêt qu'il a porté à mon travail . Qu'il en soit remercié .

Cette thèse m'a donné la chance de faire la connaissance de Peter BURROUGH, professeur à la faculté de géographie physique d'Utrecht. J'ai beaucoup apprécié les suggestions et les encouragements qu'il m'a prodigués en tant que rapporteur de ce travail. J'espère que ces premiers contacts se prolongeront dans l'avenir par de fructueuses collaborations.

Ce travail m'a également offert le privilège de partager avec Pierre BRABANT, directeur de recherche ORSTOM, une passionnante discussion lors de la remise de son rapport. J'aimerais la poursuivre, voire la concrétiser, cette fois-ci sur le terrain.

Des discussions passionnantes, j'en ai eu aussi avec Jean Paul CHEYLAN, chargé de recherche au GIP RECLUS. Bien qu'éloigné thématiquement, il m'a toujours manifesté son intérêt et sa bienveillance. Sa présence à mon jury est un nouveau témoignage de son ouverture d'esprit. Merci.

Yves ESCOUFIER, professeur de biométrie et vice-président de l'université de Monpellier II, a consacré une partie de son temps pour juger ce travail. Son avis est pour moi précieux et je voulais lui exprimer ma reconnaissance.

Michel Claude GIRARD, professeur de sciences du sol à l'INA Paris Grignon, a également accepté de participer à mon jury. Il témoigne ainsi de son intérêt toujours vivace pour un thème qu'il a beaucoup défendu par le passé. Je le remercie donc vivement.

Après avoir remercié les membres du jury, je tiens à manifester toute ma gratitude et ma sympathie à tous ceux qui ont participé de près ou de loin à la réalisation de ce travail.

Susan HOLMES, chargée de recherche au laboratoire de biométrie de l'INRA Montpellier, m'a initié à la méthode de segmentation. Elle a mis également à ma disposition le logiciel correspondant. Pour tout celà, je tenais à la remercier.

Je suis également reconnaissant à C.DEPRAETERE, chargé de recherche ORSTOM, pour m'avoir fait bénéficier de son expérience et de ses logiciels en matière d'acquisition et de traitement des MNT.

Patrick ANDRIEUX, Michel BORNAND et Marc VOLTZ ont été pour moi un soutien permanent, depuis la pose des premiers jalons du travail jusqu'à la relecture du document écrit et la préparation de l'oral. Grâce à eux, le terme "équipe de recherche" prend toute sa signification. Ils avaient déjà mon amitié, ils ont en plus ma reconnaissance.

Christophe LEDREUX a partagé bien des péripéties de ce travail. Je garde le souvenir de sa compétence et aussi de sa patience à toute épreuve. Merci Christophe, amitiés et ... à bientôt pour de nouvelles aventures informatiques!

Je ne saurais oublier mon "professeur de terrain" Robert BOUZIGUES. Avant d'entreprendre cette thèse, je me suis copieusement nourri de son expérience, en particulier en matière de cartographie des sols. Je lui exprime donc toute mon amitié et ma reconnaissance.

Enfin ces remerciements sont une occasion de témoigner à tout le personnel du laboratoire de science du sol de l'INRA Montpellier ma gratitude pour l'ambiance de travail agréable et pour l'aide de tous les jours que j'ai trouvée auprès de chacun. Je tiens à remercier en particulier Jean CORNET et François MAZZELLA pour leur contribution efficace à la réalisation des illustrations de cette thèse.

INTRODUCTION GENERALE

La couverture pédologique doit être considérée comme une **ressource non renouvelable** dont la **gestion raisonnée** s'impose de plus en plus face aux sollicitations, aux agressions et conflits d'usage dont elle est l'objet.

Cette gestion doit tenir compte de la **variabilité spatiale de la couverture pédologique**. A cause de cette variabilité, une référence établie en un lieu donné (mesure de propriété du sol, expérimentation, résultat d'enquête, modélisation, décision raisonnée,...) ne peut être automatiquement **transposée** en un autre lieu sans qu'au préalable ait été vérifiée la faisabilité de cette transposition et qu'en soient raisonnées éventuellement les modalités.

Or, produire une référence représente une opération généralement longue et coûteuse. Par conséquent, les sites en bénéficiant sont, par essence, **rares**. De ce fait, ces références sont destinées le plus souvent à être utilisées en de nombreux autres sites, sur de vastes territoires et à grande distance de leur lieu de production. Dans ce contexte, le recours à **une cartographie pédologique détaillée et systématique** (échelle > 1/10000, densité d'observation > 1/ha) représenterait un intéressant compromis entre fiabilité des transpositions et coût d'étude. Chaque référence ponctuelle serait rattachée à une unité cartographique, **l'unité de sol**, dont le contenu est supposé homogène vis à vis du comportement concerné. Le contenant de cette unité délimiterait alors la zone géographique où la référence en question pourrait être utilisée.

Pourtant, même cette solution reste encore trop coûteuse pour être généralisable. En France, elle n'a été envisagée et financée que dans trois départements (Aisne, Mayenne, Ile et Vilaine) et le taux de couverture du territoire français par des cartes pédologiques détaillées est actuellement inférieur à 10%. Au niveau mondial, mis à part quelques rares pays (Belgique, Hollande, Pologne, Cuba,...), la cartographie détaillée des sols est peu développée. Compte tenu du contexte économique actuel, il est peu probable que les collectivités acceptent de supporter l'investissement d'étude nécessaire pour combler ces manques.

Dès lors, il convient de rechercher de **nouvelles solutions** permettant d'assurer, à moindre coût, la **transposition** des rares références disponibles, vers des lieux dépourvus de toute cartographies pédologiques systématiques.

C'est dans cette perspective que s'inscrit le travail de recherche qui va être présenté dans ce mémoire. La voie choisie a pour cadre général la "**méthode des secteurs de référence**" (FAVROT 1981). Cette méthode de terrain, conçue pour satisfaire l'urgent besoin en références concernant le drainage des sols, a été appliquée dans 70 petites régions naturelles françaises entre 1980 et 1985 (voir description détaillée en première partie).

Dans ce cadre méthodologique, l'objectif général du travail consistera à étudier la **prédiction**, au sein d'une **petite région naturelle** donnée, des unités de sol présentes au moyen de leurs **lois de distribution**. Ces lois, traductions de l'organisation de la couverture pédologique, peuvent être dégagées à l'occasion de l'étude pédologique d'un périmètre limité mais représentatif de la petite région naturelle: le secteur de référence. Elles sont ensuite utilisées sur l'ensemble de la petite région naturelle au cours des cartographies suivantes, appelées "**retours à la parcelle**". Rarement (et jamais complètement) explicitées dans les documents remis aux utilisateurs, ces lois apparaissent cependant en filigrane sur la carte des sols. Deux catégories de lois de distribution peuvent être distinguées:

- Les lois fondées sur les relations entre les unités de sol et les autres éléments du milieu naturel,
- Les lois fondées sur les relations de voisinage entre unités de sol.

Le fait qu'elles puissent être utilisées pour prédire les unités de sol à l'extérieur du secteur de référence implique de supposer **la stabilité** de ces lois sur la petite région naturelle. Cette hypothèse devra donc être testée au cours du travail entrepris.

Dans cette perspective, le premier travail consistera à **analyser puis formaliser mathématiquement** le **raisonnement cartographique** utilisé lors d'un retour à la parcelle. A cette occasion, sera également défini un formalisme général des lois de distribution des unités de sol alimentant ce raisonnement. Dans un deuxième temps, ces lois seront **extraites automatiquement** à partir d'un secteur de référence d'une petite région naturelle prise comme exemple (Moyenne Vallée de l'Hérault). La mise en commun des résultats des deux étapes se traduira par la construction d'un **outil informatique** destiné à **automatiser** l'opération de retour à la parcelle dans la petite région naturelle considérée. Cet outil permettra ainsi de tester les lois de distributions des sols extraites à partir du secteur de référence sur des lieux non encore cartographiés.

La présentation du mémoire suit globalement la démarche proposée ci-dessus tout en séparant les deux types de lois de distribution présentées. Trois grandes parties sont distinguées.

La première partie présente les enjeux scientifiques et pratiques du travail de recherche. Elle aborde ensuite l'analyse et la formalisation du raisonnement cartographique utilisé lors du retour à la parcelle ainsi que les solutions informatiques adoptées pour l'automatiser. Enfin, c'est également dans cette partie qu'est décrit le cadre expérimental de ce travail: la petite région naturelle "Moyenne Vallée de l'Hérault", son secteur de référence et les secteurs de validation qui permettront de juger de la qualité des cartes fournies par le modèle construit.

La seconde partie traite spécifiquement de l'extraction automatique puis de l'utilisation des lois de distribution fondées sur les relations sol-paysage: Les modalités d'acquisition des données nécessaires à leur extraction automatique (utilisation d'un SIG) sont décrites ainsi que la méthode d'analyse de donnée utilisée (segmentation). Ensuite, les modalités d'interprétation des résultats de cette analyse sont développées en tenant compte du caractère géographique des données manipulées. Enfin, on évalue la qualité des prédictions sur les secteurs de validation. Ceci permet de tirer des conclusions quant aux perspectives d'utilisation de prédictions fondées sur de telles lois.

La troisième partie représente le pendant de la précédente pour, cette fois-ci, les lois fondées sur les relations de voisinage entre unités de sol. L'algorithme conçu spécifiquement dans le but d'extraire ces lois d'une carte des sols est décrit. La méthode permettant de faire la synthèse des prédictions qu'elles fournissent en un point donné est justifiée. Dans un deuxième temps, les résultats de prédictions sont présentés et discutés dans la perspective de leur utilisation future. Un couplage avec les lois précédentes, est ensuite tenté. Enfin, à titre exploratoire, Il est tenté une stratégie de choix de l'emplacement des sondages à partir desquels fonctionnent les lois de distribution fondées sur les relations de voisinage.

A la fin de tous les chapitres divisant chaque partie, le lecteur trouvera des conclusions (écrites en caractère gras) résumant l'essentiel à retenir. Chacune des parties fait également l'objet d'une synthèse générale.

PREMIERE PARTIE:

OBJECTIFS, OUTILS ET METHODES D'APPROCHE DE L'AUTOMATISATION DE LA CARTOGRAPHIE PEDOLOGIQUE A PARTIR D'UN SECTEUR DE REFERENCE

La nécessité d'engager un travail de recherche sur le thème de l'automatisation de la cartographie pédologique à partir d'un secteur de référence est apparue à l'occasion d'une expérience d'encadrement de nombreuses études de sol menées sur l'ensemble du territoire français. Ce thème recouvrait en fait de nombreux sous-problèmes difficiles à traiter exhaustivement dans une seule thèse. Il était donc nécessaire de définir des priorités d'étude, ce qui a été fait en choisissant de s'intéresser plus particulièrement aux lois de distribution des unités de sol. Pour autant, le thème général initialement choisi ne devait pas être perdu de vue.

Cette partie répond à ce souci. Elle présente les objectifs de ce thème et les méthodes et outils choisis pour l'aborder. Elle permet donc de positionner vis à vis des recherches sur l'automatisation de la cartographie pédologique le travail sur les lois de distribution des unités de sol présenté dans les deux parties suivantes.

Dans un **premier chapitre**, sera présenté le problème général que ce travail doit contribuer à résoudre. Il s'agit d'utiliser la connaissance acquise sur un secteur de référence pour de nouvelles cartographies au sein de la petite région naturelle que ce secteur est censé représenter. Les aspects pratiques et scientifiques de ce problème seront successivement évoqués.

Le **deuxième chapitre** présentera la méthode envisagée pour simuler la démarche cartographique mise en oeuvre par le pédologue au cours du retour à la parcelle. Il s'agit en fait d'analyser, de formaliser puis d'automatiser le raisonnement permettant de prédire les unités de sol à partir de leurs lois de distribution. Les choix effectués dans ce chapitre détermineront le formalisme des lois de distribution des unités de sol.

Le dernier chapitre décrira la petite région "Moyenne Vallée de l'Hérault", son secteur de référence et les secteurs choisis pour une première validation du modèle de démarche cartographique, le tout constituant le milieu expérimental du travail de recherche.

CHAPITRE 1

ORIGINE ET DEFINITION DES OBJECTIFS

le travail de recherche entrepris a pour objet de compléter et de conforter la démarche inhérente à la méthode des secteurs de référence. En outre, il aborde, au niveau scientifique, le problème plus général de l'étude des sols à un niveau de perception "régional" (> 10 km^2).

1. L'ASSISTANCE AU RETOUR A LA PARCELLE DANS LE CADRE DE LA METHODE DES SECTEURS DE REFERENCE

La méthode dite des "secteurs de référence" a été mise au point dès la fin des années 70 au laboratoire de science du sol de l'INRA de Montpellier (FAVROT et al, 1978; FAVROT, 1981; 1989). Cette méthode a été en particulier appliquée entre 1980 et 1985 sur 70 petites régions naturelles françaises (LAGACHERIE, 1987) dans le cadre de l'opération "secteurs de référence drainage", financée par l'ONIC et le Ministère de l'Agriculture (cf carte figure 1). Le laboratoire de Science du Sol de l'INRA Montpellier et la Division Drainage du CEMAGREF d'Antony ont assuré conjointement le suivi technique de cette opération. Ce suivi avait pour but d'assister les maitres d'ouvrages dans le choix des secteurs de référence, l'orientation et le contrôle de la qualité des études, confiées à des chargés d'études pédologiques.

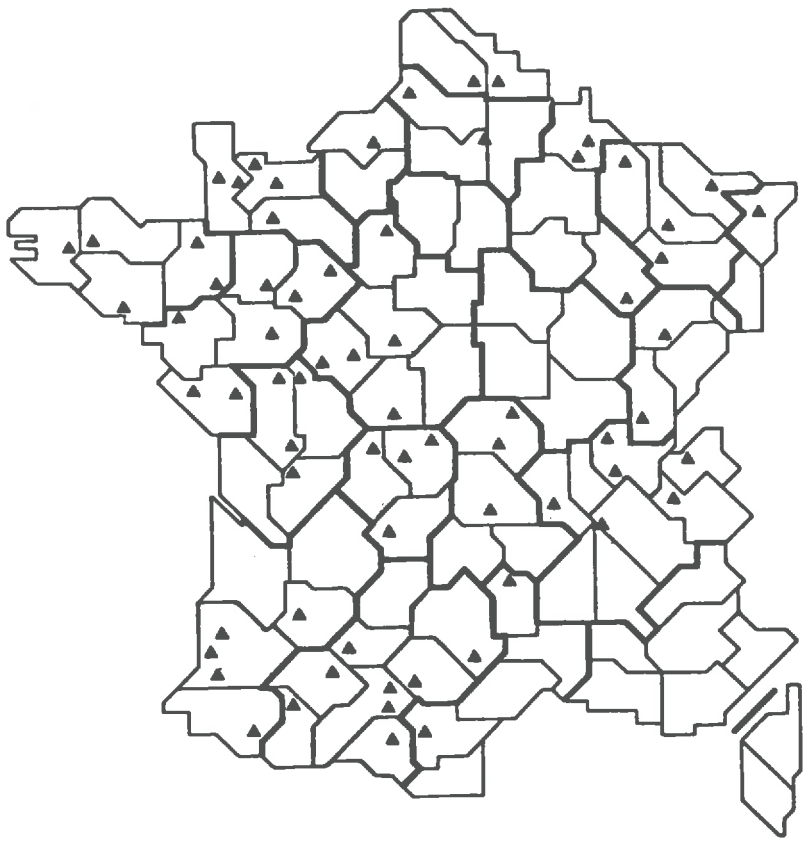


***Figure 1**: Localisation des secteurs de références étudiés au cours de l'opération drainage ONIC Ministère de l'Agriculture (d'après LAGACHERIE, 1987)*

1.1. La méthode des secteurs de référence: objectif et démarche

L'objectif de la méthode des secteurs de référence appliquée au drainage était de fournir, pour toute parcelle à drainer, des recommandations sur les caractéristiques du réseau de drainage projeté (écartement des drains, nature du remblai, type d'outil de pose,...). Ces recommandations devaient s'appuyer sur une connaissance détaillée des différents types de sol de chaque parcelle. Cet objectif, particulièrement ambitieux, ne pouvait évidemment pas être atteint par une démarche de cartographie des sols "classique". Elle aurait en effet nécessité une cartographie systématique et détaillée (échelles 1/5000, 1/10000) du territoire national dont la mise en oeuvre n'était pas envisageable compte tenu de l'investissement requis.

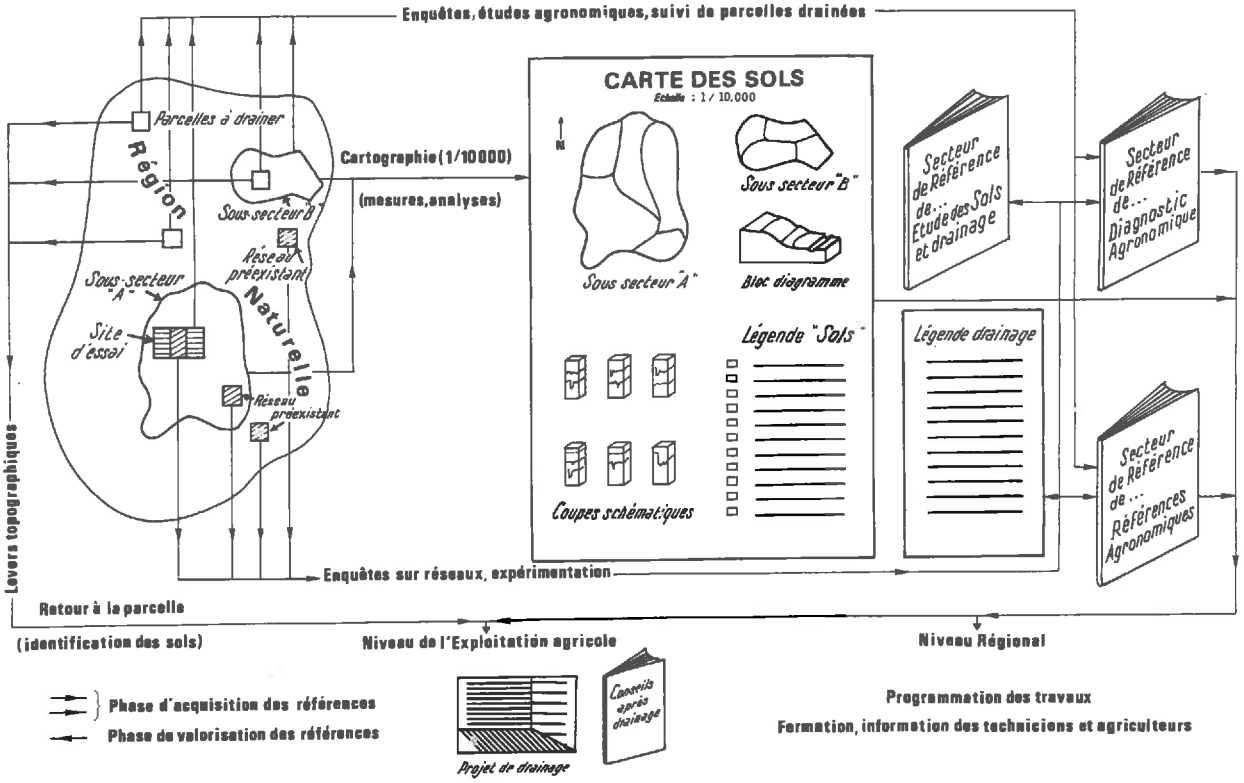


Figure 2: Les principes et étapes de la méthode des secteurs de référence (d'après FAVROT, 1989)

Alternative à cette démarche classique, la méthode des secteurs de référence (figure 2) propose d'aborder l'étude des sols par la définition préalable de "petites régions naturelles considérées comme des entités géographiques bien individualisées sur le plan géo-pédologique (voir paragraphe 2). L'étude des sols de chacune de ces petites régions naturelles est menée suivant deux étapes successives (FAVROT, 1989).

La première étape consiste en une cartographie pédologique détaillée (échelle 1/10000) et une caractérisation approfondie (morphologique, physico-chimique, hydrodynamique) des sols d'une aire échantillon représentative (le "secteur de référence") de superficie limitée (de 1 à 5% de l'ensemble de la petite région naturelle). Cette première étude hydropédologique est complétée par

une analyse critique des résultats d'expériences régionales (examen de réseaux de drainage préexistants) et, si nécessaire, par l'installation puis le suivi de dispositifs expérimentaux. L'ensemble de ces actions se concrétise par l'élaboration d'un fichier de sols régionaux avec, pour chacun d'eux, un ensemble de références associées qui prennent la forme de recommandations technologiques et agronomiques utilisables à la fois pour la conception des plans projets puis pour la conduite des terres aménagées. Ceci se traduit concrètement par l'édition d'une carte des sols avec une double légende (sol et drainage, cf figure 2) et d'un rapport technique associé.

Ensuite, pour toute parcelle de la petite région naturelle pour laquelle un agriculteur envisage un drainage, une prospection de terrain "allégée" permet le zonage et le rattachement des sols rencontrés aux unités répertoriées dans l'aire échantillon. Ceci permet d'appliquer les références préalablement élaborées selon le principe "à unités de sol semblables, recommandations semblables". Cette opération dite de **retour à la parcelle** est, comme le souligne FAVROT, " facilitée par les connaissances déjà acquises sur la nature et le mode de distribution des sols". Elle doit théoriquement être réalisable par un intervenant, pédologue ou non, distinct du chargé d'étude initial du secteur de référence.

Ainsi présentée, la méthode des secteurs de référence présente trois intérêts majeurs qui persisteraient quel que soit le type d'application envisagé:

- elle permet des économies d'échelles; la caractérisation approfondie de chaque unité de sol et la production des références est réalisée une seule fois lors de l'étude du secteur de référence et retranscrite clairement dans les documents fournis à l'utilisateur;
- elle offre la possibilité de mieux organiser les interventions sur une petite région naturelle donnée; le chargé d'étude du secteur de référence transmet sa connaissance à des tiers, pédologues ou non, susceptibles d'effectuer eux-même le retour à la parcelle. Cet avantage était déterminant car le faible effectif (200) de " pédologues draineurs" constaté à l'époque de la mise en oeuvre de la méthode (CESTRE et Al, 1984) aurait limité les possibilités d'intervention face à une demande massive (et pressée!) des agriculteurs;
- elle suit (et ne précède pas) la demande des agriculteurs; l'investissement en étude est mieux valorisé puisque seules les parcelles pour lesquelles un aménagement est effectivement projeté font l'objet d'une étude ("pas d'études sans travaux, pas de travaux sans études").

1.2. Un problème identifié à l'issue de l'opération ONIC-Ministère de l'Agriculture: la mise en oeuvre du retour à la parcelle

Un des enjeux de la réussite de la méthode est constitué par la bonne articulation entre les deux étapes évoquées précédemment. L'opération dite de "retour à la parcelle", pour être efficace, doit en effet mobiliser au maximum l'expérience acquise lors de l'étude du secteur de référence. A ce titre, le bilan de 5 années de suivi est mitigé. Il tient en deux points:

- le retour à la parcelle devient effectivement une opération légère (uniquement réalisée par sondages à la tarière) et "de routine" dans le cas où l'intervenant est le chargé d'étude du secteur de référence; elle peut également le devenir pour un autre pédologue, sous réserve d'une "mise en route" de quelques jours sur le secteur étudié par son collègue;
- par contre, le retour à la parcelle s'avère difficile pour un non pédologue si on en juge par le faible nombre de cas où, pour cette opération, un technicien drainage ou un conseiller agricole prend le relai du chargé d'étude du secteur de référence.

Ce constat limite en partie l'intérêt de la méthode, en particulier le troisième point évoqué plus haut. De fait, certains secteurs de référence ont été sous-utilisés voire totalement inexploités en l'absence de compétences locales capables de les valoriser.

Un diagnostic peut être avancé: le fait que le pédologue auteur du secteur de référence réussisse mieux le retour à la parcelle que les autres intervenants (y compris ses collègues pédologues) laisse penser qu'une partie de sa connaissance n'a pas été transmise au travers des documents qu'il remet à l'utilisateur à l'issue de son étude. Effectivement, aucun des documents remis aux utilisateurs ne fournissent un "mode d'emploi" permettant à ceux-ci de faire eux-même la carte des sols d'une parcelle à drainer alors que, pourtant, le pédologue auteur de l'étude de secteur de référence considère cette opération comme de la "routine".

Il convient donc de formaliser ce savoir faire cartographique de façon à permettre une plus large utilisation des secteurs de référence passés (environ une centaine en 1992) et à venir (opérations INRA-Ministère de l'agriculture Inventaire Gestion et Conservation des Sols, secteurs de référence irrigation, ...).

Dans cette perspective, cette formalisation doit être réalisée de telle sorte qu'elle puisse servir de base à la construction d'un outil informatique permettant d'envisager, à terme, une automatisation complète ou partielle du retour à la parcelle (notion de "retour à la parcelle assisté par ordinateur")

2. L'ETUDE DE LA STABILITE DES LOIS DE DISTRIBUTION DES SOLS AU SEIN D'UNE PETITE REGION NATURELLE

La méthode des secteurs de référence a été élaborée à l'issue d'une expérience de terrain de plusieurs années de l'équipe de cartographes de l'INRA de Montpellier. Elle s'appuie sur des hypothèses forgées au cours de cette expérience. Ces hypothèses sont en fait utilisées et admises par d'autres équipes, françaises et étrangères, ayant pratiqué la même activité même si les formulations peuvent être différentes et plus ou moins explicites.

Ces hypothèses de terrain sont au nombre de deux et peuvent être formulées comme suit:

- Une référence en matière de nature ou de comportement du sol, établie ponctuellement sur une unité de sol du secteur de référence, peut être généralisée à l'ensemble de celle ci. Ceci revient à supposer que la variabilité interne de cette unité est suffisamment faible pour que cette généralisation s'effectue sans conséquences fâcheuses pour l'utilisateur (en l'occurrence ici l'agriculteur)
- Les unités de sol sont en nombre fini et présentent des lois de distribution stables sur l'ensemble de la petite région naturelle pédologique. De plus, la majorité d'entre-elles (80 à 90% selon JC FAVROT, 1981) peut être rencontrée dans un périmètre limité (le secteur de référence) pourvu qu'il soit "bien choisi".

La première hypothèse n'est pas spécifique à la méthode des secteurs de référence. Elle est implicite pour l'ensemble de la méthodologie de cartographie des sols et a fait et fait encore l'objet de nombreux essais de validation. Au vu de l'ensemble de ces essais il est clair que cette hypothèse ne peut être vérifiée pour chaque propriété du sol prise individuellement. Il existe en effet de grandes différences de variabilité spatiale entre les propriétés impliquées et d'autre part, à propriété

identique, une forte diversité des milieux d'étude abordé. Néanmoins, dans une étude bibliographique détaillée tenant compte notamment des précédentes synthèses sur ce sujet (BECKETT et al, 1971; WILDING et DREES, 1983), VOLTZ (1992) considère qu'en moyenne la cartographie pédologique permet une stratification significative, ce qui va dans le sens de l'hypothèse formulée. En définitive, on peut conclure que, même si l'hypothèse n'est pas totalement vérifiée, les différentes expériences tentées montrent qu'elle l'est au moins partiellement. Cette hypothèse sera donc supposée vérifiée, sa validation sortant du champ d'investigation de ce travail.

La deuxième hypothèse se retrouve également dans la littérature avec des formulations variées selon les auteurs mais, en général, moins hardies que celle formulée ci-dessus. Elle apparaît à l'occasion d'articles où sont abordés des concepts proches ou équivalents à la notion de "petite région naturelle".

Ainsi SMECK et al (1983) définissent dans un article de synthèse le concept de "pedological province": "part of a region, isolated and defined by its climate and topography and characterized by a particular group of soil". Dans cette définition, apparaît l'affirmation d'un macro-système de sols délimitable dans l'espace et possédant un nombre fini de types de sol.

ASTLE et al (1969) s'appuyant sur des expériences plus anciennes encore affirment "...in any one landscape there are only a few kind of terrain. ..(they) recur in association with one another in the landscape to give a more or less regular pattern always in the same interrelations. .. A new lanscape is recognized where there is a change either in the terrain types or in the relation between them...". Cette affirmation, plus ambitieuse que la définition précédente, est très proche de la première partie de l'hypothèse formulée puisque l'idée de stabilité de la distribution des sols est explicitement présente.

Cette idée de distribution stable des unités de sols est également reprise par HOLE (1977) qui met en évidence, sur la base de cartes de sol publiées, des systèmes sols-paysage ("soilscapes") qu'il décrit par des critères quantitatifs traduisant l'arrangement des unités de sol tel qu'il apparaît sur les cartes. A l'aide de plusieurs exemples, il montre que ces systèmes peuvent avoir un déterminisme identifiable et spécifique ("soilscape fabric").

L'existence de systèmes de sol est également affirmée et utilisée par les pédologues tropicaux français. Ainsi, s'appuyant sur une longue expérience collective, BRABANT (1989) propose une démarche cohérente de cartographie des sols tropicaux. Elle consiste à identifier dans un premier temps des "systèmes sol" sur la base de cartes topographiques, géologiques et géomorphologiques ("études de niveau 1"). Ils sont ensuite caractérisés par l'étude détaillée de "sites représentatifs" ("études de niveau 2") valorisés ensuite dans les études ultérieures portant sur l'ensemble de ces systèmes-sol (études de niveaux 3,4,5). Bien que les superficies citées pour chaque entité et les méthodes de caractérisation détaillées soient différentes, la démarche proposé par BRABANT sous tend sensiblement les mêmes hypothèses d'organisation de la couverture pédologique que celles utilisées dans la méthode des secteurs de référence.

Ces quelques exemples ne constituent pas une étude bibliographique exhaustive concernant ce sujet. Cependant, choisis pour leur diversité de formulation et de communauté scientifique d'origine, ils montrent que l'hypothèse de terrain sous-tendant et caractérisant la méthode des secteurs de référence est partagée par nombre de pédologues confrontés à la nécessité de comprendre la variabilité de la couverture pédologique à une échelle régionale. Il faut noter cependant que les essais de validation de cette hypothèse sont très rares, vraisemblablement à cause de la difficulté et la lourdeur d'une telle entreprise. Deux travaux peuvent néanmoins être cités.

Le premier (POURGATON, 1977), utilise des cartes existantes pour apprécier a posteriori la surface minimale qu'il aurait fallu prospecter pour rencontrer toutes les unités de sol d'une petite

région naturelle. Les résultats, concernant trois milieux pédologiques différents, montrent qu'il semble effectivement possible d'échantillonner la petite région naturelle de façon à rencontrer une majorité des unités présentes sur son ensemble. Cependant, les pourcentages utiles pour réaliser cet échantillonnage restent supérieurs à ceux qui seront par la suite préconisés dans la méthode des secteurs de référence (1 à 5% de la petite région naturelle). Ceci s'explique en partie par le fait que les résultats de M.POURGATON ne tiennent pas compte de "l'art" de choisir un échantillon de territoire permettant d'optimiser le rapport nombre d'unité englobées/ surface prospectée.

Le deuxième travail sur ce sujet a été réalisé par FAVROT et al (1987) à la fin de l'opération "secteur de référence drainage" ONIC-MINAGRI. Il s'agit d'un test d'évaluation de la représentativité de 5 secteurs de référence caractérisant des petites régions naturelles variées. Les résultats portent au total sur 2300 ha de retours à la parcelle effectués sur ces petites régions à la suite des études de secteurs de référence. Ils semblent globalement confirmer une partie de l'hypothèse de départ puisque les superficies cumulées des nouvelles unités de sol rencontrées lors de ces retours à la parcelle ne dépassent jamais 10% quelles que soient les petites régions naturelles envisagées.

Il convient de noter que ces rares essais de validation ne concernent que l'aptitude à rencontrer dans un périmètre restreint la quasi totalité des unités de sols potentiellement cartographiables sur l'ensemble de la petite région naturelle. Par contre, rien dans ces essais ne permet de déterminer dans quelle mesure la stabilité supposée des lois de distribution des sols correspond à une réalité sur le terrain.

Pour espérer progresser dans cette recherche, l'outil informatique reproduisant l'opération de retour à la parcelle, objectif finalisé défini au sous-chapitre précédent, peut se révéler précieux. En effet, le retour à la parcelle utilise naturellement les lois de distribution des sols telles qu'elles apparaissent à l'intérieur du secteur de référence. Si, sur des zones situées à l'extérieur, la simulation informatique génère de faibles erreurs de cartographie, il faudra conclure à la stabilité, au moins sur ces zones, des lois de distribution des sols. En effet, l'outil informatique construit ne pourra, à la différence du pédologue, être soupçonné de facultés d'adaptation permettant en cours de démarche de modifier les lois de distribution des sols utilisées. Dans le cas contraire, la stabilité des lois de distribution des sols sera infirmée.

En résumé, L'objectif général de ce travail est de formaliser et d'automatiser les "retours à la parcelle". Ce terme désigne les nouvelles études pédologiques réalisées à la suite de l'étude d'un secteur de référence et bénéficiant de l'expérience cartographique acquise à son issue. Cet objectif revêt un double enjeu:

- **sur le plan finalisé, il s'agit de mieux valoriser les secteurs de référence existants et à venir,**
- **Sur le plan cognitif, il s'agit d'évaluer la validité de l'hypothèse de stabilité des lois de distributions des sols.**

CHAPITRE 2

LA DEMARCHE CARTOGRAPHIQUE UTILISEE LORS DU RETOUR A LA PARCELLE: ANALYSE, FORMALISATION et AUTOMATISATION

Dans son acception la plus générale, la démarche cartographique pédologique, désignée par GIRARD (1983) par le néologisme "cartogénèse", consiste selon ce dernier, à "modéliser la couverture pédologique après l'avoir analysée afin d'en déterminer les différents volumes et les délimiter spatialement". Cette opération est rarement décrite dans tout ses détails, même par ceux qui la pratiquent couramment. L'étude bibliographique menée par GIRARD (1983) témoigne bien de cet état de fait, résumé dans cette phrase de SCHELLING (1970): "research in the field of soil survey is among the undeveloped area of soil science. This is all the more remarkable when it is considered that soil survey have been carried out for a very long time by a great many people".

Cette démarche est généralement une opération intellectuelle complexe. Il s'agit en effet de concilier deux objectifs toujours plus ou moins contradictoires:

- obtenir des unités cartographiques dont le contenu est le plus homogène possible;
- obtenir des unités cartographiques suffisamment vastes et individualisées dans le paysage pour:
  - + ne pas fournir une image trop complexe de la couverture pédologique et, de ce fait, inexploitable par l'utilisateur,
  - + être capable, sur le terrain, de les délimiter aisément (sans trop de sondages) grâce à des lois de distributions sûres.

La manière de résoudre ce compromis diffère selon les différentes "écoles" de cartographie, selon les milieux d'études rencontrés, selon l'objectif assigné à la cartographie et même selon les unités cartographiques d'une même étude.

Des tentatives de formalisation informatique de la démarche cartographique ont été entreprises, notamment en France, par GIRARD (1983) et KING (1985). Elles n'ont été possibles qu'au prix d'une simplification évitant en particulier la prise en compte des lois de distribution des sols.

La démarche cartographique utilisée lors du retour à la parcelle, objet d'étude du présent travail, représente en fait un cas particulier de la démarche évoquée ci-dessus. La différence est que la prospection ne remet plus en cause la nature et les lois de distribution des unités de sol puisque celles-ci ont été préalablement recherchées et établies au cours de l'étude préalable du secteur de référence. La démarche cartographique est donc dans une sorte de "régime de croisière" où les nouvelles observations n'ont plus de conséquences sur les lois de cartographie adoptées par le pédologue.

L'étude de cette démarche est abordée suivant trois étapes successives qui feront l'objet des trois sous-chapitres suivants:

- la première est l'**analyse** des connaissances acquises au cours d'une cartographie des sols et de leur disponibilité dans les documents remis aux utilisateurs; appliquée au cas des secteurs de références, cette analyse permettra de proposer un "**schéma de**

fonctionnement" décrivant comment le pédologue mobilise sa connaissance du secteur pour réaliser les "retours à la parcelle";

- la deuxième est une **formalisation mathématique** du retour à la parcelle réalisée à partir du schéma de fonctionnement défini. Elle permettra de définir sous quel formalisme doivent être exprimées les lois de distribution des unités de sol.
- la troisième concerne les options prises en matière informatique pour tenter d'**automatiser** le retour à la parcelle. Les outils informatiques employés seront présentés succintement, leur mise en oeuvre effective étant par ailleurs évoquée tout au long des parties suivantes.

1. ANALYSE DE LA DEMARCHE CARTOGRAPHIQUE MISE EN OEUVRE LORS DU RETOUR A LA PARCELLE

La cartographie d'un secteur de référence, permet au pédologue d'acquérir de nouvelles connaissances. Une partie d'entre elles lui sont utiles pour réaliser plus efficacement les retours à la parcelle ultérieurs. Certaines de ces connaissances figurent effectivement dans les documents remis aux utilisateurs. Cependant, le mode de diffusion actuel de l'information (carte, rapport,...) semble insuffisant pour transmettre à un tiers le "savoir faire cartographique" (chapitre 1)

Cette situation se retrouve en fait dans de nombreux domaines (médecine, industrie,...). Elle a justifié l'émergence d'une activité nouvelle, le "génie cognitif" (VOGEL, 1988) destiné à traduire, en vue de leur informatisation, des connaissances d'experts, jusque là non exprimées et non utilisables par des tiers. C'est à cette activité que peut se rattacher le travail présenté dans ce chapitre. Deux étapes seront distinguées:

- rassembler et classifier les éléments constitutifs de l'expertise du pédologue acquise à la suite d'une cartographie de secteur de référence,
- associer ces éléments au travers d'une succession d'opérations simulant la réalisation d'une nouvelle carte (schéma de fonctionnement).

A ce stade de la réflexion, aucune contrainte de faisabilité liée à la formalisation mathématique ou à l'informatisation n'est envisagée.

1.1. La connaissance acquise à l'issue d'une cartographie de sol

De manière schématique, il est possible d'identifier quatre éléments constitutifs de l'expertise du pédologue:

- une typologie d'**unités de sol** susceptibles d'être reconnues grâce à des **lois d'identification des unités de sol** reposant sur des critères d'accès aisé (au moyen d'une tarière à main),
- des relations entre l'occurrence sur le terrain de ces unités de sol et des critères extrinsèques (descripteurs de la topographie, de la géologie, de la végétation,...) ou "de surface" (couleur, aspect du labour,...). Ces relations seront appelées par la suite **relations sols-paysage** selon la terminologie adoptée couramment dans la bibliographie ("soil-landscape relations"). Elles seront utilisées au cours de la cartographie sous forme de lois appelées dans la suite **lois sols-paysage**,
- des relations entre la présence d'une unité et celle des autres, traduites par la suite sous le terme **relations de voisinage** (entre unités), utilisées sous forme de **lois de voisinage**,

- une connaissance portant sur la délimitation des unités de sol. Celle ci peut concerner aussi bien les limites tracées au vu de la morphologie du sol que celles déduites de l'existence de critères particuliers aisément observables sur la zone de transition entre deux unités. De même que précédemment pourraient être définies des **lois de tracé de limites.**

1.1.1. La typologie des unités de sol et les lois d'identification des unités de sols associées

La démarche cartographique vise à construire une typologie d'unités de sol spécifique au milieu étudié, au moins dans le cas des approches à grande échelle comme les secteurs de référence. En effet comme le souligne PEDRO (1989), les approches actuelles de cartographie ne font plus appel comme par le passé à une taxonomie de référence, qu'elle soit génétique (exemple CPCS, 1967) ou basée sur une collection de types préétablis définis par des critères quantitatifs (ex Soil Taxonomy). Cette souplesse retrouvée permet de s'adapter à la diversité des milieux d'études et de s'attaquer à la résolution optimale du compromis cartographique évoqué plus haut.

Ainsi, comme l'évoque BAIZE (1986), des unités de sols sont créées à partir "d'un découpage intellectuel du continuum (pédologique). ..selon certains critères que l'on sélectionne, que l'on hiérarchise en utilisant des seuils". Dans la démarche cartographique classique du type de celle adoptée au cours des études de secteur de référence, ces unités de sols sont schématisées par des profils types.

Au cours de l'opération ultérieure de retour à la parcelle les échantillons de couverture pédologique observés par sondage à la tarière seront rattachés à l'une ou l'autre des unités de sol distinguées. Pour celà, sont mises en oeuvre des lois d'identification basées sur un nombre limité de critères morphologiques du sol accessibles à la tarière et dont la conjonction d'apparition sur un même sondage, à certaines profondeurs significatives, est révélatrice, pour le pédologue, de la présence d'une des unités de sol du secteur de référence.

Ces lois sont rarement apparentes dans les documents remis classiquement aux utilisateurs (légende de la carte et rapport). En effet, les véritables critères d'identification des unités de sol utilisés sur le terrain par le pédologue ne sont généralement pas soulignés au milieu de l'ensemble des descripteurs caractérisant les unités de sol au sein du profil type.

1.1.2. Les relations sols-paysage et leurs lois associées

Afin de réduire le nombre de sondages nécessaires pour délimiter les unités de sols, le pédologue tente, au cours de la cartographie (en particulier de celle du secteur de référence), d'établir les lois lui permettant de prévoir l'apparition des unités de sol au moyen de critères extrinsèques au sol (relief, roche-mère, végétation) ou "de surface" (couleur en surface, pierrosité etc...).

Bien qu'il existe, à la suite de JENNY (1941), des tentatives de formulation de lois générales exprimant les relations entre le sol et certains critères comme le relief (HUGGET, 1975; CONACHER et DARLYMPE, 1977), il est plus courant dans la littérature de voir dégager des lois ayant une portée locale au moyen d'un raisonnement bâti "sur mesure" pour le milieu étudié. Une abondante bibliographie pédologique témoigne de l'existence de telles démarches, en particulier lorsqu'elles mettent en jeu le déterminisme de formation des sols. Ainsi, en France les travaux de JP LEGROS (1975) dans le Massif Central et G.CALLOT (1977) en région Nord-Aquitaine offrent des

exemples de construction et d'application de telles lois, même si les moyens employés dépassent ceux utilisés dans une cartographie normale.

Dans le cas particulier du retour à la parcelle, ces lois, désormais non remises en cause, sont évidemment mobilisées. Cependant, les rapports d'études pédologiques ne les présentent pas de façon explicite (encore moins la démarche ayant permis de les établir), même si des indications sur les positions topographiques les plus probables des unités de sol sont fréquemment données dans les descriptions d'unités. Une lecture "experte" de la carte permet parfois d'en dégager certaines (relations sols-relief). L'adjonction de coupes ou blocs diagramme, systématisé au cours de l'opération "secteur de référence ONIC-Ministère de l'agriculture", permet également une meilleure perception des lois sols-paysage par les utilisateurs (LAGACHERIE, 1987).

1.1.3. Les relations de voisinage entre unités de sols et leurs lois associées

L'examen d'une carte pédologique permet de mettre en évidence des relations entre les unités distinguées. Certaines sont systématiquement associées dans un même lieu, d'autres, au contraire, semblent se repousser. Sur certaines cartes, apparaissent même des motifs d'unités présentant des géométries spécifiques. L'étude générale de ces motifs et de leur déterminisme, présentés sous le vocable "soil-combination" est due principalement à FRIDLAND (1972).

Ces relations de voisinage peuvent être utilisées sous forme de "lois de voisinage" au cours de la démarche cartographique de retour à la parcelle, au même titre et en même temps que les précédentes, pour limiter les observations directes de la couverture pédologique. Une unité de sol donnée étant identifiée en un point, les relations qu'elle entretient avec les autres permettent de prévoir l'occurrence ou la non-occurrence de telle ou telle unité (y compris elle même) dans l'environnement de l'observation réalisée. C'est la géométrie et le déterminisme du motif d'unités de sol auquel appartient l'unité reconnue qui oriente les prévisions. Par exemple, dans le cas, fréquent dans les secteurs de référence, de séquences (ou "combines" selon FRIDLAND) déterminées par le relief (toposéquences), l'unité de sol prédite en un lieu donné ne sera pas la même suivant que ce lieu se situera plus haut ou plus bas que le point où une unité de sol a été reconnue.

Il est très rare de trouver mentionnées dans les rapports des relations de voisinage. Seule une lecture experte de la carte des sols révèle éventuellement l'existence de telles relations. Là encore, surtout lorsqu'il s'agit de toposéquences, les coupes et blocs diagrammes fournissent une aide précieuse.

1.1.4. La connaissance sur les limites d'unités de sol

Le tracé des limites entre unités de sol s'appuie sur trois types de connaissances:

- Une connaissance de la morphologie de la couverture pédologique de ces zones de transition; très longtemps, cet aspect a été passé quasiment sous silence dans les ouvrages traitant de cartographie; seul, VINK (1963) explique en détail la démarche permettant de tracer une limite en cas de transition continue; l'un des mérites de l'analyse structurale (BOULET, 1982) a été de faire de la connaissance de ces zones de transition un objectif à part entière du pédologue cartographe; "il faut prendre le temps de travailler au limites" (A RUELLAN et Al, 1989);

- l'identification de points singuliers concernant des critères extrinsèques ou de surface ayant une signification directe ou indirecte en terme de variation de la couverture pédologique (exemple: point d'inflexion de pente, limite de terrasse géologique,...);
- la prise en compte de la forme géométrique que doit prendre la limite, souvent en relation étroite avec le contenu de l'unité de sol (FEENY, 1988).

Ainsi, dans la démarche de retour à la parcelle, tracer une limite entre deux zones que les observations et prévisions antérieures rattachent à deux unités distinctes relève souvent d'un raisonnement complexe. Il faut en effet concilier vraisemblance de position au vu des sondages effectués, nécessité de respecter une discontinuité visible (le plus souvent de façon intermittente) sur le terrain ou sur photo aérienne et souci de tracer une limite ayant une forme "plausible" (par exemple sans angles trop "vifs"). Peu d'informations existent sur cette phase dans les documents remis aux utilisateurs. En effet, les cartes de sols ne comportent pas encore, en dépit des propositions de GIRARD (1983), de formalismes permettant au lecteur de reconnaître une quelconque sémantique attachée à la limite (par exemple, classement en fonction de son caractère plus ou moins abrupt). Cette carence n'est que rarement compensée dans les rapports par une explicitation de ce que signifie exactement chaque limite et par la fourniture d'une méthode pour la retrouver sur le terrain. Les lois de tracé de limite sont donc très difficiles à dégager.

Quatre sources d'informations distinctes, issues de l'étude préalable du secteur de référence alimentent donc la démarche cartographique mise en oeuvre lors des retours à la parcelle. Le constat de semi-échec évoqué au chapitre 112 trouve un début d'explication. Il est clair, au vu des descriptions ci-dessus, que le pédologue auteur du secteur de référence est a priori mieux à même de mobiliser ces informations dans la mesure où les utilisateurs ne les retrouvent pas toujours présentées sous une forme explicite permettant leur appropriation effective.

1.2. Proposition d'un schéma de fonctionnement représentant le retour à la parcelle

Dans la perspective de préparer la formalisation mathématique du retour à la parcelle, il convient maintenant d'analyser en détail comment sont mobilisés, au cours de cette opération, les différents aspects de la connaissance acquise à l'issue de l'étude du secteur de référence.

L'objectif ultime poursuivi lors du retour à la parcelle va être de prédire en tout point de l'espace l'unité de sol présente, celle-ci ne pouvant être, suivant l'hypothèse utilisée, qu'une de celles du secteur de référence. Le formalisme adopté reprend l'idée d'une boucle, idée avancée notamment par GIRARD (1983) pour présenter sa vision de la démarche cartographique. Cette boucle est présentée sur la figure 3, Elle ordonne de façon chronologique les opérations nécessaires à la production finale d'une carte.

Dans un premier temps, le pédologue prédit, pour chaque zone de la parcelle à cartographier, la présence des unités de sols. Cette prédiction s'appuie sur les **lois sols-paysage** (cf 1.1.2.). Elles utilisent les critères extrinsèques et de surface qu'il est possible d'observer sans sondage.

Ensuite, une série de sondages à la tarière est effectuée. Pour chaque sondage, le pédologue suit une démarche en trois phases successives, la dernière étant facultative:

- le sondage est rattaché, sur la base de sa description, à l'une des unités de sol connues du secteur (cf 1.1.1) grâce à l'application de **lois d'identification**;

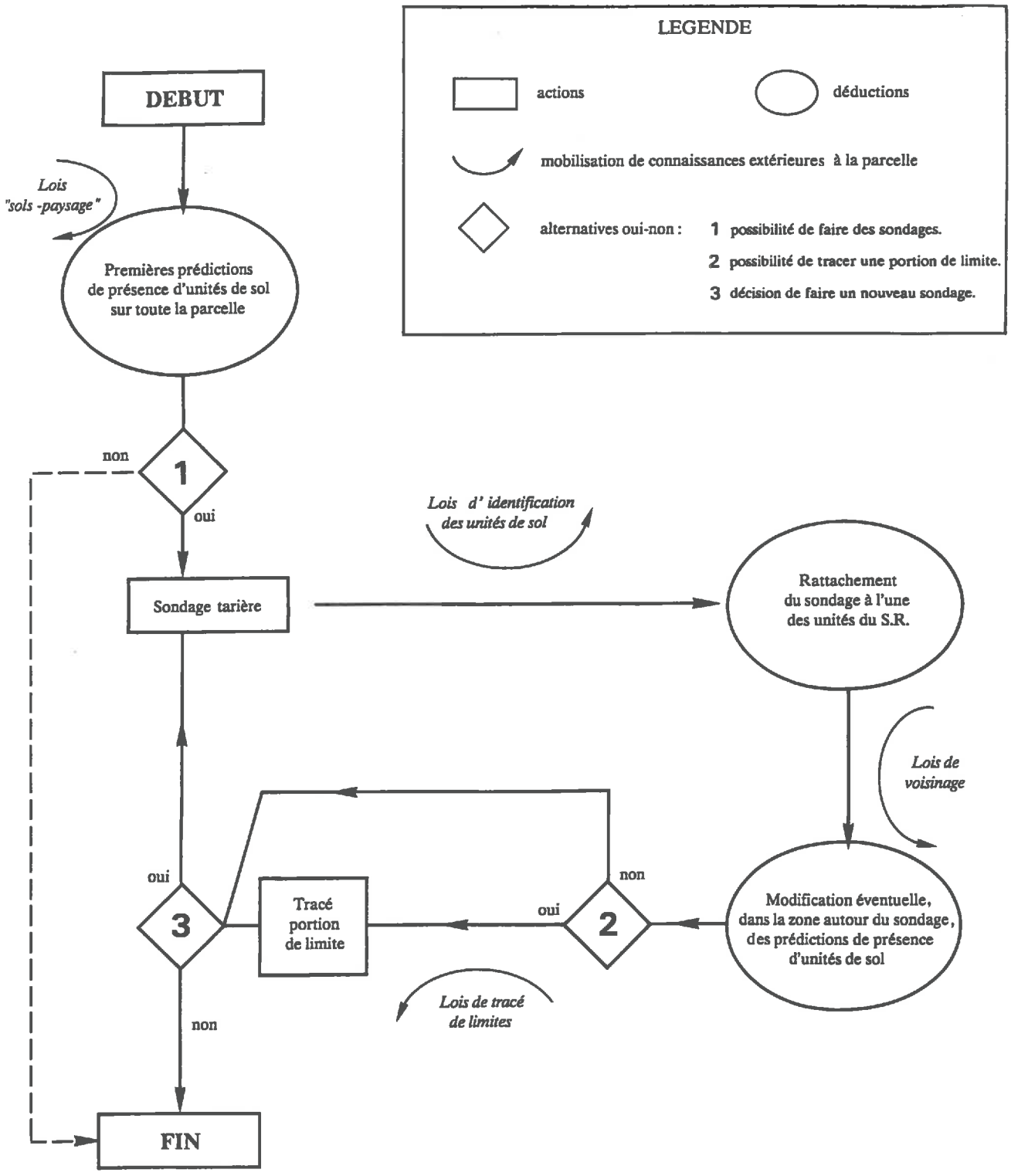


Figure 3: *schéma de fonctionnement représentant la démarche cartographique mise en oeuvre au cours du retour à la parcelle*

- à la suite de ce rattachement, le pédologue, ayant acquis une information nouvelle ("l'unité U_j occupe ce point"), mobilise sa connaissance sur les relations de voisinage entre unités de sol (cf 1.1.3); il applique donc des **lois de voisinage** qui lui permettent, pour la zone située autour du sondage, de recouper, de préciser ou éventuellement de modifier ses prédictions antérieures[1];
- dans le cas où les nouvelles prédictions produites conduisent à isoler avec une certitude que le pédologue juge suffisante deux zones appartenant à deux unités de sol distinctes, la limite entre ces deux unités peut être tracée au moyen de **lois de tracé de limites** dérivant de la connaissance de l'aspect de leur zone de contact (cf 1.1.4).

Ce raisonnement se reproduit jusqu'à ce que le pédologue décide d'arrêter les sondages, estimant en avoir réalisé suffisamment pour que les unités de sols soient délimitées avec une précision conforme à son objectif.

Pour être complet, il faut souligner que la localisation des sondages réalisés tout au long du processus est soumise également à des lois caractérisant la stratégie du prospecteur. Bien que ces lois ne soient, elles non plus, jamais explicitées, il est possible d'en évoquer quelques unes:

- ne pas dépasser un nombre de sondages fixé en début de prospection,
- observer une distance minimale entre les observations,
- sonder à l'endroit où l'incertitude est maximale,
- sonder pour permettre de tracer la portion de limite la plus longue possible...

Bien que ce soit plus difficile à exprimer de façon aussi structurée que précédemment, il est vraisemblable que la stratégie adoptée est également influencée par la connaissance acquise préalablement sur le secteur de référence. Ce point sera un peu développé au cours de la troisième partie de ce mémoire.

Parmi les règles possibles, une stratégie un peu particulière (flèche en pointillé sur le schéma figure 3) peut être intéressante à distinguer. Elle correspond au cas où le pédologue décide de ne pas faire de sondages. Dès lors, seule la première phase de la boucle est conservée ce qui correspond à une extrapolation pure des lois sols-paysage issues du secteur de référence. Ce cas particulier a fait l'objet récemment d'essais de formalisation dans la perspective d'une cartographie prédictive (SKIDMORE et Al, 1991). Bien qu'une telle stratégie ne corresponde plus à un retour à la parcelle au sens premier du terme, elle sera envisagé au cours de ce travail. Elle peut représenter en effet une voie intéressante pour valoriser les secteurs de référence en permettant la réalisation de cartes à moyenne ou petite échelle.

La boucle de la figure 3 est le résultat final d'un analyse des possibilités d'utilisation des connaissances issues du secteur de référence. Il n'est pas intervenu jusqu'à présent la moindre simplification de cette démarche suite à d'éventuelles considérations mathématiques ou informatiques. Ce schéma de fonctionnement représentant la démarche cartographique correspondra donc à un "idéal" que devra respecter le plus possible tout travail de formalisation et

---

1 Dans le cas général (où il ne s'agit pas du premier sondage) ces estimations antérieures correspondent aux résultats de l'application des règles de voisinages suite aux sondages effectués précédemment. Dans le cas particulier du premier sondage, il s'agit de préciser les premières estimations données par des lois fondées sur les relations sols-paysage.

d'informatisation. A ce titre, la boucle présentée pourrait servir de base à de futurs travaux utilisant d'autres concepts et outils que ceux présentés ci-après.

2. FORMALISATION MATHEMATIQUE DU RETOUR A LA PARCELLE

Le schéma de fonctionnement proposé dans la figure 3 (chapitre précédent) établit une distinction très nette entre deux entités, composant ensemble la démarche cartographique utilisée lors d'un retour à la parcelle:

- **le raisonnement cartographique** du pédologue matérialisé par l'enchaînement des différentes actions et déductions nécessaires pour produire une carte des sols; ce raisonnement peut être considéré comme indépendant du secteur de référence traité; sa formalisation mathématique sera donc applicable à l'ensemble des secteurs de référence, au delà de leurs spécificités propres; elle fera l'objet du paragraphe 2.1.;
- **les différentes lois mobilisées par ce raisonnement** (lois sols-paysage, d'identification des unités de sol, de voisinage et de tracé de limites); elles sont, au contraire, spécifiques à un secteur de référence donné et doivent donc être reformulées lorsque l'on change de petite région naturelle. Cependant, la formalisation du raisonnement cartographique détermine un formalisme et des modalités d'utilisation qu'il convient d'ores et déjà de préciser.

2.1. Formalisation mathématique du raisonnement cartographique utilisé lors du retour à la parcelle

L'action de cartographier une parcelle revient à déterminer, pour tout point de cette parcelle, l'unité de sol du secteur de référence à laquelle sera affecté ce point. Exprimé sous une forme mathématique, celà revient à ce qu'in fine soit vérifiée la proposition:

$$\forall x, \exists U_j \in U / u(x) = U_j \quad [1]$$

avec $U = \{U_1,...,U_j,...,U_v\}$ l'ensemble des v unités du secteur de référence

Dans le détail, la recherche d'une unité de sol pour chaque point correspond à un processus par étapes. On définira donc une série d'**états intermédiaires** décrivant, à une étape donnée, les possibilités d'affectation du point étudié à chaque unité de sol et les incertitudes correspondantes. Le premier de ces états est établi grâce à l'application des lois sols-paysage. Ensuite, de nouveaux états apparaissent chaque fois qu'une unité de sol peut être reconnue avec certitude sur un point particulier de la parcelle grâce à un sondage (observation directe du sol). Pour celà, conformément au schéma de fonctionnement défini précédemment, le sondage est rattaché à une unité grâce à une loi d'identification puis une nouvelle loi de voisinage est mobilisée pour produire un nouvel état intermédiaire.

Chaque état intermédiaire est caractérisé par la valeur indiquant l'étape qu'il décrit, cette valeur se modifiant à chaque nouveau sondage considéré (t_1 à t_f). Pour une étape t_k donnée, comprise entre t_1 et t_f, un état intermédiaire sera défini, pour chaque point, par la série des probabilités d'apparition, en ce point, de chaque unité du secteur de référence. Autrement dit, un état intermédiaire donné peut être formalisé par un vecteur $e(x,t_k)$ à v dimensions (le nombre

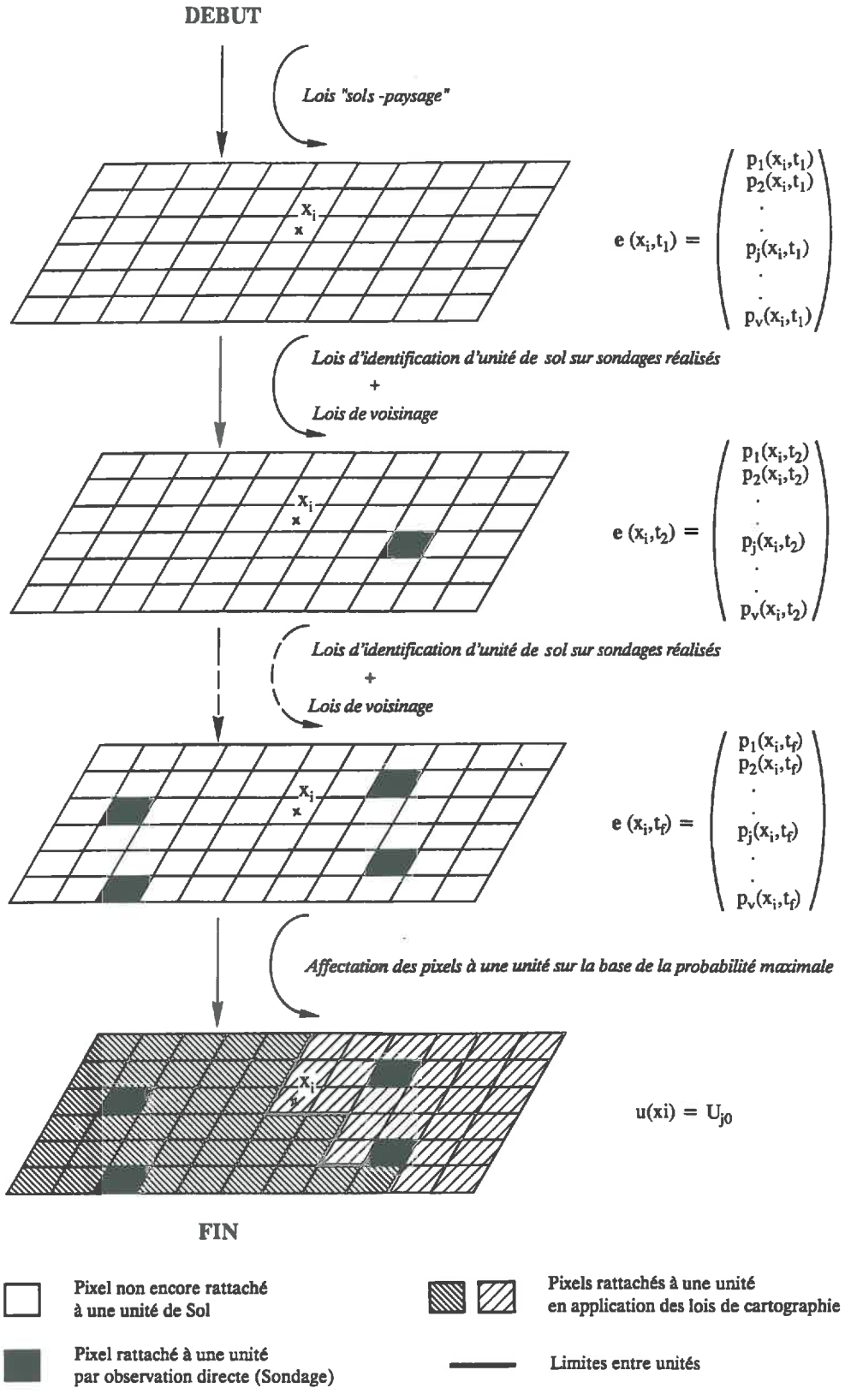


Figure 4: Formalisation mathématique de la démarche cartographique mise en oeuvre au cours du retour à la parcelle

d'unités de sol) dont les coordonnées représentent les probabilités d'affecter chacune des v unités au point x. Ceci est traduit par la définition suivante:

$$e(x,t_k) = (p_1(x,t_k),\ldots,p_j(x,t_k),\ldots,p_v(x,t_k)) \quad [2]$$

avec $p_j(x,t_k) = P[\, u(x) = U_j$ à l'étape $t_k\,]$

L'évolution normale de ces états intermédiaires serait que, sous l'influence des lois de voisinage successivement appliquées, les distributions des $p_j(x,t_k)$ atteignent à l'étape t_∞, après un nombre suffisant de sondages introduits, la situation idéale suivante: toutes les unités ont des probabilités d'apparition nulles, sauf une qui se trouve être celle prédite avec certitude (probabilité = 1). Ceci peut se résumer par la proposition:

$$\forall\, e(x,t_\infty),\ \exists\, U_L /\ p_L(x,t_\infty) = 1 \text{ et } p_j(x,t_\infty) = 0,\ \forall\, j \neq l\ (L \in \{1,\ldots,v\} \quad [3]$$

Cette situation idéale ne peut cependant jamais être atteinte compte tenu du nombre forcément fini de sondages effectués sur la parcelle. En conséquence, à l'étape t_f, consécutive à l'introduction du dernier sondage, une unité de sol doit être affectée en tout point. L'unité choisie correspondra à celle obtenant la probabilité maximale d'apparition en ce point, ce qui se traduit par la règle suivante:

$$\max_{(j)} p_j(x,t_f) = p_L(x,t_f) ===> u(x) = U_L,\ L \in \{1,\ldots,v\}\ [4]$$

Un risque d'erreur re(x) est associé à la fonction u(x) affectant les unités de sol aux points. Il correspond à la somme des probabilités d'apparition des unités autres que l'unité choisie, ce qui revient à:

$$re(x) = 1 - p_L(x,t_f) \quad [5]$$

Pour mettre en oeuvre la formulation mathématique présentée ci-dessus, il faut nécessairement disposer d'un ensemble fini de points. Pour un premier essai de représentation, on considèrera donc un ensemble de points régulièrement répartis selon un maillage carré. Chaque point représente ainsi une surface élémentaire également carrée, le **pixel**, dont il est le centre. La figure 4 représente la formulation mathématique décrite ci-dessus appliquée à un découpage de l'espace en pixels. Les unités de sol du secteur de référence sont désormais affectées aux pixels de la parcelle, chacun de ceux-ci étant représenté par son point central.

La formalisation mathématique proposée n'est qu'une représentation, parmi d'autres possibles, du schéma de fonctionnement avancé précédemment. A ce titre, il convient de revenir sur certains des choix qui ont motivé sa conception et dont les conséquences seront importantes pour la suite du travail:

1) Cette formulation ne permet pas, en l'état, d'introduire de connaissances spécifiques sur le tracé des limites. Il s'agit d'une simplification délibérée compte tenu du fait que, parallèlement, la formalisation d'éventuelles lois de tracé de limites n'a pas été envisagée dans ce travail. Dans le cas du découpage de l'espace en pixel, La règle [4] revient en fait à

considérer comme portion de limite tout côté commun à deux pixels affectés à des unités distinctes.

2) Le choix du cadre probabiliste pour représenter l'incertitude concernant les états intermédiaires n'était pas le seul possible. Il existe en effet, en particulier dans le domaine de l'Intelligence Artificielle, de nouveaux formalismes permettant de représenter une connaissance incertaine. Ces formalismes ont été récemment répertoriés et résumés par MARTIN CLOUAIRE (1992). Cependant, s'agissant d'un premier essai de formalisation, il a semblé plus confortable d'opter a priori pour les probabilités, quitte à remettre en cause ce choix à la lumière de cette première expérience.

3) L'affectation finale d'une unité de sol à un pixel n'exploite pas la totalité de l'information contenue dans le dernier état intermédiaire. Une réduction drastique de cette information est réalisée puisque on rend compte d'un histogramme de répartition au moyen de son maximum. Aussi, la carte des sols prédite doit être considérée comme une des représentations possibles du "pré-document" que constituerait l'ensemble des pixels renseignés par leurs derniers états intermédiaires. D'autres documents peuvent en effet être produits en utilisant les coordonnées des vecteurs représentant ces états: carte des risques d'erreur re(x) (qui sera exploitée au cours de ce travail) ou carte de probabilités d'occurrence $p_j(x,t_f)$ d'une unité de sol U_j donnée....

Par ailleurs, le choix de découper l'espace en pixels semble naturel puisqu'il a l'avantage d'être neutre et conforme à un mode de représentation de l'espace bien connu, le mode "raster". Ce choix impliquera cependant des contraintes antagonistes concernant d'une part les performances et, d'autre part, la précision finale de l'outil informatique automatisant le retour à la parcelle. En effet:

- le mode raster nécessite une capacité de calcul d'autant plus importante que le nombre de pixels choisi pour représenter une surface donnée est important; en conséquence, la taille des pixels ne peut être réduite à l'infini sans s'exposer à des limitations quant aux temps de réponse des systèmes informatiques;
- d'un autre côté, plus la taille des pixels est importante, plus l'imprécision des limites produites est grande puisqu'elles sont contraintes de suivre les contours de pixels; donc, une exigence de précision donné se paiera en nombre de pixels traités.

2.2. Formalisation des lois alimentant le raisonnement cartographique

La formalisation mathématique du raisonnement cartographique proposée ci-dessus fait intervenir des lois qui représentent l'expérience tirée du secteur de référence. Il convient maintenant de formaliser ces lois en conséquence.

Pour celà, un formalisme commun sera choisi. Il s'agit de **règles si (premisse) alors (conclusion)**. Un tel formalisme présente l'avantage de permettre l'utilisation directe de données qualitatives, très nombreuses dans le problème étudié (voir 2ème et 3ème partie). Au delà de ce formalisme commun, chaque type de loi apporte des contributions spécifiques au processus de retour à la parcelle. Leur formalisation sous forme de règles seront donc envisagées par ordre chronologique d'intervention, tel qu'il apparait dans la figure 4.

2.2.1. Les règles sols-paysage

Les règles sols-paysage, appliquées chacune à des ensembles de points x, doivent permettre d'établir le premier état intermédiaire $e(x,t_1)$ en tout point. En conséquence:

- la premisse d'une règle sols-paysage doit permettre de sélectionner les points sur lesquels la règle s'applique en utilisant les critères disponibles;
- la conclusion doit être nécessairement sous forme compatible avec celle de l'état intermédiaire qu'elle est censée établir soit une série de probabilités intéressant chaque unité de sol.

Ainsi, par exemple, une règle sols-paysage se présentera sous la forme suivante:

si	altitude en x $\leq$ 17 m	
et	rive du fleuve en x = gauche	
alors	$\pi_1(x),\ldots\pi_j(x),\ldots\pi_v(x)$	[6]

L'extraction et l'utilisation des règles sols-paysage seront abordées au cours de la deuxième partie.

2.2.2. Les règles d'identification d'unités de sols

Ces règles interviennent à chaque nouveau sondage pour rattacher celui-ci à une unité du secteur de référence. A la différence de précédemment, une conclusion de règle d'identification doit désigner une unité de sol et une seule, la prémisse s'appuyant sur des critères morphologiques de sol accessibles à l'observation à la tarière (texture, couleur,...).

On a choisi de ne pas aborder la formalisation et l'utilisation de ces règles. En conséquence, l'unité de sol sera, dans la suite du travail, supposée connue sur chaque sondage. Il faut noter que cette formalisation a fait l'objet de nombreux travaux:

- dans le cadre du thème abordé par ce travail de recherche (BAILLE et al, 1988; LAGACHERIE et LEDREUX, 1991);
- dans des tentatives visant à rattacher automatiquement un individu sol à une unité taxonomique de la classification américaine (KOLLIAS,1988; FISHER et al, 1989)

Auparavant, un autre formalisme que des règles avait été utilisé pour rattacher des sondages à des unités de sol d'une carte pédologique (PAVAT, 1986; SIMMONNEAUX, 1987). Il s'agissait d'une approche numérique utilisant une mesure de distance entre sondage et profil type d'unité établie suivant les travaux de GIRARD (1983) et KING (1985).

2.2.3. Les règles de voisinage

Les règles de voisinage sont sollicitées à la suite de l'identification d'une unité de sol sur un sondage. Elles doivent permettre, à chaque étape, de recalculer, pour les ensembles de points qu'elles concernent chacune, les probabilités d'apparition des unités de sol. Elles établissent de cette

manière les états intermédiaires des points pour les étapes t_2 à t_f. Le problème que pose l'emploi de ces règles est double:

- sous quel formalisme doivent-elle être fournies?
- comment modifient-elles un état intermédiaire préalablement établi?

2.2.3.1. Définition d'un formalisme pour les règles de voisinage

Comme pour les règles sols-paysage, une prémisse de règle de voisinage doit sélectionner l'ensemble de points que cette règle est susceptible de concerner. Cette sélection tiendra compte, d'une part, de l'unité reconnue sur le sondage traité et, d'autre part, de critères caractérisant la position relative du sondage traité vis à vis des autres points. Ces critères seront recherchés au cours de la troisième partie de ce mémoire.

Pour pouvoir être compatible avec les états intermédiaires qu'elle établit, la conclusion d'une règle de voisinage sera également constituée d'une série de probabilités $\pi_j(x)$. En définitive, une règle de voisinage se présentera sous la forme schématisée suivante:

si	l'unité reconnue sur un sondage est U_j	
si	position relative vis à vis du sondage = $(c_1, \ldots c_w)$	
alors	$\pi_1(x),\ldots,\pi_j(x),\ldots\pi_v(x)$	[7]

avec: $(c_1,\ldots c_w)$, critères traduisant la position relative du sondage vis à vis des autres points.

L'algorithme permettant d'extraire des lois de voisinage à partir du secteur de référence selon ce formalisme sera décrit au cours de la troisième partie de ce mémoire.

2.2.3.2. modalités d'utilisation d'une règle de voisinage

A chaque étape, un point donné est touché par une règle de voisinage. Après plusieurs étapes, un point aura donc été concerné par plusieurs règles. Les probabilités d'apparition des unités de sol établies en ce point devront tenir compte des probabilités données dans les conclusions de toutes ces règles. Ainsi, établir l'état intermédiaire $e(x,t_k))$ suite à la dernière règle déclenchée Rg_k reviendra en fait à recalculer chaque probabilité d'apparition d'unité de sol en un point x de la façon suivante:

$$p_j(x,t_k) = f(\pi_{j1}(x),\ldots,\pi_{jk}(x)) \qquad [8]$$

avec $\pi_{j1}(x),\ldots,\pi_{jk}(x)$: probabilités d'apparition en x de l'unité U_j donnée par les règles $Rg_1,\ldots,$ Rg_k déclenchées aux étapes $t_1,\ldots,t_k$.

De cette façon, les probabilités d'apparition des unités de sol constituant les états intermédiaires seront bien réactualisées à chaque étape puisque le nouveau calcul tiendra compte chaque fois d'un terme $\pi_{jk}(x)$ de plus.

La fonction f combinant les conclusions de chaque règle sera définie et justifiée au cours de la troisième partie du mémoire. Dans le détail, deux fonctions distinctes seront à considérer:

- l'une combinant entre elles les probabilités données par les seules règles de voisinage;
- l'autre permettant de combiner le précédent résultat avec la règle sols-paysage (étape t_1).

3. AUTOMATISATION DU RETOUR A LA PARCELLE: LES OUTILS INFORMATIQUES UTILISES

Le recours à l'outil informatique pour représenter une réalité touchant la cartographie pédologique ne constitue pas une première. Ainsi, en France, BERTRAND et Al (1979) réalisent l'informatisation des descriptions ponctuelles de sol. GIRARD et KING, déjà cités, proposent dans leurs thèses des constructions informatiques simulant la démarche cartographique. LEGROS (1982) utilise également l'informatique pour rendre compte de mécanismes pédogénétiques. Enfin, plus récemment, apparaissent des bases de données permettant de décrire les unités de sol des études pédologiques (ASTER, 1990; GAULTIER, 1991). Ce nouveau recours à l'informatique s'inscrit donc dans une (déjà) longue histoire.

L'outil informatique qu'il faut construire pour automatiser le retour à la parcelle sera réalisé sur la base de la formalisation proposée au chapitre précédent. Celle-ci présente deux particularités.

La première est qu'elle sépare raisonnement et lois de cartographie mobilisées par ce raisonnement. Au niveau informatique, ceci correspond à la définition des **Systèmes à Base de Connaissance**[2] (HATON et Al, 1991), famille d'outils informatiques utilisés dans le domaine de l'Intelligence Artificielle.

La deuxième tient au caractère géographique des données traitées. Les objets manipulés sont des pixels localisés dans l'espace par des coordonnées x,y. Il faut pouvoir:

- les renseigner par des variables utilisées dans les prémisses des règles (altitude,...); ceci revient à déterminer la position relative de pixels soit entre eux (distance,...) soit vis à vis d'autres objets géographiques contenus dans des cartes intéressant la même parcelle (inclusion géographique, distance,...);
- les trier sur la base de ces variables (par exemple pour connaître les points touchés par une règle donnée);
- les représenter afin de produire la carte des résultats de prédictions d'unités de sol.

Ces opérations sont en fait réalisées par des logiciels spécifiques, les **Systèmes d'Information Géographique (SIG)**.

Le logiciel idéal permettant d'automatiser le retour à la parcelle correspondrait donc à un outil satisfaisant à la fois les deux types de besoins exprimés ci-dessus. De nombreux articles évoquent, pour d'autres besoins, la nécessité de disposer d'un tel logiciel (BUISSON, 1989; WEBSTER, 1990; BURROUGH, 1992). Cependant, au moment où débutait ce travail de recherche, aucun logiciel de ce type n'était disponible.

---

2 Selon HATON et Al. le concept de "système à base de connaissance", plus général, est destiné à remplacer le terme de "système expert" dont il semble impossible maintenant de donner une définition à la fois précise et unique, sous l'influence d'une médiatisation exagérée.

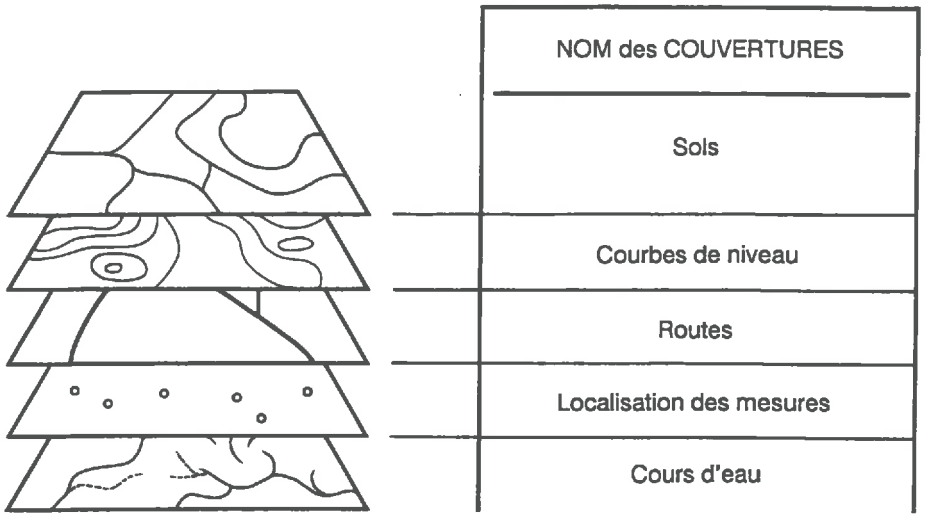


Figure 5: un exemple de décomposition des données géographiques en couvertures ARC/INFO

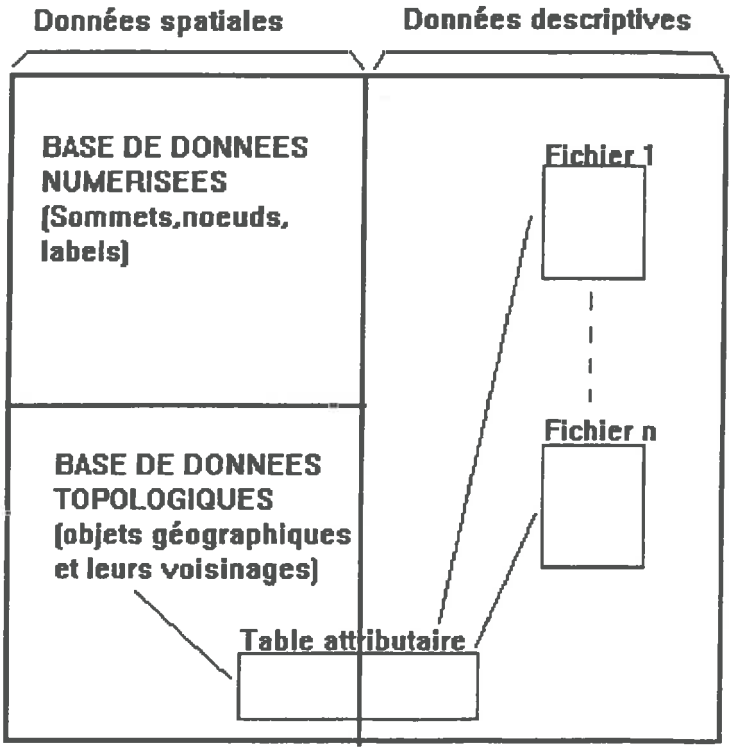


Figure 6: organisation des données au sein d'une couverture ARC/INFO (modifiée d'après (BUISSON 1989)

En conséquence, deux stratégies parallèles et complémentaires ont été poursuivies:

- recherche d'une automatisation la plus complète possible sur la base d'une association entre un Système d'Information Géographique et un Système à Base de Connaissance; il s'agit d'un travail de recherche informatique en tant que tel (LAGACHERIE et LEDREUX, 1991; LEDREUX, 1992); le résultat de ces recherches[3] n'a pas pu être exploité directement par ce travail, l'outil ayant été produit trop tardivement et s'avérant trop lent pour permettre de réaliser les différents essais inclus dans ce mémoire;
- automatisation partielle des prédictions d'unités de sol impliquant les lois de distribution des unités de sol (lois sols-paysage et lois de voisinage), plus particulièrement étudiées dans ce mémoire; il s'agit de programmes écrits en FORTRAN, exploitant le fichier des points extrait d'un SIG et fournissant en retour au SIG les résultats des prédictions, qui seront présentés sous forme de cartes. Ces programmes restent "expérimentaux" et n'ont pas l'ambition d'être mis, en l'état, entre les mains de futurs utilisateurs.

Le Système d'Information Géographique utilisé tout au long de ce travail sera ARC/INFO (ESRI, 1989). Il constitue certainement le plus célèbre et le plus répandu des Systèmes d'information Géographique. Produit par Environmental Systems Research Institute (ESRI), il est commercialisé depuis environ 15 ans au USA et seulement depuis 1988 en France. Les choix conceptuels permettant de satisfaire les besoins requis par la manipulation de données géographiques peuvent se résumer en trois principes fondamentaux.

1) Un codage de l'information géographique en mode vecteur.
2) La décomposition de l'ensemble des données géographiques en couvertures (Figure 5). Il s'agit d'un ensemble de fichiers relatifs à un seul thème (sol, géologie,...) et à un seul type d'objet géographique (point,ligne ou polygone); ainsi, pour être intégré dans ARC/INFO, une carte rassemblant des informations diverses (comme par exemple la carte topographique) sera décomposée au préalable en une série de couches monothématiques qui constitueront autant de "couvertures" au sein du SIG;
3) Une organisation des données cloisonnée (Figure 6) en trois entités (BUISSON, 1989):
   - la base de donnée descriptive (INFO) fonctionne sur un modèle d'organisation de données précurseur du modèle relationnel adopté par les SGBD du commerce (ORACLE,PARADOX,DBaseII);
   - la base de donnée digitalisées permet le stockage des données brutes intégrées au système au cours de la phase de digitalisation;
   - la base de données topologiques représente les fichiers décrivant la géométrie et les relations de voisinage des objets géographiques reconstitués à partir des données brutes; elle représente la source d'information permettant l'ensemble des analyses spatiales annoncées par le concepteur; elle est reliée à la base de données descriptives (et à l'utilisateur) par un fichier INFO appelé "table attributaire" qui identifie chaque objet géographique par un index commun aux fichiers des deux bases.

Il convient de remarquer que le choix d'ARC/INFO pose un problème: l'information géographique y est en effet codée en mode vecteur alors que le mode "raster" a été privilégié pour

---

3 SAPRISTI a permis de montrer qu'une association système expert-SIG est réalisable et les jalons d'une prochaine recherche informatique, associant désormais des chercheurs et industriels du domaine, ont été posés (LAGACHERIE et LEDREUX 1992)

formaliser le retour à la parcelle. En fait, ARC/INFO possède des fonctionnalités de passage entre les deux modes qui permettent à l'utilisateur de travailler en mode raster. De plus, aucun SIG en mode raster ne possédait, à l'époque du choix, les fonctionnalités requises pour permettre la formalisation de la connaissance issue du secteur de référence telle qu'elle sera abordée dans les parties suivantes.

Par ailleurs, outre son utilisation directe dans l'automatisation du retour à la parcelle, les fonctions d'ARC/INFO seront largement sollicitées pour l'extraction des règles sols-paysage (2ème partie) et de voisinage (3ème partie).

Ce deuxième chapitre a permis d'aborder la démarche cartographique utilisée dans l'opération de "retour à la parcelle" (suite à l'étude préalable de secteur de référence). La particularité de cette opération tient au fait que le pédologue s'appuie sur une expérience préalable de cartographie qu'il considère suffisamment sûre pour ne plus la remettre en cause. Cette stabilité laisse espérer une formalisation plus aisée. La méthode adoptée comporte trois étapes:

La première constitue une analyse de la teneur de l'expérience acquise à la suite de l'étude de secteur de référence. Elle permet d'élaborer un premier schéma de fonctionnement représentant le retour à la parcelle (figure 3). Selon ce schéma l'expérience acquise à la suite de l'étude du secteur de référence se présente sous forme de lois (lois sols-paysage, lois d'identification des unités de sol, lois de voisinage et lois de tracé de limites) qui sont utilisées tour à tour par le raisonnement cartographique.

La deuxième propose, sur la base du travail précédent, une formalisation mathématique du retour à la parcelle. Celle-ci représente les différentes étapes du raisonnement cartographique sous forme d'une série de probabilités d'apparition d'unités de sol en chaque point de l'espace. Les valeurs de ces probabilités évoluent sous l'action de différentes règles appliquées en chaque point. Ces règles ("si (prémisse) alors (conclusion)") constituent le formalisme choisi pour représenter les lois de cartographie. En fin de processus, l'unité prédite, pour un point donné, est celle obtenant la plus forte probabilité d'apparition. Pour mettre en oeuvre cette formalisation, un découpage de l'espace en pixels est choisi.

Enfin, La troisième envisage la possibilité d'automatiser, au moins partiellement, le retour à la parcelle grâce au recours à l'informatique. La solution choisie dans ce travail est constituée par une série de programmes FORTRAN associés au Système d'Information Géographique ARC/INFO.

Au terme de ce chapitre, il est possible de revenir sur les objectifs du travail pour les préciser. Des choix de formalisation du retour à la parcelle ont été effectués. Par ailleurs, les lois d'identification des unités de sol et les lois de tracé de limite ont été exclues du champ d'investigation de ce travail. En conséquence, le problème est maintenant de formaliser la connaissance acquise sur le secteur de référence en matière de lois de distribution des sols. Le formalisme général est défini: il s'agit de règles du type "si (prémisse) alors $(\pi_1(x),...,\pi_j(x),...\pi_w(x))$**", chaque** $\pi_j(x)$ **représentant une probabilité d'apparition d'une unité de sol du secteur de référence. Ces règles peuvent être obtenues selon deux voies distinctes:**

-demander à l'auteur du secteur de référence d'écrire lui même, selon le formalisme exigé, les règles traduisant les lois de distribution des sols;

- extraire ces règles, par des algorithmes adaptés, à partir de la carte des sols du secteur de référence dans laquelle elles apparaissent en filigrane.

C'est cette dernière voie qui sera privilégiée et qui fera l'objet des 2ème et 3ème parties. Elle présente en effet des avantages décisifs par rapport à la première:

- Elle allège la contribution personnelle du pédologue auteur du secteur de référence;
- Elle évite les problèmes de subjectivité inhérents à la formulation d'un avis d'expert et assure une exhaustivité plus grande dans l'exploitation de l'information disponible (évite les oublis conscients ou inconscients);
- Elle permet surtout une transposition plus facile d'une petite région naturelle à une nouvelle dans la mesure où les outils d'acquisition des règles ne dépendent pas du milieu d'étude abordé. L'ensemble du travail gagne donc de l'indépendance vis à vis du terrain d'étude.

CHAPITRE 3

PRESENTATION DU MILIEU EXPERIMENTAL

Afin d'évaluer la pertinence des choix réalisés en matière de formalisation du retour à la parcelle, il est nécessaire d'expérimenter l'outil informatique résultant sur un terrain d'étude concret. Le choix s'est porté sur la petite région naturelle **Moyenne Vallée de l'Hérault** (34). Cette région constitue depuis 1988 le milieu expérimental commun à l'équipe "Spatialisation" du Laboratoire INRA science du sol de Montpellier. A ce titre, elle est et sera dans l'avenir le théatre de nombreux travaux scientifiques de cette unité. (SIMON, 1990; LEENHARDT, 1991; MOLLET, 1991; INRA, 1992; LOVELAND, 1992;...). La Moyenne Vallée de l'Hérault est également couverte par deux études de sol à moyenne et petite échelle, permettant d'avoir ainsi une perception d'ensemble de sa couverture pédologique:

- carte des sols au 1/100.000 feuille de Lodève (BONFILS, 1992);
- carte des unités pédo-paysagères au 1/250.000 du Languedoc-Roussillon (BORNAND et Al, 1992).

Enfin, pour les besoins de travaux antérieurs (LEENHARDT, 1991), un secteur de référence a été cartographié (LEFAY et Al, 1990). Ce secteur constitue également un secteur pilote dans le cadre de l'opération nationale "Inventaire, Gestion et Conservation des sols". Le but poursuivi est de transposer la méthode des secteurs de référence, initialement appliquée au drainage des sols, à d'autres thèmes importants pour l'agriculture méditerranéenne: reconversion du vignoble et diversification des cultures.

La présentation du milieu expérimental "Moyenne Vallée de l'Hérault" comprendra trois sous-chapitres:

-une description de la petite région naturelle, en particulier de ses sols, sur la base des cartes existantes,

-une présentation du secteur de référence "Moyenne Vallée de l'Hérault" à partir duquel seront extraites les lois de distribution des unités de sol,

-une présentation des trois secteurs de validation sur lesquels l'outil informatique simulant le retour à la parcelle sera testé.

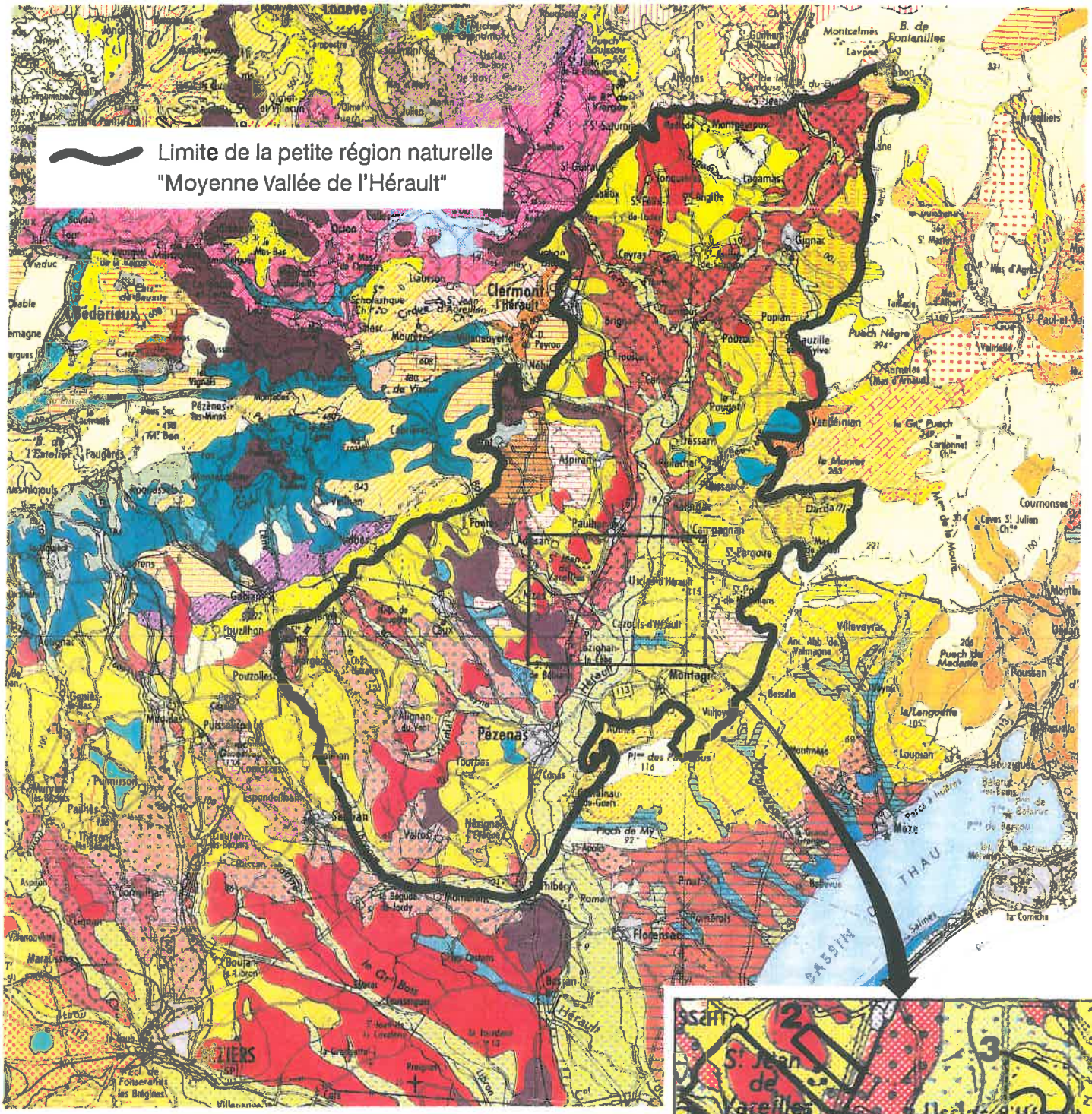
1. LA PETITE REGION NATURELLE "MOYENNE VALLEE DE L'HERAULT"

Dans le langage courant, la "Moyenne Vallée de l'Hérault" correspond au territoire traversé par le fleuve Hérault entre le "Pont du Diable" (situé entre St Guilhem le désert et Aniane) et St Thibery. Avant d'envisager une description de cette petite région naturelle, une définition et une délimitation plus précise s'imposent. Celles-ci peuvent se fonder sur la carte des pédopaysages au 1/250000, destinée en particulier à cet usage (planche 1).

Selon cette carte, la petite région naturelle Moyenne Vallée de l'Hérault peut être définie par l'association de 4 unités pédopaysagères majeures:

- l'unité 173A (sols de la plaine alluviale récente de l'Hérault);

Planche 1: Définition du cadre expérimental *



1 - Secteur de référence

2 - Secteur de validation de la Roubiaire

3 - Secteur de validation de Montmau

4 - Secteur de validation de Lézignan la Cèbe

* *Support : carte des pédopaysages du Languedoc-Roussillon au 1/250 000è, (BORNAND et AL - 1992).*

- Les unités 309A et B (sols des niveaux de terrasses d'alluvions anciennes Wurm et Riss de l'Hérault);
- l'unité 309V (sols des lambeaux de la haute terrasse villafranchienne occupant les sommets de buttes en rive droite);
- l'unité 552T (sols développés sur molasse miocène, substratum principal de la vallée).

A côté de ces 4 unités, d'autres, plus locales, apparaissent dans cette petite région naturelle (cf 1.3.).

A partir de cette définition du contenu de la "Moyenne Vallée de l'Hérault", une délimitation peut être proposée. Celle-ci s'avère plus ou moins facile à mettre en oeuvre, les limites avec les autres petites régions naturelles étant plus ou moins nettes selon les cas. Ainsi, les limites sont bien marquées, par des discontinuités de paysage et d'occupation du sol:

- à l'ouest (avec l'extrémité du Massif Central),
- au Nord (avec le Causse du Larzac),
- à l'Est (avec le Causse d'Aumelas et l'Arc de Villeveyrac".

Elles sont moins nettes par contre au Sud où l'on passe de façon plus graduelle (cf territoire entre Peyne et Thongue) à la Basse Plaine de l'Hérault caractérisée par l'emprise importante des matériaux villafranchiens, l'apparition des placages pliocènes de Servian (unité 372P), la disparition de la terrasse moyenne et l'élargissement notable de la plaine d'alluvions récentes.

Ces difficultés de délimitation rappellent celles rencontrées à l'occasion de la cartographie des sols à grande échelle. Ceci souligne bien la convergence de nature entre les concepts d'unité de sol et de petite région naturelle, au delà des niveaux de perception différents.

Le territoire ainsi délimité correspond à une superficie de 40 000 ha ce qui placerait la Moyenne Vallée de l'Hérault un peu au dessus de la superficie moyenne des petites régions naturelles françaises caractérisées par les secteurs de référence ONIC-MINAGRI (30 000 ha selon JC FAVROT, 1987).

La présentation de la Moyenne Vallée de l'Hérault comprendra trois volets:

- le premier présentera les traits généraux en matière de climat, de relief et d'occupation du sol,
- le second s'intéressera à l'histoire géologique et aux roches-mères des sols,
- le troisième évoquera la couverture pédologique telle que les cartes à petite et moyenne échelles permettent de l'appréhender.

1.1. Généralités

La Moyenne Vallée de l'Hérault est une composante de la plaine viticole de l'Hérault. A ce titre son agriculture est dominée par la monoculture de la vigne imprimant au paysage sa spécificité. Récemment, des essais de reconversion vers les grandes cultures (blé dur, semences de maïs,etc...) ont été entrepris. Ils restent cependant trop limités pour avoir une influence marquée sur le paysage, à l'exception notable de la plaine alluviale où les potentialités des sols favorisent cette reconversion. La monotonie des vignes est uniquement tempérée par de petits bosquets de pin d'Alep et par des garrigues couronnant les "puechs", petits sommets locaux liés à la présence de matériaux résistant à l'érosion et peu propices au défoncement des sols préalable à la mise en culture.

Le relief d'ensemble est mou, avec cependant des contrastes révélateurs des matériaux géologiques sous-jacents: pentes quasi nulles sur les épandages d'alluvions étagées en terrasses successives, long glacis entrecoupés de talus sur matériaux sableux marins du Tertiaire, pentes fortes, voire petites falaises sommitales au niveau des puechs. L'ensemble est en position dépressionnaire par rapport aux petites régions naturelles voisines, à l'exception de la Basse Plaine de l'Hérault. Les altitudes s'étagent entre 10 et 120 mètres avec une pente générale Nord-Sud d'environ un pour mille.

Le climat régional est de type méditerrannéen humide. Les précipitations s'étagent selon un gradient globalement Sud-Nord de 600 à 850 mm. Les températures moyennes annuelles sont de 14°. Les moyennes mensuelles de la station de Gignac (tableau 1) permettent de préciser ces données générales.

	Jan	Fev	Mar	Avr	Mai	Juin	Juil	Aout	Sep	Oct	Nov	Dec	An
P	68	63	76	54	61	42	27	44	71	116	64	70	758
T	6.4	7.4	9.2	12.0	15.4	19.7	22.5	22.1	18.9	14.4	9.7	7.3	13.8

***Tableau 1:** Station de Gignac; moyennes mensuelles des températures (T) et précipitations (P); période 1891-1979 (ASENCIO, 1984)*

Les hivers sont relativement doux (nombre moyen de jours de gelée = 43 entre octobre et avril), le mois le plus froid étant Janvier (6,4°). Les étés sont chauds (juillet: 22,5°). Les précipitations sont irrégulières dans l'année, avec une forte pointe en octobre (116 mm) et une pluviométrie estivale faible (112 mm seulement).

L'hydrographie régionale est marquée par la présence du fleuve côtier Hérault qui prend sa source au mont Aigoual et se jette en Méditerranée au niveau d'Agde. Ce fleuve est caractérisé par des crues fréquentes qui affectent sa plaine alluviale, inondant les cultures. De nombreux affluents de l'Hérault parcourent et compartimentent la petite région naturelle étudiée. Les véritables petites rivières prenant leur source dans les contreforts du Massif Central en rive droite (Dourbie, Lergue, Boyne, Peyne, Thongue), contrastent avec les ruisseaux intermittents de la rive gauche (Rouvière, Rieutord,...).

Cette petite région naturelle, comme l'ensemble de la plaine du Languedoc Roussillon, est confrontée à une grave crise de surproduction viticole du fait de la baisse de consommation de vin de table et de la concurrence des autres pays de la CEE. Deux voies sont explorées pour résoudre cette crise:

- amélioration de la qualité des vins produits par restructuration du vignoble (introduction de cépages améliorateurs);
- diversification vers de nouvelles activités agricoles (grandes cultures, production de semences, arboriculture, maraîchage de plein champ,...).

Quelle que soit l'alternative étudiée, la connaissance des potentialités des sols régionaux et, en particulier, de leur fonctionnement hydrique s'impose comme un préalable obligé. C'est dans cette perspective qu'a été lancée, dans le cadre de l'opération Inventaire, Gestion et Connaissance des sols (IGCS) du Ministère de l'Agriculture, l'idée d'un secteur de référence "agriculture en transformation".

1.2. Géologie de la Moyenne Vallée de l'Hérault: Dynamique de mise en place des roches mères des sols (BRGM, 1981)

La géologie représente, en particulier en zone méditerranéenne, un déterminant essentiel permettant d'expliquer la distribution des sols. La compréhension de la dynamique de mise en place des matériaux géologiques fait donc partie intégrante du travail de cartographie. Ceci justifie une description détaillée.

Les évènements géologiques ayant conditionné la nature des affleurements présents dans la Moyenne Vallée de l'Hérault débutent au milieu du Tertiaire, suite à l'effondrement de l'axe montagneux Pyrénéo-Provençal reliant les Pyrénées actuelles et les massifs provençaux. Suite à cet effondrement, une transgression marine affecte la zone étudiée, laissant par contre exondées les zones correspondant aux petites régions naturelles voisines (à l'exception de la Basse Plaine de l'Hérault). Au cours de cette transgression, datée de l'Helvétien (environ - 10 millions d'années), se déposent les sédiments marins qui formeront le substratum de base de la petite région naturelle: d'abord marnes bleues, recouvertes, à la faveur d'une réduction de profondeur du golfe, par des dépôts plus grossiers sablo-gréseux (unité M2a de la carte géologique au 1/50000 BRGM). Ceux-ci sont systématiquement entrecoupés par des bancs de calcaires coquillers dont les plus épais correspondent à la présence des puechs. A la périphérie du golfe Helvétien en cours de réduction, s'installe une zone lagunaire responsable de la présence actuelle de calcaires durs et de "molasses à dragées de quartz" (unité m2b).Cet épisode transgressif marque la fin des influences marines sur la région.

Au Pliocène, d'importants dépôts continentaux affectent l'ensemble de la région. Cependant, alors qu'ils affleurent sur de grandes surfaces en Basse Plaine de l'Hérault, ces dépôts seront remobilisés en quasi-totalité au niveau de la Moyenne Vallée de l'Hérault.

Le Quaternaire (- 1 million d'années) est dominé par l'influence fluviatile. L'Hérault et ses affluents creusent leur lit dans le substratum sablo-gréseux, charriant des alluvions plus ou moins caillouteuses. A la faveur des diverses glaciations, ces alluvions s'étagent selon différents niveaux de terrasses, regroupés en 3 principaux:

- terrasses villafrachiennes (Fv, 60-90 m);
- terrasses du Pléistocène moyen, subdivisées en 2 sous-niveaux Fya (10-20m) et (Fyb (8,10m), ce dernier n'étant pas toujours discernable tout au long de la vallée;
- Alluvions récentes (Fz).

Plus localement, apparaissent des lambeaux de terrasses du Pléistocène inférieur (Fx), répartis sur plusieurs niveaux, entre Fv et Fy (Fx dans la région d'Adissan notamment). D'autres niveaux sont identifiés entre Fy et Fz (terrasses de l'Hérault et de la Lergue). Ces différents niveaux erratiques, pas toujours représentables sur la carte au 1/50.000, introduisent un facteur de complexité au sein de la couverture pédologique.

Au cours de ce même Quaternaire, un volcanisme actif provoque les épandages basaltiques traversant la région. Son influence est particulièrement visible dans le paysage (plateau de Nizas, dépression de Péret,...).

Enfin, la période actuelle est marquée par des phénomènes de colluvionnement actifs dont le résultat est de mélanger, à leurs frontières, les différents matériaux déposés: recouvrements caillouteux villafranchiens sur molasse sablo-gréseuse de l'Helvétien, colluvions de molasse sur terrasses Wurmiennes (rive droite) ou alluvions récentes (rive gauche).

Ainsi, la moyenne vallée de l'Hérault possède une histoire géologique spécifique et différente de celle des régions voisines. Au moins dans ce cas particulier, il est donc possible d'avancer une définition "déterministe" de la petite région naturelle: c'est le théatre géographique d'une succession spécifique d'évènements géologiques responsable de la disposition actuelle des roches mères les unes par rapport aux autres.

1.3. Les principaux types de sols de la Moyenne Vallée de l'Hérault

Consécutivement à la mise en place des matériaux géologiques, s'est différenciée une couverture pédologique particulière dont la description fait l'objet de ce paragraphe. Seuls les sols des 4 unités pédopaysagères majeures (cf début du paragraphe 1.) seront présentés en détail, en utilisant comme source d'information la carte au 1/100.000, feuille de Lodève (BONFILS, 1992). Les autres ne seront que brièvement évoqués en fin de chapitre. Le bloc diagramme (figure 7), construit par P.BONFILS, permet de comprendre la distribution des unités de sol décrites brièvement dans ce chapitre.

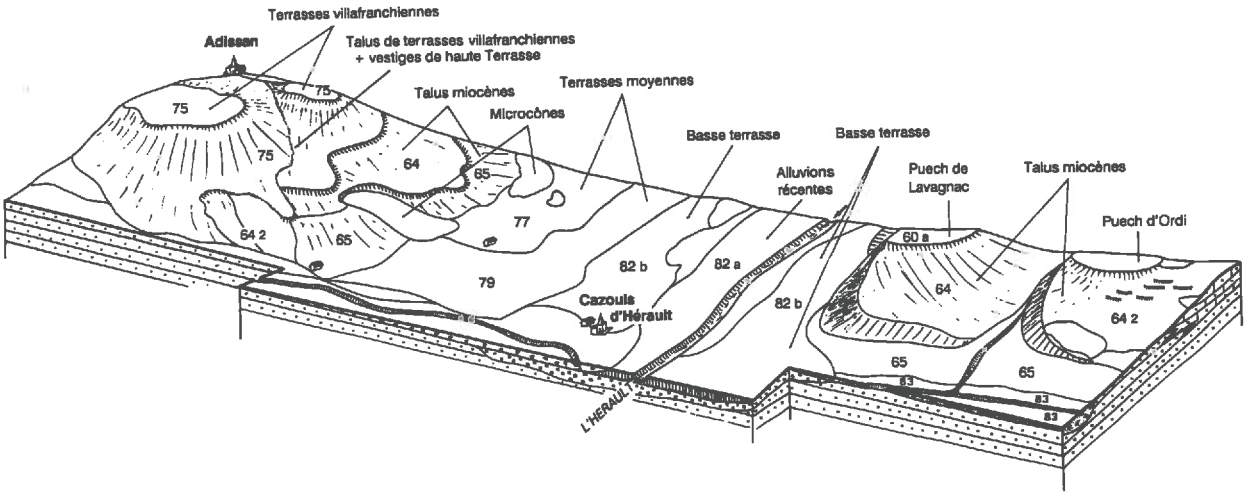


Figure 7: Les principaux sols de la Moyenne Vallée de l'Hérault et leur position dans le paysage (d'après BONFILS, 1992)

1.3.1. Les sols de la plaine alluviale (unités 82a,82b)

Ce sont des sols alluviaux peu évolués (Fluvisols), développés sur les alluvions récentes de l'Hérault. Ils sont profonds, calcaires, généralement brun-rouge (75YR) et peu chargés en éléments grossiers. La nappe phréatique de l'Hérault les affecte en profondeur, contribuant ainsi à l'alimentation hydrique des cultures. Ce caractère, qui les distingue des autres types de sol, explique le plus grand succès des tentatives de diversification observées dans ces sols (MOLLET, 1991).

Selon une logique de sédimentation fluviatile, leur texture varie progressivement avec la distance au fleuve: sableuse en bordure de l''Hérault (bourrelet alluvial), elle devient argilo-limoneuse au contact de la basse terrasse. La présence de bancs caillouteux erratiques, de niveaux intermédiaires de terrasse enfouis et d'alluvions d'affluents introduisent un facteur de complication vis à vis de l'organisation de la couverture pédologique de cette unité pédopaysagère.

1.3.2 Les sols des terrasses du Pléistocène moyen (unités 77,78,79)

La couverture pédologique est beaucoup moins homogène que dans l'unité paysagère précédente et peu de caractères généraux peuvent être dégagés. Sa différenciation est soumise à deux facteurs principaux, quoique d'importance inégale.

1) l'intensité du colluvionnement issus des niveaux miocènes voisins. En l'absence de ce colluvionnement (unités 78,79), les sols se développent directement sur les alluvions anciennes. Ils sont donc caillouteux dès la surface (50 à 80%), de texture moyenne sableuse et toujours calciques. Ils sont caractérisés par de faibles réserves hydriques.

Les zones colluvionnées (unité 77) se différencient par la présence, en surface, d'horizons limoneux, limono-argileux ou argilo limoneux, calcaires et pauvres en éléments grossiers. Ces horizons, d'épaisseur variable (50 à 100cm), surmontent les niveaux caillouteux d'alluvions anciennes. Les réserves hydriques sont sensiblement plus élevées que dans les sols précédents.

Globalement, il existe une logique de distribution de ces deux pôles (figure 7), la présence de colluvions étant localisée au contact des niveaux du Miocène. Dans le détail, c'est beaucoup plus complexe et la profondeur des horizons limoneux subit des variations importantes et souvent aléatoires.

2) L'âge des alluvions. BONFILS distingue les sols développés sur niveaux Wurm (unité 78) et Riss (unité 79) par l'apparition sur ce dernier de caractères minéralogiques témoins d'un début d'évolution fersiallitique sans que, pour autant, la morphologie et les potentialités soient différentes. Comme par ailleurs le niveau Wurm n'est pas continu sur la vallée, la différenciation sur le terrain de ces deux unités s'avère délicate.

A ces facteurs de variations "majeurs", se surimposent la présence de variations locales de teneur en éléments grossiers et l'apparition d'anciennes cuvettes de décantation générant des sols à texture lourde sur matériaux palustres. Le tout forme ainsi un ensemble assez complexe dans le détail.

1.3.3. Les sols de la terrasse villafranchienne (unité 75).

Ces sols apparaissent sur les surfaces témoins villafranchiennes qui forment des points hauts dans le paysage. Développés sur alluvions anciennes caillouteuses, leur morphologie est marquée par l'évolution fersiallitique qu'ils ont subie: horizons BT rouges (5YR) à forte teneur en argile (35 à 50%) et à revêtements argileux, pH acide, charge caillouteuse exclusivement composée de galets quartzeux. Les caractères de ces sols sont relativement constants sur toute l'unité. Cependant, une hydromorphie temporaire apparaît sur certains profils. D'autre part, sur la périphérie de l'unité

(pentes moyennes et fortes), au contact avec la molasse miocène, ces sols, remaniés par une action colluviale,sont moins rouges et recalcifiés.

Comme pour les précédents, les réserves hydriques des sols de terrasse villafranchienne sont limitées ce qui cantonne leur mise en valeur à la viticulture.

1.3.4. Les sols sur "molasse" de l'Helvétien (unités 60, 64,65)

Le terme molasse recouvre en fait des roche mères différentes responsables de la différenciation des unités. Ainsi, l'unité 60a se localise sur des faciès molassiques de grès calcaire dur ou de conglomérats caillouteux, souvent coquillers. Il s'ensuit des sols peu profonds, très calcaires, classifiés en lithosols et rendzines par BONFILS. Le plus souvent résistants au défoncement, ces sols sont généralement laissés à la végétation naturelle. Il est cependant possible de les trouver en zone cultivée, soit à la périphérie des puechs, soit en milieu de parcelle sur des zones très localisées (affleurements ponctuels).

Les unités 64 et 65 se développent, quant à elles, sur les sédiments sableux. Ces sols bruns calcaires présentent une texture équilibrée (limono-sablo-argileuse à limono-argilo-sableuse), une charge faible en élément grossiers (souvent coquillers). La couleur dominante est brun-jaune, avec, à la faveur de variations sédimentologiques, des inclusions de sols plus brun rouge (75YR). La différence principale évoquée par BONFILS entre les unités 64 et 65 concerne la profondeur du sol: moyenne (entre 50 et 100 cm) pour l'unité 64, plus importante (> 100cm) pour l'unité 65, à la faveur d'une position topographique plus dépressive. Cependant, dans le détail, la logique déterminant la profondeur d'apparition des sédiments non pédogénisés semble soumise à d'autres facteurs difficiles à percevoir, dont un facteur anthropique non négligeable.

Un autre facteur de variation au sein des unités de molasse est constitué par la proximité de terrasses villafranchiennes. Sur la rive droite de l'Hérault, l'unité 64 présente un léger colluvionnement villafranchien induisant une charge caillouteuse de surface plus importante.

Les perspectives de diversification sur ces sols, à réserves hydriques pourtant élevées, semblent cependant limitées en l'absence d'irrigation (MOLLET, 1991)

1.3.5. Les autres sols de la petite région naturelle:

A la différence des précédents, les sols présentés ci dessous n'occupent que des zones de faible étendue et/ou situées aux limites de la petite région naturelle:

- sols sur formations volcaniques: sols bruns andiques caillouteux et peu profonds (unité 52) et sol bruns colluviaux (unité 51) au niveau des coulées, vertisols (unité 54) et régosols sur tufs (unité 50) dans la dépression de Péret, comblée par des matériaux volcaniques;
- sols calcaires, graveleux, moyennement profonds (unité 61d) des reliefs tabulaires sur calcaires lacustres de l'Helvétien entre Péret et Adissan;
- sols sur molasse de l'Aquitanien (unités 66 et 66a) au Nord Est de la petite région naturelle. Cette formation détritique ne se différentiant de la molasse de l'Helvétien que par une teneur en argile généralement supérieure, les unités 66 et 66a peuvent être considérées comme des variantes plus argileuses des unités 64 et 65

- sols des niveaux de terrasses intermédiaires (entre alluvions récentes et Wurm) de la Lergue et de l'Hérault (unité 37), proches morphologiquement des sols des basses terrasses dont ils ne se distinguent que par leur caractère calcaire.
- sols bruns calciques ou calcaire, rougeâtres, graveleux des dépôts pliocènes. Ces sols sont surtout étendus dans la Basse Plaine de l'Hérault. Néanmoins, compte tenu du découpage adopté, ils apparaissent aux marges sud de la petite région, entre Peyne et Thongue.

En conclusion, la couverture pédologique de la moyenne vallée de l'Hérault présente une complexité moyenne: une organisation générale visible dans le paysage et compréhensible grâce à la connaissance de la mise en place des différents matériaux, mais aussi des variations plus difficiles à analyser et représenter compte tenu de leur caractère aléatoire ou progressif. Un facteur anthropique non négligeable vient par ailleurs compliquer, dans cette région, la tâche du cartographe. Il se manifeste par les aspects suivants:

- existence de terrasses de cultures perturbant la distribution initiale de la profondeur des sols;
- défoncements systématiques avant encépagement, provoquant un mélange des horizons originellement différenciés entre 0 et 50 cm.

2. LE SECTEUR DE REFERENCE DE LA MOYENNE VALLEE DE L'HERAULT

La carte du secteur de référence de la vallée de l'Hérault constitue le document à partir duquel vont être formulées les lois de distribution des sols susceptibles d'alimenter la représentation informatique du retour à la parcelle (chapitre 2).

Ce secteur, d'une superficie de 940 ha a fait l'objet d'une cartographie à grande échelle réalisée par LEFAY et al (1991) selon une démarche classique (855 sondages réalisés soit 1 sondage pour 0.9 ha). Les unités de sol ont été ensuite caractérisées finement dans le cadre de la thèse de LEENHARDT (1991).

Il constitue également le principal périmètre du secteur de référence "agriculture en transformation" réalisé dans le cadre de l'opération Inventaire, Gestion et Conservation des sols du Ministère de l'agriculture. Il s'étend sur les territoires des communes d'Adissan, Cazouls d'Hérault et Montagnac (planche 1).

Le choix du périmètre d'étude a été effectué avec le souci de recouper, sur un minimum de surface, les 4 unités paysagères majeures évoquées précedemment. Ceci explique sa forme générale, allongée selon un axe perpendiculaire à celui de la vallée. Ce choix limite cependant l'ambition de représentativité de ce périmètre. En effet:

- les unités "mineures" (paragraphe 1.3.5.) ne sont pas représentées;
- l'axe de différentiation régional Nord-Sud n'est pas pris en compte par le secteur du fait de son orientation générale.

Ainsi, pour espérer représenter l'ensemble de la petite région naturelle, des sous secteurs complémentaires seraient nécessaires comme celà a été souvent le cas au cours de l'opération secteur de référence drainage ONIC-Ministère de l'Agriculture".

La carte initiale du secteur de référence comprend 28 unités et sous unités. Ce chiffre correspond exactement à la moyenne du nombre d'unités trouvée sur les 70 secteurs de références de l'opération drainage ONIC-Ministère de l'agriculture (FAVROT, 1989). Cette carte fait l'objet de l'annexe 1. Elle n'a pas été utilisée directement. En effet:

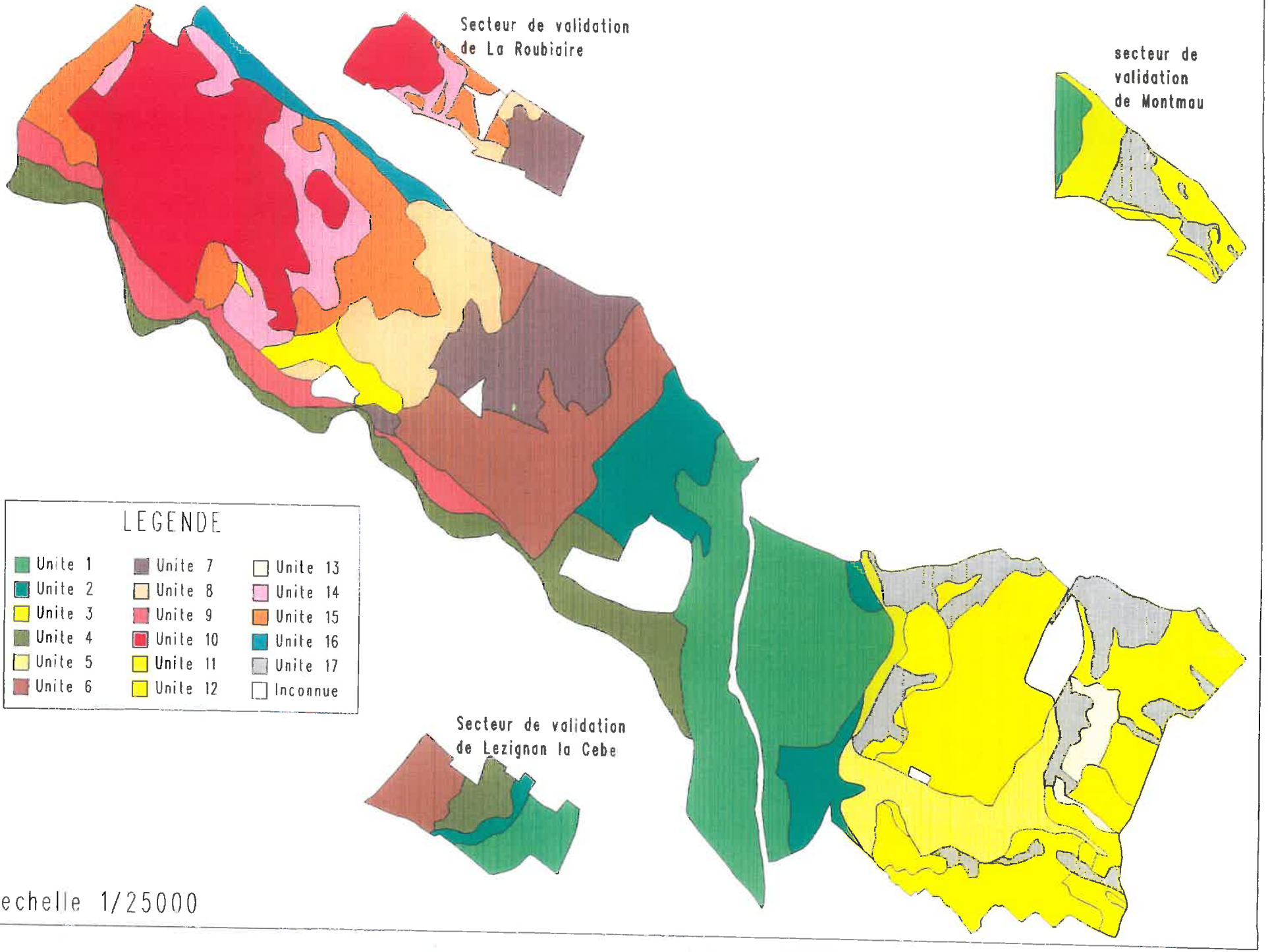
- Le nombre de 28 unités était trop élevé pour les possibilités informatiques disponibles, compte tenu de la lourdeur des algorithmes mis en oeuvre;
- nombre de ces unités occupaient des superficies trop limitées pour que des lois de distribution suffisamment générales et stables puissent être extraites dans le but de prédire ces unités.

En conséquence, à la suite de LEENHARDT (1991), un regroupement préalable d'unités a été effectué sur des bases géographiques et taxonomiques en éliminant autant que faire se peut les "petites" unités. Il en résulte une carte du secteur de référence simplifiée, comprenant 17 unités de sol. Ce regroupement est présenté dans le tableau 2 avec les les unités pédopaysagères de la carte au 1/250.000 correspondantes:

N° d'unités de sol après simplification	N° d'unité de sol du secteur de référence	Unité(s) pédopaysagère(s) correspondante(s)
1 2 4	1,2 3,4 6	173A
5	7	171E
3	5	transition 173A-552T
6 7 9	8a, 8b 9, 10 13	309A
8	11, 12	transition 552T-309A
10	14, 15	309V
14	21	transition 309V-552T
11	16	
12	17a, 17b, 18	
13	19	552T
15	22	
16	23a, 23b	
17	20a, 20b, 24	

Tableau 2: *Regroupements d'unités de sol effectués et correspondance des unités du secteur de référence avec celles de la carte des unités pédopaysagères au 1/250.000.*

***Planche 2**: Carte des sols du secteur de référence de la Moyenne Vallée de l'Hérault et des secteurs de validation*



Le résultat de ce regroupement constitue la carte figurant sur la planche 2 (aux côtés des secteurs de validation présentés plus loin). C'est sur cette nouvelle carte que seront réalisées les procédures d'extraction des lois de distribution des sols. Il convient de noter que LEENHARDT montre qu'un tel regroupement ne dégrade que faiblement l'information sur la variabilité du bilan hydrique des sols de cette petite région naturelle.

2.1. Les unités de sols de la carte simplifiée du secteur de référence

La légende de la carte du secteur de référence figurant en annexe 1 correspond à des descriptions succinctes des unités différenciées. Par soucis de concision, ces unités ne seront pas décrites plus en détail ici, de façon à éviter les redites par rapport au chapitre précédent. La description des unités de sol créées après regroupement, objet de ce chapitre, a pour but essentiel de relater la logique qui a prévalu pour représenter par une carte à grande échelle la variabilité pédologique du milieu.

Il est intéressant de présenter ces unités à la lumière du découpage en unités pédopaysagères opéré dans la carte au 1/250.000. Ainsi, deux catégories d'unités de sols se distinguent:

- les unités créées pour rendre compte de la variabilité interne de ces unités pédopaysagères,
- les unités créées pour décrire les transitions entre unités pédopaysagères.

Ces deux catégories seront envisagées successivement.

Les quatres grandes unités paysagères majeures de la petite région naturelle se retrouvent représentées directement par certaines des unités de la carte (tableau 2):

- **les sols de la plaine alluviale (unité 173A)** correspondent à 3 unités (1,2,4); la différenciation des deux premières (initialement chacunes subdivisées en deux) correspond à la prise en compte du gradient textural lié à la sédimentation fluviatile; l'unité 4 correspond aux sols développés sur les alluvions de la Boyne, affluent de l'Hérault, plus chargées en éléments grossiers;
- **les sols de la basse terrasse (unité 309A)** sont représentés par 3 unités (6,7,9); les 2 premières correspondent sensiblement aux unités 79 et 77 de la carte au 1/100000, les subdivisions ayant été éliminées par les regroupements[4]; l'unité 9 correspond aux sols développés sur les alluvions anciennes de la Boyne;
- **les sols de la haute terrasse (unité 309V)** sont décrits au sein de l'unité 10, initialement subdivisée uniquement sur le facteur pente;
- **les sols sur molasse du Miocène (unité 552T)**, très variés, sont détaillés grâce aux unités 5,11,12,13,15,16 et 17. Les distictions réalisées sont liées à la nature de la roche mère:
  + Les unités de sol sur colluvions (5,16) occupent les rives gauche et droite de l'Hérault; elles correspondent à des sols profonds développés sur des colluvions de molasse qui occupent les bas de pentes et thalwegs perpendiculaires à l'Hérault (unité 65 décrite précédemment);

---

4 Ces subdivisions concernaient:

- pour l'unité 6, la distinction de deux sous unités de superficies inégales selon la charge en cailloux,
- pour l'unité 7, l'isolation d'une unité de sol correspondant à des dépots palustres d'extension limitée (moins de 2 ha)

+ les unités de sol sur molasse "en place" (11,12,13) se distinguent encore en fonction de variations lithologiques au sein de cette molasse; L'unité 12 correspond sensiblement à la définition de l'unité 64 mais présente des profondeurs d'apparition du sédiment en place éminemment variables du fait de la présence de terrasses anthropiques; les autres (11 et 13) correspondent à des variantes de l'unité 64, particulièrement bien représentées sur le secteur:
  * l' unité 13 très argileuse dès la surface se développe sur le pôle marneux du dépôt molassique,
  * L'unité 11, développée sur des dépôts riches en coquille d'huitres se distingue de l'orthotype par sa couleur brun-rouge et sa texture légèrement plus argileuse sur l'ensemble du profil;

+ l'unité de sol sur molasse recouverte par des colluvions superficielles issues du cailloutis villafranchien (15), est localisée en rive droite et se distingue au sein de l'unité 64 par ses horizons superficiels riches en graviers et galets quartzeux;
+ les affleurements de molasse (unité 17) sont regroupés dans la carte simplifiée; ils se traduisent par l'association de sols sur grès calcaire et de sols caillouteux. Cette unité correspond sensiblement à l'unité 60a décrite dans le chapitre précédent.

Aux côtés des unités décrites précédemment, d'autres apparaissent pour représenter les zones de transition entre unités pédopaysagères:

- l'unité 3 représente, en rive gauche de l'Hérault, la transition entre alluvions récentes et matériaux tertiaires; elle est ainsi formée de colluvions de ces derniers reposant directement sur les alluvions;
- de même, l'unité 14, développée sur des colluvions de cailloutis villafranchiens recarbonatées au contact de la molasse proche constitue le terme de passage entre terrasse villafranchienne et molasse;
- enfin, l'unité 8, située aux confins des unités de molasse et de basse terrasse, traduit la couverture pédologique particulièrement complexe de cette zone du fait de l'interpénétration de ces deux milieux, de la présence de colluvions caillouteuses de bas de pente d'origine villafranchienne et de la rémanence de lambeaux de terrasses intermédiaires (initialement isolés sur la carte originale).

2.2. La carte du secteur de référence:

Finalement la carte du secteur de référence, après regroupement, comprend 17 unités de sol réparties en 55 plages cartographiques. A partir du tableau 3 (page suivante), deux remarques peuvent être faites:

- il persiste encore des unités de très faible extension impossibles à fusionner avec l'une de leurs voisines du fait d'une trop grande spécificité; c'est le cas des unités 3, 13 et 16;
- les unités sont en général concentrées sur un nombre limité de plages cartographiques; la lecture de la carte en est simplifiée; il est en effet facile de saisir l'organisation des unités les unes par rapport aux autres ainsi que l'axe de variation principal perpendiculaire à la vallée mis en évidence par le cartographe; les unités 12,13 et 17 font exception à cette

règle; la zone qu'elles occupent (rive gauche de l'Hérault) apparaît de ce fait beaucoup plus complexe.

N° unité	Aire (ha)	Aire (%)	Nombre de plages	Aire moyenne d'une plage (ha)
1	138	14.7	2	69
2	56	6.0	3	19
3	6	0.6	1	6
4	51	5.4	1	51
5	39	4.1	1	39
6	75	8.0	2	37
7	51	5.4	2	25
8	41	4.4	2	21
9	31	3.3	3	10
10	11	11.8	2	56
11	83	8.8	6	14
12	91	9.7	11	8
13	10	1.1	2	5
14	36	3.8	4	9
15	56	6.0	3	19
16	12	1.3	1	12
17	53	5.6	11	5

***tableau 3:** Principales caractéristiques géographiques des unités de sol retenues dans la carte simplifiée du secteur de référence.*

3. - LES SECTEURS DE VALIDATION

Pour tester la validité de la formalisation du retour à la parcelle et des règles de distribution d'unités de sol extraites du secteur de référence, il faut disposer, dans la petite région naturelle étudiée et à l'extérieur du secteur de référence, de périmètres déjà cartographiés, les **secteurs de validation**. Leurs cartes de sols sont présentées sur la planche 2 aux côtés de celle du secteur de référence (la position relative des secteurs de validation vis à vis du secteur de référence a été respectée). Deux soucis principaux ont sous-tendu leur choix.

1) Il fallait disposer, sur un minimum d'espace, d'un maximum d'unités de sols du secteur de référence afin d'étudier l'aptitude du modèle à les cartographier. Pour celà, les secteurs de validation occupent des positions de "charnières" entre différentes unités paysagères.
2) Il fallait ensuite être en mesure d'englober un maximum de contacts entre unités pour étudier comment ils seront détectés par l'outil informatique simulant le retour à la parcelle. Compte tenu du fait que les plages cartographiques sont généralement vastes, il est

nécessaire que chaque secteur ait une aire suffisante pour en englober plusieurs. Ceci explique pourquoi la solution consistant à choisir un semis de parcelles agricoles a été écartée. Cette solution aurait pourtant eu le double avantage de satisfaire avec plus d'efficacité le premier souci et d'être plus proches de l'utilisation prévue du modèle. Elle présentait cependant l'inconvénient rédhibitoire de ne pas permettre d'englober les contacts entre unités, les limites des parcelles agricoles coïncidant souvent avec ceux-ci.

En définitive, trois secteurs de validation ont été définis:

- **le secteur de "La Roubiaire"** (45 hectares) destiné à représenter la rive droite de l'Hérault avec l'étagement des terrasses d'alluvions anciennes;
- **Le secteur de "Lézignan la Cèbe"** (42 hectares) localisé sur la plaine d'alluvions récentes et son contact avec la terrasse du Pléistocène moyen,
- **Le secteur de "Montmau"** (44 hectares) choisi pour représenter la rive gauche de l'Hérault, occupée par les formations d'âge tertiaire.

Leurs caractéristiques font l'objet du tableau 4:

	Aire (Ha)	Nbre de sondages	densité	Nbre d'unités de sol
La Roubiaire	45	92	2.1	5
Montmau	44	54	1.3	5
Lézignan	42	68	1.5	4
Total	131	214	1.6	13

***tableau 4:** description des secteurs de validation*

Les 3 secteurs ont fait l'objet d'un retour à la parcelle. Les unités de sols ont été reconnues et délimitées au moyen de sondages à la tarière avec une densité moyenne de 1.6 sondages par hectare, adaptée à la complexité des différents milieux (cf tableau 4 et plans de sondages en annexe 2).

Au total, 14 unités de sol (sur les 17 du secteur de référence) ont été rencontrées (cf cartes des sols). Elles correspondent, en superficie, à 94% du secteur de référence. Une unité nouvelle a été identifiée sur 2.5 hectares (soit 2% de la surface totale de validation) au sein du secteur de La Roubiaire. Elle correspond à un résidu de niveau de terrasse non représenté sur le secteur de référence.

Ainsi, grâce à la diversité des situations englobées par ces trois périmètres, il sera possible:

- de vérifier si l'outil de prédiction des unités de sol construit peut être utilisé à l'extérieur du secteur de référence (sans préjuger par contre de l'extension du domaine d'utilisation);
- d'identifier ses dysfonctionnements face à des milieux diversifiés et représentatifs de la petite région naturelle.

La petite région naturelle Moyenne Vallée de l'Hérault a été choisie comme cadre expérimental du travail de recherche. Les règles sols-paysages et de voisinages seront donc extraites à partir d'une carte légèrement simplifiée de son secteur de référence (940 ha et 17 unités de sol distinguées).

Le dispositif expérimental est complété par trois secteurs de validation (130 ha au total) où un retour à la parcelle conventionnel a été réalisé. Ces secteurs feront parallèlement l'objet d'une prédiction automatique grâce à l'application de l'outil informatique simulant le retour à la parcelle, alimenté par les règles. La confrontation des deux cartes produites permettra un premier test des hypothèses et des options avancées dans ce travail.

CONCLUSION DE LA PREMIERE PARTIE:

Au cours de cette première partie, le problème de terrain qui est à l'origine du travail de recherche entrepris a été présenté. Il se résume à une constatation et une question. Un pédologue auteur d'un secteur de référence détient un savoir faire qui lui permet de réaliser plus rapidement les cartographies ultérieures effectuées dans la même petite région naturelle ("retours à la parcelle"). Comment formaliser ce savoir faire et l'intégrer à un outil informatique, de façon à aider un nouvel intervenant à réaliser lui-même ces cartographies(notion de "retour à la parcelle assisté par ordinateur") ?

Au delà de cet objectif très finalisé, il s'agit également de vérifier la validité de l'hypothèse de terrain sous tendant la méthode des secteurs de référence, soit: "l'organisation de la couverture pédologique est telle que la nature et la distribution dans le paysage des unités de sols identifiées dans un secteur de référence sont stables sur l'ensemble de la petite région naturelle que ce secteur est censé représenter".

Les objectifs étant posés, l'opération de retour à la parcelle est analysée en s'appuyant sur la bibliographie existante en matière de méthodologie cartographique. Il ressort de l'analyse que l'expertise du pédologue auteur d'un secteur de référence peut se représenter selon 4 catégories de lois:

- les lois d'identification des unités de sol,
- les lois fondées sur les relations sols-paysage (lois sols-paysage),
- les lois fondées sur les relations de voisinage entre unités de sols (lois de voisinage),
- les lois permettant de tracer les limites entre les unités de sol.

Puis une formalisation mathématique du retour à la parcelle est proposée. Ses principes généraux sont les suivants:

- considérer que le retour à la parcelle revient à rechercher en chaque point de cette parcelle l'unité de sol du secteur de référence à laquelle il appartient;
- définir, avant l'affectation finale d'un point à une unité, différentes étapes correspondant chacunes à l'introduction d'un nouvelle information (sondage); pour tout point, une étape est décrite par un vecteur dont les coordonnées représentent les probabilités de présence de chaque unité de sol en ce point;
- considérer que ces probabilités évoluent d'étapes en étapes sous l'action de règles si(premisse) alors (conclusion); ces règles sont la traduction des lois identifiées préalablement et qu'il faut donc formaliser en conséquence.

La mise en oeuvre de la formalisation du retour à la parcelle par un outil informatique est ensuite envisagée. Pour celà, l'espace est découpé en pixels de 50 m de côté. Dans un premier temps, une automatisation partielle est prévue grâce à des programmes FORTRAN associés au SIG ARC/INFO.

A ce stade du travail, compte tenu des acquis précédemment décrits et des limites posées au présent travail, l'objectif poursuivi dans les deux prochaines parties se précise. Il s'agira

- d'extraire automatiquement des règles traduisant les lois de distribution des sols à partir d'une carte de secteur de référence;

- d'intégrer ces règles à l'outil informatique simulant le retour à la parcelle défini lors de cette première partie
- d'appliquer l'outil ainsi renseigné à l'extérieur du secteur de référence pour vérifier la qualité des cartes de sol produites et juger, au travers des résultats obtenus, de la pertinence des hypothèses et des choix avancés dans ce travail.

Pour celà, un milieu expérimental sera nécessaire: celui-ci est présenté en fin de première partie. Il est constitué par la petite région naturelle "Moyenne Vallée de l'Hérault", son secteur de référence et trois secteurs de validation sur lesquels seront évalués les résultats obtenus.

Chaque catégorie de lois composant les lois de distribution des unités de sol sera envisagée dans une partie spécifique. La deuxième partie concernera les lois sols-paysage, la troisième partie s'intéressera aux lois de voisinage.

DEUXIEME PARTIE

FORMALISATION ET UTILISATION DES LOIS SOLS-PAYSAGE

L'objectif de cette deuxième partie est d'extraire, à partir de la carte du secteur de référence, des règles traduisant les lois de cartographie fondées sur les relations sols-paysage ("lois sols-paysage"). Intégrées à l'outil informatique construit lors de la première partie, ces règles ont pour fonction de délivrer, en chaque point de l'espace, une première prédiction sur la nature des unités de sol présentes. Concrètement, plusieurs questions se posent:

- quelles données d'accès aisé peuvent rendre compte des critères à partir desquels le pédologue prédit les unités de sol?
- Comment, à partir de ces données et de la carte des sols, retrouver les relations sols-paysage utilisées et les exprimer sous formes de règles respectant le formalisme général choisi dans le chapitre 2?
- Que peut-on espérer des prédictions résultantes, en termes de qualité et de perspectives d'utilisation ?

Ces problèmes seront traités au cours de cette deuxième partie qui comprend trois chapitres.

Le chapitre 4 est consacré à la présentation des données d'accès aisé qui ont été sélectionnées a priori pour remplacer les critères utilisés par le pédologue. Est également décrite la méthode permettant de mettre en relation ces données avec la carte des sols du secteur de référence.

Le chapitre 5 traite de la segmentation, méthode d'analyse de données utilisée pour élaborer les règles d'estimation des unités de sol selon le formalisme choisi a priori. Le problème de l'adaptation de cette méthode au caractère géographique des données manipulées est, à cette occasion, développé.

Le chapitre 6 présente la mise en oeuvre de la méthode décrite lors du chapitre précédent pour la Moyenne Vallée de l'Hérault. La qualité des prédictions est évaluée sur les 3 secteurs de validation présentés au chapitre 3. Cette évaluation débouche sur une discussion des modalités possibles d'utilisation de telles prédictions.

CHAPITRE 4

LES CRITERES D'ACCES AISE SELECTIONNES ET LEUR INTEGRATION AU SEIN D'ARC/INFO

La prédiction de la présence des unités de sol lors des retours à la parcelle doit faire appel, comme l'indique le chapitre 2, à deux sources d'information différentes. La première source concerne les critères d'accès aisé, relatifs à d'autres aspects du milieu naturel (topographie,géologie,...) ou concernant la surface du terrain. La deuxième correspond aux observations directes du sol. Plus précise et significative sera la première, plus limitée pourra être la seconde. Une grande attention doit donc être portée au choix des variables destinées représenter les critères d'accès aisé utilisé par le pédologue.

Face à la multiplicité de ces variables, il convient d'effectuer une première sélection permettant d'aboutir à un fichier de données gérables avec les moyens informatiques disponibles et utilisables dans le cadre de l'analyse de données projetée. Cette sélection a été réalisée sur la base des règles de choix suivantes:

- coûts d'acquisition des données inférieurs à ceux des données sol; ce critère se conçoit aisément puisqu'il serait peu démonstratif de réaliser la prédiction des sols au moyen de variables dont l'acquisition serait plus coûteuse qu'une carte pédologique!
- capacité à représenter l'un des critères utilisés par le pédologue pour prédire la présence des unités de sol; l'objectif envisagé dans ce travail est limité uniquement à la reproduction de la cartographie pédologique; en conséquence, on ne recherche pas de nouveaux critères qui dépassent les capacités habituelles d'observation d'un pédologue (traitements d'image SPOT par exemple).
- erreur minimale d'estimation (dans le cas où deux variables représentant la même réalité seraient en concurrence).

L'application de la première règle de choix limite déjà considérablement le champ d'investigation. Les variables utilisées ne peuvent être extraites que de deux sources: soit de documents cartographiques déjà publiés pour lesquels il existe des facilités d'utilisation, soit de données satellitaires dont le coût d'acquisition à l'hectare est faible. Ces dernières n'ont cependant pas été retenues compte tenu des autres règles de choix. Ainsi, les variables utilisées sont extraites à partir des deux documents cartographiques les plus précis dans leur domaine respectif et disponibles sur l'ensemble du territoire français:

- la carte géologique au 1/50.000 produite par le Bureau de Recherches Géologiques et Minières (BRGM),
- La carte topographique au 1/25 000 produite par l'Institut Géographique National (IGN)

Ces documents constitueront donc, avec la carte des sols du secteur de référence déjà décrite au cours du chapitre 3, la base de données support de la recherche des règles sols-paysage.

Très concrètement, construire cette base de données revient à produire un fichier de points répartis aux noeuds d'une grille à maille carré. Dans ce fichier les points doivent être renseignés par des variables issues des cartes topographique et géologique et, s'ils se trouvent inclus dans le périmètre du secteur de référence, par l'unité de sol à laquelle ils appartiennent. Ce choix de représentation géographique de l'information facilitera grandement l'utilisation des différentes

fonctions d'analyse spatiale mises en oeuvre au cours de cette démarche (croisements avec des cartes "en plages", interpolations, calcul de distances).

Par ailleurs, il n'était pas possible de couvrir l'ensemble de la petite région naturelle et ce avec une précision illimitée. En conséquence:

- un rectangle de 7.450 km (axe Est-Ouest) sur 5.750 km (axe Nord-Sud) a été délimité, de façon à englober le secteur de référence et les secteurs de validation (planche 1),
- le pas de la maille a été fixé à 50m pour représenter cette zone, ce qui constitue un ensemble de 17035 points.

Le présent chapitre décrit la constitution du fichier des 17035 points correspondant, renseignés par des variables issues des 3 documents cartographiques utilisés.

1. LA CARTE TOPOGRAPHIQUE: PRESENTATION ET ACQUISITION DES VARIABLES RETENUES

Le secteur de référence de la Moyenne Vallée de l'Hérault se situe sur la feuille au 1/25.000 de Pézenas. La carte topographique correspondante (IGN, 1983) a été réalisée d'après des levés photogrammétriques complétés[5] sur le terrain en 1952-53. Elle a fait l'objet d'une révision en 1981.

Parmi les nombreuses couches d'informations disponibles (routes, toponymie,...), seules celles concernant les cours d'eau et les courbes de niveau ont été utilisées. L'hypothèse sous-tendant ce choix est que la distribution des sols est influencée, dans la Moyenne Vallée de l'Hérault comme dans beaucoup d'autres petites régions naturelles, par le relief et la géométrie du réseau hydrographique. En revanche, la végétation, présente aussi sur la carte topographique, n'a pas été retenue. Elle ne constitue pas en effet un critère pertinent a priori dans la mesure où la région fait l'objet actuellement d'une quasi monoculture de vigne.

Les courbes de niveau et les cours d'eau ont été numérisés manuellement sur la zone d'extrapolation, constituant ainsi sous ARC/INFO deux couvertures d'arcs. Il convenait dès lors d'en extraire des variables pertinentes et de les transformer en couvertures de points afin de se conformer au formalisme choisi pour les données d'entrée du modèle.

1.1. Le traitement de la couverture des cours d'eau

A partir de l'examen de la géométrie des unités de sols de la carte, deux variables ont été choisies puis extraites de la couverture des cours d'eau. Il s'agit des variables "rive" (notée rv(x)) et "distance au cours d'eau le plus proche"(notée di(x))).

La variable rv(x) a été obtenue en affectant chaque point de la grille à l'un des deux polygones d'une couverture délimitant les rives droite et gauche du fleuve Hérault. Cette couverture dérive de la couverture des cours d'eau par une succession d'opérations[6] réalisées sous ARC/INFO.

---

5 Le complètement d'une carte topographique représente la dernière étape de sa fabrication. L'opération consiste d'une part à corriger les levés photogrammétriques par une vérification terrain sur les sites délicats à cartographier (terrain plat, clancs sur les prises de vue,...) et, d'autre part, à compléter la carte par des informations invisibles sur photo aérienne (exemple: toponymie)

6 Dans le détail ces opérations sont les suivantes:

- sélection, au sein de la couverture des cours d'eau, des arcs correspondant au tracé du fleuve Hérault;

La variable di(x) a été pour sa part renseignée grâce à une mesure pour chaque point de la grille, sous ARC/INFO, des distances aux arcs les plus proches de la couverture des cours d'eau.

1.2. Le traitement de la couverture des courbes de niveau

Cinq variables caractérisant le relief en chaque point ont été produites à partir de la couverture des courbes de niveau. Il s'agit de l'altitude (z(x)), du dénivelé par rapport au fleuve (dz(x)), de la pente (pt(x)), de la courbure moyenne du relief (cm(x)) et du degré d'encaissement (ec(x)). A la différence de précédemment, la filière d'obtention de ces variables est extérieure à ARC/INFO. Elle utilise des logiciels mis au point par C.DEPRAETERE (1990; 1991). Deux étapes peuvent être identifiées:

- la production du Modèle Numérique de Terrain (MNT), fichier des altitudes pour chaque point de la grille;
- la dérivation des autres variables à partir de ce premier fichier.

1.2.1. La production du MNT

Le MNT est réalisé par interpolation des courbes de niveau. Bien qu'il existe d'autres filières d'obtention (figure 8), celle-ci représentait a priori le meilleur compromis au vu des critères de précision et de coût définis en début de chapitre 1. En effet:

- la précision sur l'altitude fournie par la filière satellitaire est à priori moins bonne (5 mètres d'erreur moyenne cité par PLANCHON (1991));
- la filière photogrammétrique représente sans aucun doute la meilleure solution en terme de niveau d'erreur, dans la mesure où elle limite les étapes intermédiaires donc les sources d'erreur; son coût de mise en oeuvre (700 à 800F/ha) est cependant trop élevé pour qu'elle puisse être envisagée à l'heure actuelle dans le cadre d'une extrapolation d'un secteur de référence.

Ainsi, le MNT a été calculé par le logiciel OROLOG (DEPRAETERE, 1990) par interpolation des courbes de niveau saisies sous ARC/INFO. La méthode d'interpolation, adaptée de travaux antérieurs (YOELI, 1986) avec quelques améliorations, est détaillée en annexe 3. En résumé, elle consiste à estimer l'altitude en un point par la moyenne pondérée des valeurs que prennent, suivant 4 axes sécants en ce point, des courbes splines cubiques ajustées sur les courbes de niveau (Figure 9). La pondération privilégie la ligne de plus grande pente puisqu'elle favorise les axes dont les 4 courbes de niveaux sécantes se trouvent en moyenne les plus proches du point considéré. Ceci a pour but de simuler au mieux un expert humain qui réaliserait un compromis entre une vision "monoaxiale" et une vision "tous azimuts".

De fait, cette méthode s'avére performante. En effet,les erreurs quadratiques moyennes obtenues expérimentalement par YOELI (1986) en utilisant une méthode très proche apparaissent

---

- transfert de ces arcs dans une couverture "rive" accueillant également les arcs délimitant la zone d'extrapolation; à ce stade, les contours des polygones caractérisant chaque rive sont saisis;
- affectation, à chaque polygone d'un identificateur ponctuel permettant de construire la couverture "rive" en vue des croisements ultérieurs.

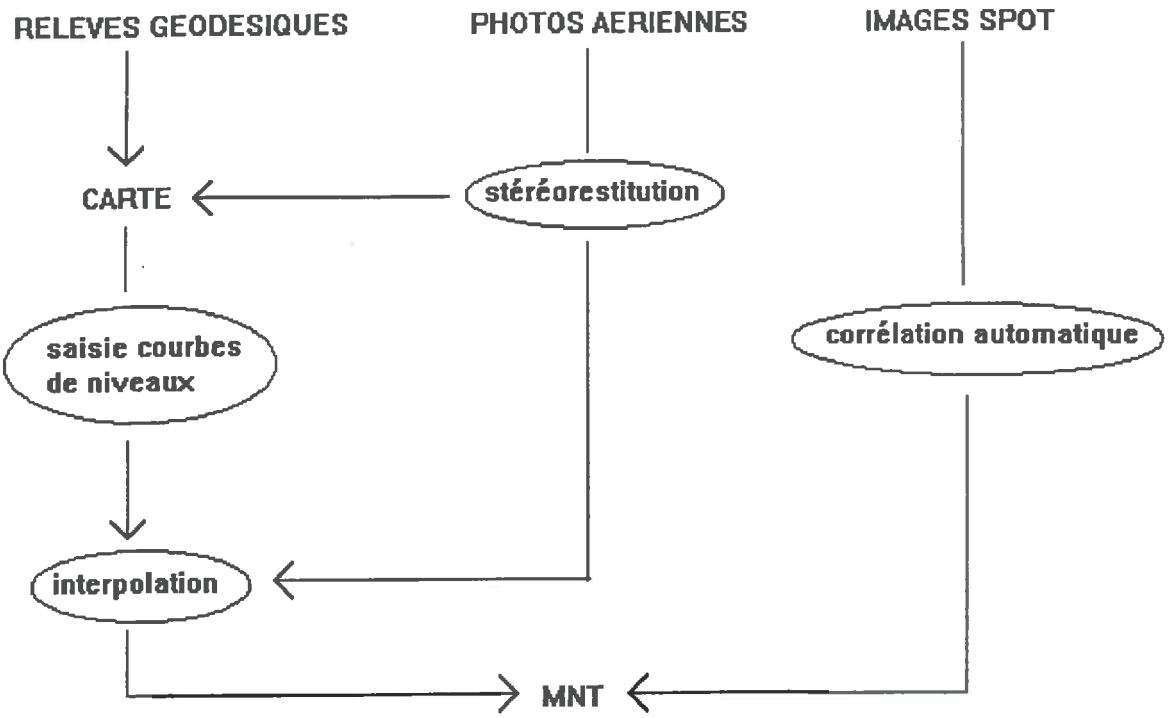


Figure 8: les différentes filières d'obtention des Modèles Numériques de Terrain (d'après DEPRAETERE, 1991)

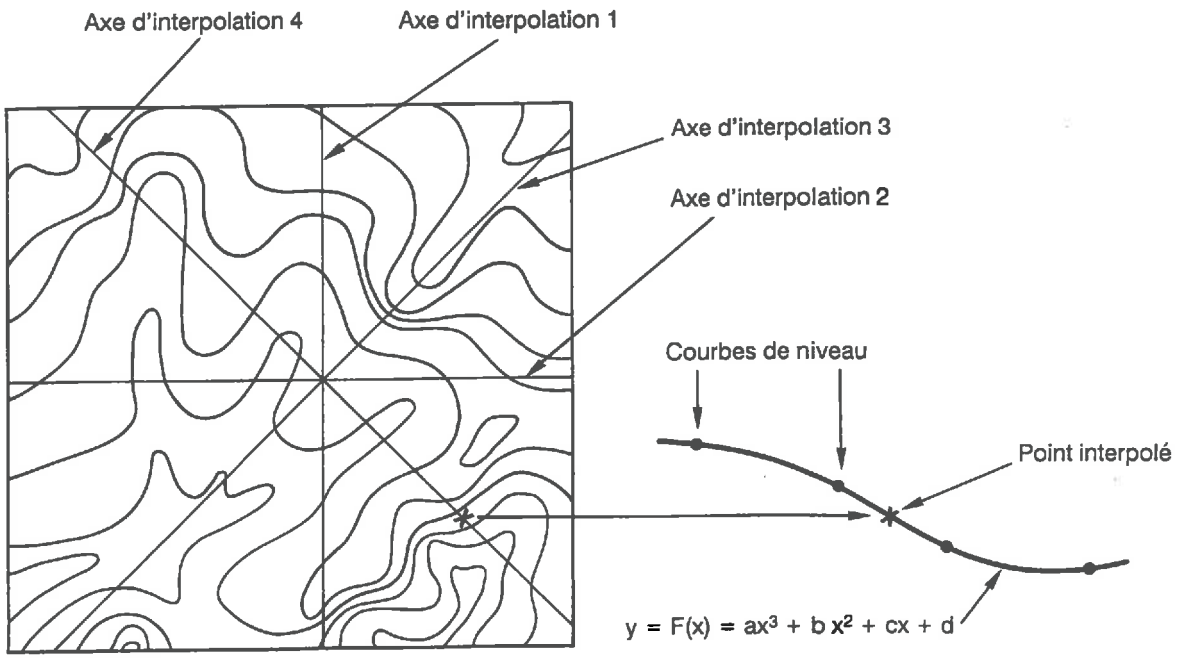


Figure 9: méthode d'interpolation des courbes de niveau

faibles quand on les compare avec les erreurs établies pour diverses méthodes d'interpolation (CLARKE et al, 1982). Cette erreur correspondrait à 5% du dénivelé entre courbes de niveau, soit 25 cm dans le cas de courbes issues de la carte au 1/25. 000 utilisée.

Cependant, ces estimations ne tiennent pas compte de deux difficultés, particulièrement apparentes dans la zone étudiée:

- l'interpolation entre courbes de niveau repose, comme toute interpolation, sur une hypothèse de continuité; celle ci n'est pas vérifiée partout sur le secteur du fait des nombreuses discontinuités topographiques créées par les talus de terrasses de culture;
- la forme des courbes splines cubiques introduit un biais dans l'estimation des altitudes au niveau de certaines formes de relief, en particulier des vallées à fond plat (figure 10); cette forme de relief occupe des superficies importantes dans la zone d'extrapolation choisie (plaine alluviale de l'Hérault); pour limiter ce biais, des courbes de niveau intermédiaires ont été rajoutées manuellement en respectant le parallélisme avec les courbes existantes et en se calant sur les points côtés visibles sur la carte.

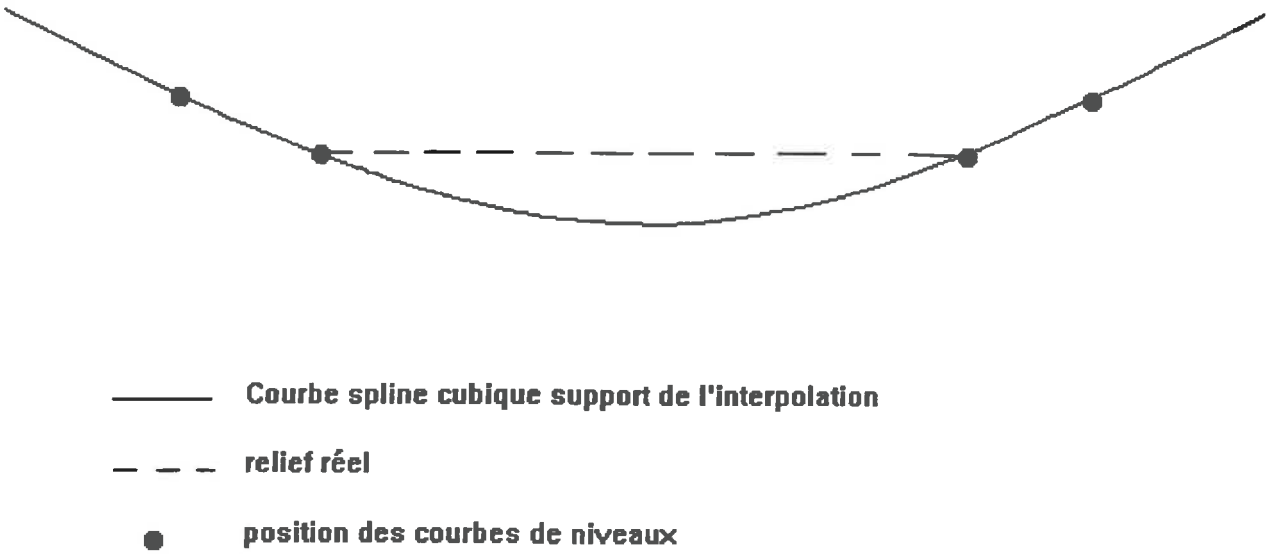


Figure 10: biais introduit par l'interpolation des courbes de niveau en cas de vallée à fond plat.

1.2.2. La production des fichiers dérivés

Les variables "pente" et "courbure moyenne" représentent les dérivées premières et secondes du champ des altitudes matérialisé par le MNT (DUPERET, 1989). Leurs calculs peuvent être réalisés suivant différentes méthodes. La démarche retenue est détaillée, pour chaque variable, en

annexe 4. Le principe général de ces calculs consiste à ajuster, sur une fenêtre carrée de 3 points sur 3, un plan continu dont l'équation utilise un développement de Taylor à l'ordre 2. Les coefficients de ce développement de Taylor pour le point central de la fenêtre sont calculés par ajustement selon la méthode des moindres carrés à partir des altitudes des 8 points voisins. L'équation du plan H étant ainsi connue pour chaque point, la pente et la courbure moyenne représentent respectivement les dérivées premières et secondes de l'expression de H.

La variable "encaissement" représente en un point la somme des dénivelés entre ce point et ses 8 voisins (DEPRAETERE, 1991). Elle représente ainsi une autre approche de la courbure du relief dont le principal intérêt est la simplicité de calcul.

Les trois variables détaillées précédemment ont été obtenues grâce au logiciel LAMONT (DEPRAETERE, 1991).

Une dernière variable, "dénivelé par rapport au fleuve", a été introduite pour être substituée à l'altitude. Cette dernière présente en effet une dérive causée par le gradient régional des altitudes d'amont en aval du fleuve. Le secteur de référence étant orienté transversalement par rapport au fleuve (cf carte 1b planche 1), ce gradient n'est pas pris en compte ce qui risquerait de perturber les prédictions utilisant directement la variable altitude. Le calcul du dénivelé par rapport au fleuve est obtenu par différence, en chaque point, entre l'altitude (z(x)) donnée par le MNT et l'altitude, au point x considéré, d'un niveau de base fictif ajusté sur le lit du fleuve. Le fleuve coulant suivant un axe sensiblement Nord Sud, ce dernier terme est estimé à partir d'une régression linéaire sur la latitude portant sur 11 points côtés identifiés le long de la vallée de l'Hérault. L'équation, obtenue avec un coefficient de corrélation de 0.98, est la suivante:

$$z_0(x) = 0.0095\ la(x) - 17303 \qquad [9]$$

avec:

$z_0(x)$: altitude en x du niveau de base ajusté sur le lit de l'Hérault
la(x): latitude au point x

Ce résultat conduit à estimer la pente du lit du fleuve Hérault à 1 pour mille, ce qui est en conformité avec les pentes longitudinales mesurées sur les moyennes vallées des rivières et fleuves entaillant, comme l'Hérault, des formations meubles détritiques du Tertiaire et du Quaternaire (BORNAND, 1972).

2. LA CARTE GEOLOGIQUE: ACQUISITION DE LA VARIABLE "UNITE GEOLOGIQUE"

Le secteur de référence de la Moyenne Vallée de l'Hérault se situe sur la carte géologique au 1/50.000, feuille de Pézenas (BRGM, 1981). 8 unités géologiques (tableau 5, page suivante) sont représentées à l'intérieur de son périmètre.

Cette carte géologique a fait l'objet d'une numérisation sous ARC/INFO. Il en résulte une nouvelle couverture de polygones. Le croisement de cette couverture avec la grille de points a permis d'ajouter aux variables topographiques une variable géologique (notée g(x)) caractérisant l'appartenance de chaque point à l'une des huit unités géologiques présentes dans la zone d'extrapolation.

g(x)	Nom de l'unité sur la carte géologique	Définition
1	Fz	alluvions récentes
2	Fya	alluvions anciennes (niveau 8-10m)
3	Fyb	alluvions anciennes (niveau 10-20m)
4	Fy	alluvions anciennes (niveau 8-20m)
5	Fx	alluvions anciennes (niveau 20-30m)
6	Fv	Cailloutis villafranchien
7	m2a	Molasse sableuse - Marne bleue
8	m2a-3	Calcaire Lumachellique

Tableau 5: définition et codification des unités géologiques présentes sur le secteur de référence

3. LA CARTE DES SOLS: EXTRACTION DE LA VARIABLE "UNITE DE SOL"

La carte des sols, présentée au chapitre 3, a été numérisée comme la précédente. Elle a été également croisée avec la grille de points support du modèle. Une nouvelle variable, "unité de sol" (notée u(x)), vient donc s'ajouter à celles présentées ci-dessus. Cependant, à la différence des précédentes, cette variable n'est renseignée que dans le cas où le point se situe à l'intérieur du périmètre du secteur de référence. Ce cas ne concerne que 3779 points sur les 17035 de la zone d'extrapolation choisie. Dans le cas contraire, elle n'est pas renseignée et prend la valeur "0".

Dans la figure 11, est présenté l'enchaînement des différentes opérations permettant de mettre en relation les unités de sol et les variables caractérisant des critères extrinsèques, en vue de la recherche de règles formalisant les lois "sols-paysage". Plusieurs fonctions d'analyse spatiale sont impliquées: recouvrement (ou croisement) de différentes couches d'information, calcul de distance, interpolation et dérivation. Le résultat de cette démarche est matérialisé par une couverture de points sous ARC/INFO.

Au niveau graphique, cette couverture se présente comme un semis de points répartis aux noeuds d'une grille à maille carrée de pas 50 mètres.

La partie descriptive (ou "sémantique") est constituée par la table attributaire ARC/INFO de points au sein de laquelle figurent 1 variable géologique et 7 variables

topographiques, choisies a priori pour leurs corrélations supposées avec la variable "unité de sol". Les 17035 points de la grille sont donc caractérisés par ces 8 variables. Par ailleurs, pour les 3779 points qui présentent la particularité d'être situés à l'intérieur du périmètre de secteur de référence, la variable "unité de sol" (u(x)) est également connue.

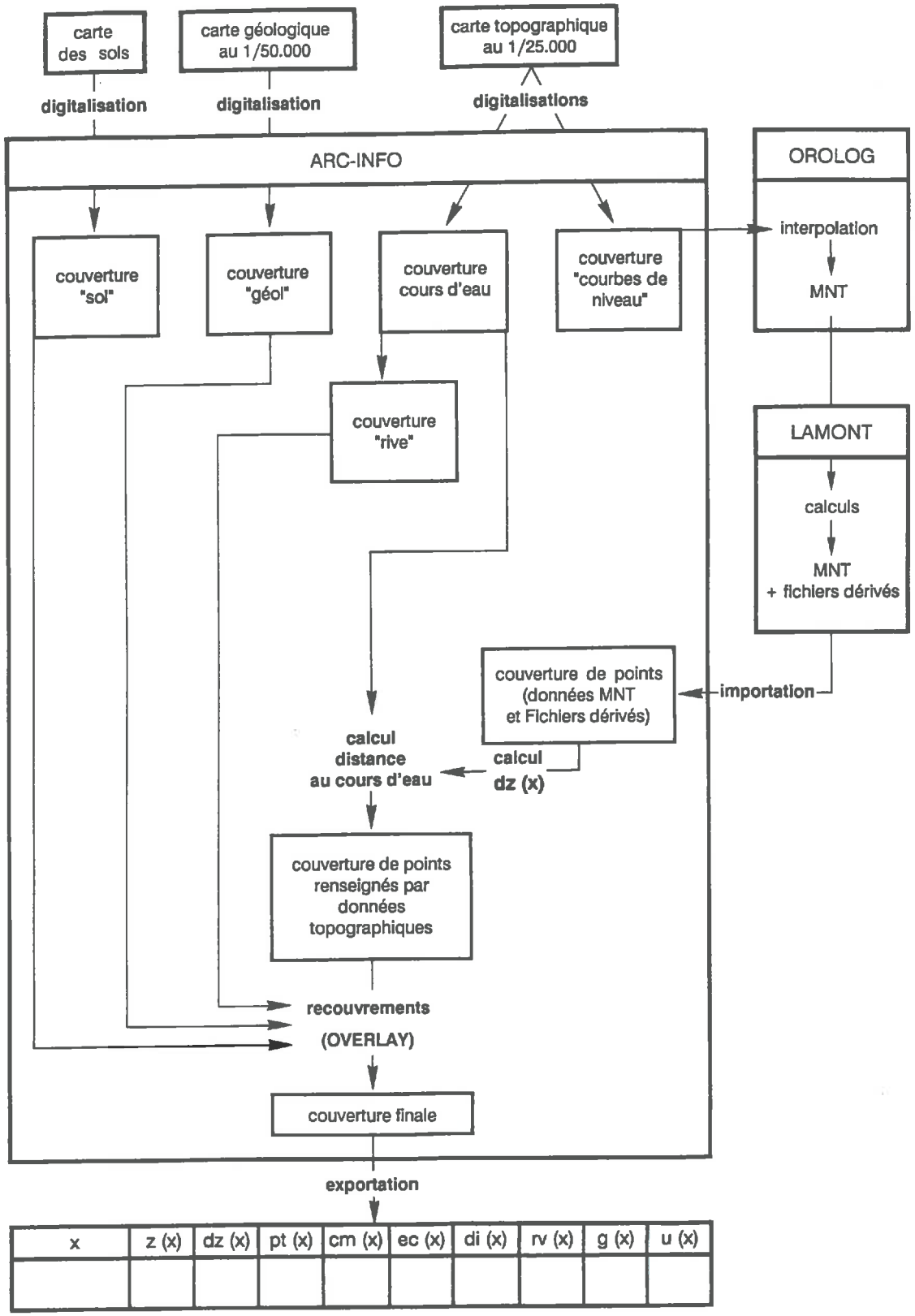


Figure 11: acquisition des variables utilisées pour formaliser les lois-sols-paysage

CHAPITRE 5

CHOIX ET ADAPTATION D'UNE METHODE D'ANALYSE DE DONNEES: LA SEGMENTATION

Le but de l'analyse de données envisagée consiste à extraire des règles permettant de prédire en tout point la variable "unité de sol" à l'aide des variables topographiques et géologiques obtenues dans le chapitre précédent. Elle porte sur le sous-ensemble des points situés à l'intérieur du périmètre du secteur de référence, pour lesquels la variable "unité de sol" est connue. Ses résultats sont destinés à être appliqués sur les points situés en dehors ce ce secteur, renseignés uniquement par les variables topographiques et géologiques.

Plusieurs exigences et contraintes permettent d'orienter le choix parmi les différentes méthodes d'analyses multidimensionnelles:

- Il s'agit de **prédire** une variable à partir d'un ensemble d'autres et non d'identifier des éventuelles corrélations entre variables;
- certaines variables sont quantitatives (altitude, pente,...) alors que d'autres sont qualitatives (unité géologique, rive,...); on ne peut donc utiliser les méthodes employées pour prédire une variable en un lieu donné à partir de variables exclusivement quantitatives; c'est le cas, par exemple, de la régression linéaire multiple (BENICHOU et LE BRETON, 1986) ou de l'analyse factorielle discriminante (LOWELL, 1991).
- Le formalisme des résultats doit pouvoir être aisément adapté à celui exigé (chapitre 2)

Un travail récent (ASPINALL, 1992) propose une approche statistique de type Bayesien, utilisant des probabilités conditionnelles. Cette voie a cependant été connue trop tardivement pour pouvoir être explorée.

Le choix s'est porté sur un groupe de méthodes d'analyses qui répondaient (au moins pour certaines d'entre elles) à ces exigences: il s'agit des méthodes de segmentation (BOUROCHE et TENENHAUS, 1970; BACCINI, 1975), connues dans la littérature anglo-saxonne sous les termes de "classification trees" (BREIMAN et Al, 1984) ou "recursive partitions" (CIAMPI et THIFFAULT, 1988) ou encore "tree based models" (CLARK et PREGIBON, 1991). Ces méthodes, relativements récentes dans la littérature (années 1960), ont été utilisées dans des domaines variés: médecine, marketing, météorologie,.....Il existe, au vu de la bibliographie consultée, un seul exemple d'application d'une méthode de segmentation sur des données spatiales. Il concerne la prédiction de la distribution spatiale d'une espèce de Kangourou (WALKER et MOORE, 1988).

Parmi les différentes méthodes de segmentation existante la méthode de BREIMAN et Al (CART) a été choisie. En effet, elle répondait le mieux aux exigences formulées plus haut. De plus, elle est actuellement la méthode de segmentation la plus répandue et la plus facile d'utilisation grâce à sa programmation réalisée sous UNIX (logiciel CART et "fonction tree" sous S).

Ce chapitre sera divisé en deux parties:

- dans un premier temps, les principes fondamentaux et les difficultés d'utilisation de la segmentation seront présentés;
- dans un deuxième temps, la prise en compte de la nature géographique et aléatoire des variables utilisées sera considérée afin de permettre de résoudre les problèmes d'interprétation des résultats d'une analyse par segmentation.

Au cours de ces deux parties est introduit un formalisme mathématique qu'il était difficile d'éviter compte tenu du domaine d'étude abordé et des sources bibliographiques servant de support aux développements présentés. Il a été tenté, autant que faire se peut, de doubler ce formalisme de commentaires, résumés et exemples permettant de mieux suivre la logique de la démarche poursuivie.

1. PRINCIPES ET PROBLEMES DE LA SEGMENTATION

Seront présentés successivement dans ce sous-chapitre l'objectif, la démarche générale et les problèmes des méthodes de segmentation.

1.1. Objectif de la segmentation: l'arbre de classification

L'objectif de toute méthode de segmentation est de fournir une prédiction d'une variable qualitative a priori inconnue sur un individu, un objet ou un lieu donné. Cette prédiction doit utiliser d'autres variables connues, qualitatives ou quantitatives, supposées corrélées avec la variable recherchée.
Ainsi, dans le cas étudié, il s'agira de prédire en tout point la variable "unité de sol" à partir de variables explicatives (altitude,unité géologique,...). Pour ce cas, cet objectif peut s'énoncer comme suit:

- on dispose de points x_i (i = 1,...,n) sur lesquels ont été observées r variables $o_1(x_i),...,o_h(x_i),...,o_r(x_i)$; soit $\mathbf{o}(x_i)$ le vecteur des observations en x_i et X l'espace de ces vecteurs;
- on sait par ailleurs que tout $\mathbf{o}(x_i)$ appartient à une classe U_j et une seule parmi w classes possibles (j = 1,...,w), l'ensemble des U_j constituant une partition de X
- Trouver une fonction $D[\mathbf{o}(x_i)]$ prédisant l'appartenance de $\mathbf{o}(x_i)$ à U_j avec un risque d'erreur minimum

La forme que doit prendre la fonction de prédiction $D[\mathbf{o}(x_i)]$ constitue l'aspect caractéristique des méthodes de segmentation (figure 12). Il s'agit d'un **arbre de classification** permettant d'affecter tout individu $\mathbf{o}(x_i)$ dans l'un des sous ensembles ("noeuds terminaux") $T_1,T_2,...T_g,...,T_s$, éléments d'une partition de l'ensemble X de départ. L'arbre doit être choisi de telle sorte que chaque population d'individus constituant un noeud terminal donné puisse être affectée globalement à l'une des classes U_j, satisfaisant ainsi l'objectif de départ.

L'arbre de classification est constitué par un ensemble de dichotomies successives divisant X, puis des sous ensembles de X notés $X_1,X_2,...,X_m,...X_t$. Chaque dichotomie est le résultat d'un tri des individus en deux sous-ensembles, selon qu'ils vérifient ou non un critère impliquant une variable explicative $o_h(x_i)$. Deux cas se présentent:

- si $o_h(x_i)$ est quantitative, le critère orientant le tri prend la forme d'une inégalité (exemple "DZ $\leq$ 17"),
- si $o_h(x_i)$ est qualitative, il s'agit d'une disjonction d'égalités (exemple "UG = 1 ou UG = 3 ou UG = 6")

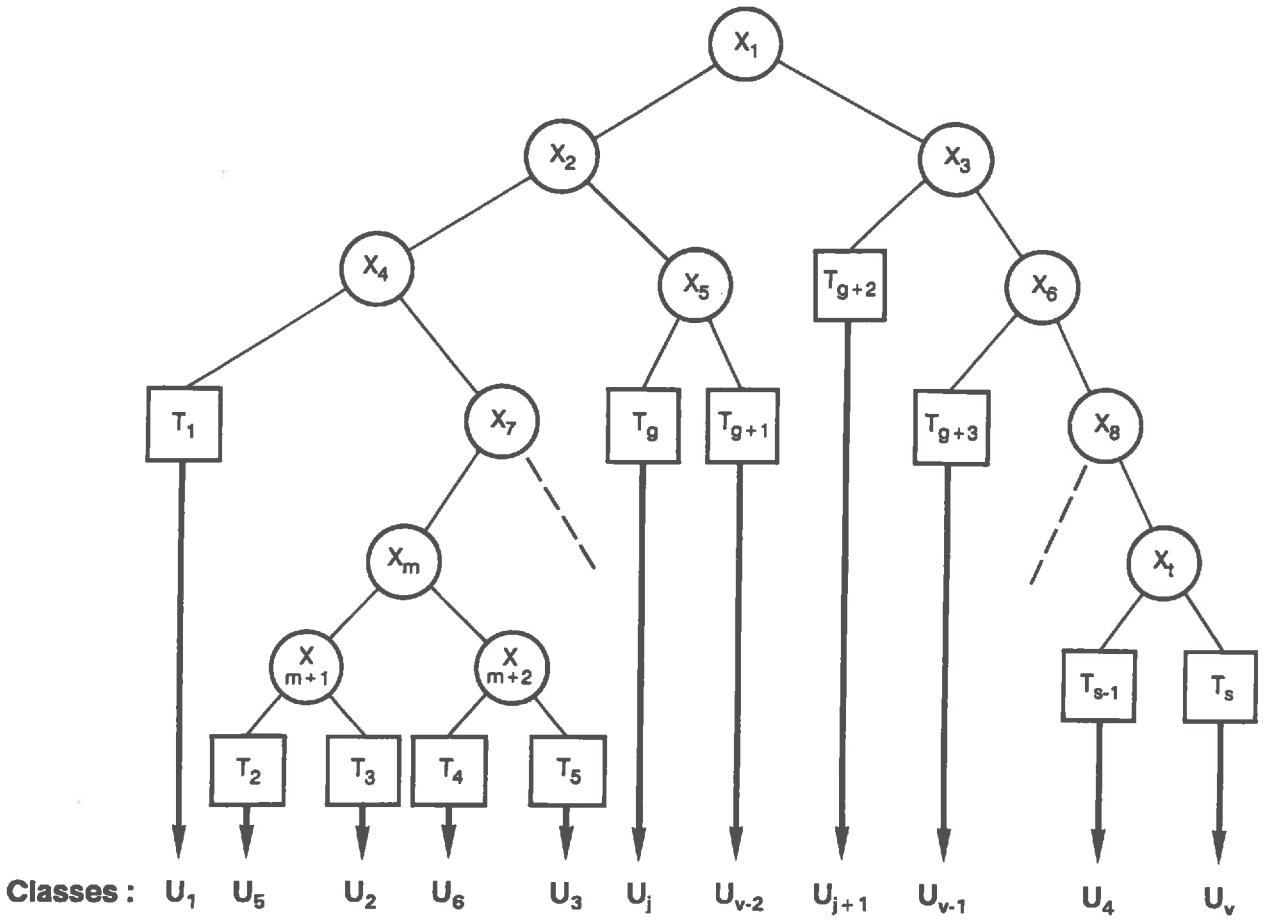


Figure 12: Arbre de classification obtenu par une méthode de segmentation

1.2. Méthode de construction de l'arbre de classification

Il existe un nombre fini mais extrèmement élevé d'arbres de classification susceptibles d'être construits selon les modalités présentées précédemment. Il y a en effet, à chaque dichotomie, autant de critères utilisables que de variables explicatives disponibles; de plus, pour une même variable, il y a aussi plusieurs façons de scinder un ensemble en deux parties suivant les valeurs impliquées. Comment construire le meilleur arbre vis à vis de l'objectif poursuivi, c'est à dire celui qui minimise l'erreur de classement des $o(x_i)$ dans les classes U_j ?

Il y a erreur de classement lorsque un individu, tombant dans un noeud terminal affecté à U_j, appartient en réalité à une autre classe. Minimiser globalement cette erreur revient donc à produire des noeuds terminaux qui soient les plus purs possible vis à vis de U_j. Autrement dit, il faut se rapprocher au maximum de l'arbre idéal où, quel que soit le noeud terminal associé à U_j, les individus qui le composent appartiennent effectivement tous à la classe U_j (erreur de classement nulle).

Pour choisir le meilleur arbre possible, toute méthode de segmentation utilise un ensemble d'apprentissage constitué d'individus $o(x_i)$ dont on connaît déjà la classe U_j à laquelle ils appartiennent. Sur cet ensemble, est appliqué une démarche pas à pas: à chaque étape, depuis X_1 jusqu'à X_t, toutes les dichotomies élémentaires possibles sont essayées. Est retenue et déclenchée celle dont le critère associé permet de distinguer des sous-ensembles "fils" les plus homogènes vis à vis de l'appartenance de leurs $o(x_i)$ aux classes U_j. La répétition de cette procédure pour chaque dichotomie permet finalement de produire des noeuds terminaux les plus purs possible, au moins vis à vis de l'ensemble d'apprentissage.

Pour choisir la meilleure dichotomie possible, il faut disposer d'une mesure d'hétérogénéité de chaque noeud vis à vis de U_j. Pour cela, il est défini un **indice d'impureté** $i(X_m)$ calculé à partir des proportions des individus de chaque unité de sol présents dans X_m. Les propriétés de cet indice peuvent s'énoncer de la façon suivante:

- soit pr_{mj} la proportion d'individus tombant dans le noeud X_m qui appartiennent à la classe U_j $(0 \leq pr_{mj} \leq 1)$
- $i(X_m)$ = maximum si $pr_{m1} = \ldots = pr_{mj} = \ldots = pr_{mv}$ (cas où toutes les proportions sont égales l'indice d'impureté est donc maximum)
- $i(X_m) = 0$ s'il existe une classe U_j telle que $pr_{mj} = 1$ (cas d'une population pure en terme d'unité de sol; L'indice d'impureté est minimum)

Il existe plusieurs indices d'impureté répondant aux propriétés énoncées ci-dessus. La méthode utilisée dans le logiciel CART utilise le "gini index":

$$i(X_m) = 1 - \Sigma_{(j=1\ldots v)}\, pr^2_{mj} \qquad [10]$$

ou, si le calcul de pr_{mj} est explicité:

$$i(X_m) = 1 - \Sigma_{(j=1\ldots v)}\, (N^2_{mj}/N^2_m) \qquad [11]$$

avec:

N_{mj}: Nombre d'individus tombant dans X_m et membres de la classe j,
N_m: nombre total d'individus dans le noeud X_m

Cet indice d'impureté ainsi défini permet de calculer, pour toute dichotomie possible sur le noeud X_m, le gain de pureté entre X_m et les deux nouveaux noeuds produits X_{m+1} et X_{m+u} :

$$\Delta[i(X_m)] = i(X_m) - [(N_{m+1}/N_m)i(X_{m+1}) + (N_{m+2}/N_m)i(X_{m+u})] \quad [12]$$

avec N_m = nombre d'individus tombant dans le noeud X_m

$\Delta[i(X_m)]$ représente en fait la différence entre l'indice d'impureté du noeud sur lequel s'applique la dichotomie et la moyenne pondérée (par le nombre d'individus) des indices d'impureté des 2 nouveaux noeuds.

La dichotomie qui permet d'obtenir un gain de pureté $\Delta[i(X_m)]$ maximum sera choisie et déclenchée. La même démarche sera appliquée de manière itérative à tous les noeuds descendants de X_m tant que ceux-ci n'auront pas été déclarés terminaux.

Le principe exposé ci-dessus est commun à la plupart des méthodes de segmentation. A partir de celui ci les variations sont liées aux facteurs suivants:

- aptitude à traiter des variables explicatives uniquement qualitatives (méthode ELISEE in BOUROCHE et TENENHAUS, 1970), uniquement quantitatives (méthode AID, SONQUIST et MORGAN, 1964) ou indifféremment l'une et l'autre (méthode CART, BREIMAN et al, 1984);
- indice d'impureté choisi; ce choix, d'après BREIMAN et Al, n'aurait qu'une importance limitée sur le résultat final de l'analyse;
- aptitude à traiter les données manquantes
- aptitude à tenir compte d'une hiérarchie dans la gravité des erreurs de classement (notion de "coût de mauvaise classification" introduite par BREIMAN et Al)
- Critère d'arrêt des dichotomies permettant de déclarer terminal un noeud (voir chapitre suivant)

En outre, certaines méthodes (ex: CART) offrent la possibilité d'utiliser des combinaisons linéaires de variables comme conditions logiques définissant une dichotomie.

1.3. Problèmes posés par les méthodes de segmentation

L'interprétation des résultats issus d'une analyse de données par segmentation soulève deux principaux problèmes:

- Comment décider qu'un noeud de l'arbre est terminal ? ou, en d'autres termes, comment choisir un arbre ayant la taille optimale capable de fournir des prédictions à la fois précises et stables ?
- Quelle règle utiliser pour affecter les noeuds terminaux à l'une ou l'autre des classes ?

1.3.1. Choix de la taille optimale de l'arbre de classification

Ce problème, évoqué par de nombreux auteurs traitant de segmentation, est illustré par un exemple cité par BREIMAN et Al (1984). A partir d'un même ensemble d'apprentissage, des arbres de taille différente ont été obtenus en arrêtant plus ou moins tôt les dichotomies. Sur ces arbres ont été calculés deux estimateurs de l'erreur de classement globale.

Le premier, noté R(T), est calculé directement sur l'ensemble d'apprentissage à partir de la démarche suivante:

- soit $T_1, T_2, \ldots T_g, \ldots T_s$ les noeuds terminaux formant la partition résultat de l'analyse,
-soit pr_{gj} la proportion d'individus tombant dans le noeud terminal T_g et appartenant à la classe U_j ,
-soit $r(T_g)$ l'estimateur de l'erreur de classement des individus tombant dans le noeud T_g

$$r(T_g) = 1 - \max_{(j)} (pr_{gj}) \quad [13]$$

L'erreur (relative) de classement donnée par la formule [13] correspond en fait à la somme des proportions des classes autres que celle à laquelle est affectée le noeud terminal.

R(T) est obtenu à partir des erreurs élémentaires $r(T_g)$ calculées sur chaque noeud en faisant leur moyenne sur l'ensemble des noeuds terminaux. Cette moyenne est pondérée par le nombre d'individus N_g tombant dans chaque noeud soit:

$$R(T) = 1/N \, \Sigma_{(g=1...s)} \, N_g.r(T_g) \quad [14]$$

Le deuxième, noté $R^{ts}(T)$, est calculé par la même formule sur un échantillon test, distinct de l'ensemble d'apprentissage, donc n'ayant pas été utilisé pour la construction de l'arbre.

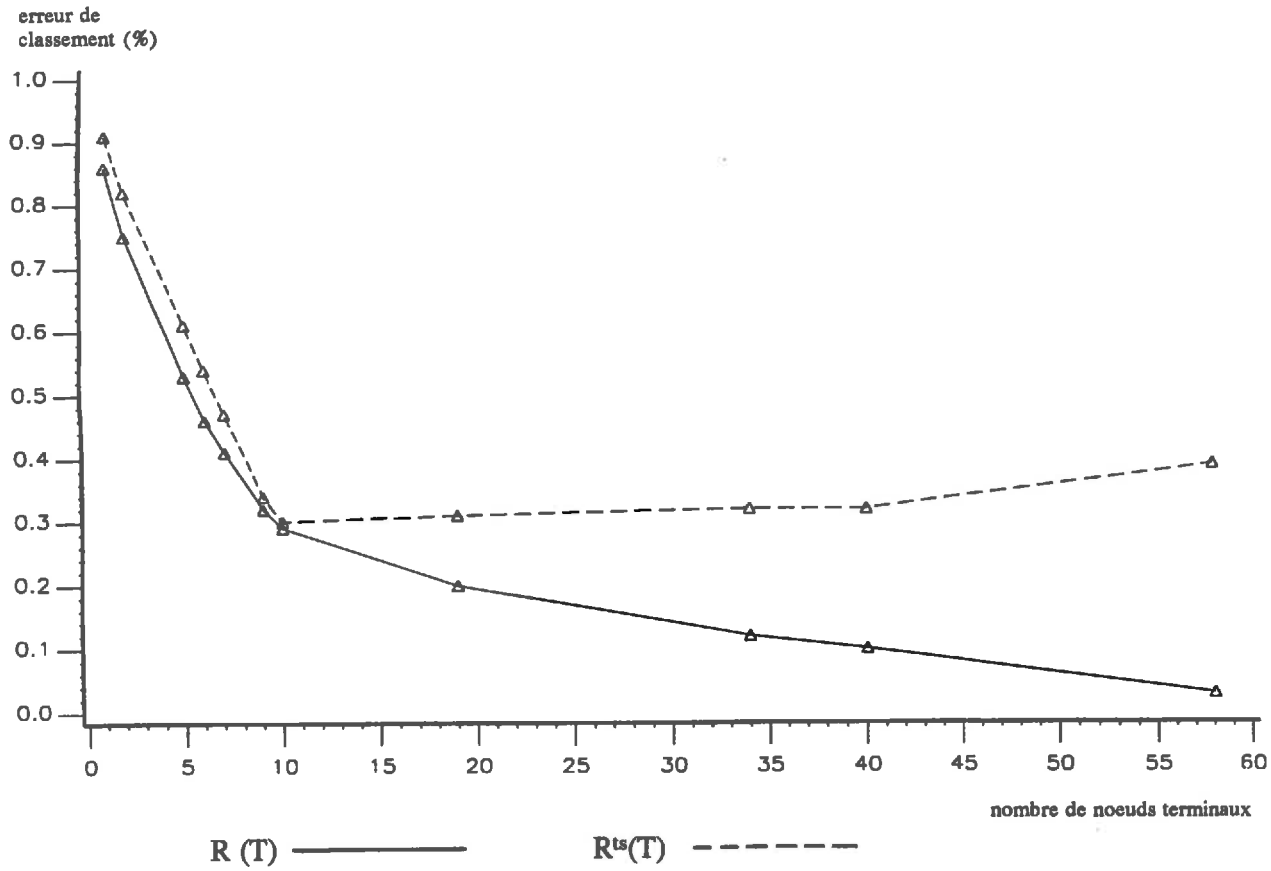


figure 13: évolution de R(T) et $R^{ts}(T)$ avec la taille de l'arbre de classification

L'examen des résultats (figure 13) révèle une grande différence entre les deux estimateurs. Le premier, R(T) décroît régulièrement avec l'augmentation de la taille de l'arbre. Le second, $R^{ts}(T)$, passe par un minimum (nombre de noeuds = 10) à la suite duquel les performances des arbres se stabilisent puis se dégradent.

Ce résultat, systématiquement observé, révèle que le gain de performance mesuré par R(T) sur les arbres dont le nombre de noeuds terminaux est supérieur à une taille critique (ici 10), est fictif puisqu'il n'est pas confirmé quand on met l'arbre en situation de prédiction sur un ensemble distinct de l'ensemble d'apprentissage. Il existe donc un réel danger de mal interpréter les arbres produits dans le cas où aucun élément objectif ne permet d'arrêter la croissance de l'arbre.

En utilisant une règle d'arrêt fixée à priori comme le propose la plupart des méthodes (BOUROCHE et TENENHAUS, 1970), le risque est double: soit l'arrêt est prématuré (dans l'exemple étudié, avant 10 noeuds) et la précision des prédictions est limitée par rapport au potentiel

offert par l'ensemble d'apprentissage; soit l'arrêt est tardif et les prédictions, apparemment très précises, ne sont pas reproductibles.

La méthode utilisée par CART propose une alternative qui consiste à laisser dans un premier temps l'arbre croître sans règles d'arrêt puis, dans un deuxième temps, à éliminer a posteriori des dichotomies sur la base d'un estimateur de l'erreur de classement calculé sur un ou des ensembles indépendants de l'ensemble d'apprentissage. Dans le même esprit, CIAMPI et THIFFAULT (1988) proposent dans le cadre de la méthode RECPAM, de choisir la taille de l'arbre sur la base de mesures de stabilité des prédictions d'un arbre appliqué à différents échantillons de données.

Ce type de solution, fort coûteuse en temps calcul, est particulièrement adaptée dans les cas où la connaissance des données a priori n'est pas suffisante pour permettre de raisonner une règle d'arrêt. Il faut cependant s'assurer que les individus constituant l'ensemble de validation, à partir duquel est calculé le nouvel estimateur, sont véritablement indépendants des individus de l'ensemble d'apprentissage utilisé pour construire l'arbre (BREIMAN et al, 1984). Cette restriction doit être considérée sérieusement car, le plus souvent, les deux ensembles sont issus du découpage d'un même ensemble de départ.

1.3.2. Affectation des noeuds terminaux à l'une des classes

La plupart des méthodes de segmentation affectent les noeuds terminaux T_g à la classe (notée U_L) dont la proportion d'individus est la plus forte du noeud,soit:

$$pr_{gl} = \max_{(j)} pr_{gj} \quad [15]$$

Cette règle pose un problème déjà évoqué en première partie à l'occasion de la reformulation du modèle de raisonnement cartographique. Elle constitue en effet une réduction drastique de l'information contenue dans chaque noeud puisqu'il s'agit de remplacer une fonction de densité estimée par son maximum. Dans le cas où, par exemple, 2 ou 3 classes se dégagent significativement au sein de la population du noeud, choisir une parmi trois sur la base d'une faible écart de population peut paraître discutable.

Dès lors, il faut s'interroger sur le niveau de pureté requis à partir duquel il est licite d'affecter un noeud terminal à une classe donnée, au risque d'accepter une (ou plusieurs) non-affectation réduisant l'intérêt de l'analyse. Les articles consultés traitant de segmentation restent cependant muets sur ce sujet, laissant au lecteur le soin de se forger sa propre règle.

2. METHODE D'INTERPRETATION DES RESULTATS DE SEGMENTATION ADAPTEE A DES DONNEES GEOGRAPHIQUES

L'objectif de la segmentation et la forme des résultats obtenus laisse penser qu'elle constitue une méthode adaptée au problème de formalisation des lois sols-paysage, les classes U_j correspondant aux unités de sol qu'il faut prédire. Pour mettre en oeuvre cette méthode, il convient cependant de ne pas négliger le fait que toutes les données manipulées sont issues de cartes, via un Système d'Information Géographique.

En effet, cette particularité introduit une difficulté supplémentaire. Alors que, dans le cas général, la segmentation s'appuie sur l'existence d'un ensemble d'apprentissage constitué de cas

connus avec certitude (ou du moins raisonnablement supposés tels), il n'en va pas de même dans le cas de points de l'espace géographique définis par les variables topo-géologiques. En effet, ici, le processus de fabrication de cas "connus" comporte en fait de nombreux risques d'erreurs. Ils apparaissent à toutes ses étapes, depuis la fabrication des cartes nécessaires jusqu'aux manipulations effectuées au sein du SIG (croisement,interpolation,...) et se propagent d'une étape à la suivante.

La présence de ces erreurs oblige à ne pas négliger l'incertitude attachée aux valeurs estimées des variables caractérisant chaque point de l'ensemble d'apprentissage. Cette incertitude se propage en effet aux proportions d'individus (pr_{mj}) utilisées dans les processus de choix et d'arrêt des dichotomies: Comment tenir compte de ce fait pour interpréter les résultats d'une segmentation, et, en particulier, pour choisir un critère d'arrêt des dichotomies ? Ce sous chapitre présente la démarche conçue pour tenter de répondre à cette question:

- dans un premier temps, sera défini un critère permettant de stopper les dichotomies; ce critère est basé sur la prise en compte de l'incertitude associée au calcul de l'indice d'impureté $i(X_m)$;
- dans un deuxième temps seront résumés les principes de calcul permettant d'estimer l'ordre de grandeur de cette incertitude; ils seront par ailleurs développés, avec les calculs correspondant, dans deux annexes spécifiques (annexe 5 et 5bis).

2.1. Nature aléatoire de $i(X_m)$ et définition d'un critère d'arrêt des dichotomies

L'examen de la formule de calcul de l'indice d'impureté $i(X_m)$ (formule [10]), qui évalue l'intérêt d'une dichotomie, indique que ce calcul est effectué à partir des termes N_m et N_{mj}, respectivement "nombre d'individus présents dans X_m" et "nombre d'individus présents dans X_m appartenant à l'unité de sol U_j".

Or, l'affectation des individus à tout noeud X_m et au sous-ensemble de X_m rassemblant les membres de l'unité U_j comporte un risque d'erreur. En effet, dans le cas où une variable topo-géologique participe à la définition d'un noeud X_m, une erreur sur sa valeur au point x_i peut entraîner l'affectation ou l'exclusion injustifiées du vecteur de ce point dans le noeud X_m. Ceci aura pour conséquence une erreur de dénombrement sur N_m et sur N_{mj}. Si, par exemple, sur un point situé réellement à l'altitude 49m on estime son altitude à 51m, il y aura une erreur de dénombrement sur le noeud défini par le critère "altitude inférieure à 50m". Si, par ailleurs, le point appartient à l'unité 8, une autre erreur de dénombrement affectera le sous-ensemble des membres de cette unité.

De plus, une erreur sur l'appartenance d'un point à une unité de sol U_j induira une erreur, cette fois seulement sur la valeur de N_{mj}. Par exemple, si, à cause d'une position de limite erronée un point, réellement dans l'unité 9, est affecté par erreur à l'unité 8, il y aura une erreur de dénombrement sur les deux sous-ensembles des membres des unités 8 et 9. Par contre, le dénombrement du noeud contenant ces deux sous-ensembles ne sera pas affecté par l'évènement.

Les erreurs concernant les points x_i se répétant systématiquement, ceci amène à abandonner l'espoir de connaitre les valeurs exactes de N_m et N_{mj}. Ces termes doivent donc être considérés comme des variables aléatoires. Ceci est d'autant plus vrai que l'on s'éloigne de la racine puisque, de nouvelles propriétés étant considérées à l'occasion de nouvelles dichotomies, elles induisent à leur tour des erreurs qui s'ajoutent aux précédentes, propagées depuis l'amont de l'arbre.

La conséquence directe de cet état de fait est qu'il faut également considérer l'indice d'impureté $i(X_m)$ comme une variable aléatoire puisqu'il est fonction de N_m et N_{mj}. Si on suppose, d'une part, que l'erreur sur $i(X_m)$ ne comporte pas de biais et, d'autre part, que la distribution de

$i(X_m)$ suit une loi normale, ce risque peut être estimé à partir de la connaissance de l'écart type d'erreur caractérisant la dispersion des valeurs de $i(X_m)$ que l'on obtiendrait s'il s'avérait possible de répéter plusieurs fois le processus permettant de le calculer (depuis la réalisation des cartes jusqu'à la formule [10]). Cet écart type d'erreur sera noté par la suite $Se[i(X_m)]$.

Le fait que $i(X_m)$ ne puisse pas être connu avec certitude amène à remettre en cause l'intérêt de certaines dichotomies et donc, permet de définir un critère d'arrêt de celles ci. Il semble en effet raisonnable de rejeter une dichotomie produisant un gain de pureté qui aurait de "bonnes" chances d'être dans le cas suivant:

$$i(X_m) - i^r(X_m) > \Delta[i(X_m)] \qquad [16]$$

avec: $i^r(X_m)$: indice d'impureté sur X_m réel
$i(X_m)$: indice d'impureté sur X_m estimé

En effet, la formule [16] présente le cas particulier où on minimiserait plus efficacement l'indice d'impureté $i(X_m)$ en recherchant sa valeur exacte plutôt qu'en lui appliquant la dichotomie proposée. On est donc dans le cas où le gain de pureté n'est pas intéressant compte tenu de l'erreur d'estimation commise par ailleurs, représentée par le premier terme.

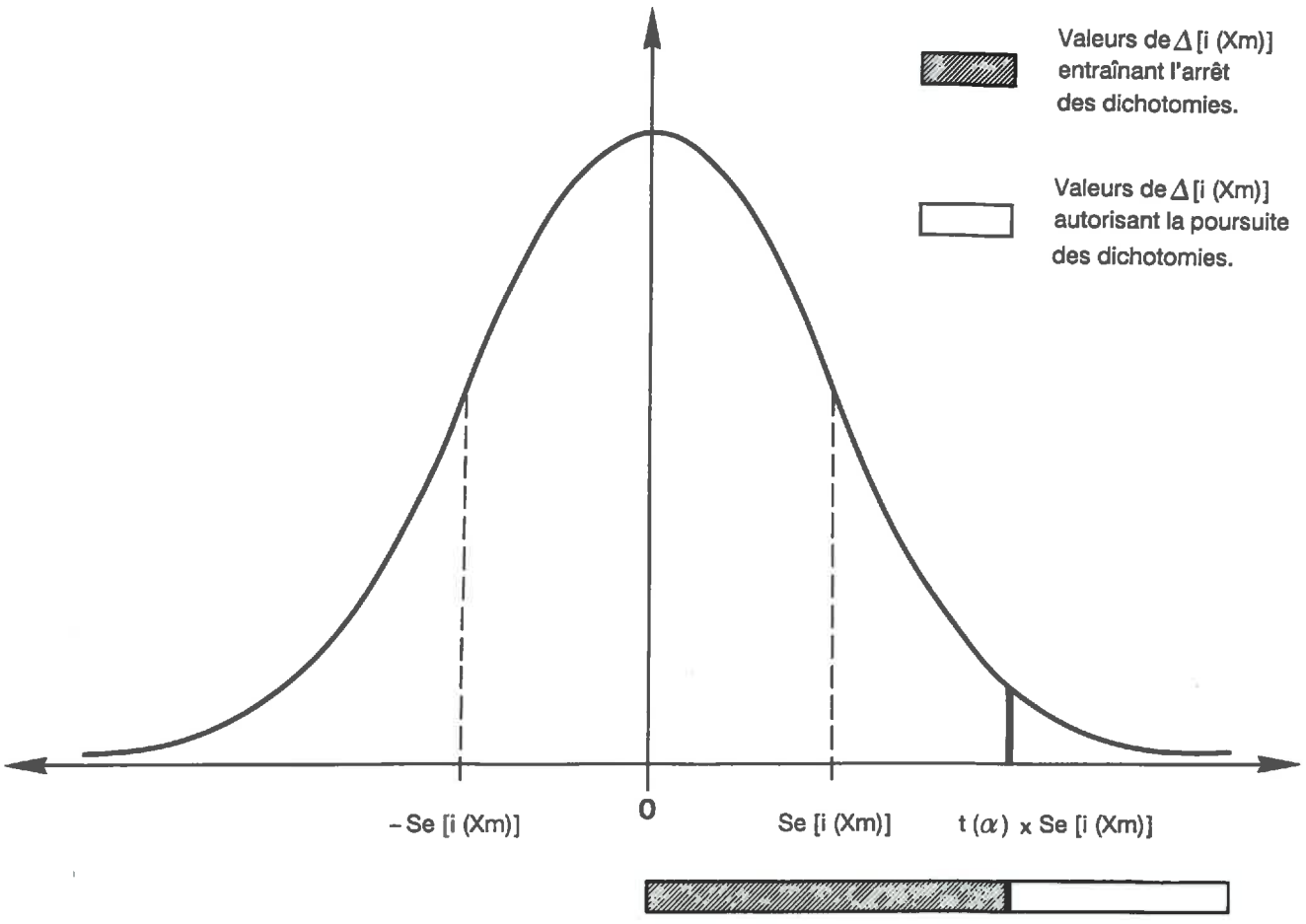


Figure 14: Définition d'un critère d'arrêt aux dichotomies.

Si l'on suppose que cette erreur est sans biais et unimodale, la loi normale centrée $N[0,Se(i(X_m)]$ constitue une approximation raisonnable de sa distribution. Il est dès lors possible d'évaluer les chances de réalisation de l'inégalité [16] et de raisonner un critère d'arrêt des dichotomies par rapport à elles (figure 14):

$$\text{Si } \Delta[i(X_m)] > t(\alpha)\ Se[i(X_m)] \Longrightarrow X_m \text{ est terminal} \quad [17]$$

Ceci revient à ne retenir que les dichotomies dont le gain de pureté est suffisamment grand pour que la situation décrite par l'inégalité [16] ait une probabilité d'apparition inférieure à α. La sévérité du critère d'arrêt peut varier suivant la valeur donnée au seuil de probabilité α, les valeurs $t(\alpha)$ étant lues à partir de la table numérique de la fonction de répartition de la loi normale.

D'autre part, il faut noter que, à α égal, plus l'incertitude sur X_m augmente, plus le gain de pureté devra être élevé pour que la dichotomie correspondante soit acceptée. Bien entendu, l'écart type d'erreur sur le terme $i(X_m)$ $(Se[i(X_m)])$ est différent pour chaque noeud, suivant les variables permettant de définir ce noeud et l'importance des erreurs associées. Le critère d'arrêt proposé par la formule [17] ne peut donc fonctionner que si l'on est capable d'estimer $Se(i(X_m)]$ pour tous les noeuds de l'arbre.

Ce critère d'arrêt suppose en outre que l'utilisateur fixe lui-même le seuil de probabilité α. En l'absence de références en la matière, il peut être intéressant d'adopter une méthode proche de celle de BREIMAN et al: dans un premier temps, construire un arbre avec un maximum de noeud terminaux puis, dans un deuxième temps, le réduire au vu des performances de chaque dichotomie. C'est la méthode qui sera appliquée dans le chapitre 6. Dans cette perspective, on caractérisa chaque dichotomie appliquée sur X_m par son quantile, appelé par la suite **risque de non pertinence** et noté α_m. Il est défini à partir de la formule précédente:

$$\alpha_m \,/\, t(\alpha_m) = \Delta[i(X_m)]/Se[i(X_m)] \qquad [18]$$

Cette opération étant réalisée pour toutes les dichotomies de l'arbre maximum, α_m représente un critère permettant de hiérarchiser les dichotomies en fonction de leur intérêt réel et d'étudier les conséquences des choix du seuil de probabilité α (voir chapitre 6).

2.2. Principes d'estimation de l'écart type d'erreur sur l'indice d'impureté $i(X_m)$

En l'absence de références bibliographiques correspondant à cet exercice, le principe adopté pour estimer l'écart type d'erreur sur $i(X_m)$ est le suivant:

- faire l'inventaire le plus exhaustif possible des différentes étapes du processus aboutissant au calcul de l'indice d'impureté;
- pour chaque étape, fournir, sur la base de la bibliographie existante en la matière, un écart type d'erreur élémentaire;
- proposer, à l'aide d'hypothèses simplificatrices, un calcul permettant de déduire, in fine, l'écart type d'erreur sur $i(X_m)$ à partir des différents écarts types d'erreurs caractérisant chaque étape de son processus de calcul.

L' annexe 5 présente et illustre la démarche avec, en particulier les références bibliographiques utilisées, les hypothèses simplificatrices avancées et la justification des choix

réalisés. Sont également développés (annexes 5 et 5bis) les calculs correspondants. Dans ce paragraphe, seuls les principes des calculs sont résumés suivant trois volets qui correspondent aux trois phases distinguées en annexe 5.

2.2.1. Estimation de l'erreur sur les individus de l'ensemble d'apprentissage

Les individus de l'ensemble d'apprentissage sont caractérisés à la fois par les variables topo-géologiques et par la variable sol recherchée (U_j). Suivant que ces variables sont quantitatives ou qualitatives, l'expression de l'erreur affectant chaque individu sera différente. Sur des variables quantitatives (variables de relief et distance aux cours d'eau), il conviendra simplement de définir un écart type d'erreur. Sur des variables qualitatives ("unité de sol", "rive", "unité géologique") ne sera considérée que l'une des formes d'erreur possible, supposée prépondérante (annexe 5). Elle correspond à l'erreur d'affectation d'un point à une unité cartographique suite à une position erronée de limite. Pour évaluer cette erreur sera donc recherché, pour chacunes des cartes numérisées sous ARC/INFO, l'écart type d'erreur sur la position des limites.

Code de la variable	Nom de la variable	Ecart type d'erreur sur variable (var. quantitative)	Ecart type d'erreur sur position limite (var. qualitative)
u (x)	unité de sol	-	entre 13 et 21 m
g (x)	unité géologique	-	entre 55 et 89 m
rv (x)	rive de l'Hérault	-	entre 9 et 17 m
z (x)	altitude	1.3m	-
dz (x)	dénivelé/Hérault	1.3m	-
di (x)	distance au cours d'eau	entre 9 et 17 m	-
pt (x)	pente	0.6°	-
cm (x)	courbure moyenne	1.5° par 50m	-
ec (x)	encaissement	11m	-

tableau 6: estimation des erreurs sur les variables utilisées dans l'analyse de segmentation

Pour calculer ces écart types d'erreur, quelle que soit leur nature, il faut tenir compte des différentes étapes ayant permis de fournir les couvertures ARC/INFO correspondantes soit, dans l'ordre:

- le travail du cartographe de terrain,
- la fabrication de la carte (dessin, impression),
- la numérisation manuelle,

- les calculs d'interpolation et de dérivation (pour les variables issues du MNT).

Les erreurs attachées aux différentes étapes seront supposées additives et indépendantes. Il sera donc possible de déduire les écarts types d'erreur recherchés en appliquant la formule simplifiée correspondant à la propagation d'erreur dans l'addition de termes indépendants. Soit:

$$Se(y) = \sqrt{[\Sigma_{(it=1...nt)} Se^2(y_{it})]} \qquad [19]$$

avec: $y = \Sigma_{(it=1...nt)} y_{it}$
y_{it}: $it^{ème}$ composante de l'expression de y (it=1,...,nt)

Les résultats de ces calculs sont fournis dans le tableau 6. Beaucoup de résultats de ce tableau sont donnés sous forme d'intervalles de valeurs. Celà correspond à des cas où les références bibliographiques ne permettent pas d'approcher avec plus de précision les écarts type d'erreur. C'est en particulier le cas dans l'appréciation de l'erreur du cartographe pour laquelle il n'existe pas de références expérimentales sur lesquelles s'appuyer.

2.2.2 Estimation de l'écart type d'erreur sur les dénombrements des noeuds (N_m et N_{mj})

Les erreurs sur les variables définissant les points se propagent aux dénombrements des populations des noeuds (N_m), et, au sein de chaque noeud, aux dénombrements des populations de chaque unité de sol (N_{mj}). L'objectif de cette phase est d'utiliser les résultats du tableau 6 pour calculer les écarts types d'erreurs sur ces dénombrements.

En l'absence de modèle de propagation d'erreur disponible dans la littérature, l'estimation de ces écarts type a été réalisée suivant une démarche particulière, détaillée en annexe 5 et résumée ici.

Pour ce qui concerne l'écart type d'erreur sur N_m, le principe général est d'estimer sa valeur en comptant les points situés sur les marges du noeud. Ces marges appelées "domaine d'incertitude" sont créées par les dichotomies successives. En effet, chacune d'elles, en séparant un noeud en deux nouveaux, crée à cette occasion une population de points pour lesquels l'affectation dans un noeud ou dans un autre est douteuse, compte tenu de l'erreur connue sur la variable ayant défini la dichotomie.

Pour dimensionner ces marges de façon à ce que leurs dénombrements puissent être assimilés aux écarts type d'erreur sur les dénombrements des noeuds X_m, elles sont définies de la façon suivante, deux cas étant à distinguer.

1) Lorsque une dichotomie contribuant à définir X_m (c'est à dire en amont de ce noeud) s'appuie sur une propriété impliquant une variable quantitative, sont considérés comme membres du domaine d'incertitude de X_m les individus appartenant à X_m et vérifiant la propriété suivante:

$$o_h^* - Se[o_h(x_i)] < o_h(x_i) < o_h^* + Se[o_h(x_i)] \qquad [20]$$

avec: o_h^*, valeur seuil de la variable quantitative $o_h(x_i)$ utilisée pour définir la dichotomie

$Se[o_h(x_i)]$, écart type d'erreur pour la variable $o_h(x_i)$ (donné par le tableau 6)

Ainsi, par exemple, si une dichotomie en amont de X_m impose que la propriété "altitude $\leq$ 50m" soit vérifiée, tous les points de X_m dont les valeurs sont comprises entre 48.7 (50m - 1.3m) et 50m (valeur seuil) seront inclus dans le domaine d'incertitude (la valeur "1,3m", lue dans le tableau 6, correspond à l'écart type d'erreur sur l'altitude). Les points dont les valeurs sont comprises entre 50m et 51.3m, n'appartenant pas à X_m, n'appartiendront pas non plus à ce domaine. Ils alimenteront par contre les domaines d'incertitude des noeuds vérifiant la propriété inverse (altitude > 50m).

2) Lorsqu'une dichotomie contribuant à définir X_m s'appuie une propriété impliquant une variable qualitative, ceci revient à sélectionner les membres de X_m suivant leur appartenance à une zone géographique correspondant à un groupe d'unités cartographiques. Dans ce cas, seront considérés comme membres du domaine d'incertitude les points de X_m situés à proximité des limites de la zone géographique en question. Pour qu'ils puissent être par la suite décomptés dans le but d'estimer un écart type d'erreur, les points "à proximité" seront définis comme suit: ensemble des points dont la distance minimale à une limite (de groupe d'unités cartographiques) est inférieure à l'écart type d'erreur de position sur les limites (donné par le tableau 6). Par exemple, si une dichotomie en amont de X_m impose que les points de X_m soient dans l'unité géologique 1 ou 2, seront membres du domaine d'incertitude tous les points situés à moins de 55m (ou 89 m suivant qu'on prend l'hypothèse haute ou basse), des limites séparant les unités 1 et 2 de leurs voisines.

L'écart type d'erreur sur N_m étant ainsi défini, celui sur N_{mj} procède de la même logique. La différence est, qu'en plus des points satisfaisant à l'une des deux conditions citées ci-dessus, seront également membres du domaine d'incertitude du sous-ensemble des membres de l'unité U_j, les points situés à moins de 13 ou 21m (cf tableau 6) d'une limite d'unité de sol.

Compte tenu du fait que les écarts types d'erreurs sur certaines variables sont donnés par des intervalles de valeurs, deux calculs parallèles pour un même noeud seront effectués, l'un prenant systématiquement les valeurs les plus fortes, l'autre les plus faibles. En conséquence, les écarts type d'erreur sur N_m et N_{mj} seront également fournis sous forme d'intervalles de valeurs.

2.2.3. Estimation de l'écart type d'erreur sur l'indice d'impureté $i(X_m)$

Le calcul de $i(X_m)$ est réalisé à partir d'une opération arithmétique impliquant N_m et N_{mj}. Connaissant les écarts type de ces termes, il est possible d'appliquer la formule générale de propagation des erreurs au travers d'une opération arithmétique. Concrètement, le calcul est effectué en deux temps:

- calcul de $Se(pr_{mj})$ (écart type d'erreur sur les proportions d'unités de sol) à partir de $Se(N_m)$ et $Se(N_{mj})$:

$$Se^2(pr_{mj}) = pr_{mj} \times \left[\frac{Se^2(N_m)}{N_m} + \frac{Se^2(N_{mj})}{N_{mj}} - \left[\frac{Se^2(N_m)}{N_m} \times \frac{Se^2(N_{mj})}{N_{mj}} \right] \right] \qquad [21]$$

- calcul de $Se[i(X_m)]$ à partir de $Se(pr_{mj})$:

$$Se\left[i(X_m)\right] = 2 \sum_{j=1}^{j=v} pr_{mj} \times Se(pr_{mj}) \qquad [22]$$

Pour effectuer ces calculs, des hypothèses simplificatrices ont du être avancées: Les termes N_m et N_{mj} ont été supposés complètement dépendants (coefficient de correlation = 1) et les différents termes pr_{mj} ont été supposés de leur côté indépendants deux à deux.

Comme précédemment, deux valeurs, minimum et maximum, de $Se[i(X_m)]$ seront calculées en prenant tour à tour les 2 valeurs minimales, puis maximales de N_m et N_{mj}. Le résultat du calcul de $Se[i(X_m)]$ se présentera donc aussi sous forme d'intervalles de valeurs.

Au cours du chapitre 5, a été choisie une méthode d'analyse de données susceptible d'utiliser le fichier des points inclus dans le secteur de référence. Cette méthode, connue sous le nom de "segmentation", permet de dégager des prédictions sur les unités de sols à partir de variables extraites de documents disponibles (carte topographique et carte géologique).

La prise en compte de l'erreur systématique sur les données issues de cartes et manipulées au sein d'un SIG a permis de proposer une solution originale au problème du choix d'un critère d'arrêt au dichotomies, principal problème soulevé par les méthodes de segmentation. La définition de ce critère s'appuie sur une estimation de l'erreur affectant l'indice d'impureté $i(X_m)$ sur lequel est basé le choix des dichotomies élémentaires de l'arbre de classification. Cette erreur est issue de la propagation des erreurs liées aux processus de fabrication et de manipulation au sein d'un SIG des documents cartographiques, puis des différents calculs au sein même de la méthode de segmentation. Cependant l'estimation de l'erreur sur $i(X_m)$ reste très approximative compte tenu de l'absence de données expérimentales et de modèles théoriques complets pouvant la conforter. Dès lors, il parait plus réaliste d'avancer des intervalles de valeur plutôt que d'essayer de choisir une seule valeur fortement discutable compte tenu du contexte de son obtention. Dans le chapitre suivant, sera présenté la mise en oeuvre de ce critère avec en particulier la gestion du formalisme en intervalles de valeurs.

CHAPITRE 6

APPLICATION DE LA SEGMENTATION A LA FORMALISATION DES LOIS SOLS-PAYSAGE EN MOYENNE VALLEE DE L'HERAULT

Au cours du chapitre précédent une méthode d'analyse de données, la segmentation, a été choisie dans la perspective de dégager des prédictions d'unités de sols à partir de variables topographiques et géologiques. Son utilisation a été adaptée afin de tenir compte de la nature géographique, donc incertaine, des données traitées. Il convient maintenant de revenir à la petite région naturelle choisie dans la première partie, la Moyenne Vallée de l'Hérault, afin de dégager, à partir du secteur de référence, des prédictions d'unités de sols suceptibles d'être appliquées sur la petite région naturelle qu'il est censé représenter. Cette application à un terrain expérimental permettra d'aborder les questions suivantes.

- Au niveau de la méthodologie d'analyse de données, la segmentation et ses adaptations au contexte géographique s'avèrent-elles pertinentes ?
- Au niveau de l'analyse du milieu expérimental, les variables choisies a priori sont elles également pertinentes? Quelle est la justesse et la précision des prédictions obtenues? Sont elles de qualité homogène sur l'ensemble du territoire?
- Quel va être l'intérêt des règles sols-paysage ainsi dégagées et leurs modalités d'utilisation au sein de l'outil informatique représentant le retour à la parcelle.?

La démarche choisie pour aborder ces problèmes peut être divisée en deux étapes.

1) Dans un premier temps, les prédictions seront élaborées à partir des données du secteur de référence. A ce niveau, différentes tailles d'arbres de classement seront retenues a priori pour être testées. Une analyse comparative en sera effectuée afin de dégager des critères de qualité perceptibles a priori et susceptibles d'orienter le choix de l'utilisateur pour l'une ou l'autre de ces tailles,
2) Dans un deuxième temps, la qualité des prédictions, pour les différentes tailles d'arbres, sera mesurée sur les trois secteurs de validation présentés au chapitre 33 de la première partie. Compte tenu de ces résultats, seront discutées et éventuellement remises en cause, les modalités d'interprétation et d'utilisation de la méthode de segmentation.

1. PRODUCTION ET SELECTION DES ARBRES DE CLASSIFICATION A PARTIR DES DONNEES DU SECTEUR DE REFERENCE

Les données utilisées dans cette analyse correspondent au fichier des 3779 points du secteur de référence obtenus selon la démarche présentée au cours du premier chapitre de cette partie. L'appartenance des points au secteur de référence permet de renseigner, pour ces points, l'unité de sol. Par ailleurs, comme tous les points de la zone d'extrapolation, ils sont renseignés par des variables topo-géologiques. L'arbre de classement généré par la segmentation sur ces points doit

donc permettre de prédire l'unité de sol (la variable à expliquer) grâce aux variables topo-géologiques (les variables explicatives).

1.1. Mise en oeuvre pratique de la segmentation

Le logiciel CART (BREIMAN et Al, 1984) est conçu pour fonctionner sous système d'exploitation UNIX. Il a pu être utilisé dans ce travail grâce à une collaboration du Laboratoire de Biométrie de l'INRA de Montpellier. Son usage est interactif grâce à des menus permettant à l'utilisateur de fournir au logiciel les informations nécessaires au bon déroulement de l'analyse (définitions des variables utilisées, spécifications des fichiers d'entrée-sortie, choix des options d'affichage de résultats,...).

Ces menus ont également pour fonction de préciser les options choisies par l'utilisateur concernant la méthode de segmentation elle-même (évoquées au chapitre 5). Il a été délibérément choisi de ne pas tester l'ensemble de ces options. En effet, elles concernent chacune un aspect particulier de l'analyse, ce qui nécessiterait le test d'un grand nombre de combinaisons d'options possibles. Ceci constituerait un sujet de recherche à part entière trop éloigné de la problématique envisagée dans ce travail. Par ailleurs, les essais ponctuels concernant certaines de ces options (changement de méthode de mesure de l'impureté des noeuds, utilisation de combinaisons linéaires de variables pour définir les dichotomies...), n'ont jamais modifié profondément les résultats obtenus. Ainsi, les analyses effectuées tout au long de ce travail ont utilisé les options standard, proposées par défaut par le logiciel.

L'utilisation de ces menus nécessite peu de temps (moins de 2mn). De plus, il existe des possibilités de mémorisation des options et spécifications retenues qui accélèrent les manipulations ultérieures. Par contre, la segmentation elle même s'avère fort coûteuse en temps calcul puisqu'il faut attendre environ 15 mn, sur station Spark1, pour que CART délivre ses résultats.

Parallèlement à l'utilisation de CART, un programme FORTRAN (annexe 6) a été élaboré dans le but d'automatiser, au moins partiellement, la recherche du risque de non pertinence des dichotomies de l'arbre α_m (formule [17]) tenant compte de la propagation des erreurs. Il permet de calculer, pour chaque noeud de l'arbre, l'écart type d'erreur moyen sur l'indice d'impureté ($Se[i(X_m)]$). L'utilisateur doit préalablement transférer manuellement la définition de chaque noeud donnée par CART. Le calcul du gain de pureté et la recherche α_m dans les tables de la loi normale restent également manuelles.

Le calcul total nécessite en moyenne 10 mn par noeud. Il serait souhaitable (et possible) de réduire ce temps par une informatisation plus complète du processus.

1.2. Conséquences de l'application des critères d'arrêt sur la taille de l'arbre obtenu

On a vu dans le chapitre précédent l'importance et la difficulté, dans le cadre de la segmentation, de choisir de façon raisonnée une taille optimale de l'arbre de classification. CART utilise une méthode proposée par BREIMAN et al pour résoudre ce problème (cf même chapitre). Par ailleurs, un critère d'arrêt alternatif a été proposé au cours de ce travail afin de tenir compte du cas, supposé particulier, de l'utilisation de données géographiques. Le premier aspect de l'interprétation des résultats sur la Moyenne Vallée de l'Hérault consiste donc à vérifier comment ce nouveau critère

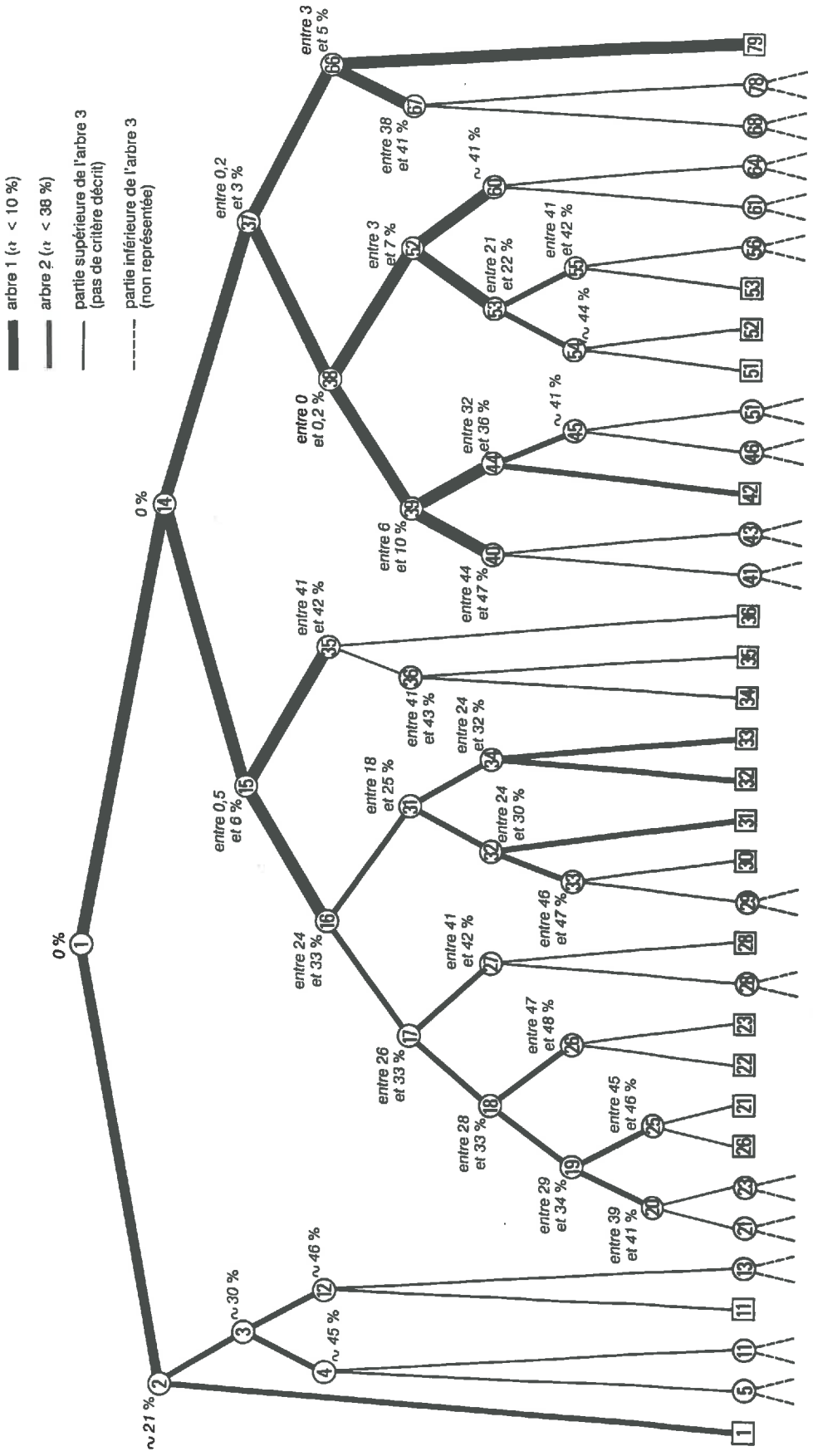


Figure 15: risques de non pertinence des dichotomies de l'arbre de classification et choix des différentes tailles d'arbre.

fonctionne et, éventuellement, modifie les résultats d'une segmentation réalisée selon la méthode de BREIMAN et al.

CART fournit un arbre présentant 79 noeuds terminaux résultant de réductions successives à partir d'un arbre maximum de 568 noeuds. Ces réductions sont effectuées sur la base de l'estimateur de l'erreur de classement $R^{ts}(T)$ calculé sur un échantillon test ("test sample") de 1224 points prélevés sur l'ensemble de départ et n'ayant pas participé à l'analyse.

Pour chaque dichotomies de cet arbre, a été déterminé le risque de non pertinence α_m. Il correspond au risque que le gain de pureté apporté par la dichotomie sur ce noeud soit trop faible vis à vis de l'erreur d'estimation de son indice d'impureté commise par ailleurs (chapitre 5). Compte tenu des incertitudes sur les estimations d'erreur de certaines variables explicatives, pour lesquelles seules des intervalles de valeurs d'écart type d'erreur peuvent être fournies (tableau 6), α_m s'exprime lui aussi sous forme d'un intervalle [7]: le calcul d'erreur utilisant systématiquement les valeurs minimums données par le tableau 6 constitue sa borne inférieure, l'utilisation des valeurs maximum fournissant de même sa borne supérieure.

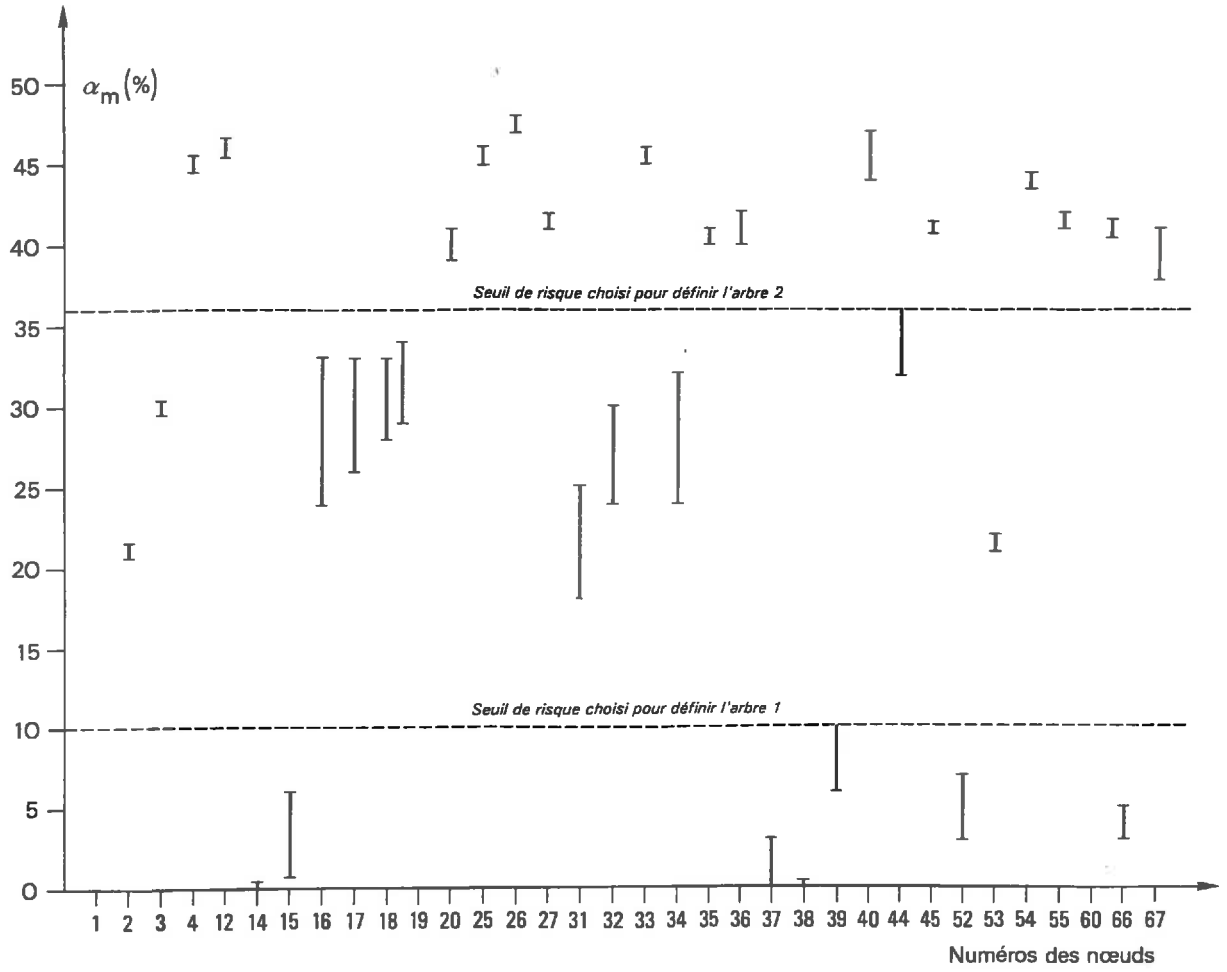


Figure 15bis: visualisation des intervalles de valeurs et mise en évidence des seuils de risque de non pertinence

7 Lorsque l'amplitude de la fourchette des valeurs de α_m est inférieure à 1%, une seule valeur de α_m est donnée

La figure 15 représente la partie supérieure de l'arbre de classement généré par CART. Pour chaque noeud, sont indiqués, dans cette figure, les intervalles de risque de non pertinence des dichotomies les affectant. Ces intervalles sont par ailleurs reportés sur un axe vertical, le numéro des noeuds figurant en abscisse (figure 15bis). Plusieurs remarques peuvent être formulées au vu de ces 2 figures.

1) Les risques de non pertinence s'accroissent au fur et à mesure que l'ensemble de départ se divise. Cette évolution est principalement liée à l'augmentation constante de l'erreur sur l'indice d'impureté. En effet, au fur et à mesure que de nouvelles conditions sont introduites pour définir les noeuds X_m, les erreurs affectant N_m et N_{mj}, deviennent plus importantes. Puisque ces termes entrent dans le calcul de $Se[i(X_m)]$, celui-ci augmente également. Comme par ailleurs l'autre terme de la formule [18], à savoir le gain de pureté, ne suit pas la même évolution, le risque de non pertinence ne peut que croître. Cette règle est cependant mise en défaut pour les dichotomies affectant les noeuds 37 et 38 du fait d'un gain de pureté particulièrement faible apporté par la première des deux. Par ailleurs il a été observé des évolutions inverses dans les dichotomies les plus éloignées de la racine de l'arbre, non représentées sur la figure 15. Il s'agit dans ce cas d'un artéfact de calcul [8] .
2) Le manque de précision des valeurs d'écart type d'erreur sur les variables (tableau 6) ne se répercute pas avec une ampleur telle qu'elle puisse géner l'interprétation des résultats. Les intervalles de valeurs résultant des doubles calculs sont en effet assez resserrés. En particulier, aux 2 extrémités de l'arbre, la détermination du risque de non pertinence de la dichotomie s'avère quasiment insensible au choix de valeurs minimales et maximales d'erreurs sur les variables. le calcul de α_m s'avère donc relativement robuste vis à vis des estimations d'erreurs nécessaires pour l'obtenir. Ce point est important car, compte tenu de la difficulté de préciser cette erreur (annexe 5), un résultat contraire aurait condamné l'emploi de ce critère d'arrêt.
3) La figure 15bis suggère l'existence de deux discontinuités d'évolution du risque de non pertinence α_m correspondant à des niveaux non recoupés par un intervalle de valeur de α_m. Ces niveaux se situent entre 10 et 18% pour le premier, 36 et 38% pour le second. Ceci permet d'identifier 2 sous arbres possibles de l'arbre proposé par CART (figure 15). Ces sous arbres seront retenus a priori pour étudier comment évoluent la nature et la qualité des prédictions suivant la sévérité du critère d'arrêt choisi. Ainsi dans la suite seront nommés:
   - **arbre 1**: arbre à 9 noeuds terminaux dans lequel toute dichotomie présente un risque de non pertinence inférieur à 10% quelles que soit les hypothèses d'erreurs formulées sur les variables;
   - **arbre 2**: arbre à 20 noeuds terminaux limité de la même façon avec un risque de non pertinence inférieur à 36%;
   - **arbre 3**: arbre à 79 noeuds terminaux correspondant à l'arbre initial proposé par CART sans tenir compte de l'incertitude sur les données traitées.

---

8 A partir d'un certain degré de complexité concernant la définition des noeuds, l'erreur calculée sur $i(X_m)$ peut ne pas augmenter. En effet, compte tenu du cumul des erreurs provoquées par les dichotomies en amont, il arrive qu'à un certain niveau, les points des noeuds produits soient tous considérés comme marginaux (chapitre 5), l'écart type d'erreur calculé sur les dénombrements plafonnant alors à 100%. Par voie de conséquence (formules [21] et [22]), l'écart type d'erreur sur $i(X_m)$ stagne également malgré l'augmentation du nombre de dichotomies effectuées. Si, par ailleurs, le gain de pureté $\Delta[i(X_m)]$ augmente entre deux dichotomies consécutives, selon la formule [18], le risque de non pertinence des dichotomies décroît.

Les arbres 1 et 2 constituent une réduction conséquente et significative du nombre de noeuds terminaux par rapport à l'arbre 3. Ainsi, l'application de ce nouveau critère d'arrêt influence fortement le résultat de l'analyse,au moins dans le cas du secteur de référence de la Moyenne Vallée de l'Hérault. Il n'est donc pas redondant avec le critère d'arrêt défini par BREIMAN et al même si, ponctuellement, certains arrêts se situent au même niveau.

1.3. Analyse comparative des arbres sélectionnés et des prédictions de sols fournies

Si, à l'évidence, la taille de l'arbre produit est fortement contingente du critère d'arrêt portant sur le risque de non pertinence, cette variation de taille induit-elle des modifications des prédictions fournies ? Quelles en sont, le cas échéant, la nature et la signification ? Par ailleurs, que retiennent ces arbres des relations sols-paysage inscrites en filigrane sur la carte des sols du secteur de référence de la Moyenne Vallée de l'Hérault? La réponse à ces questions recquiert un examen attentif des résultats de chaque arbre.

Les figures 16 et 17 présentent les arbres 1 et 2 retenus précédemment. L'arbre 3, difficile à représenter compte tenu de sa taille, ne figure pas dans ce mémoire. La lecture des prédictions formulées à partir d'un arbre s'effectue comme suit: les branches successives reliant la racine de l'arbre à chaque noeud terminal fournissent les conditions sur les variables explicatives définissant le noeud terminal. De plus, pour chaque noeud terminal, sont indiquées, par unités de sol, les proportions de points tombant dans ce noeud (exprimée en %).

Si l'on assimile les proportions d'unités de sol à des probabilités, l'arbre permet de fournir des règles sols-paysage dont le formalisme répond aux exigences présentées au chapitre 5. Ainsi, à partir du noeud terminal n°9 de l'arbre 1, il serait possible d'écrire la règle:

$$
\begin{array}{ll}
\text{si} & 2.05\text{m} \leq dz(x) < 29.9\text{m}, \\
\text{si} & rv(x) = \text{"droite"} \\
\text{et si} & g(x) = 4 \text{ ou } 6 \\
\text{alors} & \\
 & \pi_{10}(x) = 0.9,\ \pi_{14}(x) = 0.07,\ \pi_{15}(x) = 0.03 \\
\text{et} & \pi_j(x) = 0, \forall\, j \notin \{10,14,15\}
\end{array}
\qquad [23]
$$

Si la source d'information "sols-paysage" était la seule utilisée pour prédire les sols, la règle [23] permettrait alors de délivrer, pour tous les points satisfaisant à sa prémisse, la prédiction suivante: "unité de sol n° 10 prédite (puisqu'elle obtient, dans le noeud n°9, la majorité relative) avec un risque d'erreur de 10%" (correspondant à la somme des probabilités d'apparition des autres unités).

	Nombre de noeuds	Nombre de variables explic.	Nombre d'unités reconnues	Erreur de classement prévue (R^{ts} (T))
arbre 1	9	3	8	46 %
arbre 2	20	5	13	35 %
arbre 3	79	7	17	28 %

tableau 7: caractéristiques principales des arbres étudiés

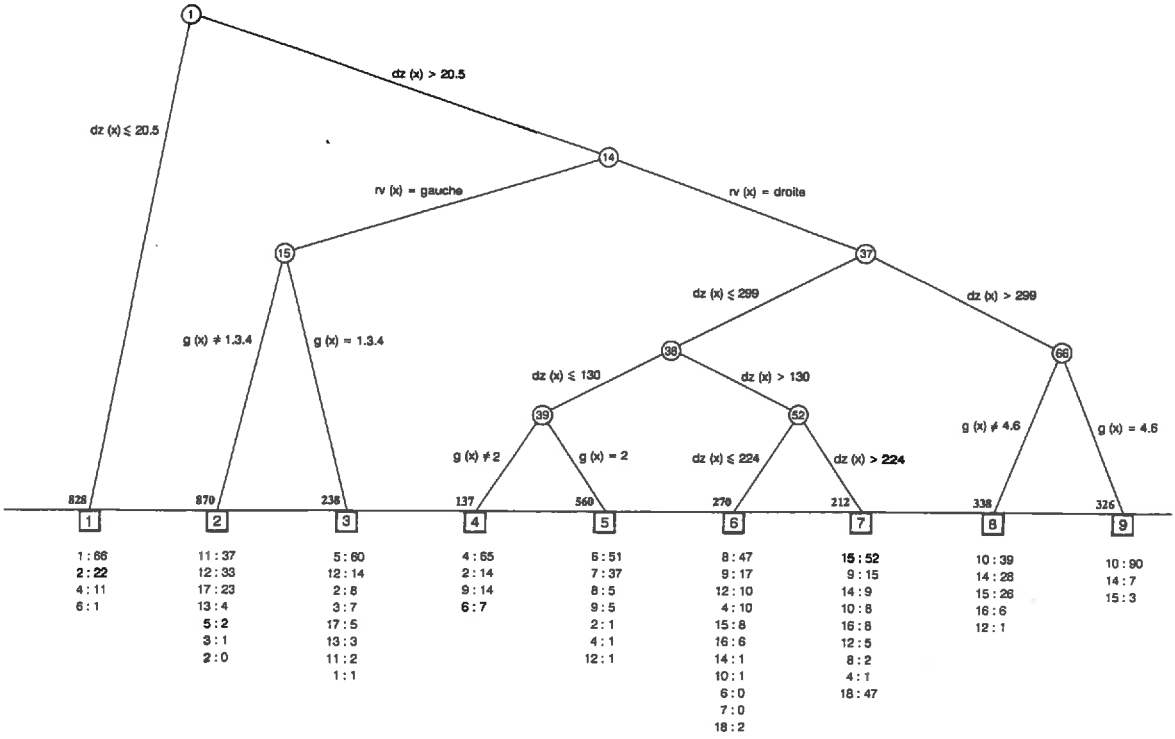


Figure 16: arbre de classification n° 1

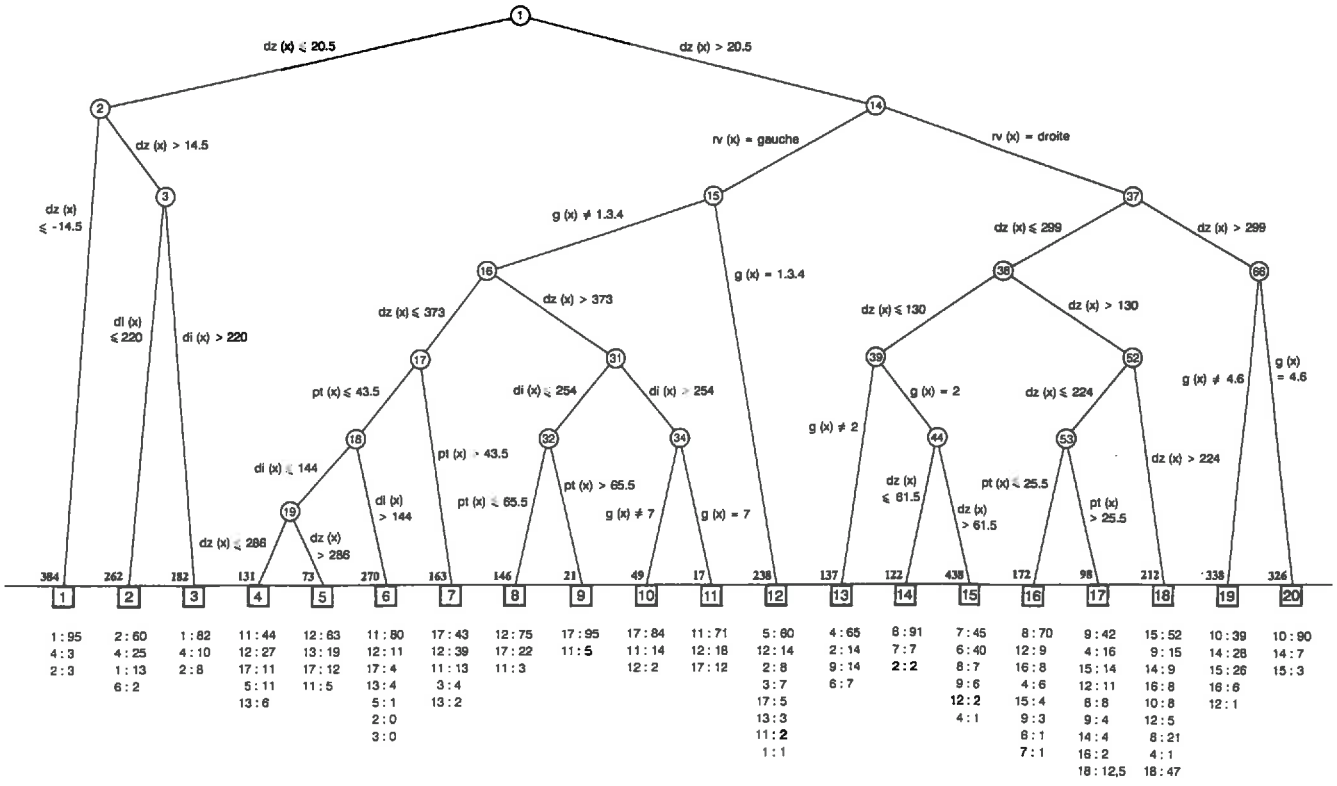


Figure 17: arbre de classification n° 2

C'est cette interprétation, la plus immédiate, de la segmentation qui sera adoptée dans un premier temps. Avant d'examiner, pour chaque arbre le détail des prédictions, il convient de s'arrêter à certains descripteurs globaux rassemblés dans le tableau 7 (page précédente).

Ces descripteurs révèlent l'opposition entre, d'une part, l'arbre 1 fournissant un ensemble de règles "sols-paysage" simple d'emploi (9 noeuds terminaux donc 9 prédictions et seulement 3 variables explicatives mobilisées) mais apparamment peu performant ($R^{ts}(T) = 46\%$, 8 unités reconnues) et, d'autre part, l'arbre 3, beaucoup plus complexe à utiliser (79 noeuds et 7 variables explicatives) mais, semble-t-il, plus efficace ($R^{ts}(T) = 23\%$, 17 unités reconnues). L'arbre 2 représente un moyen terme entre ces 2 extrèmes.

Dans le cadre d'un examen plus poussé des règles sols-paysage et pour compléter la lecture des arbres, des cartes du secteur de référence visualisant les prédictions des règles fournies par chaque arbre ont été réalisées sous ARC/INFO (planche 3). En surimpression, apparaissent également les limites et numéro d'unités de la carte des sols du secteur de référence. il est donc possible, grâce à ce type de document, de confronter la carte réelle et les prédictions émises et de visualiser ainsi les erreurs affectant ces dernières. En celà, la méthode proposée rejoint les démarches d'analyse des erreurs introduites récemment dans le domaine des SIG (CHRISMAN, 1991; JAMET, 1991). Elle en diffère par le fait que la classification des erreurs ne correspondra pas à une classification couramment utilisée dans ce domaine. En effet, même la plus élaborée d'entre elle (JAMET, 1991) s'avère difficile à appliquer au cas étudié.

Les 3 arbres sélectionnés seront examinés en partant du plus simple (arbre 1) au plus complexe (arbre 3). A l'occasion de l'examen du premier arbre, sera fixée la classification des erreurs utilisée dans toute l'analyse.

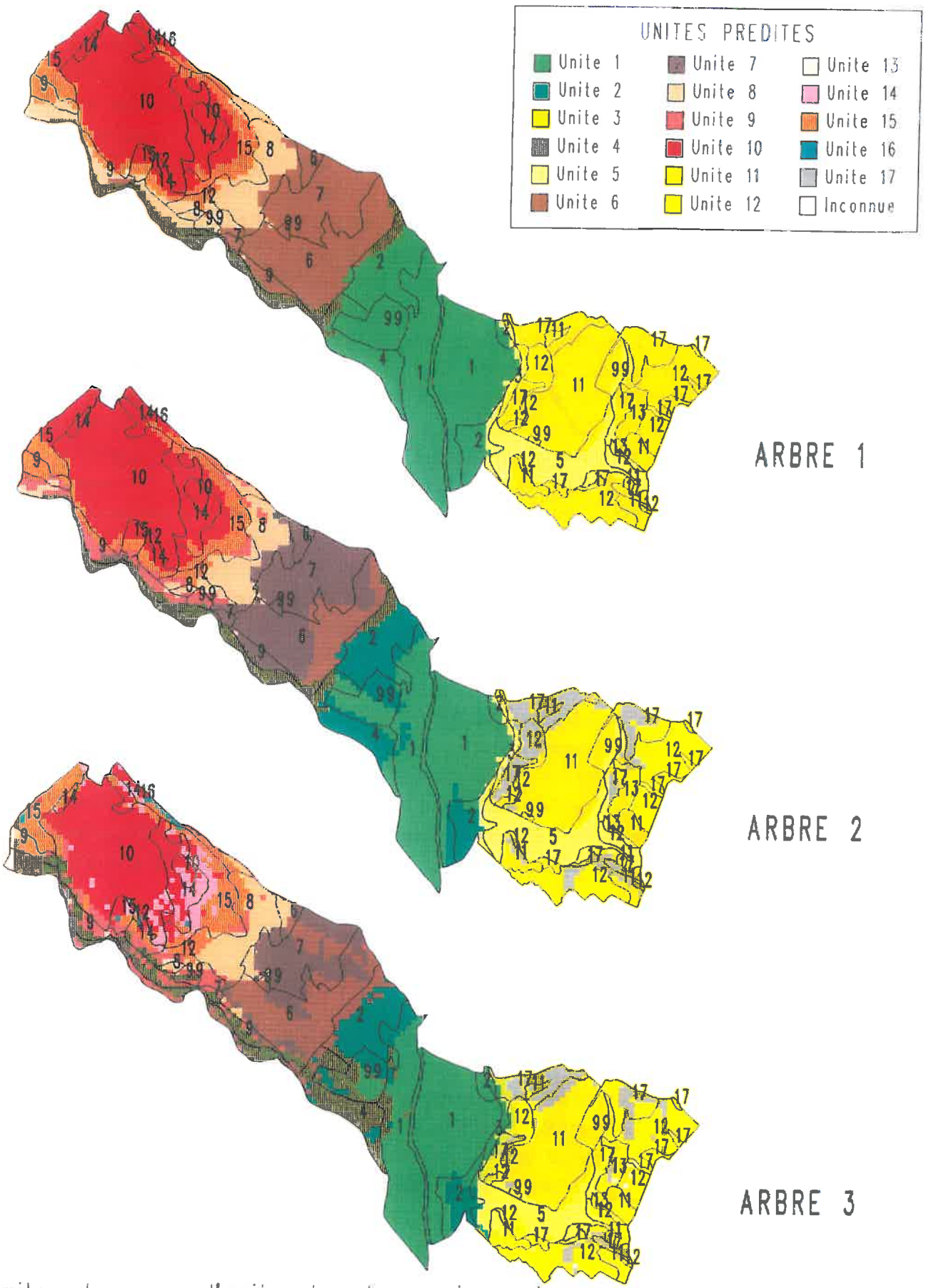
1.3.1. Examen des prédictions de l'arbre 1

Les 9 prédictions de l'arbre 1 utilisent uniquement 3 variables explicatives: rive (rv(x)), altitude relative par rapport à l'Hérault (dz(x)), et unité géologique (g(x)). Ces trois variables, utilisées donc prioritairement par la segmentation, peuvent être considérées comme les plus pertinentes pour ce qui concerne la discrimination des unités de sol de ce secteur de référence. Les deux premières traduisent directement les deux grands traits d'organisation ressortant d'une lecture rapide de la carte: étagement des sols avec l'altitude, typique d'une organisation de vallée fluviatile, et dissymétrie des rives de l'Hérault. La variable g(x), pour partie redondante avec dz(x), est utilisée lorsque la partition au moyen de la carte géologique s'avère plus précise que celle utilisant la carte topographique seule.

Au vu de la carte, l'erreur de classement prévue, (cf tableau 7) se présente sous trois formes bien différentes qu'il convient de définir pour préciser l'analyse.

1) **l'erreur de discrimination.** Elle tient au fait que certaines unités de sol, ne sont jamais prédites car elles se trouvent, dans les noeuds, dominées par des unités plus importantes en superficie. Il est important de noter que la plupart des unités dominées ne le sont que par une seule. Ainsi, tout se passe comme si l'arbre suggérait des regroupements d'unités de sols suivant un découpage global. Dans le cas étudié, celui-ci s'avère cohérent puisqu'il isole des unités de paysage bien individualisées: alluvions récentes de l'Hérault (unité 1 dominant 2, noeud 1), collines molassiques (unité 11 dominant 12,17 et 13, noeud 2), bas de pentes et bas-fonds de colluvions de molasse (unité 5 dominant 3, noeud 3),

Planche 3: *Application au secteur de référence des prédictions fournies par les arbres de classification. Confrontation avec la carte réelle.**



: limites et numeros d'unites de sol en surimpression

basse terrasse (unités 6 dominant 7, noeud 5), cailloutis villafranchien (unités 10 dominant 14, noeud 8). Les 2 unités de sols dont les points se dispersent dans plusieurs noeuds (9 et 16) présentent la double particularité d'être peu représentées et liées à des affluents de l'Hérault, c'est à dire discordantes par rapport au sens de variation général des sols.

2) **l'erreur de délimitation.** A la différence de la première, seul un examen de la carte permet de la mettre en évidence. Elle correspond aux cas où les limites réelles de la carte des sols et celles proposées par l'arbre ne se superposent pas mais sont par contre globalement parallèles entre elles. Dans le cas étudié l'erreur de délimitation (soit le décalage entre les limites) dépasse rarement 4 pixels (c'est à dire 200 m) ce qui correspond à une précision requise pour une carte des sols au 1/50000 (GEPPA, 1967). Très logiquement, elle est nettement plus importante pour les unités non regroupées (8, 15 et 4) que pour les groupes d'unités. Ces derniers correspondent vraisemblablement à un objectif de précision plus en rapport avec la qualité des données topo-géologiques utilisées.

3) **L'erreur de topologie.** Comme la précédente, elle est d'ordre géographique mais d'une nature différente et plus préoccupante dans la perspective d'une extrapolation future. Elle correspond en effet à des délimitations ne respectant pas la disposition réelle des unités de sol les unes par rapport aux autres. Ceci se manifeste par l'apparition de plages cartographiques prédites en un lieu totalement "déconnecté" de la position réelle de l'unité faisant l'objet de la prédiction. C'est le cas pour ce qui concerne l'unité 4 (alluvions de la Boyne, affluent de l'Hérault) qui fait l'objet d'une prédiction sur une zone éloignée de cet affluent située entre les alluvions et la basse terrasse. Quantitativement peu importante (environ 30 points soit, au plus, 8 hectares), elle révèle, contrairement aux précédentes, une discordance entre le modèle et la démarche cartographique réelle.

   Un examen approfondi permet de mettre en évidence la cause de cette discordance. L'isolement de l'unité 4, est justifié, suite à la dichotomie sur le noeud 39, par son appartenance à l'unité géologique 3 (Fya, alluvions anciennes 8-10m) plutôt qu'à l'unité géologique 2 (Fyb, alluvions anciennes 10-20m). Or, dans la réalité, l'unité 4 devrait être localisée sur aucune de ces deux unités géologiques puisqu'elle correspond aux sols développés sur alluvions récentes d'un affluent de l'Hérault. La distinction de l'unité 4 ne repose donc sur aucun déterminisme réel. Ainsi l'erreur de topologie est révélatrice d'un problème de représentation des véritables lois sols-paysage.

   Il est à noter enfin que le noeud 39, à l'origine de la distinction de l'unité 4, est celui pour lequel le risque de non pertinence est le plus élevé (6 à 10%). En cas de suppression, l'arbre obtenu aurait 8 noeuds terminaux et correspondrait à un risque de non pertinence dont la valeur médiane serait inférieure à 5%.

1.3.2 examen des prédictions de l'arbre 2

Les 20 règles de prédictions de l'arbre 2 utilisent 2 variables supplémentaires, la pente (pt(x)) et la distance par rapport à l'exutoire le plus proche (di(x)). Elles permettent de distinguer l'unité 9 et, surtout, des unités préalablement regroupées dans un même noeud: séparation des unités d'alluvions récentes 1 et 2, des unités de collines molassiques 11,12 et 17, des unités de basse terrasse (6 et 7). L'unité 13 remplace, auprès de l'unité 16, l'unité 9 dans l'ensemble des petites unités négligées par le modèle, c'est à dire affectables à aucun noeud. L'augmentation de

performance du modèle induit un déplacement dans l'importance respective des 3 types de distortions géographiques identifiées dans l'analyse précédente.

L'erreur de discrimination devient moins importante puisque seuls subsistent les groupes de sols "bas de pente et bas fonds de colluvions molassiques " et cailloutis villafranchiens", 13 unités sur 17 étant désormais isolées.

Les erreurs de délimitations, au contraire, s'amplifient: c'est surtout le cas entre les unités 6 et 7 où l'absence de pente augmente encore la difficulté d'isoler 2 unités par DZ. Ce type d'erreur affecte également la délimitation des unités 11,12,17. Dans ce dernier cas, compte tenu du fort morcellement des unités de la carte initiale, la forme même des plages cartographiques s'en trouve grandement modifiée.

De nouvelles erreurs de topologies apparaissent: existence d'un "couloir" d'unité 1 au sein de l'unité 2, inclusions d'unités 9 (terrasses de la Boyne) entre les unités 15 et 8 et inclusion d'unité 17 dans 11 ou 12 et vice-versa. Les deux premières sont, comme précédemment, liées à des discriminations sans fondement déterministes, la troisième étant liée au caractère aléatoire de l'apparition des unités de sols en milieu molassique.

1) Pour séparer les unités 1 et 2, la distance à l'exutoire est employée à l'inverse de toute les lois de sédimentation fluviatile selon lesquelles l'unité de sol la plus argileuse (dans ce cas unité 2) devrait être la plus éloignée du cours d'eau; ce problème résulte vraisemblablement de la présence de cours d'eaux secondaires venant perturber le calcul de la variable "distance au cours d'eau le plus proche" (DI).
2) L'emploi de la pente, pour séparer les unités 8 et 9 semble également dépourvu de toute logique: en effet, l'unité 9, unité de terrasse, donc de pente faible est isolée de l'unité 8, unité de glacis colluvial par une pente plus forte! L'explication de ce phénomène apparemment paradoxal vient du fait que le modèle numérique de terrain présente un pas trop élevé pour rendre compte de l'unité 9 dont il ne retient que l'environnement, caractérisé effectivement par des pentes fortes (bas de l'unité 10).
3) Dans le troisième cas, il ne semble pas y avoir a priori de problèmes du type de ceux identifiés dans les deux cas précédents. Cependant, le caractère aléatoire de l'organisation lithologique de la molasse rend difficile l'établissement de règles sols-paysage ayant un caractère général. En particulier, il semble que l'extrême sud du secteur soit mal représenté par les règles fournies par l'arbre 2, la production (par l'arbre) de ces dernières étant surtout influencée par la zone située au Nord et à l'Est de l'unité 5 de la carte des sols, d'où l'apparition d'inclusions non justifiées.

1.3.3. Examen des prédictions de l'arbre 3

Les 79 prédictions de l'arbre 3 utilisent encore 2 variables supplémentaires associées aux 5 précédentes. Il s'agit des variables "courbure moyenne" (CM) et "encaissement" (EC). A l'issue de ces prédictions, l'erreur de discrimination est totalement éliminée puisque toute unité à ce niveau est majoritaire dans au moins un noeud ce qui lui assure une prédiction individualisée. Compte tenu de la complexité extrême de l'arbre, il devient plus difficile d'examiner le détail des prédictions. Il est cependant possible de suivre l'évolution des deux autres types d'erreurs évoqués précédemment.

Les erreurs de délimitation sur les nouvelles unités distinguées s'avèrent inégalement importantes. Elles sont faibles pour l'unité 3 et l'unité 14, l'apparition de cette dernière permettant de préciser sa limite avec l'unité 15 par rapport aux prédictions précédentes. Par contre, les erreurs de délimitations sont importantes pour 16 et surtout 13. La limite entre 14 et 10 reste floue avec de

nombreuses inclusions de part et d'autre. Ceci traduit de nouveau la difficulté de différencier ces 2 unités au moyen de variables trop imprécises pour cet objectif. Par contre la précision de certaines limites s'améliore localement: c'est le cas entre 15 et 10, 6 et 7, 17 et 12

Les erreurs de topologie se multiplient sans qu'il soit possible de les inventorier toutes. Elles se caractérisent par l'apparition de petites populations de pixels isolés (1 à 10) témoins de la non généralité de certaines distinctions. Outre les unités précédentes, ces erreurs touchent les unités 13, 16, 6 et 7. Par contre, l'erreur de topologie concernant l'unité 4 commise par l'arbre 1 et perdurant dans l'arbre 2 semble résorbée. La combinaison de conditions correspondantes doit cependant avoir une efficacité prédictive douteuse en l'absence d'un déterminisme réel sous-tendant cette combinaison.

En résumé, l'analyse détaillée des arbres produits lors du chapitre précédent révèle les aspects suivants.

1) Il apparaît une nette hiérarchie dans l'utilisation des variables explicatives: la variable dz(x), souvent associée avec g(x), structure les arbres de manière dominante, quelles que soient les zones géographiques traitées. Ceci apparaît normal compte tenu de l'étagement des sols dans un système de vallée traduit par ces 2 variables. Par contre, les variables issues du traitement du MNT ("encaissement", "courbure moyenne", et dans une moindre mesure "pente") semblent reléguées aux dichotomies terminales. Le pas et l'imprécision du MNT utilisé sont vraisemblablement en cause dans cet état de fait. La variable rv(x), à deux modalités, ne peut être utilisée qu'une fois. Cependant, sa position au sommet de l'arbre démontre a posteriori son utilité. Le rôle de la variable di(x) est important à partir de l'arbre 2.
2) Un premier examen global indiquait que la taille de l'arbre et l'erreur de classement évoluaient a priori en sens inverse quels que soient les estimateurs choisis. La confrontation, dans l'espace géographique, des prédictions de chaque arbre avec la carte réelle permet d'affiner ce premier diagnostic en décomposant, au vu des cartes, l'erreur de classement en trois types d'erreur:
   - une **erreur de discrimination**, liée au fait que les unités de faible superficie ne sont pas différenciées par une prédiction et sont, le plus souvent, absorbées par une unité voisine (au sens géographique et sémantique) plus vaste qu'elle; cette erreur est naturellement en constante diminution au fur et à mesure que le nombre de noeuds terminaux augmente jusqu'à s'annuler pour l'arbre 3;
   - une **erreur de délimitation** liée à un décalage entre les limites réelles et les limites des unités ou groupes d'unités prédites; elle ne semble pas diminuer significativement dans le même sens que la précédente; au contraire, les nouvelles unités isolées semblent délimitées moins précisément ce qui participe à l'augmentation de l'erreur;
   - une **erreur de topologie** liée à une mauvaise restitution, par le modèle de l'organisation des unités entre elles; les quelques exemples étudiés dans le détail ont montré qu'il y avait à l'origine de ces erreurs soit une absence de signification des prédictions en termes de déterminisme de différenciation des sols, soit une perte de généralité de la distinction (une amélioration sur une zone localisée du secteur de référence crée à d'autres endroits, des plages cartographiques fictives). Ce phénomène s'amplifie nettement de l'arbre 1 à l'arbre 3.
3) La distinction de ces 3 types d'erreur est importante dans la mesure où elles ne présentent pas la même signification dans la perspective des prédictions futures. La première,

affectant surtout l'arbre 1, suggère un changement du contenu des prédictions dans le sens d'un regroupement d'unités. En d'autres termes, il est prédit à l'utilisateur une erreur forte s'il ne limite pas ses ambitions en terme de précision des prédictions. C'est donc une erreur connue et contrôlable. Par contre, l'erreur de topologie laisse craindre un échec des prédictions qu'elle affecte, dans la mesure où, déjà sur le secteur de référence, elles ne correspondent pas à des lois sols-paysage identifiables. Il semblerait donc souhaitable de produire un arbre dépourvu de ce dernier type de distortion.

4) Dans cette perspective, il est satisfaisant de constater que sa croissance (de l'arbre 1 à l'arbre 3) est en cohérence avec l'évolution du risque de non pertinence des prédictions. Cependant, aucun des deux seuils de risque de non pertinence choisis au chapitre précédent ne correspond au seuil d'apparition de cette distortion. Celui-ci se situe à 5%, c'est à dire légèrement plus sévère que celui de l'arbre 1 et nullement marqué par un niveau identifiable a priori. Ce nouvel arbre (figure 18), appelé arbre 0, sera conservé pour les validations ultérieures.

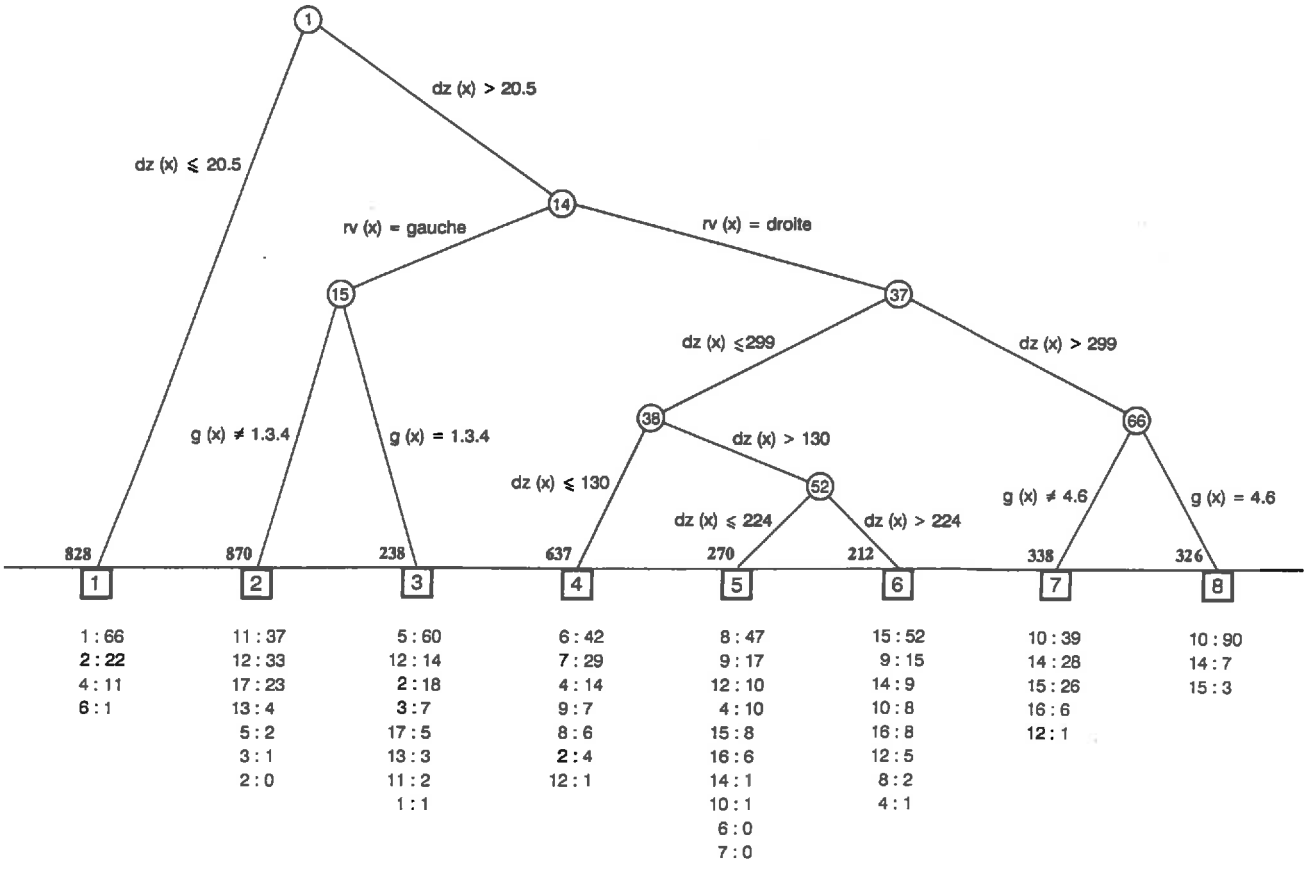


Figure 18: arbre 0 (risque de non pertinence des dichotomies inférieur à 5%)

Les caractéristiques générales de l'arbre 0 sont consignées dans le tableau 7bis.

	Nombre de noeuds	Nombre de variables explic.	Nombre d'unités reconnues	Erreur de classement prévue
arbre 0	8	3	7	48 %

Tableau 7bis: caractéristiques principales de l'arbre 0.

2. QUALITE DES PREDICTIONS ET PERSPECTIVES D'UTILISATION

Dans le sous-chapitre précédent, des prédictions de sol ont été élaborées à partir des données du secteur de référence de la Moyenne Vallée de l'Hérault. Elles ont été analysées en détail dans le souci de faire ressortir des critères qui permettraient à l'utilisateur de mieux interpréter les résultats d'une segmentation appliquée à des données géographiques en utilisant des outils et concepts adaptés. En particulier, la confrontation des prédictions avec la carte réelle permet une analyse qualitative des erreurs de prédiction qui complète le calcul du risque de non pertinence avancé dans le chapitre 5. Cette analyse s'avère-t-elle pertinente ?, Quelles prédictions, assorties de quelles modalités d'utilisation, sont en définitive les plus valides ?

La réponse à ces questions nécessite d'appliquer les diverses prédictions à l'extérieur du secteur de référence. Ceci a été réalisé sur le rectangle englobant le secteur de référence sur lequel les données topo-géologiques étaient disponibles (chapitre 4). Toutes les cartes résultantes, correspondant à chaque arbre testé, ne figurent pas dans ce document. A titre d'exemple, la carte correspondant à l'arbre 0 sera présentée sur la planche 4.

Les résultats de chaque arbre de classification choisis seront évalués sur les 3 secteurs de validation présentés en fin de chapitre 3:

- le secteur de La Roubiaire (noté RB dans les tableaux), recoupant les différents niveaux de terrasse de la rive droite et représentant 180 points de validation;
- le secteur de Montmau (noté MT), situé sur la rive gauche (collines et colluvions molassiques) et représentant 177 points de validation;
- le secteur de Lézignan la Cèbe (noté LZ) situé au contact entre les alluvions récentes et la basse terrasse de l'Hérault et représentant 164 points de validation.

Ils représentent au total 521 points de validation qui recoupent l'essentiel des unités de sol du secteur de référence et se répartissent dans un grand nombre de noeuds terminaux différents. Ceci permet ainsi de tester la majorité des règles sols-paysage produites par chaque arbre, exception faite de l'arbre 3 (cf tableau 8).

	Nbre noeuds terminaux	Nbre noeuds intéressant les SV
arbre 0	8	8
arbre 1	9	9
arbre 2	20	16
arbre 3	79	35

Tableau 8: nombre de noeuds terminaux des arbres concernés par les secteurs de validation ("S.V.")

Dans un premier temps, seront présentés les résultats bruts de cette validation, puis ces résultats seront repris et discutés dans deux perspectives distinctes:

- production d'une carte à petite ou moyenne échelle, dans le cas particulier où aucun sondage ne vient compléter cette information;
- utilisation des règles dans le cadre de l'automatisation du retour à la parcelle.

2.1. Analyse de la qualité des prédictions sur les secteurs de validation

Sur chaque secteur de validation est mesurée **une erreur de prédiction.** Elle correspond au pourcentage des pixels pour lesquels une des règles sols-paysage lue à partir de l'arbre étudié prédit une unité autre que celle existant réellement d'après la carte manuelle. Les résultats globaux de ces calculs sont présentés dans le tableau 9, pour les 3 secteurs et les 4 arbres retenus.

	RB	MT	LZ	TOT
Arbre 0	56	95	34	62
Arbre 1	56	95	57	69
Arbre 2	54	71	74	66
Arbre 3	68	64	71	67

***Tableau 9:** erreurs de prédiction mesurées (en %) sur les secteurs de validation*

Ils révèlent un niveau général d'erreur très élevé quel que soit le secteur de validation. Ce résultat médiocre n'est pas surprenant compte tenu de la différence de précision existant entre la carte à prévoir (carte des sols au 1/10.000) et les cartes fournissant les variables à expliquer (carte topographique au 1/25.000 et la carte géologique au 1/50.000).

Les erreurs observées sur les secteurs de validation sont systématiquement plus élevées que les erreurs de classement estimées à priori (tableaux 7 et 7bis). Ceci se traduit par un écart compris entre 14% (pour l'arbre 0) et 44% (pour l'arbre 3). Cet écart peut être interprété comme étant une mesure du degré de représentativité du secteur de référence vis à vis des secteurs de validation, pour ce qui concerne les règles sols-paysage produites.

La diminution d'erreur entre l'arbre 1 et l'arbre 3, prévue lors de l'élaboration des prédictions, ne se confirme pas. Au contraire, c'est l'arbre 0, a priori encore moins efficace que l'arbre 1 qui obtient l'erreur la plus faible (62%). Ceci tendrait à valider la démarche consistant à choisir comme critère d'arrêt un risque de non pertinence suffisamment bas pour écarter toute erreur de topologie. Cependant, ce résultat est à nuancer si l'on examine les secteurs séparément: pour le secteur de Montmau, l'arbre 2 apporte une amélioration substantielle même si celle-ci ne permet pas d'atteindre des niveaux d'erreurs acceptables. Ce résultat est inversé dans le cas de Lézignan la Cèbe où, au contraire, les niveaux d'erreurs augmentent régulièrement de l'arbre 0 à l'arbre 3. Ceci permet d'identifier un effet milieu qui jouerait un rôle dans la stabilité des prédictions et mettrait localement en défaut le nouveau critère d'arrêt, appliqué globalement sur l'arbre.

L'augmentation d'erreur entre l'arbre 0 et l'arbre 1 est uniquement localisée sur le secteur de Lézignan la Cèbe. Les prédictions de l'arbre 1 en cause correspondent à celles pour lesquelles avaient été diagnostiquée, dans le chapitre précédent, une "erreur d'organisation de sol" (dichotomie appliquée au noeud 39 de l'arbre 1). Dans ce cas précis, le pronostic de non fonctionnement de la prédiction s'est donc avéré exact.

Sur la base des matrices de confusion confrontant, unités de sol par unités de sol, les effectifs de points estimés et mesurés pour les 3 secteurs, il est possible de préciser cette première analyse en présentant les résultats de façon à isoler des milieux contrastés vis à vis de leur difficulté de cartographie qui ne recoupent pas totalement la différenciation par secteur. les résultats sont présentés dans le tableau 10 (page suivante):

	Fond de vallée	Rive droite	Rive gauche
Arbre 0	55	39	96
Arbre 1	75	39	96
Arbre 2	82	43	53
Arbre 3	80	54	53

Tableau 10: erreurs de prédiction (en %) et milieux d'études

Les 3 milieux différenciés ne sont pas égaux quant aux performances des prédictions. C'est le milieu "rive droite" (molasse et haute terrasse, unités 10,15,14) qui obtient les meilleurs résultats tous arbres confondus. Il correspond à une zone où les unités de sols s'ordonnent bien dans le paysage.

Le milieu "rive gauche" obtient des résultats inégaux suivant les arbres. L'erreur particulièrement élevée observée sur les arbres les plus simples (0 et 1) correspond en fait à une grosse erreur de discrimination (88%) liée au fait que les unités 12 et 17, présentes en majorité, sont dominées dans la prédiction les concernant,par l'unité 11. Dans ces cas, il est donc difficile d'établir des comparaisons avec les autres milieux.

Si l'on fait exception des cas précédents, le milieu "fond de vallée" (alluvions récentes colluvions et basse terrasse, unités 1 à 7) présente les plus mauvais résultats quels que soient les arbres. Ceci est vraisemblablement causé par l'imprécision des variables explicatives, qui est maximum dans ce milieu. En effet, d'une part, l'absence de relief accroît les erreurs d'estimation des variables issues du MNT; d'autre part, les risques d'erreur concernant la carte géologique sont élevés dans ce milieu quaternaire récent où, en l'absence de critères visibles sur le terrain, les limites entre les différents niveaux d'alluvions reposent uniquement sur une interprétation de la carte topographique.

Par ailleurs, le contraste entre milieux concernant l'évolution des performances des prédictions en fonction des arbres, déjà perçu précédemment, se trouve ici confirmé. Si, pour les milieux "fond de vallée" et "rive droite" les performances se dégradent avec la croissance de l'arbre, ce n'est pas le cas pour le milieu "rive gauche" pour lequel l'arbre 2 apporte un gain substantiel par rapport aux arbres les plus simples. Même si, dans ce cas, la nature de l'erreur observée sur les arbres les plus simples est un peu particulière, le gain observé est réel. Celà démontre les limites de la définition d'un critère d'arrêt ne prenant pas en compte la spécificité des milieux.

En résumé, l'application, sur des secteurs de validation, des prédictions de sol issues de segmentation permet de tirer des enseignements vis à vis de l'utilisation de ces prédictions dans la Moyenne Vallée de l'Hérault et, sur un plan plus méthodologique, sur la pertinence de la démarche utilisée pour interpréter les résultats.

1) Compte tenu des niveaux d'erreur atteints, il n'est pas possible d'utiliser ces résultats en l'état. En effet, même en prenant l'erreur minimale (62%), on se situe très loin des exigences de pureté d'unités cartographiques formulées pour une carte des sols détaillée (15% d'impuretés admises selon les normes GEPPA, 1967). Il faut donc, dans ce cas,

remettre en cause la règle classiquement adoptée en segmentation qui consiste à affecter un noeud à une seule unité. Cette démarche fera l'objet du chapitre suivant.

2) La nécessité d'un critère d'arrêt plus sévère que celui proposé par CART semble claire compte tenu des résultats obtenus par l'arbre 3, consécutif à son application. Globalement, le calcul d'un risque de non pertinence des dichotomies puis le choix, grâce à un examen cartographique sous SIG, d'un seuil qui respecterait le déterminisme d'organisation des sols apparaissent pertinents. Cependant, des dysfonctionnements de cette méthode apparaissent lorsque sont distingués des milieux qui diffèrent quant à la difficulté de les cartographier. Ainsi, il semblerait intéressant d'introduire, dans cette méthode, la prise en compte de la spécificité de ces milieux.

2.2. Conséquences sur les modalités d'utilisation des prédictions de sols issus de segmentation en Moyenne Vallée de l'Hérault

Compte tenu des résultats obtenus au cours des chapitres précédents, il convient de limiter les ambitions initiales des prédictions. Deux voies sont possibles en fonction du contexte dans lequel s'insère leur emploi:

- dans le cas où les prédictions par les règles sols-paysage représentent la seule source d'information disponible, il convient de limiter les ambitions concernant le contenu des unités prédites; en d'autres termes, il faut regrouper les unités du secteur de référence en unités de paysage cohérentes suceptibles d'être prédites avec des niveaux d'erreurs acceptables de façon à proposer une carte à petite ou moyenne échelle issue de l'extrapolation du secteur de référence;
- dans le cas où, comme dans ce travail, ces prédictions sont destinées à être complétées par d'autres informations (sondages), l'objectif final de prédire une unité de sol du secteur de référence doit être gardé. Cependant, il n'est pas besoin de choisir, au vu des seules règles sols-paysage, une seule unité puisque il y aura, par la suite, confrontation avec d'autres sources d'information susceptibles de préciser le choix. Par conséquent, une certaine imprécision peut être introduite dans le résultat des prédictions issues des règles sols-paysages (prédictions "floues").

Ces deux voies seront exploitées successivement dans ce chapitre:

2.2.1. Production d'une carte à moyenne ou petite échelle

Une analyse détaillée (paragraphe 1.3.) a permis de montrer que les arbres successifs, à l'exception de l'arbre 3, suggèrent des regroupements d'unités qui s'avèrent conformes à une réalité pédologique. Dans la perspective de limiter au maximum l'erreur de la carte issue des prédictions, il semble donc logique de suivre ces suggestions. Concrètement, la règle adoptée pour regrouper des unités est la suivante: une unité de sol, minoritaire dans un noeud, est rattachée à l'unité majoritaire de ce noeud si elle remplit deux conditions:

- avoir dans ce noeud la majeure partie (>90%) de sa population de points,
- ne pas être majoritaire dans un autre noeud.

Le résultat de cette démarche et sa validation sont résumés, dans le tableau 11 (page suivante), en 5 données:

- nombre de groupes d'unités établis selon la règle édictée précédemment;
- nombre et superficie (en % de la superficie du secteur de référence) des unités négligées; les unités négligées représentent les unités minoritaires ne satisfaisant pas aux conditions de la règle de regroupement;
- erreur de classement prévue (en %): elle résulte de l'application de la formule de calcul de R(T) (formule [13]) en considérant, à la place des unités, les groupes d'unités;
- erreur de prédiction (en %): elle est issue de l'application des prédictions de groupes d'unités de sols aux secteurs de validation.

	Nbre groupes	Unités négligées		Erreur prévue	Erreur de prédiction
		Nombre	Aire (en % du total)		
Arbre 0	7	3	10 %	21 %	26 %
Arbre 1	8	2	5 %	18 %	33 %
Arbre 2	13	2	2 %	30 %	54 %

tableau 11: résultats des regroupements d'unités.

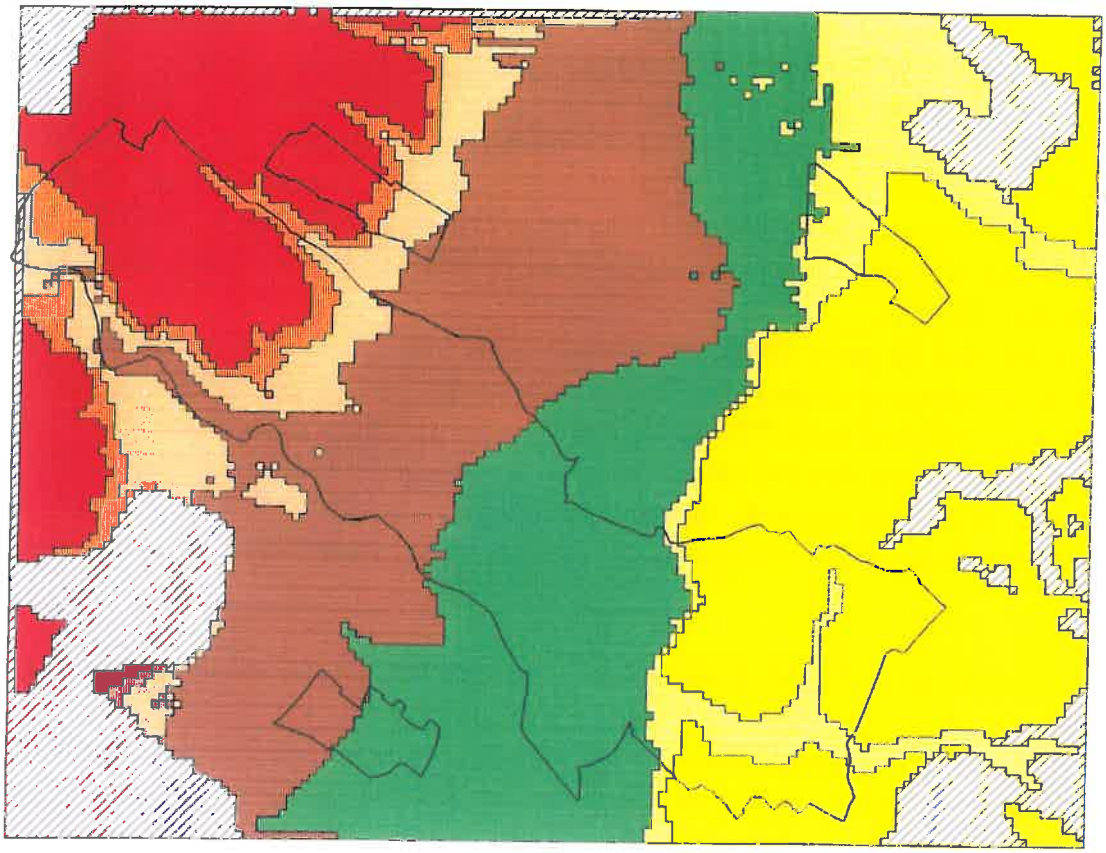
A titre de comparaison, sont également présentées (tableau 11bis) les caractéristiques des cartes au 1/100 000 et 1/250 000 de la zone, ainsi que leurs erreurs mesurées sur les secteurs de validation (les correspondances entre les unités du secteur de référence et les unités 1/100 000 et 1/250 000, présentées en annexe 7, ont été établies par recherche d'emprise géographique commune et vérification de concordance de description).

Cartes utilisées	Nombre d'unités (dans rectangle autour SR)	Erreur sur les secteurs de V
(1/100000) (BONFILS, 1992)	11	40 %
1/250000 (BORNAND et Al, 1992)	5	31 %

***Tableau 11bis:** description des performances des cartes à petite et moyenne échelle réalisées sur la zone.*

L'analyse comparée des deux tableaux révèle que les ratios erreur mesurée/nombre d'unités différenciées produits par les prédictions sont cohérents avec ceux obtenus par les cartes à petite et moyenne échelle de la région. Les erreurs de prédiction obtenues sont en deçà, ou, au pire des cas,

***Planche 4**: Résultat de l'application des règles sols-paysage fournies par l'arbre 0*



GROUPES D UNITES PREDITS

- sols d 'alluvions récentes (unités 1,2,ou 4)
- sols de colluvions de molasse (unités 3 ou 5)
- sols sur différents faciès de molasse (unités 11,12,13 ou 17)
- sols sur basses terrasses de l'Herault (unités 6 ou 7)
- sols colluviaux des zones de transition entre molasse et basses terrasses (unité 8)
- sols sur molasse recouverte par des colluvions de haute terrasse (unité 15)
- sols de la haute terrasse villafranchienne (unités 10 ou 14)
- autres non connus dans le secteur (ex: a cause d une unite geologique nouvelle)
- limites du secteur de reference et des secteurs de validation

Echelle 1/50000

du même ordre que les normes de pureté des unités de sol édictées par le GEPPA (1967) pour les études de reconnaissance (50% d'impuretés acceptées). Les prédictions générées par l'arbre 2 semblent produire une carte dont l' échelle serait voisine de celle de la carte au 1/100 000. L'application des arbres 0 et 1 permettrait de son côté d'établir une carte dont le niveau de précision se situerait entre celui des cartes au 1/250 000 et
1/100 000, avec une erreur remarquablement réduite pour une carte de cette échelle. De plus, les prédictions de l'arbre 0 sont assorties d'une erreur estimée peu différente de l'erreur mesurée.

A titre d'exemple, la carte résultant de l'application des règles sols-paysage de l'arbre 0 est présentée planche 4. Si l'on examine les unités négligées par cette carte, il apparait une différence avec les cartes à petite échelle de la région (cf carte 1/250000, planche 1). Parmi les 3 unités négligées, 2 correspondent à celles liées à la Boyne, affluent de l'Hérault. Or, ces unités ne sont pas distribuées selon l'axe principal de différenciation des sols de la vallée (orienté en fonction de l'Hérault). De leur côté, les pédologues auteurs des cartes au 1/250 000 et 1/100 000 n'ont pas négligé cet axe secondaire en représentant au moins l'une des unités concernées (alluvions récentes d'affluent de l'Hérault).

Ainsi, tout se passe comme si, en étant guidé par le seul souci de réduire l'erreur de prédiction, la méthode proposée n'était capable de représenter que les unités dont la distribution dans le paysage est conforme au modèle dominant sur la carte du secteur de référence. Un tel problème pourrait éventuellement être corrigé au niveau de la confrontation des prédictions des règles sols-paysage avec la carte du secteur de référence (paragraphe 1.3). En effet, il serait possible à ce niveau de forcer la prédiction d'une unité donnée en prenant le risque d'augmenter l'erreur de prédiction.

2.2.2. Utilisation des règles sols-paysage dans le cadre de l'automatisation du retour à la parcelle

Les règles sols-paysage élaborées par segmentation représentent la formalisation des lois sols-paysage permettant au pédologue d'exploiter les relations existantes entre unités de sol et critères extrinsèques. On a vu que ces règles, utilisées seules, n'étaient pas suffisamment sûres pour espérer prédire les unités de sol. Ceci est conforme à la réalité où l'on peut supposer que les lois sols-paysage sont en fait surtout utilisées pour éliminer de l'éventail des possibilités certaines unités de sol. En effet, la cartographie précise n'intervient que par la suite, avec la réalisation de sondages où sont exploitées les relations de voisinage entre unités de sol (chapitre 2).

Il faut donc trouver un mode de combinaison entre les deux sources d'information (sols-paysage et voisinage) permettant de représenter au mieux cette démarche. Une première étape à cette recherche est de transformer les prédictions issues des règles sols-paysage selon un formalisme qui permettra ultérieurement (chapitre 9) de mettre en oeuvre ce mode de combinaison.

Ceci suggère une utilisation différente des résultats de segmentation: au lieu de ne conserver qu'une unité prévue par noeud comme celà est pratiqué classiquement en segmentation, il est possible d'utiliser les proportions obtenues dans chaque noeud terminaux pour éliminer des unités. Cette élimination doit être effectuée avec une règle claire assortie d'un risque éventuellement modulable. Cette règle peut s'énoncer mathématiquement comme suit:

- soit ß un niveau de risque d'erreur vis à vis de l'élimination d'unité, consenti a priori.
- soit E_g, sous-ensemble du noeud terminal T_g, l'ensemble des ve unités U_j éliminées du noeud T_g ($j=1,\ldots,v$ et $ve<v$)
- soit pr_{mj}, la proportion d'individus membres de U_j et appartenant au noeud T_g

$$E_g = \{U_j / \Sigma_{(Eg)} pr_{mj} \leq ß \text{ et } ve = \max(1,....,v)\} \qquad [24]$$

Autrement dit, sont éliminées autant d'unités de sol que possible, pour peu que le total de leurs proportions reste en dessous du seuil de risque choisi. Par exemple sur le noeud n°3 de l'arbre 1, dont les proportions d'individus sont rappelés dans le tableau 12a, la règle édictée s'appliquerait, suivant le risque consenti, de la façon suivante, résumée dans le tableau 12b:

Unités	5	12	2	3	17	13	11	1
Proportions (en %)	60	14	8	7	5	3	2	1

tableau 12a: description du noeud terminal n°3 de l'arbre 1 (rappel)

Risque β	N° unités éliminées
0	4, 6, 7, 8, 9, 10, 14, 15
5	les mêmes + 1, 11
10	les mêmes + 1, 11, 13
20	les mêmes + 1, 11, 13, 17, 3

tableau 12b: unités de sol éliminées pour le noeud N°3 de l'arbre 1 en fonction du risque consenti.

Cette règle a été mise en oeuvre sur les 4 arbres retenus au cours du chapitre 31 en modulant ß selon 4 valeurs (0%, 5%, 10% et 20%). Il en résulte ainsi 16 (4^2) ensembles de prédictions "floues" dont il est possible d'évaluer les performances en les appliquant aux secteurs de validation. La comparaison des résultats de ces prédictions, d'un genre différent que précédemment, doit tenir compte de deux facteurs:

- l'erreur de prédiction; elle correspond désormais au pourcentage de points sur lesquels les unités de sol réellement présentes ont été, à tort, éliminées par les prédictions;
- l'imprécision; elle s'exprime en nombre moyen d'unités de sol prévues par point; son calcul correspond à la moyenne des v-ve (nombre d'unités retenues) caractérisant chaque prédiction, pondérée par le nombre de points qu'elle concerne.

	Arbre 0	Arbre 1	Arbre 2	Arbre 3
$\beta = 20$ %	31 % (2.9)	38 % (2.8)	51 % (2.4)	51 % (1.6)
$\beta = 10$ %	13 % (3.8)	20 % (3.5)	25 % (3.3)	40 % (2.1)
$\beta = 5$ %	13 % (4.4)	13 % (4.3)	14 % (3.1)	34 % (2.5)
$\beta = 0$ %	2 % (6.2)	2 % (6.0)	2 % (5.2)	25 %(3.5)

Tableau 13: erreur de prédiction et nombre moyen d'unités retenu (chiffres entre parenthèses)

Les niveaux d'erreur et de précision observés sur le tableau 13 varient considérablement (respectivement de 2% à 51% et de 1.6 à 6.2 unités conservées en moyenne) suivant les options choisies, tant en matière de critères d'arrêt que de risques d'élimination. Les sens de variations observés sont ceux attendus, même s'il existe des paliers: l'erreur décroit au fur et à mesure que le critère d'arrêt de l'arbre devient plus sévère et que l'élimination devient moins sélective.

Il y a, au vu du tableau, des convergences de résultats entre options différentes (exemple arbre 2, ß = 10% et arbre 3, ß = 0% ou arbre 0, ß =20% et arbre 3, ß = 5%). Ceci ouvre de nouvelles perspectives quant à l'utilisation de la segmentation. Dès lors qu'est abandonnée la règle d'affectation d'un noeud à une seule unité, il convient désormais de rechercher un équilibre satisfaisant entre sévérité du critère d'arrêt des dichotomies et sévérité d'élimination des unités, éventuellement arbitrée par la simplicité d'utilisation (c'est à dire,en fait, priorité au premier critère en cas d'égalité).

Ainsi, certaines options semblent proches de ce compromis. En effet, sans préjuger de leurs résultats futurs (chapitre 9), les compromis précision/erreur obtenus représentent bien les performances d'un pédologue prédisant, avant de faire des sondages, les unités de sols sur une nouvelle parcelle (erreur $\leq$ 13%, 3.8 à 4.4 unités conservées).

CONCLUSION DE LA DEUXIEME PARTIE

Au cours de cette deuxième partie, a été abordée la première étape du retour à la parcelle qui vise à prédire les unités de sol en un lieu non prospecté grâce à des "règles sols-paysage". Ces règles formalisent les lois de distribution des unités de sol fondées sur les relations sols-paysage qui apparaissent au cours de la cartographie préalable du secteur de référence.

Pour obtenir ces règles sols-paysage, une démarche spécifique a été construite. Elle associe l'utilisation d'un SIG et une méthode d'analyse de données appelée segmentation. Elle a été appliquée à la petite région naturelle "Moyenne Vallée de l'Hérault" qui constitue le milieu expérimental choisi pour ce travail.

Les résultats et enseignements qu'il est possible de tirer à l'issue de cette première phase se situent nettement sur deux plans distincts:

- un plan pédologique concernant la qualité et les perspectives d'utilisations de prédictions de sols basées sur les relations sols-paysage;
- un plan méthodologique relatif à l'utilisation (surtout conjointe mais aussi séparée) d'un SIG et de l'analyse par segmentation.

Malgré les réserves d'usage à formuler compte tenu de la faible superficie d'ensemble des secteurs de validation, l'exploitation des relations sols-paysage conduit à des prédictions présentant un ratio erreur /précision compatible avec deux utilisations possibles.

1) la production d'une carte à moyenne ou petite échelle. Ceci est à replacer dans la stratégie générale d'étude de sols proposée par l'IGCS. Selon celle-ci, l'esquisse au 1/250000 précède le secteur de référence et permet de le positionner et de définir sa zone de représentativité. Au vu des résultats obtenus, il apparaît que le secteur de référence pourrait, en retour, améliorer la qualité de l'esquisse au 1/250000 par l'application de prédictions sur des groupes d'unités de sol identifiés dans ce secteur de référence.
2) L'élimination raisonnée d'unités de sol préalable à la prospection par sondage dans le cadre du retour à la parcelle. Les résultats montrent qu'il est possible de réaliser une première sélection des unités susceptibles d'être présentes, à l'image de la démarche du pédologue cartographe décrite en première partie. Une incertitude demeure sur l'intérêt d'une telle sélection: élimine-t-elle suffisamment d'unités pour avoir un impact significatif sur la diminution du nombre de sondages nécessaires à la production d'une carte à grande échelle? Cette question fera l'objet du chapitre 9.

Au delà de ces perspectives d'utilisation, il faut souligner les inégalités d'aptitude à la généralisation suivant les milieux pédologiques recoupés par le secteur de référence. A ce titre, le milieu "fond de vallée" (alluvions récentes et basses terrasses) est défavorisé en raison de l'incertitude, plus importante en ce milieu, sur les variables décrivant le milieu naturel et intéressantes pour la cartographie des sols.

Enfin, les variables pressenties pour formaliser les relations sols-paysage ne s'avèrent pas également pertinentes. En particulier sur ce secteur, les variables décrivant la courbure du relief à partir du MNT (courbure moyenne, encaissement) semblent ne présenter qu'un intérêt limité dans le milieu étudié.

Bien que l'objectif premier de cette deuxième partie ait été la recherche de règles sols-paysage pour prédire des unités de sol, la démarche construite à cette occasion comporte des aspects méthodologiques débordant largement ce strict cadre pédologique:

L'idée directrice fondant la démarche proposée a été d'associer étroitement (et non simplement de juxtaposer) un système d'information géographique et l'analyse de données par segmentation. Il semblait en effet séduisant d'utiliser de concert ces deux outils pour extraire automatiquement des règles sols-paysage à partir des relations existant entre variables issues de couches géographiques différentes. Mais, compte tenu de l'incertude inhérente au caractère géographique des données manipulées, le risque potentiel, souligné par nombre de biométriciens, de mal interpréter les résultats de la segmentation devenait considérable.

Pour contrôler ce risque, l'association des deux outils a consisté à mobiliser les connaissances en matière d'analyse d'erreur existant dans le domaine des SIG pour assister l'interprétation des résultats de la segmentation. Concrètement, un nouveau critère d'arrêt des dichotomies de l'arbre de segmentation a été défini dans le but de fournir des prédictions conciliant stabilité et précision. Ceci a été réalisé en deux temps:

1) Calcul, pour chaque dichotomie de l'arbre, d'un "risque de non pertinence" basé sur la prise en compte et l'estimation des erreurs sur les variables géographiques manipulées et de leurs conséquences en matière de choix d'une dichotomie. Cette première phase permet de choisir des sous-arbres à partir de l'arbre maximal proposé par la méthode de segmentation sur la base d'un seuil fixé à partir du risque de non pertinence;
2) Affinage de ce premier choix par confrontation, au sein du SIG, entre la carte réelle du secteur de référence et la carte traduisant les prédictions issues des premiers arbres sélectionnés. Cette confrontation permet une analyse des distortions géographiques introduites par les prédictions dont certaines ("erreurs de topologie") hypothèquent le succès futur des prédictions et remettent éventuellement en cause les seuils choisis dans la première phase.

Sur la base des validations effectuées, il semble que cette démarche se justifie dans la mesure où les meilleurs résultats en termes d'erreurs (cf 2.1.), ou bien en terme de ratio erreur/précision (cf 2.2.) ne sont jamais obtenus par l'arbre fourni avec le critère d'arrêt standard fourni par le logiciel de segmentation.

Elle est cependant imparfaite dans la mesure où un doute persiste en particulier sur le fait que la première selection isole bien les meilleurs arbres. En effet, le fait que les erreurs de prédictions varient pour un même arbre suivant les différents milieux géographiques de la Moyenne Vallée de l'Hérault, laisse supposer qu'on gagnerait à prendre en compte les spécificités de ces milieux. D'autre part, seul le modèle d'organisation de sol dominant dans la petite région naturelle est pris en compte, ce qui pénalise la prédiction d'unités de sols qui suivent une logique de distribution différente (unités d'affluents de l'Hérault). Pour répondre à ces problèmes, deux voies d'amélioration seraient à explorer.

1) Une composante "milieu" serait à introduire dans le calcul du risque de non pertinence des dichotomies. Elle permettrait d'arrêter ces dernières soit plus tôt (au niveau des branches explorant un milieu "difficile"), soit plus tard (lorsque les branches concernent un milieu plus "facile"). Cette composante pourrait s'appuyer a priori sur deux éléments:
   - un degré de "flou" caractérisant la nature et la lisibilité dans le paysage des limites de sols; en d'autre termes, il s'agirait de permettre au pédologue qui a

fait la carte d'introduire un estimateur de la confiance qu'il accorde à l'extrapolation à d'autres secteurs de chaque limite qu'il a tracée.
- un calcul d'erreur tenant compte d'une répartion spatiale du risque d'erreur (exemple: altitude moins précise en situations plates)

2) Une deuxième voie pour améliorer la démarche consisterait à réaliser plus souvent des aller-retour entre les 2 phases de sélections d'un arbre présentées plus haut de façon à tester plus de seuils. Ceci suppose une interactivité totale des deux outils permettant à l'utilisateur une visualisation immédiate des conséquences de ses choix en terme de critère d'arrêt.

Par ailleurs, la démarche adoptée dans cette deuxième partie a permis d'aborder des problèmes intéressant séparément les deux outils utilisés:

- pour ce qui concerne les SIG, le besoin de calculer un risque de non-pertinence à nécessité une estimation puis un calcul de propagation d'erreurs des cartes sources tout au long des procédures utilisées dans le SIG; ce calcul a essayé d'être exhaustif et le moins simplificateur possible;
- pour ce qui concerne la segmentation, la règle classique d'affectation d'un noeud à une unité de sol prévue unique a été remise en cause compte tenu de la mauvaise qualité des résultats initiaux; le choix d'une nouvelle règle (correspondant au choix d'un risque d'élimination) peut jouer le même rôle que la définition d'un critère d'arrêt dans le sens d'un meilleur contrôle de la stabilité des prédictions données par les arbres de classification.

TROISIEME PARTIE

FORMALISATION ET UTILISATION DES LOIS DE VOISINAGE

La formalisation du retour à la parcelle, présentée au chapitre 2, indique que de nouvelles prédictions d'unités de sol, s'appuyant sur une série d'observations ponctuelles et rapides (sondages tarière), viennent étayer, préciser et souvent modifier les premières prédictions traitées dans la partie précédente. Ces prédictions sont formulées à partir des relations de voisinage entre unités et sont définies au cours de la cartographie préalable du secteur de référence. Ainsi, connaissant l'unité de sol à laquelle se rattache, en un lieu donné, un sondage donné, le cartographe est en mesure de prédire les unités de sol possible "aux alentours" de ce lieu, la définition exacte des "alentours" restant à préciser. A chaque sondage réalisé correspond une nouvelle prédiction, un pixel pouvant être naturellement concerné par plusieurs sondages. De ce fait, l'affectation finale d'un pixel à une unité donnée est le résultat d'une synthèse des différentes prédictions consécutives aux différents sondages réalisés dans la zone étudiée.

La présente partie traite de l'automatisation de cette démarche,le rattachement du sondage à l'une des unités du secteur de référence étant, il faut le rappeler, exclu de cette étude. Concrètement, il s'agira:

- de formaliser, à partir de la carte du secteur de référence, des lois de cartographie issues des relations de voisinage entre unités de sol (appelées lois de voisinage) selon un formalisme de règles "si (prémisse) alors (conclusion)", tel qu'il a été défini au cours du chapitre 2;
- d'utiliser ces règles au sein de l'outil informatique représentant le retour à la parcelle selon les modalités définies au cours du chapitre 2; pour celà doivent être abordés les problèmes de combinaison de prédictions forcément plurielles (prédictions à partir des règles "sols-paysage" + une prédiction nouvelle à chaque sondage réalisé); il faut également envisager le choix de la densité et de la position des sondages à réaliser pour optimiser la prédiction finale;
- d'appliquer l'outil ainsi construit sur des secteurs où la carte est déjà faite afin de vérifier la qualité des prédictions fournies.

Comme précédemment, le secteur de référence de la Moyenne Vallée de l'Hérault et les secteurs de validation, définis au cours du chapitre 3, constitueront le cadre expérimental de l'étude envisagée. La démarche adoptée fera l'objet de trois grands chapitres.

Le chapitre 7 s'intéressera à l'exploitation préliminaire des données du secteur de référence: choix d'un algorithme permettant d'extraire des règles de voisinage à partir de la carte du secteur de référence et choix des modalités de combinaisons de ces règles.

Le chapitre 8 traitera de l'application, dans un premier temps isolée, des règles de voisinage sur les secteurs de validation. A cette occasion, seront appréciées la qualité des prédictions résultantes et mises en évidence les limites des options choisies,

Le chapitre 9 verra l'utilisation des règles de voisinage dans le cadre général d'un retour à la parcelle: d'une part, sera établi le mode de combinaison, puis recherchées les éventuelles synergies avec les règles issues des relations "sols-paysage"; d'autre part, sera définie, à titre exploratoire, une stratégie de prospection pédologique visant à reproduire les décisions du pédologue soucieux de rentabiliser au mieux les sondages réalisés.

CHAPITRE 7

UTILISATION DES DONNEES DU SECTEUR DE REFERENCE POUR FORMALISER LES LOIS DE VOISINAGE ET DEFINIR LEUR MODALITES DE COMBINAISON

Comme pour la partie précédente, il s'agit d'utiliser les données du secteur de référence pour construire des règles qui, incluses dans l'outil informatique simulant le retour à la parcelle, seront suceptibles de délivrer une prédiction sur les sols.. La démarche adoptée sera parallèle à celle présentée pour construire les règles sols-paysage.

1) Dans un premier temps, un algorithme sera choisi et mis en oeuvre pour construire les règles. A la différence de précédemment, où un point donné ne pouvait faire l'objet que d'une seule prédiction, plusieurs prédictions affecteront un même point (par exemple, un point situé à mi-chemin entre deux sondages réalisés pourra faire l'objet de deux prédictions, éventuellement contradictoires). Donc, il faudra également trouver un autre algorithme capable de faire la synthèse entre prédictions différentes.
2) Dans un deuxième temps, les règles ainsi produites et combinées seront appliquées sur les données du secteur de référence de la Moyenne Vallée de l'Hérault qui ont servi à les générer. Il s'agira de "caler" certains paramètres de combinaison pour lesquels il n'existe pas de critère de choix a priori.Par ailleurs, celà permettra de disposer d'un essai de référence afin d'apprécier, dans le chapitre suivant, la dégradation éventuelle de la qualité des prédictions lorsqu'elles sont "exportées" à l'extérieur du secteur de référence.

1. LES ALGORITHMES D'EXTRACTION ET DE COMBINAISON DES REGLES DE VOISINAGE

Les problèmes d'extraction de règles de voisinage d'une part et la recherche d'une combinaison de ces règles dans la perspective de fournir une prédiction finale d'autre part, constituent deux aspects bien distincts qui feront l'objet d'autant de sous-chapitres.

1.1. Extraction des règles de voisinage à partir de la carte des sols du secteur de référence

A la différence de la partie précédente, où il avait été possible de choisir et d'adapter une méthode d'analyse de données issue d'un domaine scientifique différent, l'examen de la bibliographie disponible n'a pas permis de trouver une méthode éprouvée et reconnue d'extraction, à partir d'une carte des sols (ou, plus généralement, d'une carte en plages), de règles permettant de prédire les unités cartographiques au voisinage d'un point. Il convient de citer cependant le travail de GRZEBYK (1991) qui, sans avoir des objectifs de prédiction, constitue le seul exemple dégagé par l'étude bibliographique, d'étude quantitative d'organisation d'unités de sol à partir d'une carte pédologique.

Une méthode spécifique d'extraction des règles de voisinage a donc été élaborée. Elle sera présentée suivant trois étapes:

- recherche de variables pertinentes permettant de caractériser la position relative d'un point par rapport à un sondage;
- découpage de l'espace autour d'un sondage, en zones homogènes vis à vis des prédictions de sol (**zones d'isoprédiction**), en utilisant les variables définies précédemment;
- calcul des probabilités d'apparition des unités de sol pour une unité reconnue donnée et une "zone d'isoprédiction" donnée.

1.1.1. Recherche de variables pertinentes caractérisant la position relative d'un point par rapport à un sondage

On considèrera dans la suite de l'exposé l'existence d'une série de sondages réalisés sur q points particuliers de la parcelle, notés $x_1,...x_k,...x_q$. Sur chaque point où un sondage est réalisé, l'unité de sol est supposée connue par identification directe au moyen d'une observation du sol.

Dès lors, il est intéressant de redéfinir la position géographique des autres points de la surface à cartographier relativement à chaque point où un sondage est réalisé. En effet, la **position relative** d'un point donné vis à vis d'un sondage va déterminer, sur ce point, la prédiction de sol émise à la suite de la réalisation du sondage. Dans cette perspective, deux variables caractérisant cette position relative semblent importantes à retenir.

> **1) le rayon de voisinage ($r_k(x)$).** Il correspond à la distance géographique entre un point x de la surface à cartographier et le point x_k objet d'un sondage. Il semble en effet nécessaire de prendre en compte cette variable puisqu'il est évident, à la lecture d'une carte, que les unités de sol varient au fur et à mesure que l'on s'éloigne d'un point particulier de la carte pris au hasard.
>
> **2) Le sens de voisinage ($s_k(x)$).** Sous ce terme sera désigné le sens de progression dans le motif d'organisation des unités de sol (équivalent d'un "soil combination" de FRIDLAND évoqué au chapitre 2) lorsqu'est parcouru le chemin du point x_k vers le point x. Par exemple, si le motif est une toposéquence, les unités de sols rencontrées varient suivant que la progression depuis le sondage s'effectue vers le haut, vers le bas ou reste à la même altitude.

Si la variable $r_k(x)$ ne pose pas de problème de mise en oeuvre particulier, la nouvelle variable $s_k(x)$ nécessite une définition précise. On propose que ce soit une variable qualitative à trois modalités dont les dénominations changeraient suivant les motifs d'organisation. Ainsi, dans le cas particulier de la Moyenne Vallée de l'Hérault où le motif d'organisation général supposé est une toposéquence, ces 3 modalités possibles s'appelleront " plus haut", "plus bas" et "même altitude". $s_k(x)$ sera donc définie comme suit:

$$
\begin{aligned}
(z(x_k)-1{,}7m) \le z(x) \le (z(x_k)+1{,}7m) \quad &==> \quad s_k(x) = \text{"même altitude"} \\
z(x) > z(x_k) + 1{,}7m \quad &==> \quad s_k(x) = \text{"plus haut"} \\
z(x) < z(x_k) - 1{,}7m \quad &==> \quad s_k(x) = \text{"plus bas"}
\end{aligned}
\qquad [25]
$$

avec $z(x)$, $z(x_k)$: altitude aux points x, x_k

La valeur 1,7 mètres représente le seuil en dessous duquel une différence d'altitude n'est pas considérée comme significative compte tenu de la qualité des données utilisées. L'évaluation de ce seuil est faite à partir de l'écart type d'erreur sur l'altitude (1.3 m), calculé à l'occasion du calcul de propagation d'erreurs de la précédente partie. A partir de cet écart type, peut être déterminé, pour un niveau de confiance que l'on choisit, l'amplitude de l'intervalle de confiance sur la valeur de l'altitude. Pour un niveau de confiance de 80%, d'après les tables de répartition de la loi normale réduite centrée, cette amplitude est de 1.28 x 1.3m soit 1.7 m. Ainsi, par exemple, faut-il interpréter la modalité "plus haut" par "il y a, compte tenu de l'erreur supposée sur z(x), 80% de chances que $z(x) > z(x_k)$".

1.1.2. Définition, autour du sondage, des "zones d'isoprédiction" (figure 19)

Pour pouvoir isoler, autour de chaque sondage, des zones considérées comme homogènes vis à vis des prédictions de sol qu'il suscite, il faut disposer d'une stratification de l'espace. Pour celà, on définira, relativement à chaque point particulier où un sondage est réalisé, des **zones d'isoprédiction**. Elles seront délimitées grâce aux 2 variables $r_k(x)$ et $s_k(x)$ définies précédemment.

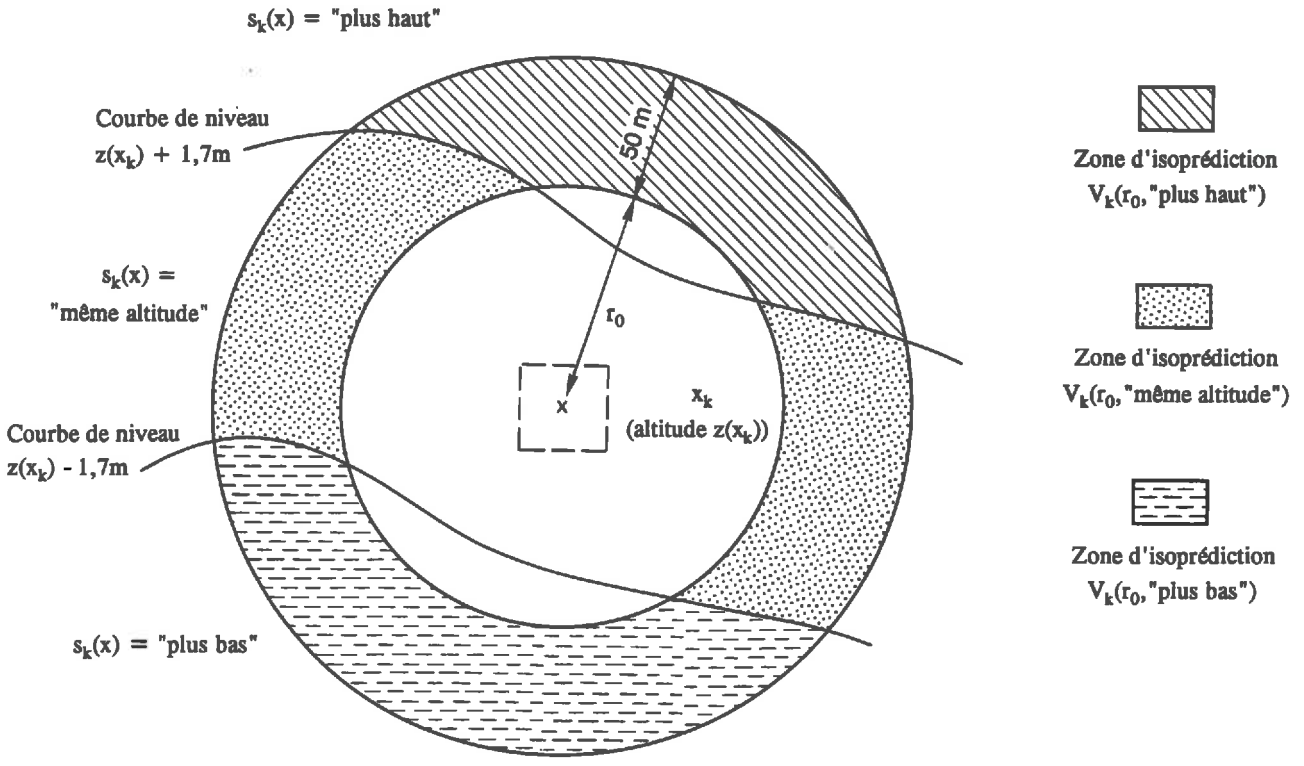


Figure 19: exemple de délimitation, autour d'un sondage, de 3 zones d'isoprédiction

Dans un premier temps, le rayon de voisinage permet de définir des couronnes concentriques autour du sondage considéré. Chaque couronne est définie par un paramètre noté r_0 qui correspond à

son rayon minimum, sa largeur étant constante et fixée à 50 mètres. On définira ainsi 30 couronnes depuis $r_0 = 0$ m ($0 \leq r_k(x) < 50$m) jusqu'à $r_0 = 1500$m ($1500\text{m} \leq r_k(x) < 1550\text{m}$) [10] .

Dans un deuxième temps, ces couronnes sont elles même subdivisées en trois zones, chacunes caractérisées par un second paramètre, noté s_0, correspondant aux trois modalités possibles du sens de voisinage.

Ainsi, tout point situé à une distance du sondage x_k comprise entre r_0 et $r_0 + 50$m d'une part et dont le sens de voisinage vis à vis de ce sondage est s_0 appartiendra à la zone d'isoprédiction notée $V_k(r_0, s_0)$. La définition mathématique correspondante d'une zone d'isoprédiction s'établit comme suit:

$$V_k(r_0, s_0) = \{ x / r_0 \leq r_k(x) < r_0 + 50\text{m et } s_k(x) = s_0 \} \qquad [26]$$

avec: $r_0 \in R$
$s_0 \in$ {"plus haut", "plus bas", "même altitude"}

Tout point peut être rattaché, grâce à cette formule, à l'une des zones d'isoprédiction caractérisées par les deux paramètres r_0 et s_0. Le nombre de zones définies est important (3x30 = 90) ce qui ne sera pas sans conséquences sur la lourdeur future du modèle. Il pourrait être, dans l'avenir, diminué par le choix d'une largeur de couronne plus importante que 50 mètres, au besoin variable avec la distance.

1.1.3. Utilisation des points situés à l'intérieur du secteur de référence pour calculer les probabilités d'apparition au sein des zones d'isoprédictions

Le concept de "zone d'isoprédiction " étant établi, il est maintenant possible d'aborder la construction des règles de voisinage à partir des données du secteur de référence. L'algorithme réalisant cette opération fonctionnera de la façon suivante.

1) Pour chacun des points du secteur de référence, considérés tour à tour comme des points de sondage (x_k), on délimite les différentes zones d'isoprédiction $V_k(r_0,s_0)$. l'indice k varie entre 1 et q, q étant, dans cette situation particulière, le nombre total de points du secteur de référence. Conformément aux options prises antérieurement, les points du secteur de référence seront pris sur une grille de pas 50m.

2) Pour chaque zone d'isoprédiction élémentaire $V_k(r_0,s_0)$ ainsi délimitée, on calcule:
   - le nombre total de points de la zone d'isoprédiction (noté n_k);
   - les nombres de points (notés n_{kj}) de la zone d'isoprédiction appartenant à chaque unité de sol U_j (j varie de 1 à v).

Ces calculs peuvent se formaliser de la façon suivante:

- soit M_j l'ensemble des points appartenant à U_j ($M_j = \{x/ u(x) = U_j, j=1,...v \}$

---

10 La valeur maximum de 1550m correspond au seul souci d'explorer un rayon de voisinage suffisamment grand pour que tout point du paysage soit concerné par chaque sondage réalisé. Par la suite, ce choix sera réexaminé à la lumière des résultats

$$n_{kj} = \text{card}\,(V_k(r_0,s_0) \cap M_j)$$
$$n_k = \text{card}\,(V_k(r_0,s_0)) \qquad [27]$$

3) On considère séparément les sous-ensembles de points x_k appartenant à la même unité U_L. Sur chacun de ces sous-ensembles, on fait la somme, pour chaque couple de paramètres (r_0, s_0), des différents n_k et n_{kj} obtenus sur chaque point. Il est alors possible de calculer la probabilité[11] $\pi_j(U_L,r_0,s_0)$ d'apparition d'une unité U_j dans la zone d'isoprédiction délimitée autour d'un sondage appartenant à une unité U_L et caractérisée par les paramètres r_0 et s_0. Cette probabilité est calculée par la formule:

$$\pi_j(U_L,r_0,s_0) = \Sigma_{(ML)}\, n_{kj} \,/\, \Sigma_{(ML)}\, n_k \qquad [28]$$

avec L,j $\in \{1,\ldots,v\}$, v étant le nombre total d'unités du secteur de référence
$k \in \{1,\ldots,q\}$, q étant nombre total de points dans le secteur de référence
M_l: sous-ensemble des points appartenant à U_l

4) Une règle Rg (U_L,r_0,s_0) s'écrira donc:

si $u(x_k) = U_L$
et si $x \in V_k(r_0,s_0)$
alors
$(\pi_1(x),\ldots,\pi_j(x),\ldots\pi_v(x))$ [29]

avec $\pi_j(x) = \pi_j(U_l,r_0,s_0)$

Chaque $\pi_j(x)$ représente en fait une probabilité d'apparition d'unité de sol calculée précédemment. En d'autres termes la règle Rg s'exprime par:
- si, sur un sondage, est reconnue l'unité U_L,
- et si un point est situé dans la zone d'isoprédiction délimitée autour de ce sondage et caractérisée par les paramètres r_0 et s_0,
-alors les probabilités d'apparition sur ce point sont $(\pi_1(x),\ldots\pi_v(x))$.

La démarche de construction des règles de voisinage présentée ci-dessus a été mise en oeuvre pour le secteur de référence de la Moyenne Vallée de l'Hérault au moyen d'un programme informatique écrit en FORTRAN (cf annexe 8). Au préalable, a été extrait à partir d'ARC/INFO le fichier des 3779 points du secteur de référence répartis sur une maille régulière de pas 50m. Il faut rappeler que, conformément au découpage de l'espace adopté (chapitre 2) ces points représentent en fait des pixels dont ils constituent les centroïdes. Le fichier obtenu contient les variables utiles à la construction des règles: latitude, longitude en Lambert mètre, altitude et unité de sol dans laquelle tombe le centroïde du pixel.

---

11 Il est d'usage de considérer comme des probabilités les fréquences obtenues sur les grands nombres. Par exemple, si l'as de coeur est tiré 28 fois sur 1000 essais, on pourra prétendre que la probabilité de tirage de l'as de coeur aux cartes est de 1/36. Ceci serait illicite si les calculs portaient sur de petits nombres. Dans le cas étudié, assimiler les fréquences aux probabilités dans les zones d'isoprédiction n'est possible que parce les populations mises en jeu sont très importantes.

Sur station SUN SPARK 1, le programme construit les règles, pour une unité de sol donnée, en 10 minutes en moyenne. Les résultats font l'objet, dans un premier temps, d'un fichier par unité de sol et par modalité du paramètre s_0. Chaque fichier constitue un tableau à double entrée rassemblant les probabilités $\pi_j(x)$ par unités de sol prédites et par paramètre r_0 (correspondant aux rayons de voisinage minimum des couronnes successivement considérées.

Dans un deuxième temps, les fichiers correspondant aux différentes unités de sol sont mis bout à bout dans la perspective de leur utilisation future. Il en résulte ainsi 3 gros fichiers résultat, chacun correspondant à une modalité du paramètre s_0.

1.2. Utilisation des règles de voisinage: recherche d'un algorithme de combinaison

Pour pouvoir aborder l'utilisation des règles obtenues à partir de la carte du secteur de référence, un changement de notation par rapport à la phase précédente est nécessaire. Conformément à la formalisation mathématique proposée au chapitre 2, on considèrera désormais les sondages $x_2,\ldots,x_k,\ldots x_f$ réalisés aux étapes $t_2,\ldots,t_k,\ldots,t_f$ du retour à la parcelle (en fait, $f = q+1$). L'étape t_1 ne fait pas l'objet de sondage puisqu'elle correspond à l'utilisation des règles sols-paysage.

Par ailleurs, les règles de voisinage déclenchées successivement suite aux sondages introduits seront notées $Rg_2,\ldots,Rg_k,\ldots Rg_f$. Chaque règle fournira une série de probabilités d'apparition d'unités de sol notées $(\pi_{1k}(x),\ldots,\pi_{jk}(x),\ldots\pi_{vk}(x))$

Dans la mesure où plusieurs règles sont déclenchées, un point donné de la surface à cartographier peut naturellement être concerné par plusieurs règles Rg_k. Ceci soulève un nouveau problème: comment, en ce point, combiner les prédictions délivrées par chacune des règles de façon à obtenir, in fine, une prédiction constituant une synthèse optimale?

Le problème ainsi posé peut être considéré comme une illustration d'un problème plus général bien connu, en particulier dans le domaine de l'Intelligence Artificielle (MARTIN CLOUAIRE, 1992): comment combiner des sources d'informations issues de diverses sources (experts différents, capteurs différents,...) pour permettre une décision finale ?
Une bibliographie abondante existe sur le sujet et plusieurs méthodes ont été proposées. Cependant, dans le cas étudié, l'éventail des solutions possibles est limité puisque le choix d'un cadre probabiliste a déjà été fait pour représenter l'incertitude des prédictions (chapitre 2). En effet, seules des méthodes relevant de ce cadre peuvent être mises en oeuvre. En particulier, des méthodes prometteuses basées sur la théorie des possibilités (ZADEH, 1978) ou sur la théorie des croyances (SHAFER, 1976) ne peuvent être envisagées.

Le choix d'une méthode de combinaison des règles de voisinage s'effectuera en deux temps:
- choix de la formule de combinaison;
- choix du système de pondération entre les prédictions dans le cadre de la formule choisie.

1.2.1. Choix d'une formule de combinaison

GENEST et ZIDEK (1986) ont réalisé une revue bibliographique très détaillée sur les différentes formules permettant de combiner des probabilités d'apparition d'un évènement donné, fournies par différents experts. Il en ressort que le choix ne peut faire l'économie d'une connaissance détaillée de la façon dont sont produites chacune des probabilités participant à la synthèse. Ainsi,

dans le cas de probabilités d'occurrence d'unités de sol dégagées d'une carte, la nature géographique des données manipulées doit être prise en considération.

De nouveau, l'erreur sur chacune des probabilités et son mode de propagation dans la formule choisie constitue un argument majeur de choix. Intervient dans ce cas l'erreur sur les limites de la carte pédologique qui se propage sur les dénombrements des points participant au calcul des probabilités.

Une estimation rapide de l'ordre de grandeur de l'erreur sur ces probabilités peut être avancé. Pour celà, est utilisé l'écart type d'erreur sur la position des limites de la carte des sols Se(l) (estimé dans la partie précédente). En multipliant ce dernier par la longueur des limites de la carte, il en résulte une aire qui correspond à une valeur globale d'écart type d'erreur sur la superficie des unités. Ramenée à la surface totale du secteur, cette valeur atteint entre 7 et 11%, suivant la valeur extrème de la fourchette d'estimation utilisée (13m ou 21m). S'il est fait l'hypothèse que cette erreur est uniformément répartie entre les unités de sol et s'il est négligé l'erreur causée par l'emploi de l'altitude, la valeur calculée constitue une estimation de l'erreur sur les probabilité d'apparition des unités de sol.

Compte tenu de l'ordre de grandeur de l'écart type d'erreur obtenu sur chaque probabilité (correspondant donc, grosso modo, à 10% de la valeur de cette probabilité), il apparaît sage de préférer une méthode de combinaison additive à une méthode multiplicative. La première limite en effet la croissance de l'erreur sur le résultat des combinaisons, lorsque le nombre des règles appliquées augmente.

C'est pourquoi, à l'étape t_f du retour à la parcelle, la probabilité de présence en x d'une unité U_j, suite à l'application de différentes règles,$Rg_2,\ldots,Rg_f$ sera une moyenne arithmétique pondérée des probabilités données par chaque règle, soit:

$$p_j(x,t_f) = \Sigma_{(k=2\ldots f)}\ w_k \cdot \pi_{jk}(x) \qquad [30]$$

avec w_k: un coefficient de pondération attaché à la règle Rg_k ($\Sigma_{(k=2\ldots f)}\ w_k = 1$)

$\pi_{jk}(x)$: probabilité d'apparition de U_j donnée par la règle Rg_k déclenchée à la suite du $k^{ième}$ sondage;

Cette formule est citée par GENEST et ZIDEK sous le nom de "linear opinion pool". Il faut souligner que ce choix va à l'encontre de ceux réalisés par des auteurs intervenant dans le domaine des systèmes d'information géographique qui se sont trouvés confrontés à des problèmes similaires (MIDDELKOOP et Al, 1989). Ces auteurs semblent en effet privilégier les approches Bayesiennes en dépit de leur caractère multiplicatif et des nombreuses hypothèses et approximations nécessaires pour fixer tous les termes de la formule de calcul d'une probabilité résultante.

1.2.2. Choix d'un système de pondération pour le calcul d'une probabilité résultante

GENEST et ZIDEK considèrent que le caractère arbitraire des pondérations constitue "un sérieux obstacle" à la mise en oeuvre des méthodes de combinaison de probabilités issues des différentes sources. Cette affirmation souligne le soin qu'il convient d'accorder à cette phase.

Le problème peut se formuler de la façon suivante: parmi les différents paramètres caractérisant une règle, lesquels sont susceptibles d'avoir une influence sur la qualité de cette règle ?

Dans le cas où cette influence est mise en évidence peut-on en tenir compte pour pondérer les résultats des prédictions délivrées ?

Comme évoqué précédemment, chaque règle est liée à trois paramètres:

- l'unité U_L dont l'identification sur un sondage déclenche la règle,
- le sens de voisinage s_0 ("+haut","+bas","même altitude"),
- le rayon de voisinage retenu pour dimensionner la couronne (r_0).

Dans une première approche, seule l'influence du dernier paramètre (r_0) sera étudiée en détail. L'hypothèse retenue est qu'elle est a priori prépondérante par rapport à celle des deux autres. Afin de justifier cette hypothèse, deux aspects seront successivement explorés:

- influence du rayon de voisinage sur la précision des prédictions des règles,
- influence du rayon de voisinage sur le risque d'erreur des prédictions liées aux effets de bordures (zones d'isoprédiction débordant du périmètre du secteur de référence)

1.2.2.1. Rayon de voisinage et précision des résultats

Comme évoqué précédemment, chaque règle Rg_k fournit, sur les points qu'elle touche, une probabilité d'occurrence de chaque unité de sol (notée $\pi_{jk}(x)$). L'ensemble de ces probabilités constitue donc la prédiction fournie par la règle $(\pi_{1k}(x),\ldots,\pi_{jk}(x),\ldots,\pi_{vk}(x))$. Il est possible de caractériser la précision de cette prédiction au moyen d'un critère (noté $i(Rg_k)$) calculé à partir des $\pi_{jk}(x)$. Ce critère est de même nature que l'indice d'impureté sur les noeuds de l'arbre de segmentation défini dans la partie précédente. En effet, il doit vérifier les propriétés suivantes:

- $i(Rg_k)$ = maximum si $\pi_{1k}(x) = \ldots = \pi_{jk}(x) = \ldots = \pi_{vk}(x)$ (imprécision maximum)
- $i(Rg_k)$ = minimum si $\exists\, U_j / \pi_{jk}(x) = 1$ (précision maximum)

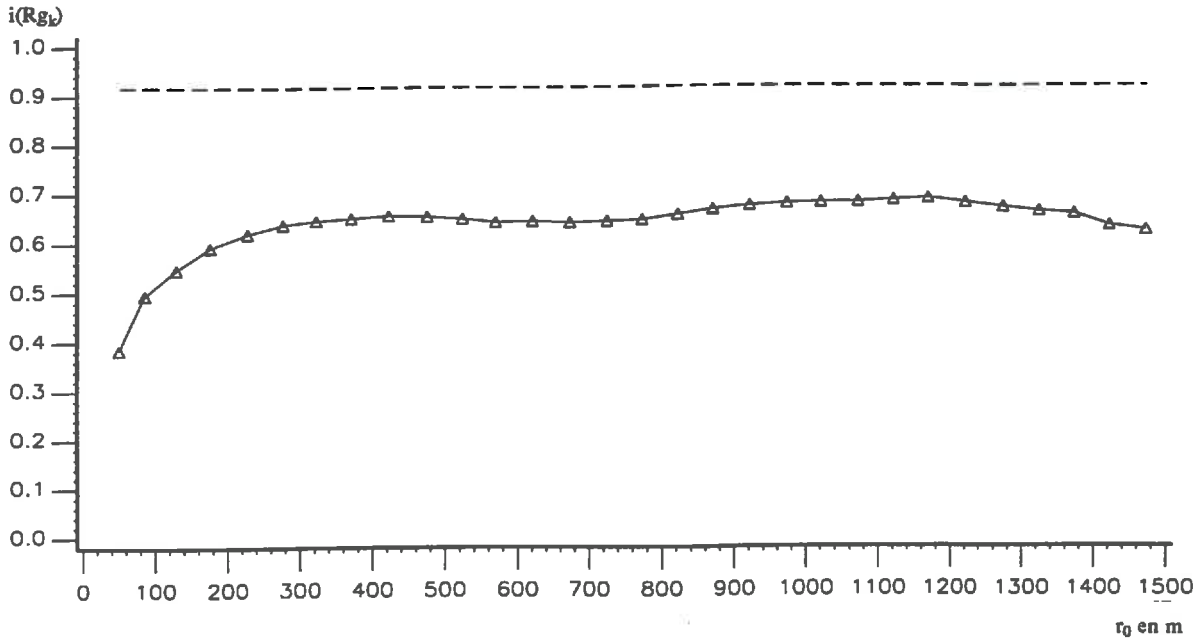


Figure 20: évolution de la précision des règles en fonction de r_0

Ce critère sera donc repris pour étudier l'influence du rayon de voisinage sur la précision des prédictions. Sa formule est la suivante(rappel):

$$i(Rg_k) = 1 - \Sigma_{(j)} \pi^2_{jk}(x) \qquad [31]$$

Dès lors, il est possible de construire la courbe $i(Rg) = f(r_0)$, i(Rg) représentant la moyenne des $i(Rg_k)$ effectuée sur une population de règles ayant le même paramètre r_0, toutes unités de sol et sens de voisinage confondus. Cette courbe est représentée sur la figure 20. Elle met en évidence une dégradation rapide de la précision des prédictions entre 0 et 300m. Après 300m, la courbe atteint un palier quasi parfait jusqu'à 1500m. Ainsi, existe-t-il une différence nette de précision de prédiction suivant la valeur du rayon de voisinage.

Ce résultat n'est pas sans conséquences sur la recherche des probabilités résultantes $p_j(x,t_k)$. La formule [29], comme toute moyenne, ayant tendance à lisser les extrèmes, doit-on laisser annuler, sur un pixel donné, une prédiction précise suite à un sondage proche, par toute une série de prédictions imprécises provenant des sondages plus lointains? En d'autres termes, Est-il licite de ne pas privilégier la décision d'un arbitre sûr de lui car situé à 10 mètres de l'action par rapport à l'avis de plusieurs collègues fortement indécis car observant le match depuis les tribunes?

Dès lors, apparaît une première raison pour pondérer la combinaison des prédictions en fonction inverse du rayon de voisinage.

L'examen de la figure 20 est également l'occasion de réexaminer le choix du rayon de voisinage maximum effectué en début de chapitre 1. Il faut noter en effet que le palier atteint par la courbe après 300m se situe nettement en dessous de l'imprécision maximale potentielle du secteur de référence obtenue en calculant I(Rg) sur les proportions de chaque unité de sol au sein du secteur de référence (tireté). C'est à dire que, même à 1500m, la prédiction est encore porteuse d'une information sur l'occurrence des unités de sol. Fixer le rayon de voisinage maximum à une valeur si élevée ne serait donc pas inutile de ce point de vue.

1.2.2.2. Rayon de voisinage et effets de bordures

Comme le souligne GRZEBYK (1991) un biais peut survenir lorsqu'est exploré le voisinage d'un objet géographique (en particulier un pixel). En effet, avec l'augmentation du rayon de voisinage, la zone explorée (correspondant à la "zone d'isoprédiction" définie précédemment) déborde de plus en plus du périmètre de la carte (figure 21, page suivante). Il s'ensuit une augmentation du nombre de points non renseignés par l'unité de sol et soustraits, de ce fait, au dénombrement par unité.

Dans le cas du secteur de référence, il a été calculé le pourcentage des points concernés par ce biais pour chaque rayon de voisinage. La courbe résultante (figure 22) révèle une croissance quasi linéaire des points situés à l'extérieur du périmètre.

En conséquence, face à la diminution relative de la zone effectivement explorée, il est de moins en moins licite de supposer cette dernière représentative de toute la zone d'isovoisinage potentielle. Ainsi, au fur et à mesure que le rayon de voisinage augmente, la "confiance" accordée aux résultats diminue. Ceci constitue un deuxième argument justifiant une pondération en fonction inverse du rayon de voisinage.

Compte tenu de ces deux types d'arguments, w_k sera donc une fonction inverse de r_0. Concrètement, si l'on considère la combinaison des f-1 règles $Rg_2,\ldots,Rg_f$, le coefficient de pondération w_k qui s'appliquera aux probabilités fournies par un des règles s'exprimera comme suit:

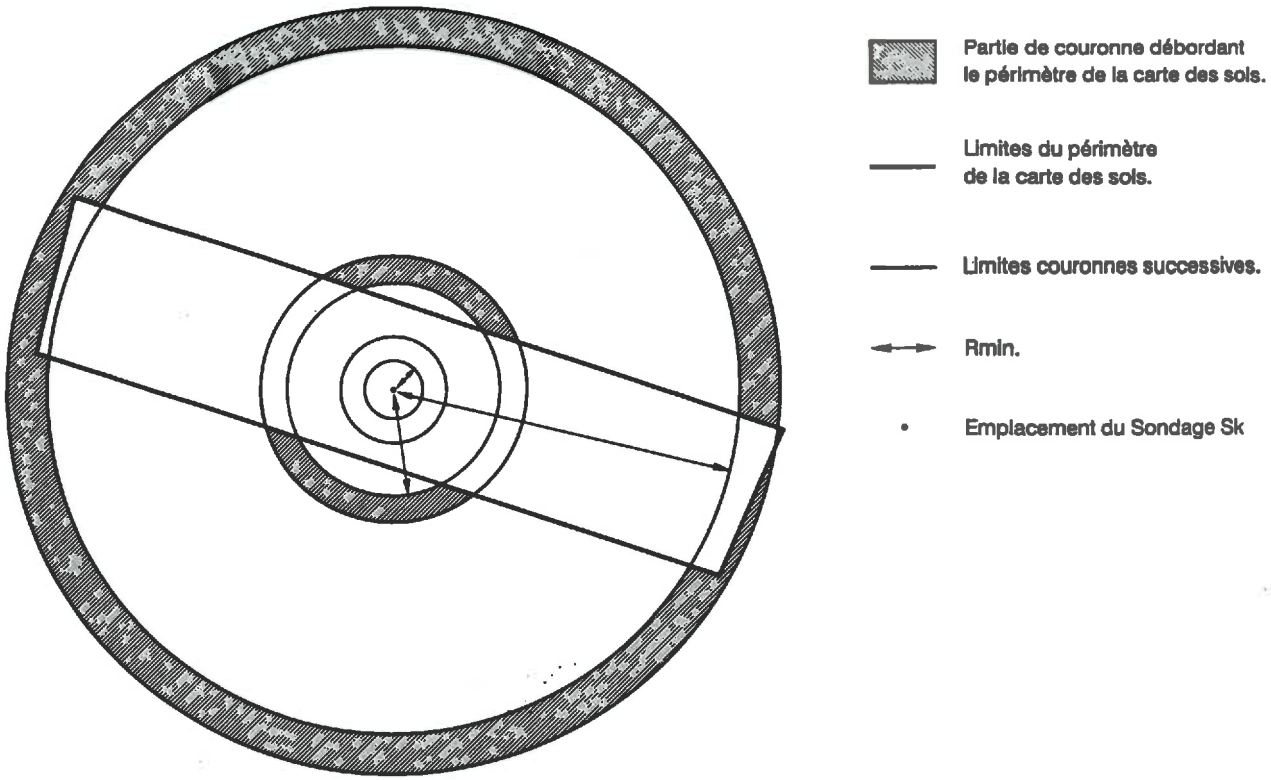


Figure 21: augmentation de superficie relative des zones situées hors secteur de référence avec le rayon minimum des couronnes (r0)

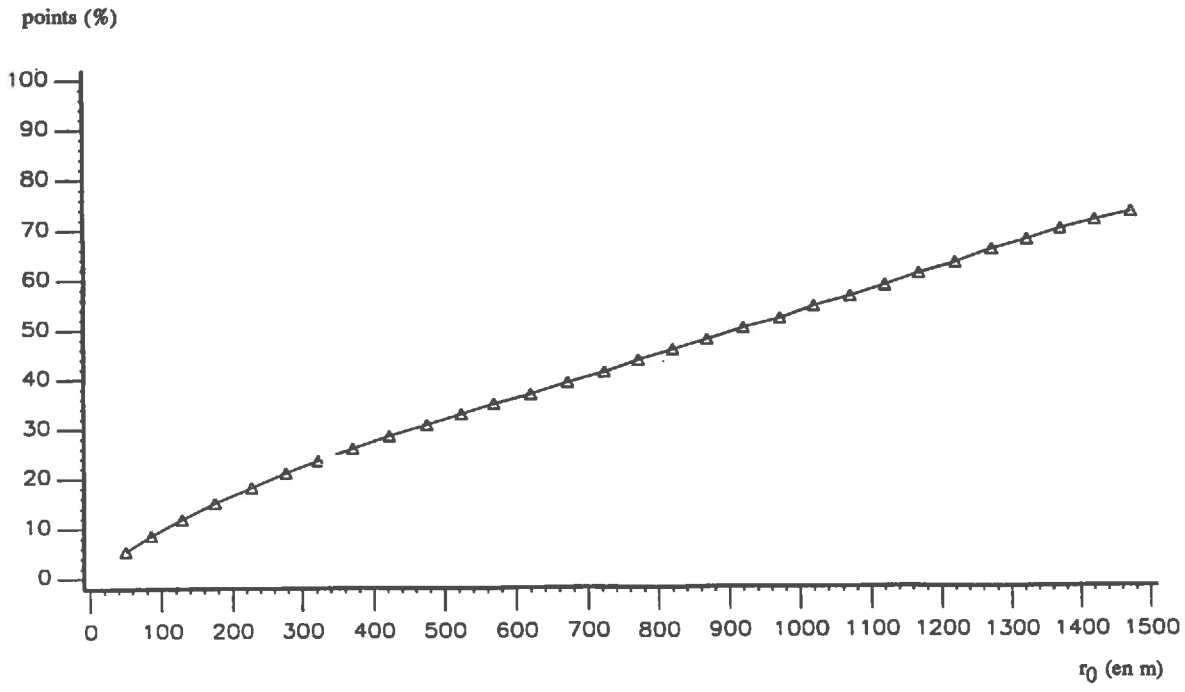


Figure 22: pourcentage de points hors périmètre avec l'augmentation de r_0

$$w_k = (1/r_{0k}^a) / (\Sigma_{(2...f)}\ 1/r_{0k}^a) \qquad [32]$$

avec: r_{0k}^a: valeur prise par le paramètre r_0 dans l'expression de la règle Rg_k
$(\Sigma_{(2...f)}\ 1/r_{0k}^a)$: facteur correctif permettant de ramener les résultats entre 0 et 1
a: constante déterminant l'intensité de la pondération inverse sur r_0.

Le facteur w_k étant désormais défini, il convient de l'introduire dans la formule [30] calculant les probabilités résultantes. Ainsi, la probabilité d'apparition d'une unité U_j au point x à l'étape t_f se calcule comme suit:

$$p_j\ (x,t_f) = (\Sigma_{(k=2...f)}\ \pi_{jk}(x)/r_{0k}^a) / (\Sigma_{(k=2...f)}\ 1/r_{0k}^a) \qquad [33]$$

Dans ce sous-chapitre, ont été construits les algorithmes permettant d'extraire les règles de voisinage puis de les utiliser dans le cadre d'un retour à la parcelle automatisé.Deux étapes ont été nécessaires.

1) **D'abord, ont été extraites de la carte du secteur de référence, des règles de voisinage "si(prémisse) alors (conclusion)." permettant de délivrer au voisinage d'un sondage donné, une prédiction dont la forme est compatible avec celle choisie lors du chapitre 2. Chaque règle de voisinage obtenue fournit une probabilité d'occurrence de chaque unité de sol applicable à un point. Il faut connaître, sur ce point, d'une part sa distance à un sondage préalablement rattaché à une unité de sol ("rayon de voisinage"), et d'autre part, sa position topographique relativement à ce sondage ("sens de voisinage"). Pour celà, un découpage de l'espace autour et par rapport au sondage ("zones d'isoprédictions") a été proposé sur la base de variables caractérisant la position relative d'un pixel par rapport à un sondage. Une de ces variables ("sens de voisinage") est définie par rapport au "motif d'unités de sol" ("soil combination" au sens de FRIDLAND) que révèle la carte.**

2) **Ensuite, dans la perspective d'avoir à faire la synthèse entre plusieurs prédictions du type précédent, une formule de combinaison a été proposée (formule [33]). Elle revient, pour une unité de sol et un point de prédiction donnés, à faire la moyenne arithmétique des différentes probabilités fournies par chaque règle, pondérée par un facteur inverse de r_0, paramètre traduisant la distance entre le point étudié et les différents sondages déclenchant les règles. Le choix de cette pondération s'appuie sur une analyse détaillée de la qualité supposée des prédictions délivrées par les règles et des facteurs susceptibles d'influençer cette qualité.**

Ces deux algorithmes permettent donc d'envisager maintenant d'automatiser le retour à la parcelle sur les bases définies au chapitre 2. Ainsi, si l'on reprend la notation utilisée dans ce chapitre, il est désormais possible de délivrer une prédiction finale, permettant, pour chaque point de la parcelle à cartographier, de fournir:

- **l'unité U_L prédite: Il s'agit de l'unité de sol obtenant la plus forte probabilité d'apparition**

- **le risque d'erreur re(x) correspondant à la somme des probabilités d'apparition des unités autres que U_L**

Avant d'appliquer les règles de voisinage à l'extérieur du secteur de référence, il convient de trouver la valeur de la pondération, fonction inverse de r_0, permettant la combinaison des règles. Cette valeur sera obtenue par calage en appliquant, dans un premier temps les règles de voisinage sur le secteur de référence.

2. APPLICATION DES REGLES DE VOISINAGE A L'INTERIEUR DU SECTEUR DE REFERENCE

Afin d'être en mesure d'évaluer la qualité des prédictions fournies par les règles de voisinage, il convient maintenant de déterminer un protocole d'utilisation de ces règles. Elles seront appliquées, dans un premier temps à l'intérieur du secteur de référence. Ce même protocole sera repris également dans le prochain chapitre pour valider les prédictions à l'extérieur du secteur de référence. Il sera présenté dans un premier paragraphe de ce sous-chapitre, puis sera mis en oeuvre avec deux objectifs précis:

- caler la pondération de la formule [33];
- évaluer l'ampleur de la dégradation de l'information entre la vraie carte du secteur de référence et l'image qu'en donne les prédictions.

2.1. Protocole d'étude des performances des règles de voisinage

Comme pour la phase d'acquisition des règles de voisinage, les points qui feront l'objet de prédictions se disposent sur une grille régulière de pas 50m. Chaque point représente en fait une surface élémentaire (le pixel) dont il constitue le centre. Pour que les règles de voisinages soient mises en oeuvre, il faut en principe des sondages, c'est à dire des points particuliers où une observation directe de la couverture pédologique au moyen d'une tarière permet de connaître l'unité de sol qui occupe ce point.

Dans ce travail, les sondages seront en fait simulés en rattachant le point concerné à l'unité de sol de la vraie carte dans laquelle il tombe. En d'autres termes, pour permettre le déclenchement d'une règle, à chaque étape, sera dévoilée une petite parcelle d'information de la vraie carte correspondant à un sondage supposé réalisé. Cet artifice permet d'éviter la mise en oeuvre de la phase de rattachement d'un sondage à une unité de sol qui, comme précisé antérieurement, n'est pas envisagée au cours de ce travail. Dans le cas où le sondage a fait l'objet d'une prédiction antérieure en sa qualité de simple point, cette dernière sera annulée au profit du résutat issus du rattachement direct. En conséquence, si un point x devient le sondage x_k rattaché à l'unité U_l, la prédiction prendra la forme suivante:

$$u(x_k) = U_L \quad ===> \quad p_L(x_k, t_k) = 1$$
$$p_j(x_k, t_k) = 0 \; \forall j \neq l \qquad [34]$$

Cette formule traduit en effet la certitude d'avoir U_L et donc l'impossibilité d'avoir d'autre unité que U_L

Les résultats étant suceptibles de varier suivant le nombre de sondages réalisés, diverses densités de sondages seront successivement testées. Pour ce qui concerne le secteur de référence, 4 densités seront définies par des sous-échantillonnages de plus en plus serrés de la grille de départ (1 sondage tous les 256, 64, 16 et 4 points):

sous-échantillonnage	1/256	1/64	1/16	1/4
Distance minimum entre sondages	800 m	400 m	200 m	100 m
Nombre théorique de sondages	15	59	236	945
Nombre réel[12]	11	60	236	931
Densité (/100 ha)	1.2	6.4	25.1	99.0
% de l'ensemble des points	0.3	1.6	6.2	24.6

***Tableau 14:** protocole d'implantation des sondages sur le secteur de référence*

Le sous-échantillonnage systématique permet (figure 23, page suivante) de comparer des résultats du modèle en des endroits différents où il faut s'assurer qu'ils ont fait l'objet de la même densité de sondage. Il est ainsi simulé une série d'études pédologiques effectuées avec des densités de sondages croissantes et selon une prospection systématique et régulière du terrain sans stratégie particulière concernant l'implantation des sondages (stratégie dite "grid survey").

Un programme FORTRAN (cf annexe 9) a été écrit pour mettre en oeuvre ce protocole. Il utilise les règles et la formule de combinaison des prédictions mises au point précédemment. Il fournit pour chaque point du secteur de référence une prédiction d'unité de sol et son risque associé.

---

12 La différence observée entre nombre théorique et nombre réel correspond en fait à l'erreur systématique apparaissant lorsqu'on estime la surface d'une plage cartographique par comptage des points d'une maille régulière couvrant cette plage. De nombreux auteurs se sont penchés sur ce problème. En particulier, FROLOV et MALING (1969) proposent, pour estimer cette erreur, la formule suivante:

$\log f = K + 1.5 \log d - 0.875 \log A$

avec:

A: aire de la plage
f: erreur relative sur l'aire de la plage
d: longueur du côté de la maille élémentaire
K: constante de forme de la plage

Selon cette formule, l'erreur d'estimation sur l'aire augmente lorsque la maille devient plus large. Donc, de la même façon, l'écart entre nombre théorique et nombre réel de sondages est d'autant plus grand que la maille de sondages est lâche.

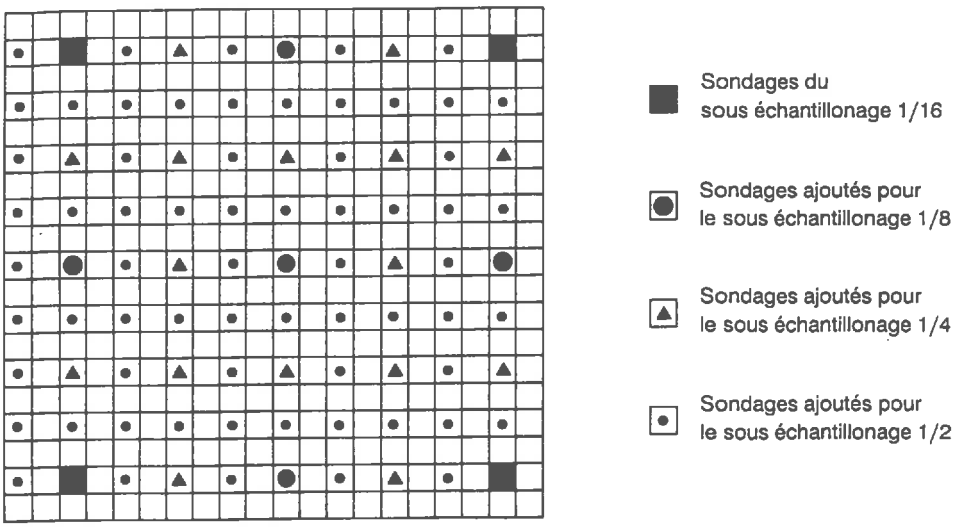


Figure 23: localisation des sondages correspondants aux différents sous-échantillonnages

Le programme, installé sur une station de travail UNIX, utilise trois fichiers sources:
- le fichier des règles de voisinage extraites selon l'algorithme présenté au chapitre précédent;
- le fichier des 3779 points centraux de pixels sur lesquels portent les prédictions (moins ceux assimilés à des sondages) renseignés par leurs coordonnées x,y,z;
- le fichier des sondages renseignés également par x,y et z, par l'unité de sol dans laquelle tombe leurs centroïdes et par une variable particulière qui identifie, en fonction de la densité traitée, les pixels assimilés à des sondages.

Les performances de l'outil informatique de prédiction des unités de sol ainsi programmé sont mesurées pour chaque densité en comparant les prédictions aux valeurs réelles. Plusieurs types d'erreur sont distingués[13]. Dans la mesure où ils sont destinés à alimenter les discussions qui suivent, dans ce chapitre et dans les suivants, une définition préalable s'impose.

1) **L'erreur apparente** (notée Ea) correspondra à la proportion, calculée sur l'ensemble des points de la zone étudiée, des points mal classés (c'est à dire affectés à une autre unité de sol que celle dans laquelle il tombe au vu de la vraie carte).
2) **L'erreur de prédiction** (notée Ep) correspondra également à la proportion des points mal classés, mais, cette fois, calculée sur l'ensemble des points diminués de ceux ayant été assimilés à des sondages. Cette erreur sera généralement préférée à la précédente dans l'analyse des résultats car elle évite un biais de l'erreur apparente qui devient génant lorsque deux densités différentes du plan de sondages sont comparées. En effet, les sondages (fictifs) étant par nature sans erreur, leur prise en compte dans le calcul favorise les fortes densités par rapport aux faibles. Naturellement, l'erreur de prédiction sera

13 Les différents types d'erreur définis dans cette partie ne correspondent pas à la classification formulée dans la partie précédente. Les objectifs et les raisonnements étant très différents entre les deux parties, elles ont nécessité chacune la définition d'une terminologie propre.

toujours plus élevée que l'erreur apparente, l'écart se creusant avec l'augmentation du nombre de sondages.

Afin d'affiner l'analyse des performances du modèle, il sera distingué deux composantes de l'erreur de prédiction:

- **l'erreur de délimitation** (notée Epd) concerne les points "mal classés" dont les pixels qu'ils représentent sont en réalité occupés en partie par l'unité prédite (figure 24); Compte tenu de la taille du pixel, cette erreur correspond en fait à une erreur de positionnement de limite inférieure à 35m (distance séparant le centroïde à l'angle du pixel) d'où son qualificatif;
- **l'erreur de prédiction vraie** (notée Epv) correspond aux points "mal classés" qui ne sont pas dans la situation décrite précédemment. C'est ce type d'erreur qui fera le plus objet d'attention dans la perspective de mieux comprendre le fonctionnement de l'outil de prédiction des sols.

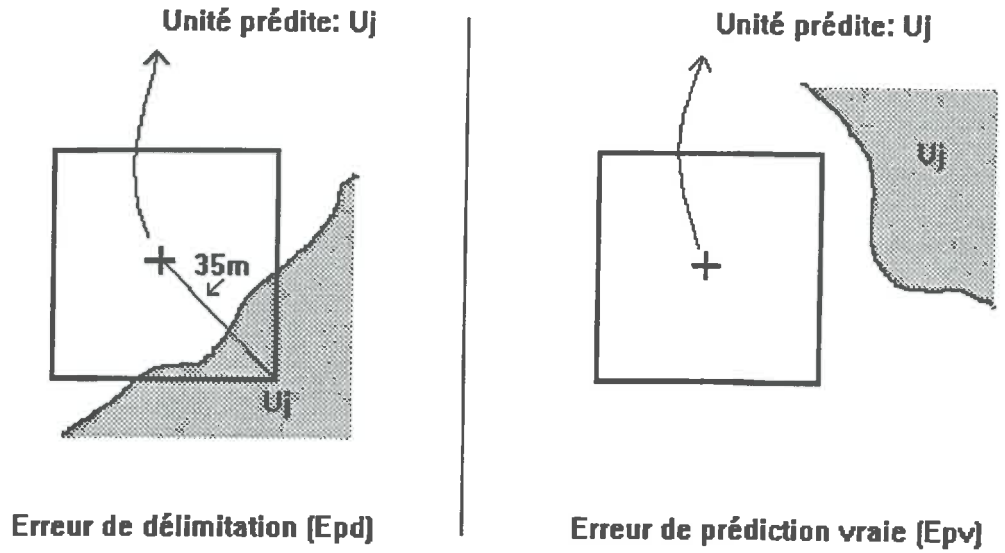


Figure 24: Les deux composantes de l'erreur de prédiction

Après chaque essai du modèle, sont introduits dans une base de donnée (PARADOX) les fichiers nécessaires au calcul de chacune des erreurs présentées ci-dessus: résultats des prédictions d'une part, vraies unités d'autre part et, pour les besoins du calcul de Epd, listes des unités occupant le périmètre de chaque pixel (obtenues par croisement préalable sous ARC/INFO).

2.2. Résultats des prédictions sur le secteur de référence

L'application de l'outil de prédiction sur le secteur de référence a pour but de caler l'intensité de la pondération entre les règles appliquées en un point et d'examiner quel degré maximum de performance peut être atteint, étant entendu que dans tout autre lieu géographique le risque d'erreur sera plus élevé. Bien entendu, ceci ne constitue pas une validation des règles de prédiction puisque dans ce cas, elles sont appliquées aux mêmes données qui ont par ailleurs servi à les générer.

Deux questions vont être successivement traitées:
- quelle type de pondération en fonction du rayon de voisinage minimum doit être appliquée ? Cette pondération intéresse la valeur de la constante a (formule [33])
- comment évoluent les différentes erreurs en fonction de la densité de sondages?

2.2.1. Calage de la pondération des prédictions issues des différentes règles de voisinage utilisées

La figure 25 représente le résultat d'essais de différentes valeurs de a (formule [33]), depuis a=0 (pas de pondération appliquée) jusqu'à a=4. Pour chaque essai, les 4 différentes densités de sondages présentées dans le chapitre précédent sont testées. Les courbes obtenues représentent l'évolution de l'erreur de prédiction avec ces densités de sondages.

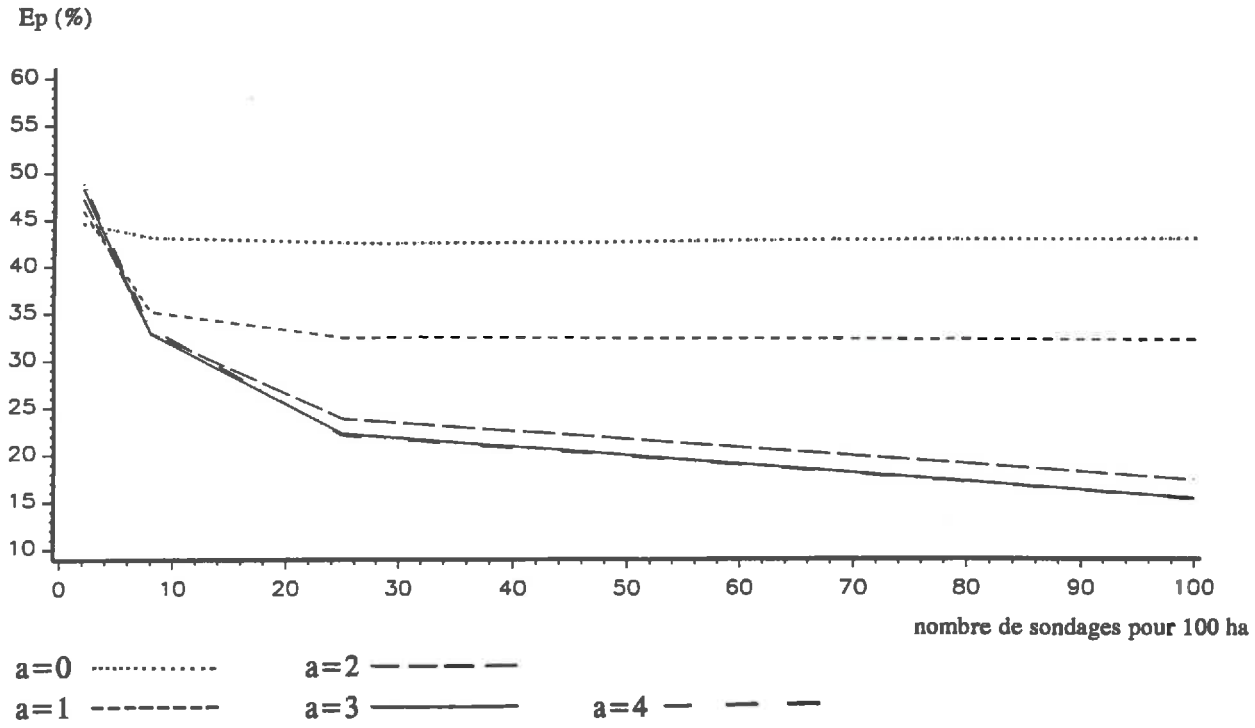


***Figure 25:** erreur de prédiction sur le secteur de référence; essais de différentes pondérations pour combiner les règles de voisinage.*

Les courbes obtenues révèlent une diminution quasi générale de l'erreur lorsque a est augmentée jusqu'à a=3. Par contre, a=4 n'apporte plus d'amélioration. Ceci confirme le caractère perturbant des prédictions issues de sondages lointains puisque leur diminution de poids entraîne une amélioration générale et sensible des performances du modèle.

L'amélioration est d'autant plus importante que la densité de sondages est forte. Ceci s'explique dans la mesure où l'augmentation de la densité des sondages augmente également les chances, pour un pixel donné, d'être proche d'un sondage. de ce fait, il y a plus de chances pour qu'une "bonne" règle de prédiction figure dans l'ensemble des règles émises sur un point. En cas de pondération, cette règle devient prépondérante, compte tenu de la proximité du sondage qui la déclenche avec le point traité.

Au vu de ces résultats une pondération inverse du cube du rayon de voisinage sera préférée. Dans la suite de l'exposé, l'examen des résultats ne s'intéressera plus qu'à ce choix. Sa pertinence sera testée ultérieurement, à l'occasion de la validation du modèle à l'extérieur du secteur de référence.

2.2.2. Evaluation de la qualité des prédictions à l'intérieur du secteur de référence

Dans ce chapitre seront présentés, mais peu discutés, les résultats des performances du modèle. Il s'agit surtout d'établir des éléments de comparaison permettant d'alimenter l'interprétation future des résultats de validation. De plus, certains aspects du comportement du modèle se retrouveront également à l'extérieur du secteur de référence. Leur analyse fine ne sera réalisée qu'après avoir obtenu tous les résultats les concernant.

Le tableau 15 regroupe l'ensemble des résultats.

Sous-échantillonnage	1/256	1/64	1/16	1/4
Dens. moy. (sondages/100 ha)	1.2	6.4	25.1	99.0
Ep (%)	48.3	32.8	22.2	15.3
Epd (%)	9.6	11.8	12.1	11.9
Epv (%)	38.7	21.0	10.1	3.4

Tableau 15: qualité des prédictions sur le secteur de référence et densités de sondages

L'erreur de prédiction diminue avec l'augmentation de la densité de sondages (de 48.3 à 15.3%). Cependant, elle semble freinée au fur et à mesure que celle-ci augmente (cf figure 25 pour a=3). Les premiers sondages introduits semblent donc avoir plus d'intérêt que les suivants. D'autre part, malgré une densité forte (1/ha, soit 1 sondage pour 4 points), les prédictions fausses représentent une valeur non négligeable.

L'analyse des deux composantes de l'erreur de prédiction révèle 2 évolutions différentes en fonction de la densité de sondage (figure 26, page suivante).

1) L'erreur de délimitation reste quasiment stable. Elle ne semble donc pas influencée par le nombre de sondages réalisés.

2) Au contraire, l'erreur de prédiction vraie diminue avec la densité des sondages. Ainsi, c'est elle qui imprime l'évolution constatée précédemment.

Il en résulte un changement dans l'importance relative des deux composantes suivant la densité de sondages considérée. Pour une faible densité, l'erreur de prédiction vraie domine largement. Par contre une forte densité voit l'erreur de délimitation devenir majoritaire, l'inversion de tendance se situant au niveau de la densité 25 sondages/100ha.

Pour la densité la plus forte, l'erreur de prédiction vraie atteint des valeurs particulièrement faibles (3,4%) ce qui traduit finalement une bonne capacité du modèle à retrouver la carte réelle, aux distortions de limites près (11.9%).

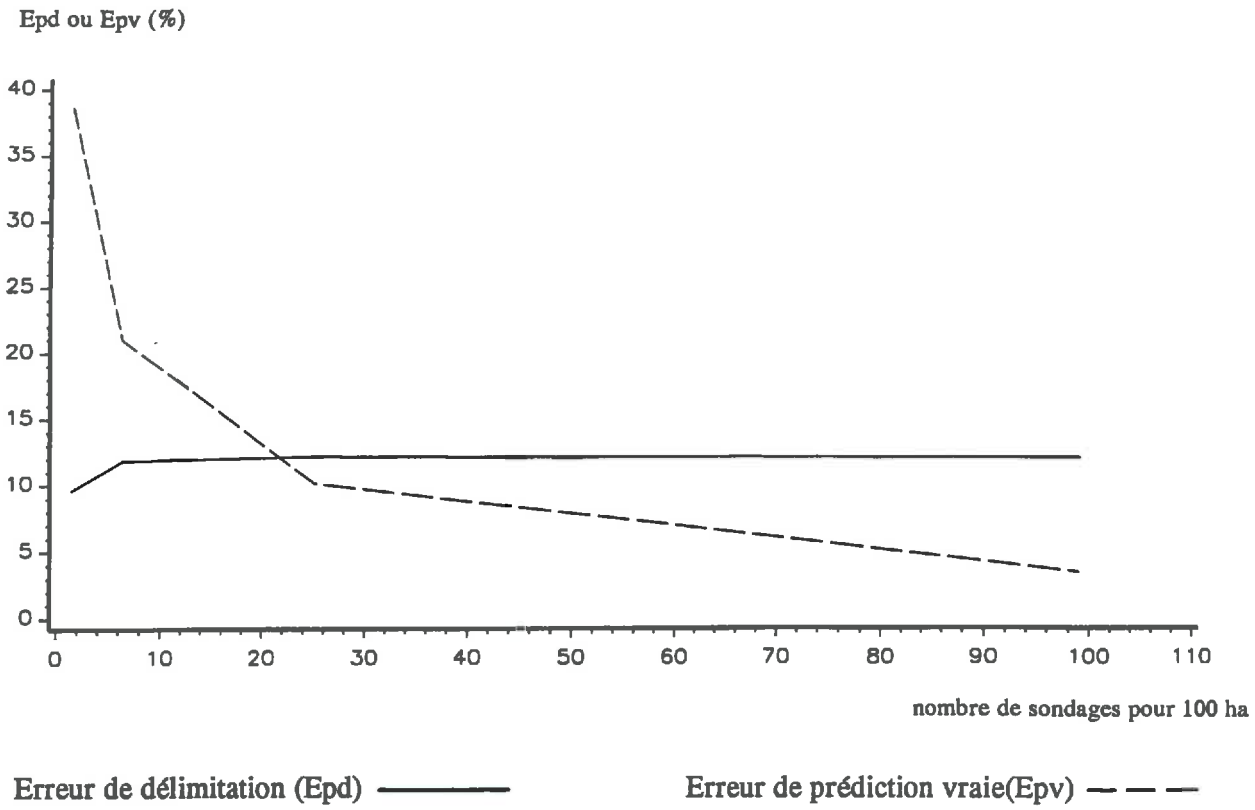


Figure 26: évolution des deux composantes de l'erreur de prédiction sur le secteur de référence

Au cours du chapitre 7, a été construit un outil de prédiction des unités de sol capable d'utiliser des observations ponctuelles de la couverture pédologique pour prédire, sur les lieux voisins non encore prospectés, l'unité de sol présente. Cet outil exploite uniquement les relations de voisinage entre unités de sol telles qu'elles peuvent apparaître sur la carte du secteur de référence. Deux aspects distincts ont du être abordés successivement.

1) l'extraction des règles de voisinage à partir des données du secteur de référence, selon le formalisme défini au chapitre 2. Un algorithme a spécifiquement été conçu à cet effet, il est résumé dans la conclusion du premier sous-chapitre.

2)la combinaison des prédictions issues de plusieurs règles appliquées en un même point. Une formule de combinaison a été choisie. Elle met en jeu un paramètre d'intensité de pondération qui a fait l'objet d'un calage grâce à un premier essai de l'outil de prédiction des sols à l'intérieur du périmètre du secteur de référence.

L'analyse des performances de l'outil de prédiction a été menée d'abord à partir des données du secteur de référence selon un protocole spécifique décrit dans ce chapitre. Elle montre que de bonnes performances sont théoriquement possibles à l'extérieur, aux imprécisions de limite près. Bien entendu, pour que ces espoirs se concrétisent, il faudrait désormais que l'hypothèse de stabilité des lois de voisinage s'avére juste. Seule l'application de l'outil de prédiction aux secteurs de validation permettra d'apporter des éléments de réponse à cette interrogation.

CHAPITRE 8

UTILISATION DES REGLES DE VOISINAGE POUR PREDIRE LES UNITES DE SOL A L'EXTERIEUR DU SECTEUR DE REFERENCE

La validation entreprise doit permettre d'évaluer les performances de l'outil de prédiction mis en situation réelle d'utilisation. L'objectif est d'analyser finement les erreurs de prédiction commises de façon à préciser l'influence de divers facteurs sur le comportement de l'outil: type de milieu pédologique, densité de prospection, localisation des sondages,.... Cette analyse critique doit également déboucher sur la mise en évidence des limites actuelles, permettant ainsi d'indiquer les perspectives d'amélioration possible.

Comme dans la partie précédente, la qualité des prédictions sera évaluée sur 3 secteurs de validation, couvrant au total 130 ha (521 pixels). Ces secteurs font l'objet d'une description détaillée au chapitre 3. Il s'agit:

- du secteur de La Roubiaire (noté RB) recoupant les différents niveaux de terrasse de la rive droite de l'Hérault;
- du secteur de Montmau (noté MT) situé sur les collines et colluvions molassiques de la vallée de l'Hérault;
- du secteur de Lézignan la Cèbe (noté LZ) localisé au contact des alluvions récentes et anciennes de l'Hérault.

Le protocole d'évaluation des performances de l'outil de prédiction, défini au cours du chapitre précédent et appliqué auparavant sur le secteur de référence, a été reconduit. Seules les densités testées ont dû être modifiées compte tenu de la taille plus restreinte des secteurs de validation. La densité correspondant au sous-échantillonnage 1/256 a été abandonnée puisqu'un seul sondage aurait été possible par secteur. Elle est remplacée par une nouvelle série de densités intermédiaires (sous-échantillonnages 1/36 et 1/9). A propos de ces nouvelles densités, il est important de souligner qu'elles correspondent à deux ensembles de sondages en grande partie disjoints de ceux de la série précédente (1/64,1/16,1/4). L'effet "localisation des sondages" devra donc être gardé à l'esprit pour expliquer d'éventuelles différences entre ces deux séries.

Toujours à cause de la petite taille des secteurs de validation, une précaution supplémentaire a été introduite pour choisir les populations de sondages: afin de ne pas favoriser une extrémité de secteur par rapport à l'autre, le premier sondage est positionné de manière à correspondre au pixel situé le plus près possible du centroïde du périmètre du secteur.

Le tableau 16 (page suivante) rappelle la densité théorique et indique la densité et le nombre (chiffre entre parenthèses) de sondages relatifs au sous-échantillonnages retenus.

Il existe des différences substantielles entre densités attendues et densités obtenues. Ces différences sont d'autant plus importantes que les densités sont faibles (voir explication au revoi de page n°12). La manifestation la plus flagrante de ce phénomène est que les densités obtenues par les sous-échantillonnages 1/64 et 1/36 sont très voisines; elles s'égalisent même en ce qui concerne le sous secteur de Montmau.

Sous-échantillonnage	1/64	1/36	1/16	1/9	1/4
Densité. Théorique.	6,25	11.1	25.0	44.4	100
Rb	6.7 (3)	11.1 (5)	24.4 (11)	44.4 (20)	100.1 (45)
Mt	8.9 (4)	8.9 (4)	24.4 (11)	46.7 (21)	100.6 (44)
Lz	7.5 (3)	10.0 (4)	25.0 (10)	47.5 (19)	107.5 (43)
Ensemble	7.7 (10)	10.0 (13)	24.6 (32)	46.2 (60)	101.5 (132)

***Tableau 16:** densités de sondages (nombre de sondages/100 ha) et nombre de sondages (chiffres entre parenthèse) des différents sous-échantillonnages réalisés.*

Le protocole ainsi décrit a été mis en oeuvre grâce au même programme FORTRAN que celui utilisé précedemment sur les données du secteur de référence. Cependant, compte tenu du faible nombre de points traités simultanément (au maximum 179), ce programme peut fonctionner sous MS-DOS avec des temps de réponse acceptables (de l'ordre de 3mn). Comme précédemment, les fichiers résultats ont été par la suite repris sous PARADOX pour permettre le calcul des différents termes d'erreur utiles pour l'analyse.

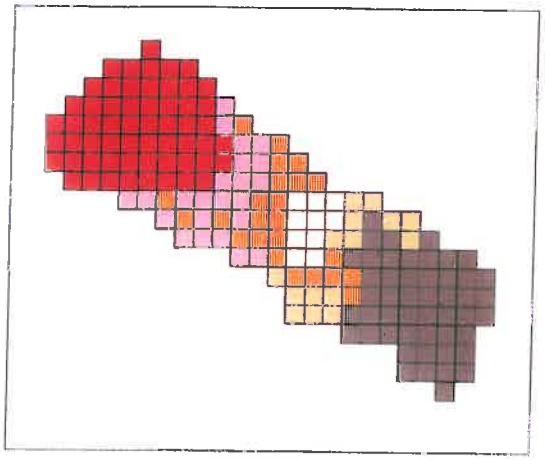
De plus, ces fichiers ont également été réintégrés sous ARC/INFO. Chaque point représentant le pixel dont il est le centre, il est possible de réaliser des cartes de prédiction d'unités de sol, pour chaque secteur de validation et chaque densité de sondages mise en oeuvre (planches 5a, 5b, 5c)

L'analyse des résultats sera présentée en trois sous-chapitres:

- dans un premier, une analyse globale, c'est à dire fondée uniquement sur l'erreur de prédiction, permettra d'une part de vérifier la pertinence de la pondération choisie et, d'autre part, de mettre en évidence les facteurs influençant les performances de l'outil de prédiction;
- dans un deuxième sous- chapitre, l'analyse sera poussée plus avant afin de mieux comprendre l'origine des différences observées; dans ce but, les deux termes "erreur de délimitation" et "erreur de prédiction vraie" seront distingués;
- Le troisième sous-chapitre s'intéressera au risque d'erreur re(x) associé à chaque prédiction d'unité de sol fournie. Il permettra d'apprécier si l'outil de prédiction des unités de sol est capable d'identifier lui-même les lieux où il commet effectivement des erreurs.

***Planche 5a**: Cartographie du secteur de validation de La Roubiaire utilisant les lois de voisinage extraites du secteur de référence*

CARTE REELLE



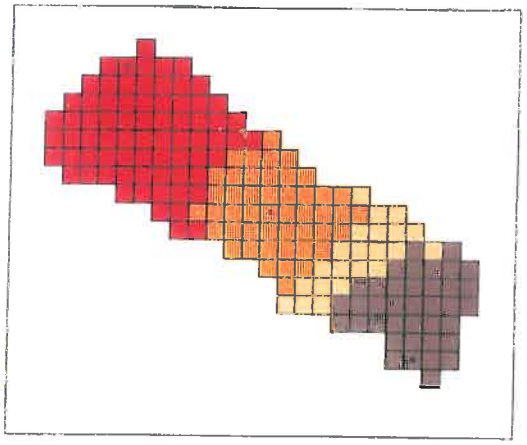
LEGENDE

Unite 1	Unite 7	Unite 13
Unite 2	Unite 8	Unite 14
Unite 3	Unite 9	Unite 15
Unite 4	Unite 10	Unite 16
Unite 5	Unite 11	Unite 17
Unite 6	Unite 12	Inconnue

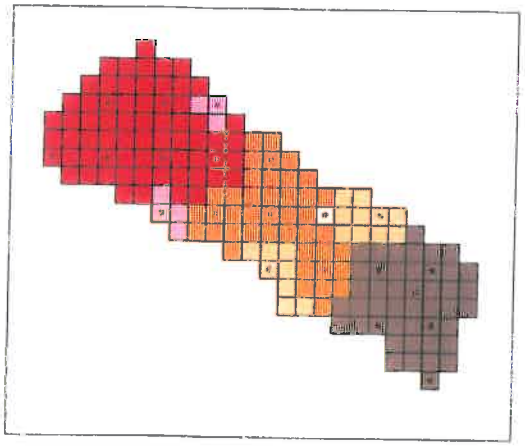
* emplacement des sondages

Echelle 1/20 000 (un pixel = 50m X 50m)

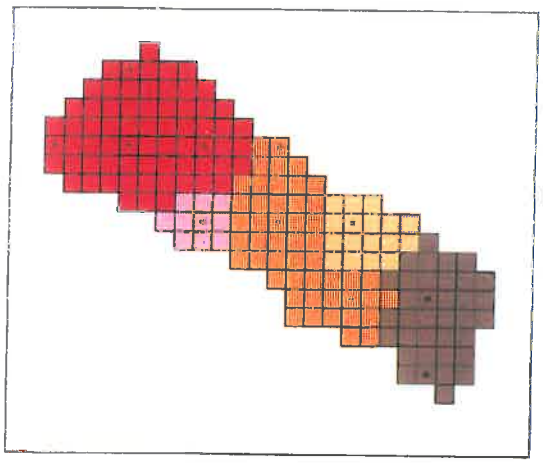
CARTES PREDITES AVEC DIFFERENTES DENSITES DE SONDAGES
strategie : semis regulier de sondages ("grid survey")



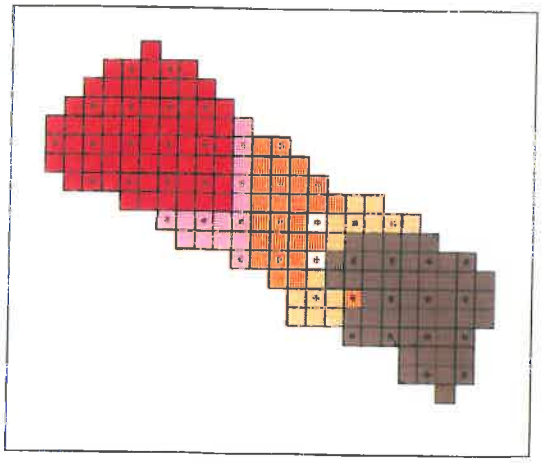
Sous echantillonage 1/8 : 3 sondages
soit 6.7 pour 100 ha



Sous echantillonage 1/3 : 20 sondages
soit 44.4 pour 100 ha



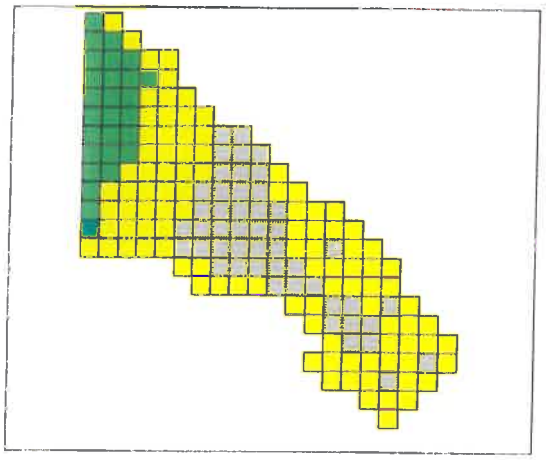
Sous echantillonage 1/4 : 11 sondages
soit 24.4 pour 100 ha



Sous echantillonage 1/2 : 45 sondages
soit 100.1 pour 100 ha

***Planche 5b:** Cartographies du secteur de validation de Montmau utilisant les lois de voisinage extraites du secteur de référence*

CARTE REELLE



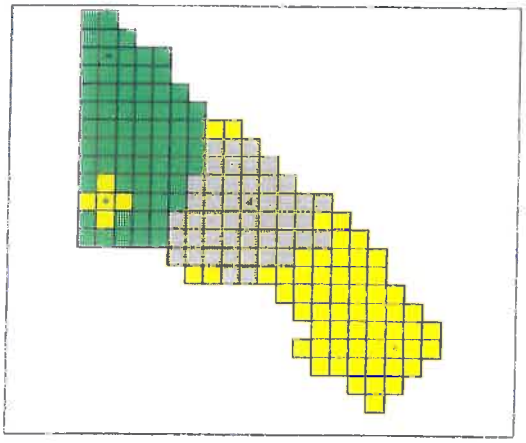
LEGENDE

Unite 1
Unite 2
Unite 3
Unite 4
Unite 5
Unite 6
Unite 7
Unite 8
Unite 9
Unite 10
Unite 11
Unite 12
Unite 13
Unite 14
Unite 15
Unite 16
Unite 17
Inconnue

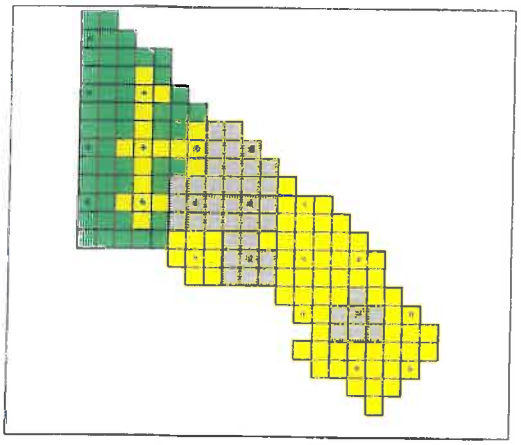
emplacement des sondages

Echelle 1/20 000 (un pixel = 50m X 50m)

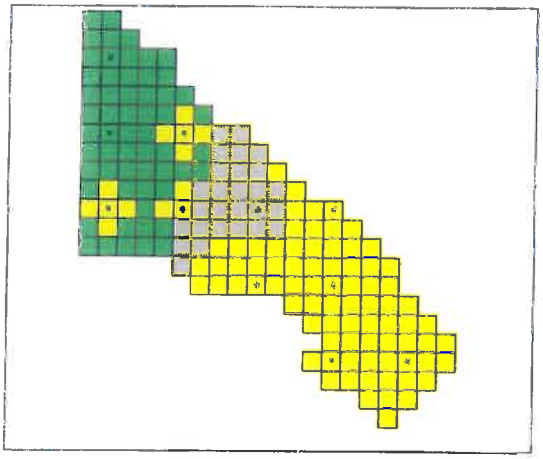
CARTES PREDITES AVEC DIFFERENTES DENSITES DE SONDAGES
strategie : semis regulier de sondages ("grid survey")



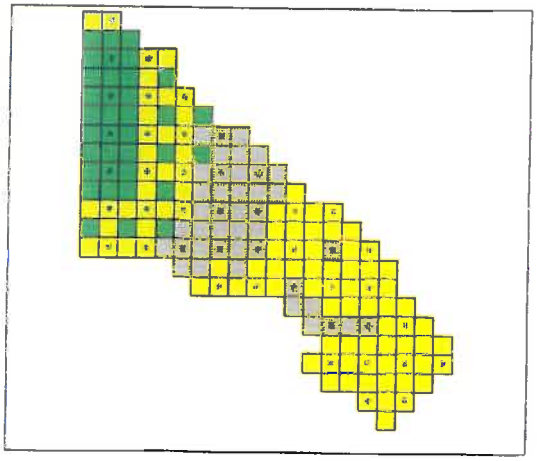
Sous echantillonage 1/8 : 4 sondages
soit 8.9 pour 100 ha



Sous echantillonage 1/3 : 21 sondages
soit 46.7 pour 100 ha



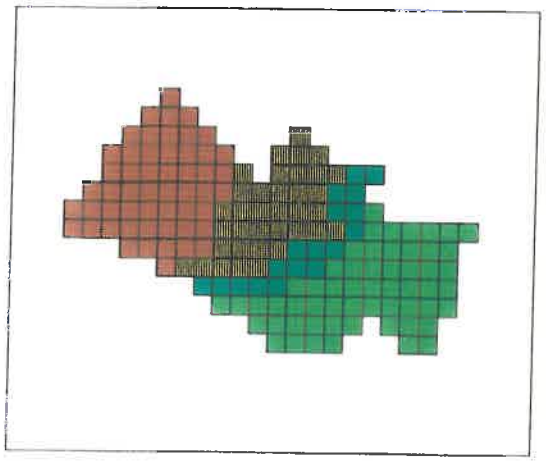
Sous echantillonage 1/4 : 11 sondages
soit 24.4 pour 100 ha



Sous echantillonage 1/2 : 44 sondages
soit 100.6 pour 100 ha

***Planche 5c:** Cartographies du secteur de validation de Lézignan la Cèbe utilisant les lois de voisinage extraites du secteur de référence*

CARTE REELLE



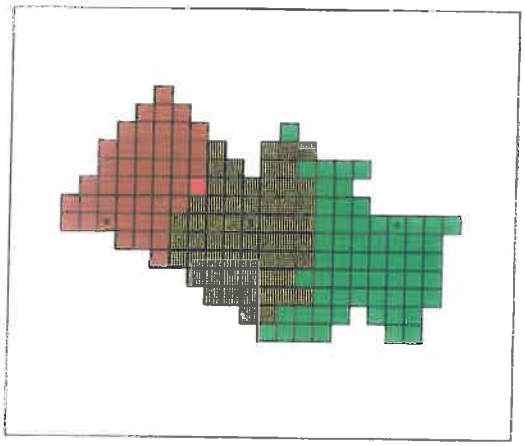
LEGENDE

Unite 1	Unite 7	Unite 13
Unite 2	Unite 8	Unite 14
Unite 3	Unite 9	Unite 15
Unite 4	Unite 10	Unite 16
Unite 5	Unite 11	Unite 17
Unite 6	Unite 12	Inconnue

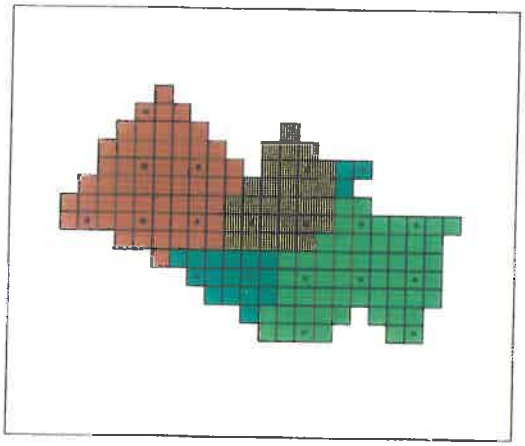
▪ emplacement des sondages

Echelle 1/20 000 (un pixel = 50m X 50m)

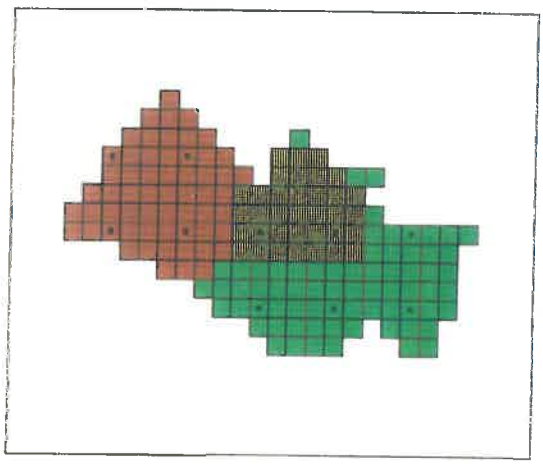
CARTES PREDITES AVEC DIFFERENTES DENSITES DE SONDAGES
strategie : semis regulier de sondages ("grid survey")



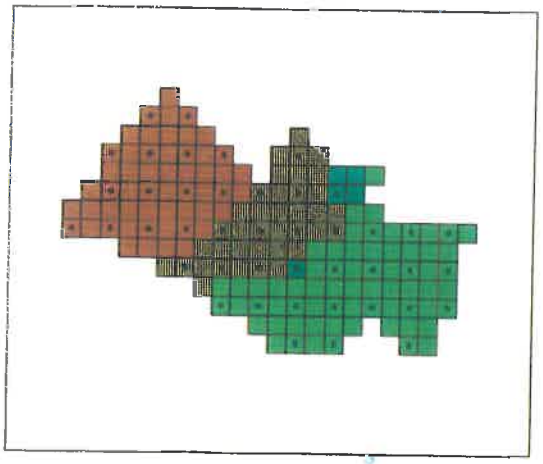
Sous echantillonage 1/8 : 3 sondages
soit 7.5 pour 100 ha



Sous echantillonage 1/3 : 19 sondages
soit 47.5 pour 100 ha



Sous echantillonage 1/4 : 10 sondages
soit 25 pour 100 ha



Sous echantillonage 1/2 : 43 sondages
soit 101.5 pour 100 ha

1. ANALYSE DE L'ERREUR DE PREDICTION OBTENUE SUR LES SECTEURS DE VALIDATION

Dans un premier temps, il convient de vérifier la validité du calage de la pondération entre les différentes règles. En d'autre termes, le choix (a=3) résultant de l'application de l'outil de prédiction sur le secteur de référence est-il encore le meilleur lorsque l'outil fonctionne à l'extérieur de ce secteur ?

A cet effet, la figure 27 montre l'erreur de prédiction, calculée pour l'ensemble des trois secteurs de validation, en fonction de la densité de sondages. Les différentes courbes correspondent aux mêmes pondérations que celles testées sur le secteur de référence.

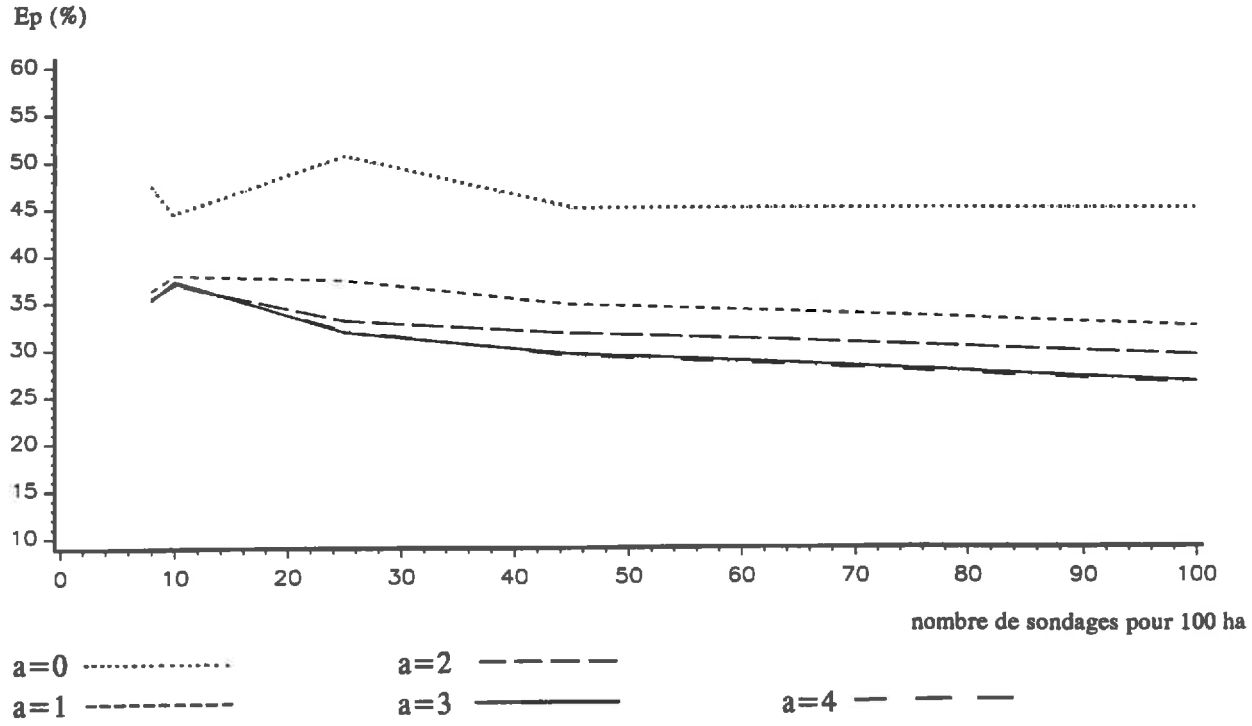


Figure 27: erreurs de prédictions sur les secteurs de validation en fonction de la densité de sondages; essai de différentes pondérations pour combiner les règles

Il n'y a pas de différences de comportement entre secteur de référence et secteurs de validation quant à l'ordre vis à vis des performances du modèle: nette différence entre absence (a=0) et présence de pondération, avec, comme pour le cas du secteur de référence, une erreur stable après a= 3.

Ceci est un premier résultat de validation puisqu'il est montré que le paramètre de calage, constitué par la constante a, et déterminé à partir des données connues du secteur de référence, peut être également utilisé à l'extérieur de ce secteur. Par la suite les résultats analysés utiliseront tous une pondération des prédictions de sol en fonction inverse du cube du rayon de voisinage.

Une première appréciation des performances du modèle est fournie par le tableau 17 (page suivante).

Sous-échantillonnage	1/64	1/36	1/16	1/9	1/4
Ea	34.6	36.4	29.8	26.1	19.6
Ep	35.3	37.3	31.8	29.5	26.3

Tableau 17: erreurs apparentes et erreurs de prédiction en fonction du sous-échantillonnage pratiqué.

L'erreur apparente est, comme prévu, systématiquement plus faible que l'erreur de prédiction. Les niveaux d'erreur obtenus sont entre 2 et 3 fois plus faibles que ceux des prédictions fondées sur les relations sols-paysage (chapitre 6). A titre de comparaison, pour le sous-échantillonnage 1/4, les résultats obtenus (19.6% d'erreur pour 1 sondage/ha) s'approchent de la qualité demandée a priori par FAVROT (1989) pour une étude de secteur de référence (entre 10 et 20% d'erreur pour une densité de 1 à 2 sondages/ha).

La figure 28 visualise l'évolution de l'erreur de prédiction déjà mentionnée dans le tableau précédent, comparée à celle obtenue auparavant dans le secteur de référence.

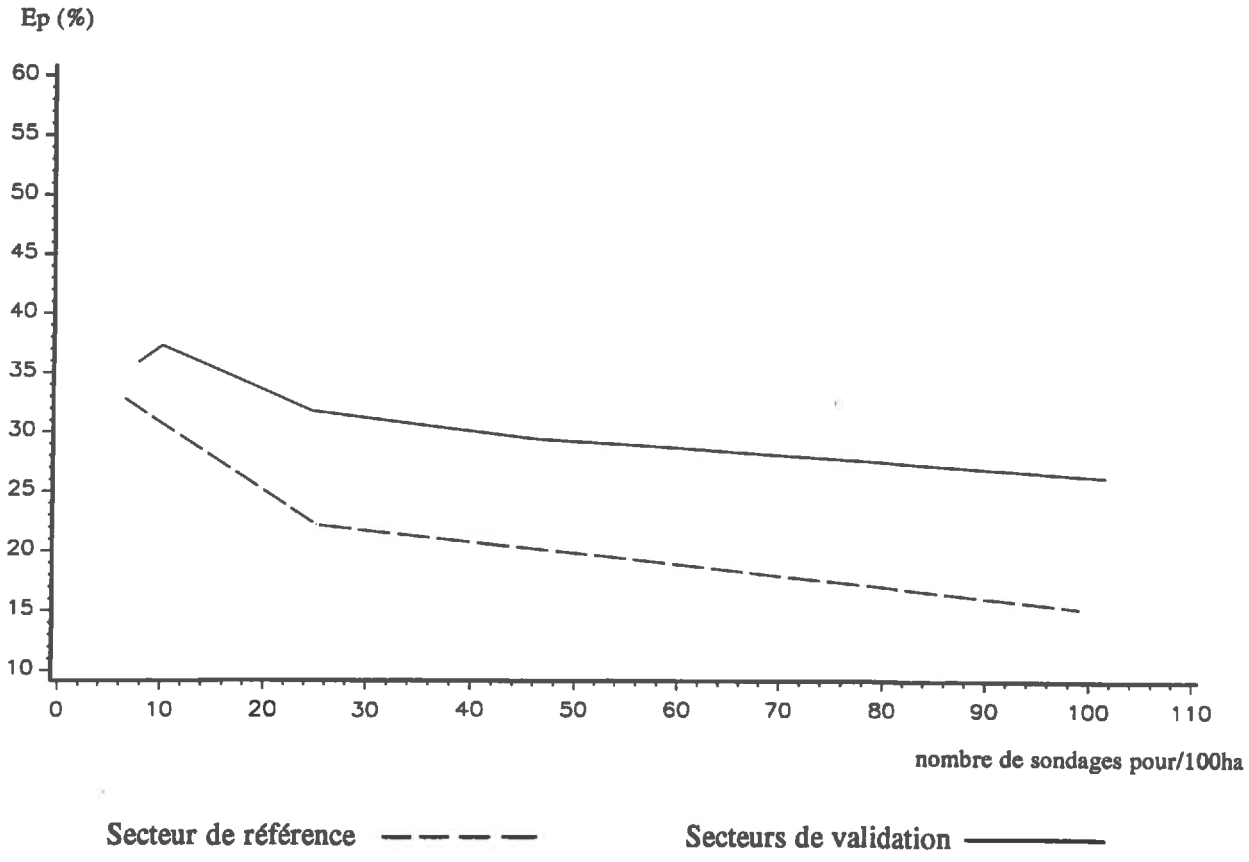


Figure 28: comparaison des erreurs de prédiction entre secteur de référence et secteurs de validation

Deux facteurs influençant les performances de l'outil de prédiction des unités de sol sont mis en évidence.

1) Le fait de mettre le modèle en situation réelle de prédiction à l'extérieur du secteur de référence dégrade naturellement ses performances. Cependant l'augmentation de l'erreur de prédiction reste relativement limitée (entre +2.5 et +11% d'erreur suivant les densités),
2) Comme pour le secteur de référence, l'augmentation du nombre de sondages a globalement pour effet de diminuer l'erreur de prédiction (- 9% d'erreur entre sous-échantillonnage 1/64 et 1/4). Cependant:
   - cette diminution est légèrement moins importante que pour le secteur de référence; le phénomène de perte d'efficacité de la densification des sondages, déjà observé sur ce dernier, est plus net encore sur les secteurs de validation; en conséquence, l'écart de performances entre secteur de référence et secteur de validation augmente légèrement avec la densité de sondages,
   - il existe un accident correspondant au sous-échantillonnage 1/36 dont l'explication sera fournie au vu de l'examen détaillé des performances des trois sous-secteurs.

Jusque là, seule l'erreur de prédiction calculée sur l'ensemble des secteurs de validation a été envisagée. Le tableau 18 donne maintenant les erreurs de prédictions mesurées séparément sur chaque secteur:

	1/64	1/36	1/16	1/9	1/4
La Roub.	31.8	31.6	29.8	26.9	25.8
Montmau	45.7	52.6	41.6	37.8	35.8
Lézignan	29.2	27.5	23.4	22.1	16.5

Tableau 18: évolution, par secteur, des erreurs de prédiction avec le sous-échantillonnage adopté.

Il met en évidence un importante variation suivant les secteurs en cause: l'écart d'erreur atteint,en moyenne, 19% entre les secteurs de Montmau et Lézignan la Cèbe.

Par ailleurs, il permet d'expliquer l'augmentation de l'erreur entre les sous-échantillonnages 1/64 et 1/36: elle provient uniquement du secteur de Montmau, l'évolution pour les autres secteurs étant normale. Pour ce premier secteur, les deux sous-échantillonnages considérés conduisent en fait à un même nombre de sondages (cf tableau 16), seul leur lieu d'implantation variant. L'importance de ce dernier facteur se trouve ainsi ponctuellement mise en évidence puisque l'écart d'erreur se trouve être non négligeable (6.9%).

En résumé, une première analyse des résultats du modèle permet d'établir deux constats.

1) **Les niveaux d'erreurs de prédiction atteints sont globalement satisfaisants dans la mesure où ils révèlent un faible écart de performances entre l'intérieur et l'extérieur du secteur de référence. Il faut rappeler qu'il n'en était pas ainsi pour les performances des prédictions fondées sur les relations sols-paysage (chapitre 6).**
2) **Plusieurs facteurs, plus ou moins attendus, semblent influencer la qualité des prédictions. Comme prévu, l'erreur décroît lorsque le nombre de sondages utilisés augmente. Les écarts de performance sont toutefois moins importants que ceux observés entre les sous secteurs de validation. Enfin, se trouve ponctuellement mise en évidence l'importance de la localisation des sondages.**

2. RECHERCHE DES MECANISMES RESPONSABLES DES VARIATIONS DE PERFORMANCES

Le chapitre précédent a mis en évidence certains faits pour lesquels il convient de rechercher des explications permettant de progresser dans la compréhension du comportement de l'outil construit et de la démarche cartographique qu'il simule.

Dans cette perspective, les deux composantes de l'erreur de prédiction (Epd et Epv) seront analysées séparément. La figure 29 montre leur part respective et leur évolution avec la densité de sondages pour l'ensemble des secteurs de validation.

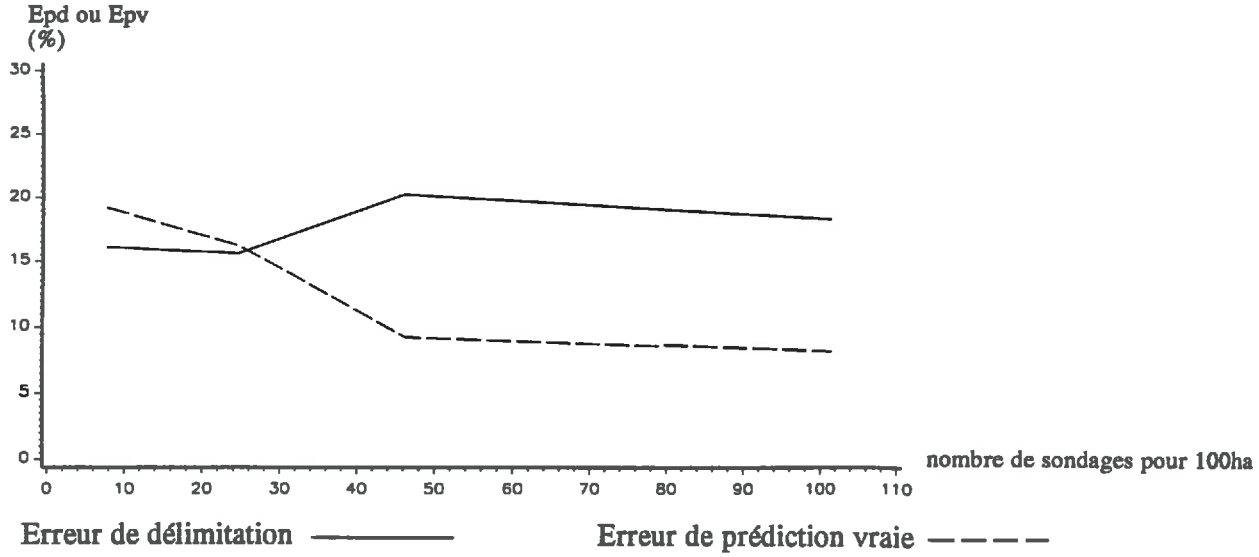


Figure 29: evolution des deux composantes de l'erreur de prédiction en fonction de la densité de sondages

Des résultats comparables à ceux observés sur le secteur de référence se dégagent.

1) L'erreur de délimitation semble peu influencée par la densité de sondages. En revanche, l'erreur de prédiction vraie diminue constamment. Elle explique à elle seule la décroissance de l'erreur de prédiction préalablement mise en évidence. (NB: sur cette figure, le sous-échantillonnage 1/36, dont on a expliqué l'accident, a été éliminé pour plus de clarté).
2) En moyenne, la part de l'erreur de délimitation dans l'erreur totale est majoritaire (57%). Mais, compte tenu de l'évolution citée ci-dessus, la part respective de chaque composante dans l'erreur totale évolue avec la densité de sondages. Ainsi, l'erreur de délimitation est largement prépondérante (environ 70 % de l'erreur) aux densités correspondant à un sous-échantillonnage fort (1/4, 1/9) et légèrement minoritaire (46%) pour la densité la plus faible explorée. Le point d'équilibre correspondrait dans ce cas au sous-échantillonnage 1/16, pour lequel les deux composantes sont quasiment égales.

Les évolutions apparemment indépendantes de ces deux composantes permettent de supposer que les processus qui les génèrent sont eux aussi indépendants. Ils seront donc recherchés séparément dans les deux sous-chapitres qui suivent.

2.1. Analyse de l'erreur de délimitation

Il faut rappeler que le terme "erreur de délimitation" désigne les erreurs qui consistent à prédire, en un point, une unité de sol qui, en réalité, occupe en partie le pixel représenté par ce point. Elle correspond en fait à une erreur de position de limite inférieure à 35m (distance séparant le centroïde du pixel de 50m de côté à l'un de ses angles).

S'il n'existe pas de tendances nettes d'évolution avec la densité de sondages, le tableau 19 montre à l'inverse que l'erreur de délimitation (ici moyenne des résultats obtenus pour les sous-échantillonnages 1/4, 1/16 et 1/64) varie suivant les secteurs considérés.

Secteurs	RB	MT	LZ	Ensemble SV	SR
Epd moyen (%)	18.5	20.0	11.8	16.6	12.0

tableau 19: erreurs de délimitation suivant les secteurs considérés (moyennes de toutes les densités confondues)

L'erreur de délimitation est, au total, plus élevée sur l'ensemble des secteurs de validation que sur le secteur de référence. Cependant, les principales différences d'erreurs s'observent entre le groupe RB, MT d'une part et LZ, SR d'autre part. En conséquence, il ne semble pas qu'il y ait une influence particulière à rechercher consécutive au passage de l'intérieur à l'extérieur du secteur de référence.

secteurs	RB	MT	LZ	SR
longueurs totales des limites (km)	3.8	4.2	1.6	49.8
densités de limites (mètres/ha)	84	94	38	53
Epd (nbre points/km)	9	8	12	9

tableau 20: erreurs de délimitation exprimées en fonction des longueurs des limites.

Le tableau 20 présente l'erreur de délimitation sous une autre forme que précédemment: l'introduction de la longueur des limites des cartes réelles dans l'expression des résultats d'erreur a pour effet d'égaliser les résultats. Seul, le secteur de Lézignan la Cèbe se distingue encore par une erreur plus forte mais qui reste cependant du même ordre de grandeur. En conséquence, tout se passe comme si l'erreur de délimitation était simplement proportionnelle à la quantité de limites réelles à tracer, exprimées ici par leur longueur totale.

Il s'agit donc d'une erreur apparamment indépendante des conditions dans lesquelles sont produites les prédictions (terrain d'étude et densité de sondages). Pour espérer la réduire, il faudrait vraisemblablement remettre en cause des choix effectués pour construire l'outil de prédiction des

unités de sol (en particulier le découpage en pixels et la taille de ces pixels). Parallèlement, il conviendrait d'introduire la phase de tracé de limite qui a été pour l'instant ignorée dans ce travail.

2.2. Analyse de l'erreur de prédiction vraie

L'erreur de prédiction vraie correspond à une erreur plus grave que la précédente. En effet, elle concerne le cas où l'unité prédite n'occupe pas du tout le pixel traité. A la différence de la précédente, elle est sensible aux conditions dans lesquelles est placé l'outil de prédiction: elle décroît avec la densité de sondages (figure 29) et croît globalement entre l'intérieur et l'extérieur du secteur de référence (+3% en moyenne, toutes densités confondues).

L'analyse de ces variations est donc importante pour comprendre le comportement de l'outil de prédiction et en vérifier éventuellement la conformité avec une démarche cartographique "humaine".

Cette analyse sera abordée en deux temps:

- dans un premier paragraphe, l'évolution de l'erreur de prédiction vraie avec la densité de sondages sera plus particulièrement traitée; l'objectif est, en particulier, de mieux comprendre pourquoi il y a perte d'efficacité progressive des sondages introduits;
- le deuxième paragraphe s'intéressera à l'erreur liée à un défaut de représentativité du secteur de référence vis à vis des secteurs de validation; dans ce but, le sous-échantillonnage 1/4 sera étudié plus en détail dans la mesure où l'erreur "irréductible" qui l'affecte encore peut être moins imputée que les précédents, soit à un manque de sondages, soit à un défaut de positionnement de ceux-ci. L'objectif est d'amorcer une réflexion sur le problème de la représentativité du secteur de référence.

2.2.1. Evolution de l'erreur de prédiction vraie avec la densité de sondages

La figure 30, tirée du tableau 21, représente, pour les trois sous secteurs de validation, l'évolution de l'erreur de prédiction vraie suivant la densité de sondages:

	1/64	1/36	1/16	1/9	1/4
La Roubiaire	14.2	12.7	13.1	6.9	6.6
Montmau	27.2	31.8	22.9	17.3	11.9
Lézignan	16.1	15.6	12.3	3.4	5.8

***Tableau 21:** erreurs de prédictions vraies (%) en fonction du sous-échantillonnage.*

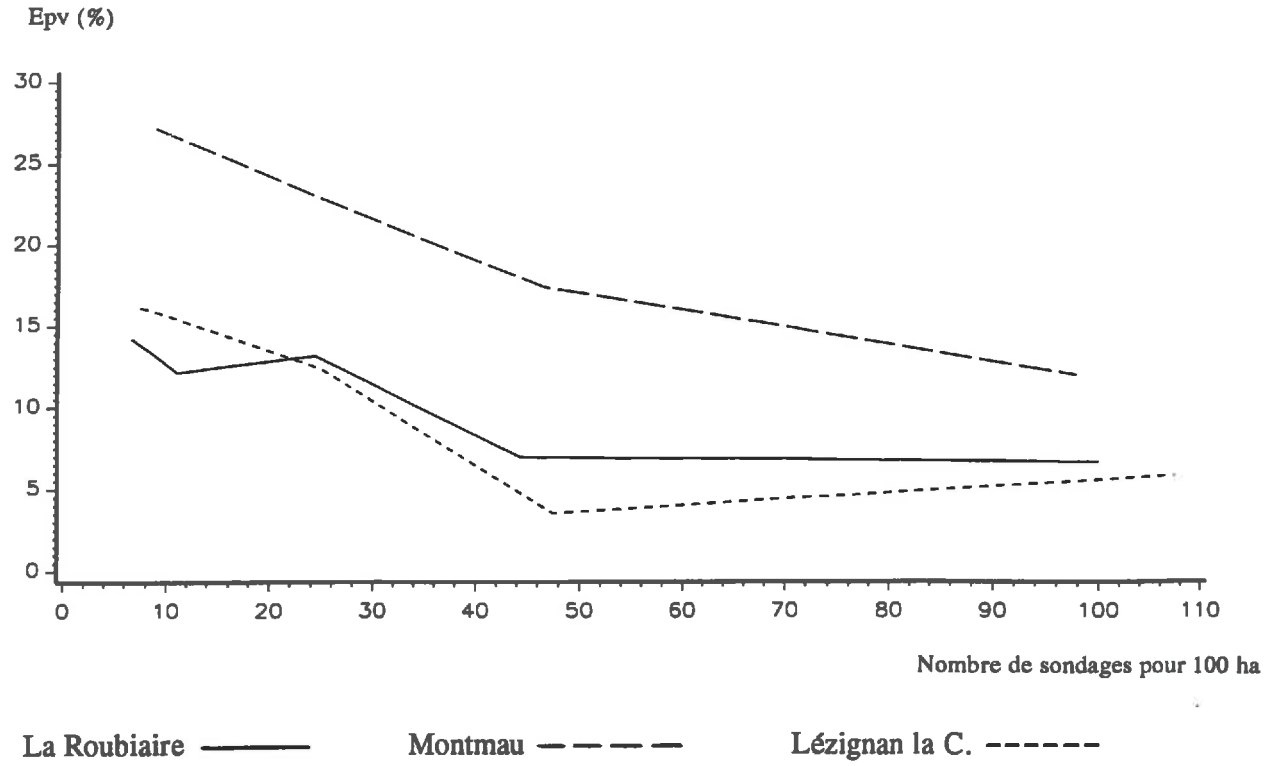


***Figure 30:** Comparaison entre les secteurs de validation des évolutions d'erreurs de prédiction vraies avec la densité de sondages*

Le secteur de Montmau se distingue nettement des deux autres par des niveaux d'erreurs systématiquement plus élevés. Cette différence est le premier révélateur d'un effet "milieu pédologique" qui sera retrouvé par la suite.

Par ailleurs, les courbes de la figure 30 révèlent que l'erreur de prédiction vraie ne diminue pas systématiquement entre deux densités de sondage consécutives: il y a même des augmentations d'erreurs en dépit (ou à cause?) de l'adjonction de nouveaux sondages. L'ensemble de ces phénomènes participent ainsi à la perte globale d'efficacité des nouveaux sondages révélée antérieurement par l'analyse globale.

Pour tenter d'expliquer ces phénomènes, Il est comparé systématiquement deux à deux, sur un même secteur, les sous-échantillonnages qui se trouvent être "voisins" en terme de densité de sondages (par exemple 1/4 et 1/9 ou 1/64 et 1/36). Pour cela, sont considérés leurs couples de valeurs d'erreur de prédiction (12 couples correspondant à 4 par secteurs multiplié par 3 secteurs). Par exemple, le couple (14.2,12.7) sera identifié puisqu'il correspond, sur le secteur de La Roubiaire, aux valeurs d'erreur des sous-échantillonnages 1/64 et 1/36 (cf tableau 21, page précédente). Pour ces couples, deux observations sont mises en relation:

- l'évolution de l'erreur entre les deux termes du couple; soit diminution, soit stagnation ou accroissement;
- l'évolution du nombre de plages cartographiques de la vraie carte touchées par les sondages des sous-échantillonnages correspondants; soit stagnation (voire diminution) soit augmentation. Ces nombres sont donnés par le tableau suivant:

secteurs	1/64	1/36	1/16	1/9	1/4
La Roubiaire	3	5	6	8	8
Montmau	4	2	5	6	7
Lézignan	3	3	3	4	4

tableau 22: nombre de plages cartographiques de la vraie carte touchées par les sondages compte tenu des différents sous-échantillonnages pratiqués

Les résultats peuvent être synthétisés par un tableau à double entrée (tableau 23) donnant le nombre de couples dans chaque situation:

Evolution de l'erreur / Nbre plages	Diminution	Stagnation ou augmentation
augmentation	6	1
stagnation ou diminution	2	3

tableau 23.: nombre de couples en fonction des évolutions conjointes de l'erreur de prédiction vraie et du nombre de plages cartographiques explorées par les sondages

Ce tableau met en évidence une relation entre les deux facteurs considérés:

- une diminution de l'erreur entre deux sous-échantillonnage consécutifs se produit majoritairement (6 contre 2) lorsque les sondages permettent de toucher des plages cartographiques nouvelles de la vraie carte des sols;
- à l'inverse, la majeure partie des stagnations s'observent dans des situations où la densité immédiatement supérieure ne permet pas de toucher une nouvelle plage cartographique.

Ce dernier résultat est éclairant vis à vis des observations antérieures qui concluaient à une diminution de l'efficacité des nouveaux sondages introduits par densification systématique. Il faudrait en effet, pour espérer optimiser la prospection, que les nouveaux sondages soient des occasions permettant d'explorer de nouvelles situations pédologiques,la découverte d'une nouvelle future plage cartographique constituant l'exemple extrême de "nouvelle situation pédologique". Or, le protocole tel qu'il a été établi correspond à une stratégie de prospection systématique ("grid survey") favorisant le risque de sondages "inutiles" (c'est à dire ne découvrant rien de neuf).

En revanche, sur le terrain, le "vrai" pédologue, soucieux d'économiser son temps, raisonne l'implantation de ses sondages suivant une stratégie tenant compte, au moins en partie, des exigences mentionnées ci dessus. Dans le prochain chapitre, un essai de simulation de cette stratégie sera tenté

ce qui permettra de revenir sur le problème de l'optimisation du rapport nombre de sondage/erreur de prédiction.

2.2.2. Erreur de prédiction vraie et défaut de représentativité du secteur de référence vis à vis des secteurs de validation

Indépendamment du phénomène découvert précédemment, l'erreur de prédiction vraie n'atteint pas, à densité de sondages égale et stratégies similaires, le niveau "plancher" (3.4%) obtenu sur le secteur de référence avec le sous-échantillonnage 1/4. Certes, l'écart est faible (+4.8%) confortant ainsi l'hypothèse de représentativité du secteur de référence, au moins vis à vis des relations de voisinage entre unités de sol. Cependant, les secteurs de validation étant petits et, de plus, très proches géographiquement du secteur de référence, il est important d'analyser cet écart, aussi minime soit-il, dans la mesure où il peut révéler des problèmes susceptibles de s'amplifier lorsque l'outil de prédiction s'appliquera sur de plus vastes zones.

Le tableau 24 détaille, par unité de sol, le nombre de points comptabilisés dans l'erreur de prédiction vraie pour le sous-échantillonnage 1/4 (densité de sondages 1/ha).

Unités	Nombre de points	Proportion de l'effectif de l'unité (en %)
2	7	37
3	10	23
6	1	2
11	1	11
14	3	13
15	4	17
17	4	9
18	2	20
Total	32	

***Tableau 24:** nombre et pourcentage de points, par unités de sol, participant à l'erreur de prédiction vraie*

Les unités 2 et 3 sont plus particulièrement concernées: d'une part, elles totalisent à elles deux plus de la moitié de l'erreur totale et, d'autre part, le pourcentage de pixels "mal prédits" est plus élevé pour ces unités que pour les autres.

Chacune d'elles illustre un problème différent, qu'il est possible d'analyser au vu des cartes de prédictions correspondantes.

L'unité 2 constitue, sur la carte du sous secteur de Lézignan la Cèbe, une plage cartographique étroite et allongée parallèlement au cours de l'Hérault (planche 5c). Sa mauvaise prédiction s'explique par l'occurrence simultanée de deux problèmes.

1) La plage est trop étroite pour que les sondages, même avec une forte densité, la prospectent sur toute sa longueur. Il y a, de ce fait, une déformation de l'image que peut avoir, de cette plage, l'outil de prédiction à partir du semis de sondage qui lui est fourni.
2) Les règles de voisinage ne peuvent corriger ce défaut dans la mesure où les zones d'isoprédictions délimitées autour des sondages prennent la forme de "couronnes" et non de bandes comme il le faudrait. En effet, la variable "sens du voisinage" qui, normalement, subdivise ces couronnes en 3 zones ("+haut","+bas" et "même altitude") n'est pas opérante dans ce cas précis puisque, du fait de la pente quasi nulle en ce lieu, tous les pixels du voisinage sont affectés à la modalité "même altitude".

Il ne s'agit donc pas, dans ce cas, d'un réel problème de représentativité du secteur de référence mais plutôt d'une perte d'efficacité, liée aux caractéristiques de l'outil de prédiction, dans la situation particulière d'une plage cartographique étroite en terrain plat.

L'unité 3 constitue, sur le secteur de Montmau, une plage cartographique dont la largeur est importante (200 mètres en moyenne). La carte de prédiction (planche 5b) met en évidence, au sein de cette plage, des erreurs matérialisées par des pixels d'unité 1 en position d'inclusion. Tout se passe comme si les prédictions d'unités de sol, déclenchées à la suite des sondages de l'unité 3, avaient tendance à surestimer, à partir d'un certain rayon de voisinage, la présence de l'unité 1 au détriment de l'unité 3. Cette tendance s'explique au vu de la carte du secteur de référence: sur celle-ci, l'unité 3 n'occupe qu'une bande très étroite (50 m) entre les unités 1 et 17. Les règles de voisinage reproduisant naturellement cette géométrie, l'erreur de prédiction est donc inévitable. Ainsi, il s'agit bien dans ce cas d'un problème de représentativité du secteur de référence.

Outre ces deux unités, il faut souligner également l'erreur attendue causée par l'unité 18, unité nouvelle rencontrée sur le sous secteur de La Roubiaire.

L'analyse détaillée des erreurs de l'outil de prédiction des unités de sol a permis de préciser et d'expliquer les problèmes observés à l'issue du chapitre précédent.

Il apparaît qu'une bonne partie (57%) de l'erreur de prédiction n'est en fait qu'une erreur de délimitation. Cette erreur ne dépend pas des conditions dans lesquelles est placé l'outil (terrain d'étude, nombre de sondages disponibles). Elle est inhérente à sa construction (taille des pixels).

L'autre composante (erreur de prédiction vraie) varie entre 19.2% et 8.2%. Ce résultat doit être interprété avec prudence compte tenu de la taille réduite des secteurs de validation. Il n'en est pas moins encourageant en termes de représentativité du secteur de référence. l'erreur de prédiction vraie subit l'influence de plusieurs facteurs et c'est elle, en fait, qui imprime les variations sur l'erreur de prédiction observées au chapitre précédent. Son analyse permet de mieux comprendre les mécanismes sous-jacents à ces variations.

1) **La décroissance de l'erreur avec l'augmentation de la densité de sondages n'est effective que si les nouveaux sondages introduits permettent d'explorer des situations pédologiques nouvelles. Ainsi s'explique la forme des courbes erreur = f(densité de sondages) observées tant sur le secteur de référence que sur les secteurs de validation. La densification systématique des sondages mise en oeuvre dans le présent protocole**

favorise l'augmentation du nombre de sondages "inutiles", c'est à dire ne remettant pas en cause une situation déjà établie. Dans le 3ème chapitre, une statégie d'implantation de sondages sera définie et mise en oeuvre afin de remédier à ce problème et de se rapprocher ainsi du comportement d'un pédologue de terrain, économe de sa peine, de son temps et donc de ses sondages.

2) La diminution d'erreur de prédiction vraie avec la densité de sondages est de toute façon bornée. L'analyse des mécanismes responsables reste forcément partielle compte tenu de la faible superficie des actuels secteurs de validation. Parmi plusieurs origines d'erreur possibles, des problèmes de représentativité du secteur de référence ressortent cependant: unité nouvelle (unité 18) ou forme nouvelle des plages cartographiques d'une unité du secteur de référence (unité 3).

Par ailleurs, l'analyse des deux composantes d'erreur précise le diagnostic concernant les différences de performances entre secteurs de validation, mises en évidence dans le chapitre précédent. En premier lieu, l'erreur de délimitation est directement influencée par la longueur totale des limites d'une carte. Ce facteur explique à lui seul l'écart de performances entre le secteur de Lézignan la Cèbe et le secteur de la Roubiaire. Par contre, il n'explique pas l'écart substantiel de performances constaté entre ces deux secteurs d'une part et le secteur de Montmau d'autre part. Un effet "milieu pédologique" est donc probable. Il sera discuté en fin de partie.

3. EVALUATION DE LA PERTINENCE DU RISQUE D'ERREUR FOURNI PAR L'OUTIL DE PREDICTION DES UNITES DE SOL

Comme évoqué au chapitre 2, la prédiction d'une unité de sol pour chaque point est fournie avec un "risque d'erreur" (re(x)). Ce risque correspond à la probabilité qu'une autre unité que celle prédite occupe le point x considéré. Il s'agit en fait d'une autre facette de l'expertise extraite du secteur de référence: être capable de localiser les zones où il existe une incertitude hypothéquant les prédictions fournies. Cette aptitude serait en particulier utile pour élaborer une stratégie alternative d'implantation de sondages.

Dans la perspective d'étudier la pertinence de ces estimations d'erreur, il peut être calculé, sur une surface cartographiée, un risque d'erreur prévu moyen (re) correspondant à l'évaluation globale, par l'outil de prédiction, de la qualité de la carte produite. Ce risque sera calculé en faisant la moyenne des différents $re(x_i)$ obtenus sur les points $x_1,\ldots,x_n$. Soit:

$$re = 100\ \Sigma_{(i=1,..,n)}\ re(x_i)/n \qquad [35]$$

$re(x_i)$: risque d'erreur sur un point
n: nombre de points
re: risque d'erreur moyen prévu (exprimé en %)

Cette valeur peut être comparée à l'erreur de prédiction (Ep) définie préalablement, pour les différentes densités de sondages retenues. Ainsi, la figure 31 (page suivante) révèle une tendance générale du modèle à surestimer sa propre erreur. Par contre, les évolutions en fonction de la densité de sondages marquent un parallélisme encourageant.

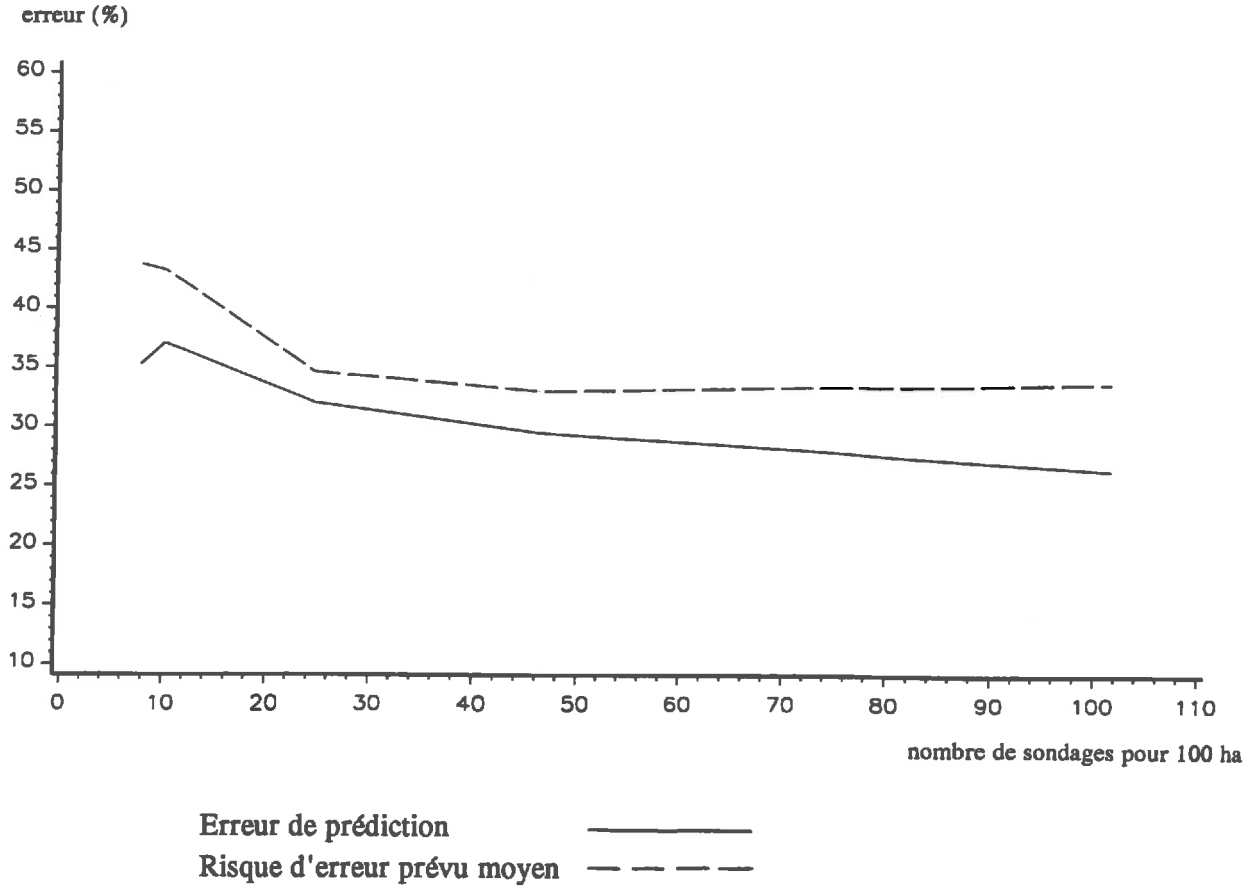


Figure 31: comparaison des évolutions, avec la densité de sondages, de l'erreur de prédiction et de l'erreur estimée par l'outil de prédiction (ensemble des secteurs de validation)

Pour préciser ce premier résultat, il convient de se demander si les plus grands risques d'erreurs annoncés se traduisent effectivement par des points mal classés. Dans cette perspective, les 2 populations de points "bien classés" et "mal classés" sont distinguées. Sur ces deux populations, sont calculées respectivement re_v et re_f, moyennes des risques d'erreurs sur chaque population. Les résultats sont présentés (figure 32) sous forme de courbes traduisant l'évolution de re_v et re_f suivant la densité de sondages et ce pour les trois secteurs considérés. il en résulte les points suivants.

1) Quels que soient la densité de sondages et le secteur de validation, les risques d'erreur prévus par le modèle sont systématiquement plus élevés sur la population de points effectivement "mal classés".
2) L'écart entre les 2 courbes, en relation directe avec l'aptitude du modèle à prévoir correctement ses erreurs, augmente avec la densité de sondages. Par ailleurs, le secteur de Montmau se distingue une nouvelle fois des deux autres par des écarts entre courbes généralement moins importants.
3) La forme des courbes est sytématiquement différente. Les risques d'erreur calculés sur les populations de points "bien classés" diminuent régulièrement. Le modèle gagne de l'assurance sur ses prédictions au fur et à mesure que de nouveaux sondages sont introduits. En revanche, à partir d'une certaine densité de sondages, les risques d'erreurs calculés sur les populations de points "mal classés" augmentent. L'ajout de nouveaux sondages fait donc perdre de l'assurance à l'outil de prédiction.

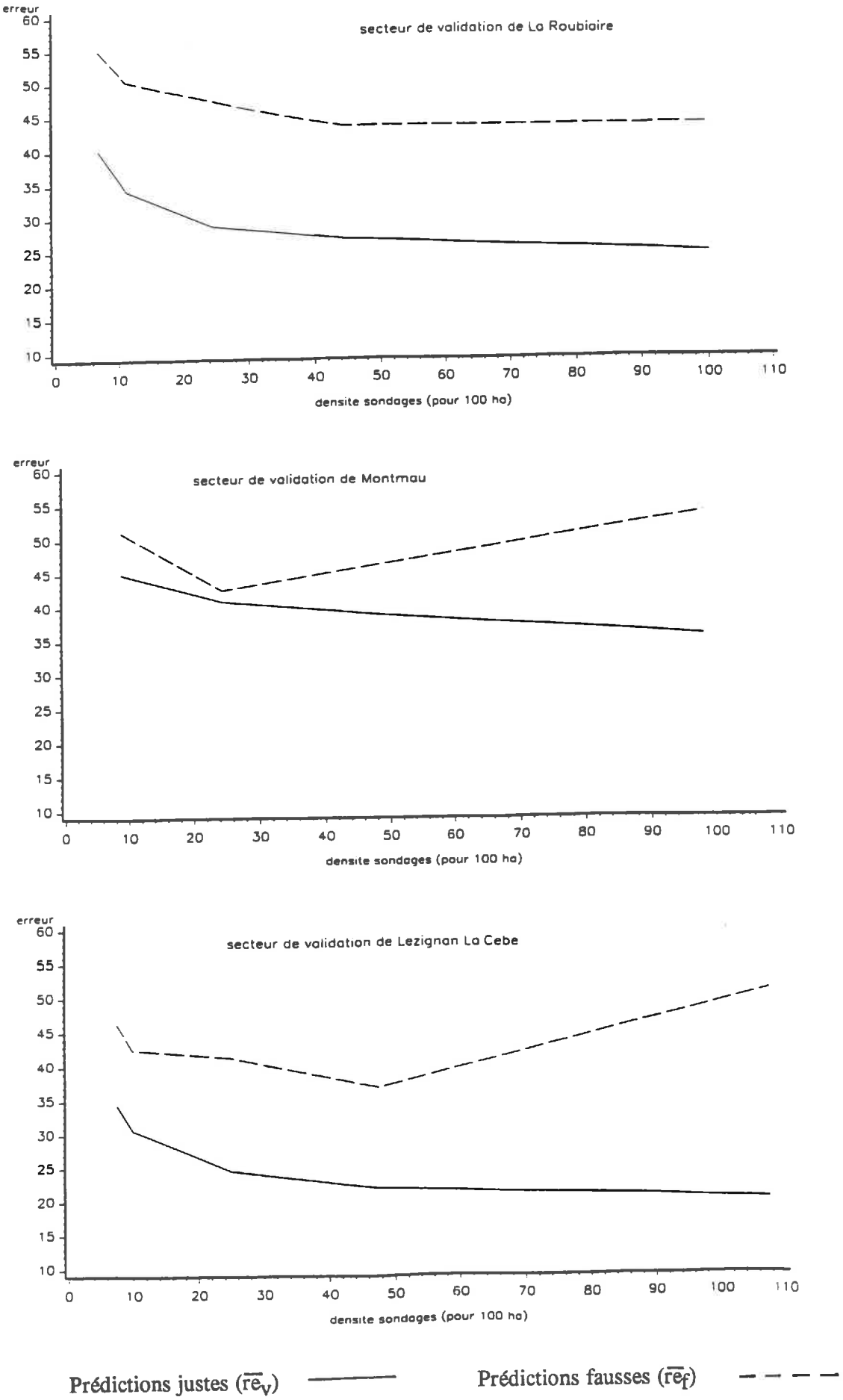


Figure 32: *comparaison des moyennes d'erreur estimées entre prédictions justes et prédictions fausses*

Tous ces faits attestent de la pertinence de l'estimation d'erreur fournie par l'outil de prédiction des unités de sol sur ses propres prédictions. Il convient de remarquer que "l'erreur d'estimation d'erreur" évolue suivant la même logique que l'erreur de prédiction vraie: d'une part, elle diminue lorsque la densité de sondages augmente et, d'autre part, elle est plus élevée en moyenne sur le secteur de Montmau.

En conséquence, il semble prometteur d'utiliser cette estimation d'erreur pour construire une stratégie de prospection plus économe en sondages. Par ailleurs, les valeurs de $re(x_i)$ présenteraient un intérêt dans la perspective de fournir à un utilisateur une estimation de la confiance accordée par le modèle à ses propres résultats.

Au cours de ce deuxième chapitre, a été testé sur 3 secteurs de validation l'outil de prédiction utilisant les relations de voisinage entre unités de sol, formalisées au chapitre précédent à partir de la carte des sols du secteur de référence.

Les résultats principaux ont déjà été synthétisés dans la conclusion du sous-chapitre 2. Ils conduisent à un constat encourageant, bien que la faible superficie des secteurs de validation interdise toute conclusion définitive.

1) **Les erreurs de prédiction sont quantitativement peu élevées et qualitativement constituées en majorité par des erreurs de délimitation. Leur ampleur semble partiellement influencée par le milieu pédologique abordé. Elles semblent plus causées par des insuffisances du modèle construit que par des problèmes de représentativité du secteur de référence.**
2) **La bonne cohérence des résultats permet d'identifier plusieurs de ces insuffisances et les voies d'amélioration possibles dont certaines seront développées en conclusion de partie. L'une de ces voies sera mise en oeuvre au cours du troisième chapitre. Elle concerne l'élaboration d'une stratégie d'implantation des sondages permettant d'économiser ces derniers tout en conservant la même qualité de prédiction. Dans cette perspective, les résultats obtenus au cours du sous-chapitre 3 montrent qu'une telle stratégie peut s'appuyer sur l'aptitude du modèle à prévoir ses propres erreurs.**

CHAPITRE 9

UTILISATION DES REGLES DE VOISINAGE DANS LE CADRE D'UNE SIMULATION DU RETOUR A LA PARCELLE

Jusqu'à présent, les règles de voisinage ont été utilisées séparément afin d'apprécier la qualité des prédictions qu'elles délivrent. Or, comme évoqué au chapitre 2, ces règles sont en fait destinées à être utilisées par un outil automatisant l'ensemble de la démarche cartographique mise en oeuvre au cours de l'opération de retour à la parcelle. La formalisation de cette opération (chapitre 2), fixe les conditions d'utilisation des règles de voisinage.

1) Elles ne peuvent être déclenchées qu'après un rattachement du sondage à une des unités de sol du secteur de référence, sur la base de critères morpho-pédologiques. Comme évoqué précédemment, la modélisation de ce rattachement sort du champ de ce travail.
2) Elles ne s'appliquent pas ex nihilo mais viennent modifier une précédente estimation obtenue grâce à l'application de lois basées sur les relations entre unités de sol et critères extrinsèques ou de surface. Le premier sous-chapitre sera consacré à l'étude des modalités et des résultats de la combinaison entre les deux sources d'information en question.
3) En pratique, la localisation des sondages ne procède que rarement d'une couverture systématique du territoire selon une grille. Une statégie sous-tend leur choix, permettant de les économiser autant que faire se peut. Un premier essai de construction d'une telle stratégie sera tenté au cours d'un deuxième sous-chapitre. Les résultats et l'intérêt par rapport à une prospection systématique seront évalués.

1. COMBINAISON DE PREDICTIONS ISSUES DES REGLES SOLS-PAYSAGE ET DES REGLES DE VOISINAGE

En fin de deuxième partie, ont été discutées les modalités d'utilisation des prédictions fondées sur des relations sols-paysage. En résumé, il résulte de cette discussion que ces prédictions ne sauraient porter directement sur l'occurrence d'une unité de sol particulière. Par contre, elles pourraient permettre de prédire l'occurrence de l'un ou l'autre des membres d'un groupe d'unités au sein duquel se trouverait l'unité réelle, celle que l'on souhaite détecter.

Dès lors, la synergie possible avec les prédictions fondées sur les règles de voisinage apparaît clairement: elle consiste à accélérer l'identification de la bonne unité sur un point donné, en limitant l'éventail des unités de sol possibles en ce point.

Cette articulation entre les deux sources d'information va être mise en oeuvre et expérimentée au cours de ce sous-chapitre. Dans le détail, deux étapes seront distinguées:

- dans un premier temps, il s'agira de définir le mode de combinaison entre les deux sources d'information de façon à reproduire la synergie définie ci-dessus;
- dans un deuxième temps, les résultats seront évalués sur les secteurs de validation.

1.1. L'algorithme de combinaison entre règles sols-paysage et règles de voisinage

Il a été montré précédemment (chapitre 6) comment formaliser, pour chaque noeud terminal de l'arbre de classification, une prédiction excluant des unités de sol de l'éventail des possibilités. Ainsi, la formule [24] associe à chaque noeud terminal T_g, un ensemble E_g des unités de sols éliminées.

Il est donc possible en affectant chaque point d'un périmètre donné à un des noeuds terminaux de l'arbre de classification, de définir quelles unités sont à éliminer sur ce point. Ceci peut se formaliser en définissant,pour chaque point x dont le vecteur d'observation tombe dans T_g, une série de variables booléennes $(e_1(x),\ldots,e_j(x),\ldots e_v(x))$ traduisant le fait d'éliminer ou non l'unité U_j de l'ensemble des unités possibles:

$$\begin{array}{lll} \forall\, x \,/\, o(x) \in T_g, & U_j \in E_g & ===> e_j(x) = 0 \\ & U_j \notin E_g & ===> e_j(x) = 1 \end{array} \qquad [36]$$

avec: $o(x)$ le vecteur des observations au point x (altitude, pente,...)
T_g: un noeud terminal de l'arbre de classification utilisé ($g = 1,\ldots,s$)
E_g: l'ensemble des unités de sol éliminées associé, par la formule [24], au noeud terminal T_g

Autrement dit, $e_j(x)$ est nulle si (et seulement si) U_j est éliminée du champ des possibilités par l'arbre de classification.

Grâce à cette reformulation, il est désormais possible de proposer une formule de combinaison permettant d'obtenir une synthèse des prédictions issues des deux sources d'informations envisagées:

- soit $(e_1(x),\ldots,e_j(x),\ldots e_v(x))$ le vecteur en x résultant de l'utilisation des règles sols-paysage selon les modalités décrites ci-dessus
- soit $(q_1(x),\ldots,q_j(x),\ldots q_v(x))$ le vecteur des probabilités d'apparition des unités de sol après synthèse des règles de voisinage déclenchées entre les étapes t_2 et t_f

$$\begin{array}{l} \Sigma_{(1\ldots v)}\, e_j(x).q_j(x) \neq 0 ===>,\ p_j(x,t_f) = e_j(x).q_j(x)/\Sigma_{(1\ldots v)}\, e_j(x).q_j(x) \\ \Sigma_{(1\ldots v)}\, e_j(x).q_j(x) = 0,\ ===>,\ p_j(x,t_f) = q_j(x) \end{array} \qquad [37]$$

avec:
$p_j(x,t_f)$: probabilité d'apparition de l'unité U_j à l'étape finale t_f

La formule [37] permet ainsi de calculer un vecteur de probabilité réalisant la synthèse de toutes les déductions faites à chaque étape de la démarche. Elle respecte bien les objectifs de combinaison définis en introduction du chapitre:

- toute unité éliminée par la source d'information "sols-paysage" perd ses chances d'être prédite puisque sa probabilité d'apparition à l'étape finale devient nulle;
- par contre, les nouvelles probabilités des unités retenues se trouvent majorées pour que leur total reste égal à 1, les écarts et rapports entre elles étant cependant conservés.

Le cas où $e_j(x).q_j(x) = 0$ correspond à un conflit entre les deux sources d'information: pour toute unité du secteur de référence, l'une au moins des sources considère comme impossible sa présence ($e_j(x)=0$ ou(et) $q_j(x)=0$), si bien qu'in fine, tous les $p_j(x,t_f)$ seraient nulles quel que soit j.

1.2. Analyse des résultats de combinaison sur les secteurs de validation

La formule de combinaison représentant la synergie prévue entre les deux sources étant établie, elle sera utilisée sur les secteurs de validation.

Avant d'utiliser cette formule, il convient de renseigner, pour chaque point, les variables $e_j(x)$. Pour cela, il faut fixer au préalable, pour chaque noeud terminal, les ensembles d'unités éliminées E_g. Comme évoqué au chapitre 6, ces ensembles dépendent finalement de deux critères liés à l'analyse par segmentation:

- un critère α définissant la taille de l'arbre et donc les noeuds terminaux à partir desquels sont définis les ensembles E_g
- un critère ß définissant le seuil d'élimination utilisé dans la formule [24]

Ainsi, dans le but de tester différentes possibilités, sont essayés, au chapitre 6, 4 critères α différents (arbres 1 à 4) et 4 seuils d'élimination ß (0, 5, 10 et 20%). La combinaison de ces deux critères conduit à 4^2 séries d'ensemble E_g, soit 4^2 modes d'éliminations d'unités de sol possibles. Les résultats de ces différents modes d'élimination, exprimés en termes d'erreur et de précision, sont indiqués dans le tableau 13. Le problème consiste à choisir le mode d'élimination dont le rapport précision/erreur est optimal dans la perspective de son utilisation combinée avec les règles de voisinage. C'est à dire:

- suffisamment précis pour avoir un intérêt; en effet, plus l'éventail d'unités de sol possibles est laissé large, moins l'aide à la prédiction est intéressante;
- suffisamment juste pour ne pas entraîner des erreurs sur la prédiction finale; compte tenu de la formule de combinaison, les erreurs notées au tableau 6 se propagent en effet directement sur la prédiction finale.

On choisira de ne tester que 3 modes d'élimination d'unités de sol possibles, représentatifs de l'ensemble (les modes d'élimination ayant des erreurs supérieures à 40% étant cependant exclus de l'échantillonnage). Ces 3 modes d'élimination sont rappelés dans le tableau 25.

	Mode d'élimination n°1	Mode d'élimination n°2	Mode d'élimination n°3
N° arbre	0	2	3
Seuil d'élimination (%)	10	0	10
Précision (nbr moy. d'un.prédites)	3.8	5.2	2.1
Erreur (%)	13	2	40

Tableau 25: les différents modes d'élimination d'unités de sol choisis

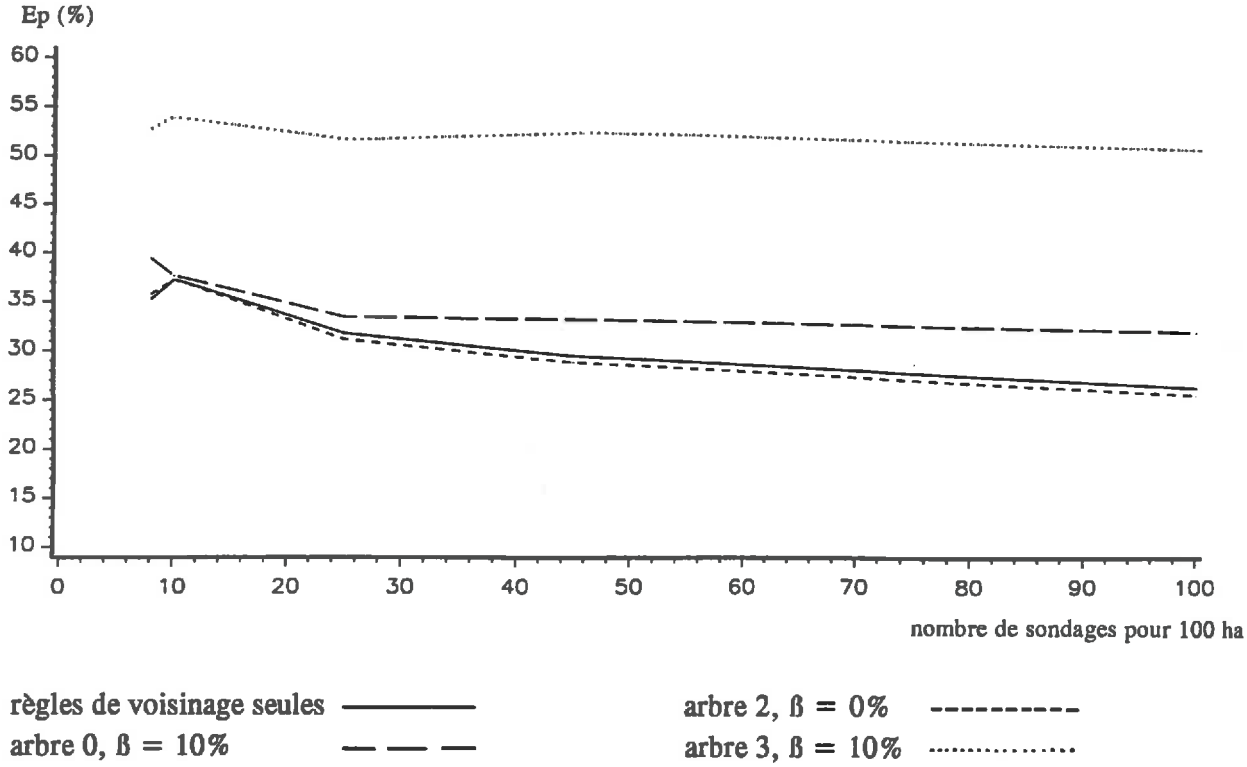


***Figure 33:** comparaison de différentes combinaisons possibles entre règles sols-paysage et règles de voisinage en termes d'évolution de l'erreur de prédiction en fonction de la densités de sondage.*

Les résultats font l'objet de la figure 33. Quel que soit le mode d'élimination testé, l'introduction des règles sols-paysage n'améliore pas la qualité des prédictions: les utilisations des modes d'élimination 1 et 3 se traduisent même par des augmentations sensibles d'erreur, le mode d'élimination 2 n'apportant de son côté aucun changement.

Ainsi, dans le cas étudié, la synergie attendue entre les deux sources d'information n'opère pas. Deux raisons peuvent être évoquées, la première liée aux options prises, la seconde aux données utilisées et au terrain d'étude:

1) La détection de conflits entre les deux sources d'information est peu opérante: aucun conflit correspondant à la définition proposée plus haut n'a pu être mis en évidence sur l'ensemble du secteur de validation. Ceci a pour principale conséquence qu'une erreur de prédiction issue de la source "sols-paysage" ne peut être détectée et corrigée au moyen de la deuxième source. Elle se propage donc à l'identique sur la prédiction finale.
2) Il existe une trop grande différence de qualité entre les deux sources utilisées compte tenu des données qui les alimentent: même la densité de sondages la plus faible sur laquelle s'appuient les prédictions issues des relations de voisinage constitue une information plus "riche" que celle des cartes topographiques et géologiques sur lesquelles s'appuient les éliminations d'unités. La densité de sondages "seuil" à partir de laquelle la prise en compte, dans les prédictions, des relations sols-paysage permettrait d'améliorer le résultat, doit donc être plus faible. Sa recherche nécessiterait de pratiquer des sous-échantillonnages plus faibles (1/256,...) sur des secteurs de validation plus vastes. Il est par ailleurs vraisemblable que cette "densité seuil" est fortement dépendante du terrain d'étude. A ce

titre, la Moyenne Vallée de l'Hérault, milieu peu contrasté en matière de relief, ne favorise pas l'exploitation des relations sols-paysage. Nul doute qu'un milieu montagnard offrirait des résultats plus favorables (une "densité seuil" plus forte) pour la manifestation de la synergie entre les deux sources.

A l'issue de ce sous-chapitre, il n'a pas été démontré de synergies entre les prédictions issues d'une part des règles sols-paysage et, d'autre part, des règles de voisinage. Ces deux sources ne semblent donc trouver leur intérêt que dans deux objectifs disjoints: production de cartes à petite échelle pour les relations sols-paysage (chapitre 6), de cartes à grande échelle pour les relations de voisinage.

Cependant, ce constat n'est peut être pas définitif et universel.

1) **Le mode de combinaison entre les deux sources d'information doit être revu dans le sens d'une meilleure détection et d'une résolution plus pertinente des éventuels conflits. Il s'agit, à la différence du problème de combinaison résolu précédemment, de deux sources sans aucun paramètre commun permettant d'ajuster une pondération. La résolution de ce problème passe par un travail de recherche plus approfondi dans le domaine de la gestion de connaissances incertaines (MARTIN CLOUAIRE, 1992).**
2) **La différence de qualité entre les deux sources limitant l'intérêt de leur association peut être réduite grâce à des améliorations sur la chaîne de production des données sources utilisées par les règles "sols-paysage",**
3) **Le milieu expérimental utilisé n'est pas très favorable à l'expression et à la mise en évidence de l'intérêt d'une éventuelle synergie entre les sources car l'écart de qualité entre elles y est maximal: d'une part, un relief mou favorise les erreurs sur les descripteurs de paysage, d'autre part l'augmentation de densité de sondages est facilitée par l'absence d'obstacles à la prospection pédologique. En milieu montagnard, où les deux conditions citées s'inversent, la nécessité de combiner ces deux sources apparaît a priori plus évidente.**

2. RECHERCHE ET UTILISATION D'UNE STRATEGIE D'IMPLANTATION DES SONDAGES

Jusqu'à présent, le protocole d'utilisation des règles de voisinage simulait une prospection pédologique systématique, dite en "grid survey". Ainsi, pour chaque densité explorée, les sondages étaient disposés régulièrement et uniformément sur le périmètre à étudier. Au cours de l'analyse des résultats, il a été avancé que cette disposition pourrait limiter la qualité des prédictions (chapitre 8). De fait, sur le terrain, les décisions du pédologue ne conduisent que rarement à une telle disposition des sondages.

L'objectif de ce chapitre sera donc d'expérimenter l'intérêt d'une stratégie d'implantation non systématique des sondages ("free survey") dans la perspective d'augmenter l'efficacité de la prospection, c'est à dire le rapport nombre de pixels bien classés/ nombre de sondages. Deux étapes seront distinguées et feront l'objet de deux sous-chapitres distincts:

- élaboration de la stratégie de prospection,
- analyse des résultats et évaluation de l'intérêt d'une telle stratégie.

2.1. Description de la stratégie utilisée

L'analyse, au chapitre précédent, de l'erreur de prédiction vraie a permis de définir un objectif possible pour une stratégie d'implantation de sondages: limiter le nombre de sondages "inutiles", c'est à dire ceux ne remettant pas en cause les prédictions antérieures.

Par ailleurs, au cours du chapitre précédent, on a montré que le modèle semblait capable de prévoir les endroits où ses prédictions sont le plus sujettes à caution (risque d'erreur re(x) élevé). Dès lors, l'application de l'objectif fixé ci dessus devient possible: un point pour lequel l'outil de prédiction donne un risque d'erreur faible ne sera pas sondé. En effet, il y a peu de chances (d'après cet outil) pour que le sondage réalisé sur ce point remette en cause la prédiction antérieure. Autrement dit, ne seront sondés que les points pour lesquels le modèle n'est pas "sûr" de sa prédiction, le seuil de risque correspondant à "sûr" restant à définir. A partir de la règle édictée ci-dessus, le protocole suivant a été choisi.

- **Première phase:** sous-échantillonnage systématique 1 point sur 64 (distance minimale entre sondages:400m); cette première phase correspond exactement au sous-échantillonnage 1/64 pratiqué dans la démarche du chapitre 8
- **Deuxième phase**: sélection de points devenant les lieux de nouveaux sondages suivant deux conditions:
  - + faire partie de l'ensemble de points correspondant au sous-échantillonnage 1/16 (distance minimale entre sondages: 200m),
  - + avoir un risque d'erreur supérieur à 50%.
- **Troisième phase:** sélection de nouveaux sondages avec les mêmes critères que précédemment sauf que le sous-échantillonnage imposé devient 1/4 (distance minimale entre sondages = 100m)
- **Quatrième phase:** dernière densification de sondages, tout point présentant encore un risque d'erreur supérieur à 50% étant sélectionné comme lieu de nouveau sondage (distance minimale entre sondages:50m).

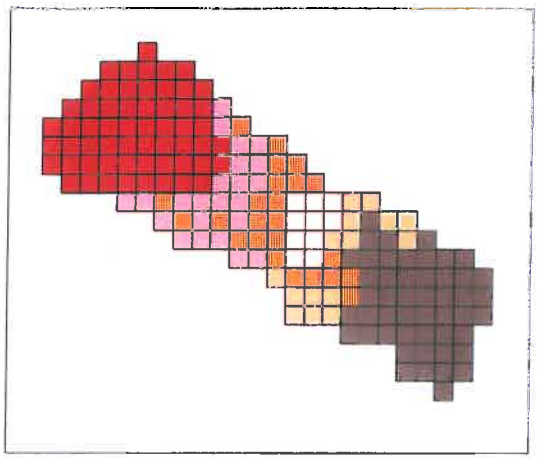
A chaque phase, les règles de voisinage sont appliquées à partir des sondages sélectionnés. Il en résulte, pour chaque secteur de validation, quatre séries de prédictions (dont trois nouvelles). Le protocole ainsi défini est appliqué au moyen d'un programme FORTRAN (cf annexe 10) fonctionnant sur le même principe que les programmes précédents, seul l'algorithme de choix des sondages le distinguant. Comme pour le chapitre précédent, les résultats sont présentés, au moyen d'ARC/INFO, sous formes de séries de cartes de prédiction (planches 6a, 6b, 6c), chacun des points traités représentant un pixel de la carte.

Pour chaque série de prédictions fournies par ce programme, il est calculé, comme précédemment, la densité moyenne de sondages utilisés et les différentes erreurs définies au chapitre précédent.

Le protocole présenté permet en fait de simuler 4 étapes de prospection qui se distinguent uniquement entre elles par la distance minimale autorisée entre sondages (et donc le nombre maximum de sondages autorisés). Ces étapes sont le pendant des densifications successives effectuées au cours du protocole appliqué auparavant. Une comparaison est donc possible entre statégie "grid survey" et "free survey" pour différents sous-échantillonnages.

***Planche 6a:** Cartographies du secteur de validation de La Roubiaire utilisant les lois de voisinage extraites du secteur de référence*

CARTE REELLE



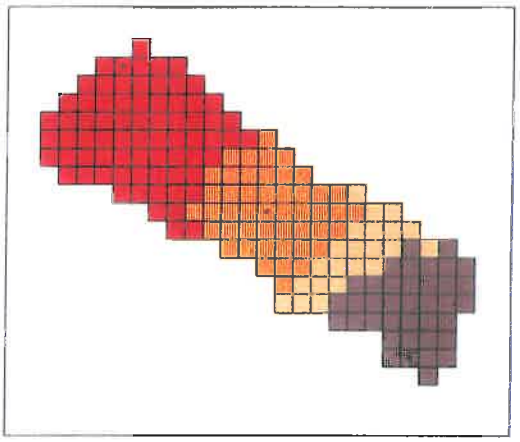
LEGENDE

Unite 1	Unite 7	Unite 13
Unite 2	Unite 8	Unite 14
Unite 3	Unite 9	Unite 15
Unite 4	Unite 10	Unite 16
Unite 5	Unite 11	Unite 17
Unite 6	Unite 12	Inconnue

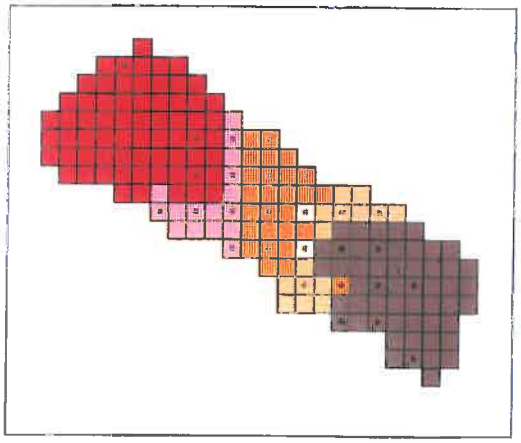
emplacement des sondages

Echelle 1/20 000 (un pixel = 50m X 50m)

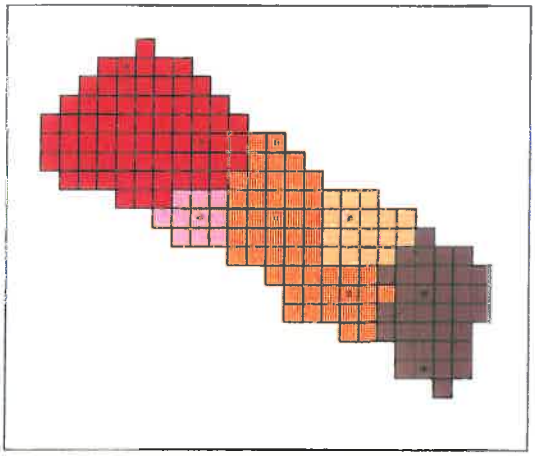
CARTES PREDITES AVEC DIFFERENTES DENSITES DE SONDAGES
strategie : sondages realises si erreur prevue > 50% ("free survey")



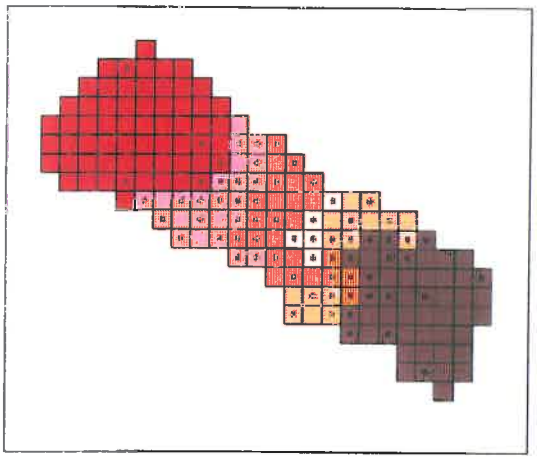
distance minimum entre sondages : 400m
3 sondages soit 6.7 pour 100 ha



distance minimum entre sondages : 100m
27 sondages soit 60 pour 100 ha



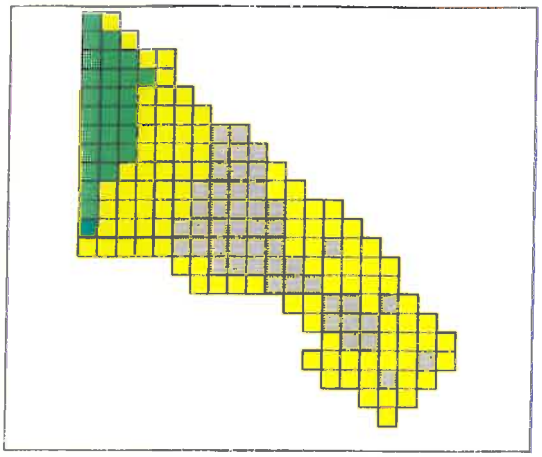
distance minimum entre sondages : 200m
9 sondages soit 20 pour 100 ha



distance minimum entre sondages : 50m
68 sondages soit 151.1 pour 100 ha

***Planche 6b**: Cartographies du secteur de validation de Montmau utilisant les lois de voisinage extraites du secteur de référence*

CARTE REELLE



LEGENDE

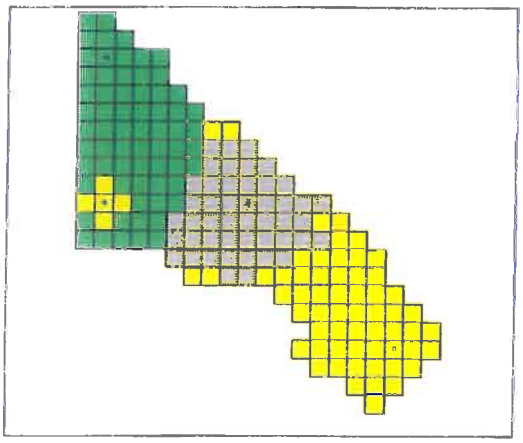
Unite 1	Unite 7	Unite 13
Unite 2	Unite 8	Unite 14
Unite 3	Unite 9	Unite 15
Unite 4	Unite 10	Unite 16
Unite 5	Unite 11	Unite 17
Unite 6	Unite 12	Inconnue

• emplacement des sondages

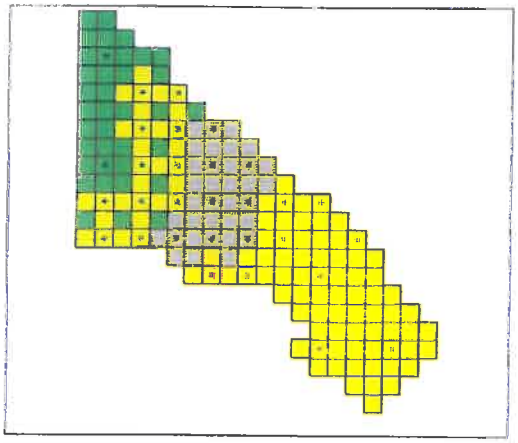
Echelle 1/20 000 (un pixel = 50m X 50m)

CARTES PREDITES AVEC DIFFERENTES DENSITES DE SONDAGES

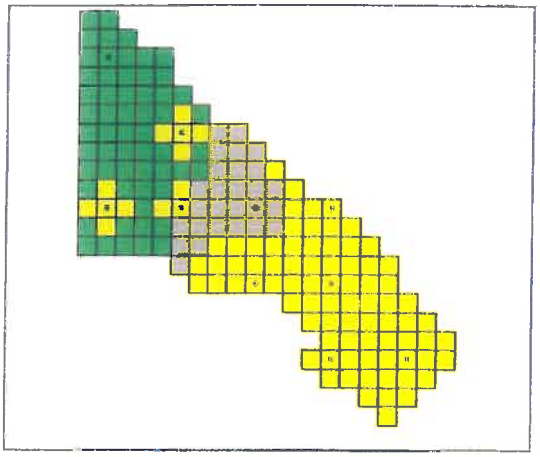
stratégie : sondages réalisés si erreur prévue > 50% ("free survey")



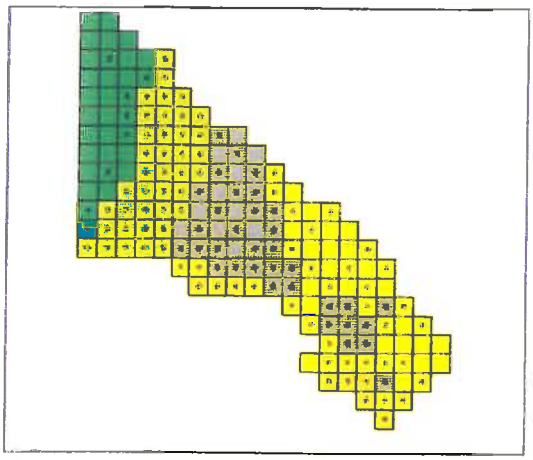
distance minimum entre sondages : 400m
4 sondages soit 8.9 pour 100 ha



distance minimum entre sondages : 100m
30 sondages soit 66.7 pour 100 ha



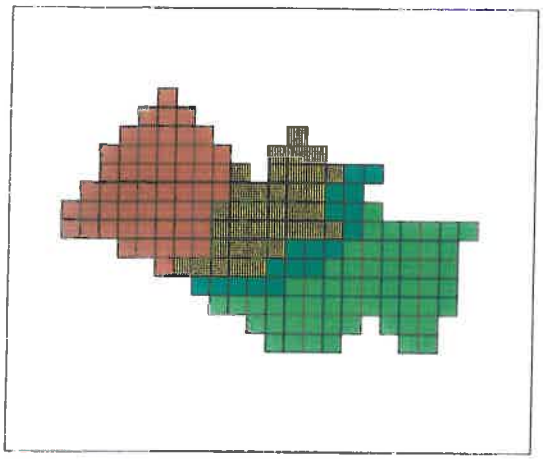
distance minimum entre sondages : 200m
10 sondages soit 22.2 pour 100 ha



distance minimum entre sondages : 50m
119 sondages soit 264.4 pour 100 ha

***Planche 6c**: Cartographies du secteur de validation de Lézignan la Cèbe utilisant les lois de voisinage extraites du secteur de référence*

CARTE REELLE



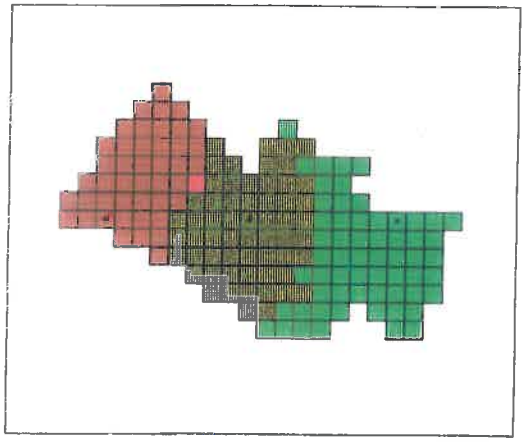
LEGENDE

Unite 1	Unite 7	Unite 13
Unite 2	Unite 8	Unite 14
Unite 3	Unite 9	Unite 15
Unite 4	Unite 10	Unite 16
Unite 5	Unite 11	Unite 17
Unite 6	Unite 12	Inconnue

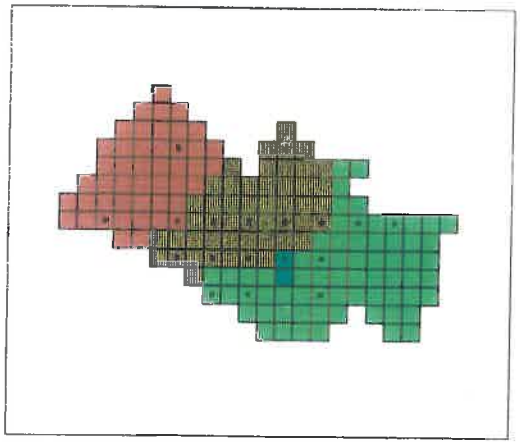
emplacement des sondages

Echelle 1/20 000 (un pixel = 50m X 50m)

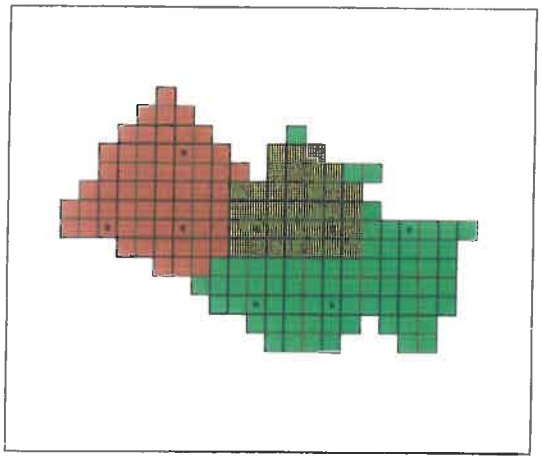
CARTES PREDITES AVEC DIFFERENTES DENSITES DE SONDAGES
strategie : sondages realises si erreur prevue > 50% ("free survey")



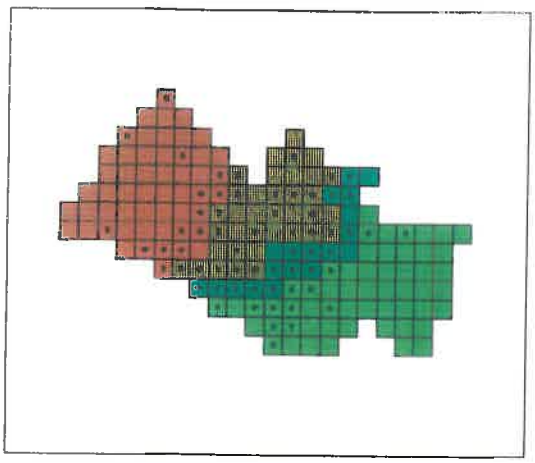
distance minimum entre sondages : 400m
3 sondages soit 7.5 pour 100 ha



distance minimum entre sondages : 100m
18 sondages soit 45 pour 100 ha



distance minimum entre sondages : 200m
8 sondages soit 20 pour 100 ha



distance minimum entre sondages : 50m
63 sondages soit 157 pour 100 ha

2.2. Analyse des résultats de la stratégie "free survey" sur les secteurs de validation

Dans un premier temps, il convient d'examiner si la stratégie de prospection suivie au cours du protocole présenté au chapitre précédent se traduit effectivement par une économie de sondages: Le tableau 26 présente, pour les quatre phases du protocole (caractérisées par les distances minimum entre sondages), le nombre de sondages sélectionnés et, entre parenthèses, le nombre de sondages économisés par rapport à une prospection systématique (c'est à dire sous-échantillonnage sans condition portant sur re(x)). Les résultats sont présentés par secteur.

Distance Mini.	RB	MT	LZ	TOT
400 m	3 (0)	4 (0)	3 (0)	10 (0)
200 m	9 (2)	10 (1)	8 (2)	27 (5)
100 m	27 (18)	30 (14)	18 (25)	75 (57)
50 m	68 (111)	119 (58)	63 (101)	250 (270)

Tableau 26: nombre de sondages effectués (économisés) par une stratégie "free survey"

L'analyse du tableau montre que l'économie de sondages ne devient effective qu'à partir de la deuxième phase du protocole, substantielle qu'à partir de la troisième. Il semble donc qu'un nombre de sondages minimum soit nécessaire pour que l'outil de prédiction ainsi construit se permette ce type d'économie. Par ailleurs, il apparaît que le secteur de Montmau se distingue une nouvelle fois des deux autres par une moindre économie de sondages. Les règles de voisinage délivrant des prédictions moins précises que sur les autres secteurs, plus de sondages sont nécessaires pour les conforter.

Dans un deuxième temps, il s'agit de s'assurer que l'intérêt de cette économie de sondages n'est pas annulé par une dégradation concommittante de la qualité des prédictions. Le tableau 27 donne, pour chaque phase du protocole et chaque secteur, les erreurs de prédictions commises par le modèle et, entre parenthèses, l'écart d'erreur avec la stratégie "grid survey" pour le même sous-échantillonnage, utilisée au chapitre précédent.

Le tableau 27 se lit, par exemple, de la façon suivante: pour la phase correspondant à une distance minimale entre sondages de 100m et pour le secteur de Lézignan la Cèbe, l'erreur de prédiction est de 21.9% soit une augmentation d'erreur de 5.4% par rapport à la prospection systématique réalisée avec la même distance minimum entre sondages (sous-échantillonnage 1/4 du chapitre précédent).

Ensemble	Distance mini	La Roubiaire	Montmau	Lézignan la C.
400 m	31.8 (0)	45.1 (0)	28.6 (0)	35.3 (0)
200 m	29.4 (-0.4)	40.7 (-0.3)	23.7 (-0.3)	31.4
100 m	21.7 (-4.0[14])	36.7 (+1.8)	21.9 (+ 5.4)	26.7 (+ 0.6°)
50 m	12.6 (?)	15.5 (?)	9.9 (?)	12.2 (?)

Tableau 27: erreurs de prédiction (en %) et évolution par rapport à la stratégie "grid survey"

Les seules comparaisons effectives qu'il est possible de faire avec la stratégie "grid survey" concernent les deuxième et troisième phases (distance minimales 200m et 100m). Dans les deux cas, les variations globales d'erreur de prédiction (colonne "ensemble") entre les deux stratégies apparaissent négligeables. Ainsi se trouve justifiée l'économie de sondages proposée par l'outil de prédiction sur la base de son estimation d'erreur. La dernière phase (distance minimum 50m) permet de diminuer encore les erreurs de prédiction d'une manière sensible, au prix cependant d'une augmentation conséquente du nombre de sondages.

Par ailleurs, les résultats varient suivant les différents secteurs. Ainsi, le secteur de Lézignan la Cèbe est le théatre d'une augmentation d'erreur de prédiction dépassant 5%. L'examen de la carte montre que le problème concerne surtout l'unité 2, non touchée par le modèle lors de la phase précédente (sous-échantillonnage 1/16). Ce fait explique que l'erreur ait été sous estimée à cet endroit et donc qu'il n'y ait pas eu de nouveaux sondages permettant de rectifier les erreurs de prédiction. Par ailleurs,le secteur de Montmau, où l'économie en sondages est moindre, ne voit pas l'erreur de prédiction rattraper les autres secteurs sauf pour la dernière phase du protocole où il bénéficie d'un nombre de sondages presque double.

Il n'en reste pas moins vrai que l'introduction de la stratégie "free survey" permet d'obtenir de meilleurs résultats de prédiction en termes de rapport qualité/ nombre de sondages nécessaires. Il faut citer en particulier la troisième phase du protocole qui, avec 75 sondages (soit une densité de l'ordre de 1 pour 2ha) limite l'erreur de prédiction à 26.7% en moyenne, ce qui correspond à une erreur apparente de 22.9% et une erreur de prédiction vraie de 7.6%. On économise ainsi, à qualité égale, plus de 40% de sondages par rapport à la stratégie "grid survey" équivalente (sous-échantillonnage 1/4).

Pour compléter l'analyse, il convient de vérifier si la courbe d'évolution de l'erreur de prédiction vraie suite à l'introduction de la stratégie "free survey" est dépourvue du défaut qui avait

---

14 Cette valeur négative ne correspond pas à une diminution réelle de l'erreur. Il s'agit plutôt d'un artéfact lié au calcul de l'erreur de prédiction. Il convient de rappeler que cette erreur correspond au rapport entre deux termes: le nombre de points mal classés (numérateur) et le nombre de points non touchés par un sondage (dénominateur). Dans le cas considéré, les numérateurs correspondant à chacune des deux prédictions comparées sont quasi égaux alors que, par contre, les dénominateurs sont différents. Il est plus élevé dans le calcul portant sur la prédiction utilisant le moins de sondages, donc, l'erreur de prédiction calculée apparaît plus faible

motivé cette introduction, à savoir la diminution de pente avec l'augmentation de la densité de sondages

La figure 34 montre que ce défaut, quoique atténué subsiste toujours. Ainsi, perdure une perte relative d'efficacité des nouveaux sondages, preuve que la stratégie adoptée n'est pas encore optimale.

L'examen de cette courbe est par ailleurs l'occasion de noter que pour des densités de sondages certes élevées, le modèle est capable de reproduire à l'identique la carte des sols, aux erreurs de délimitations près (Epv = 0).

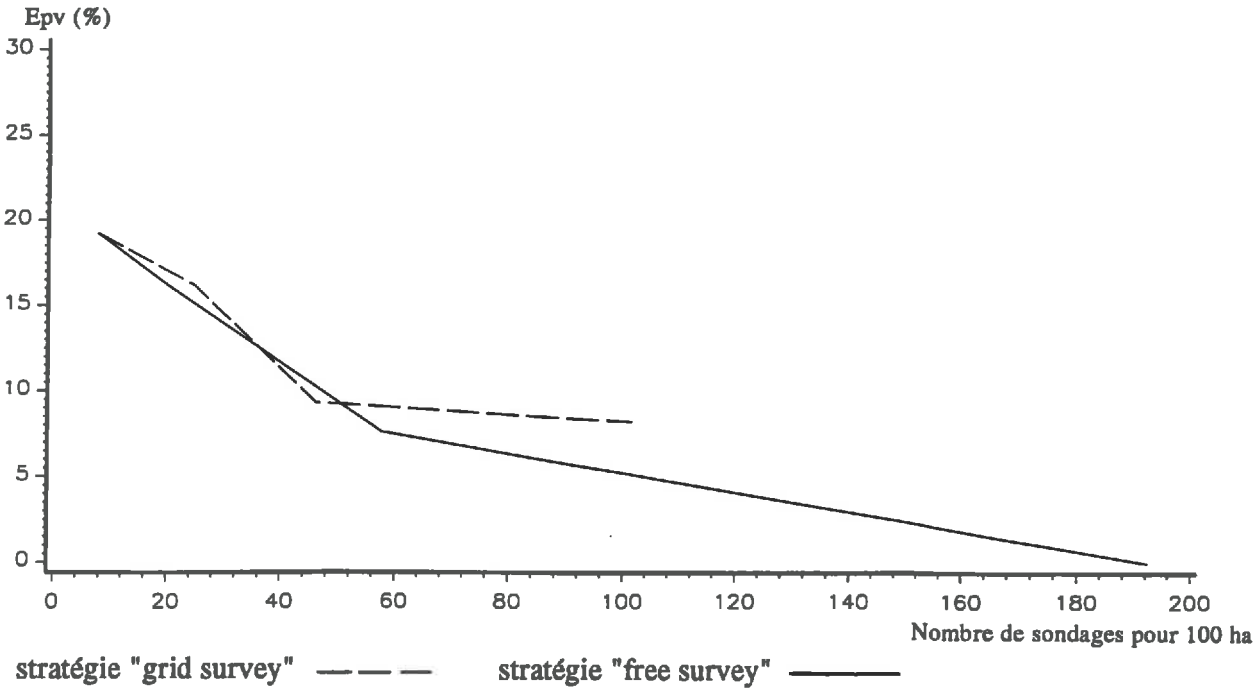


Figure 34: Comparaison des stratégies "grid survey" et "free survey" (en termes d'Epv)

Jusqu'à présent, le plan de sondages utilisé par les prédictions était fixé par des sous-échantillonnages systématiques des points disponibles. Au cours de ce chapitre, cette stratégie de prospection systématique (ou "grid survey") a été remplacée par une nouvelle stratégie ("free survey") qui permet à l'outil de prédiction d'économiser des sondages sur les points pour lesquels il est le plus sûr de ses prédictions (re(x) > 50%).

L'analyse des résultats obtenus par cette nouvelle stratégie révèle une amélioration sensible des performances de l'outil.

Ce fait est important car il permet d'illustrer l'intérêt pratique de l'estimation d'erreur associée à chaque prédiction, préalablement analysée au chapitre précédent. Il est ainsi montré que la connaissance extraite d'une carte des sols peut non seulement servir à prédire ponctuellement une unité de sol suite à l'introduction d'un sondage, mais aussi orienter avec pertinence le choix de l'implantation des sondages dans le but d'alléger la prospection.

La définition de la stratégie "free survey" adoptée au cours de ce chapitre constitue une première tentative. Elle serait en effet susceptible d'être perfectionnée en utilisant encore plus la connaissance issue du secteur de référence. Entre autres, deux voies semblent possibles.

1) **Pour choisir les premiers sondages, phase pour laquelle la stratégie actuelle est inopérante, il convient d'explorer des stratégies utilisant les règles sols-paysage extraites du secteur de référence. Celà pourrait permettre en effet d'accéder plus rapidement à l'ensemble des situations pédologiques contrastées de la zone à étudier.**
2) **Pour choisir les derniers sondages, phase pour laquelle la stratégie actuelle est perfectible, il conviendrait de simuler la phase où le pédologue suit sa limite par des sondages bien positionnés. Ceci passerait par une formalisation préalable de la connaissance sur les limites entre unités de sol dont l'intérêt a déjà été souligné.**

De telles améliorations doivent également s'accompagner du passage à un système expert afin de permettre de programmer le plus librement et le plus aisément possible ce type de stratégie.

CONCLUSION DE LA TROISIEME PARTIE

Au cours de cette troisième partie, ont été formalisées les lois fondées sur les relations de voisinage entre unités de sol. Les règles de voisinage résultant de cette formalisation ont été ensuite utilisées sur les secteurs de validation pour tester leur aptitude à prédire les unités de sol. Pour celà, un outil de prédiction des unités de sol a été construit sur la base de la formalisation de l'opération de retour à la parcelle (chapitre 2).

L'outil de prédiction fonctionne à partir de sondages: à chaque sondage introduit, une règle de voisinage est déclenchée et fournit, sur les autres points, une prédiction correspondant à une série de probabilités de présence pour chaque unité de sol. Comme plusieurs sondages sont réalisés, plusieurs règles sont également déclenchées. En conséquence, il est nécessaire de réaliser sur un point donné, la synthèse des différentes prédictions issues de chaque règle. Ceci constitue un élément nouveau et une difficulté supplémentaire par rapport à la partie précédente.

En l'absence de travaux antérieurs sur la formalisation et l'utilisation des lois de voisinage, il a été mis en oeuvre une démarche originale, déjà résumée au cours du chapitre 7 (conclusion du sous-chapitre 1.). Les aspects nouveaux de cette démarche sont représentés par:

- la variable "sens de voisinage" qui caractérise la position relative d'un point vis à vis d'un sondage en tenant compte de l'arrangement général des unités de sol (toposéquences);
- la stratification de l'espace relativement au point de sondage qui permet d'isoler des zones géographiques homogènes vis à vis des prédictions de sol consécutives au sondage;
- la méthode de pondération qui permet de combiner, sur un point, les prédictions de plusieurs règles de voisinage; elle utilise une fonction inverse de la distance entre le sondage déclenchant la règle et le point considéré.

L'analyse des résultats des prédictions du modèle ainsi construit permet d'ores et déjà de tirer des enseignements sur le plan pédologique.

1) Les prédictions du modèle, à l'intérieur même du secteur de référence, permettent de reproduire à l'identique avec suffisamment de sondages (aux erreurs de délimitation près) la carte initiale. Ce fait atteste de la réalité des lois de cartographie fondées sur les relations de voisinage entre unités de sol qui sous-tendent les règles élaborées. Jusqu'ici ces lois était présentes mais non mises en valeur par le graphisme de la carte des sols. Elles pouvaient être perçues de manière intuitive ou, au mieux, décrites sous forme qualitative (FRIDLAND, 1972). Ces lois sont désormais formulables sous forme de règles permettant des prédictions quantifiées.
2) L'application, à l'extérieur du secteur de référence, des règles de voisinage, révèle des niveaux d'erreurs limités (surtout concernant les erreurs de prédictions vraies) et, de toute façon, en augmentation faible par rapport à leur application à l'intérieur du secteur de référence. Ces lois seraient donc stables au sein d'une région naturelle donnée. Trois restrictions doivent cependant être apportées.
    a) L'éventuelle stabilité de ces lois n'est vérifiée pour l'instant qu'aux alentours immédiats du secteur de référence. La faible superficie et la proximité des

secteurs de validation par rapport au secteur de référence sont trop faibles pour tirer des conclusions sur l'ensemble de la petite région naturelle.

b) Les lois élaborées semblent plus ou moins stables et efficaces suivant les milieux pédologiques abordés. Ainsi, le secteur de validation de Montmau, représentant la rive gauche molassique de l'Hérault, présente des erreurs de prédictions sensiblement plus élevées que les deux autres. Si d'éventuelles et futures modifications de l'outil de prédiction construit (voir plus loin) ne changeaient pas ces résultats, il conviendrait d'invoquer l'inégalité foncière des milieux d'étude quant à leur "aptitude à être cartographiés" [15], réalité bien connue des praticiens de la cartographie.

c) Des problèmes de représentativité apparaissent ponctuellement lorsque l'unité de sol à prédire occupe en réalité des plages cartographiques de géométrie trop différente de celles rencontrées dans le secteur de référence (unité 3). Ce fait amène à réfléchir sur les orientations techniques devant guider le choix du secteur de référence. Aux deux exigences classiques ("périmètre le plus limité possible" et "nombre d'unités le plus grand possible") il faut en ajouter une troisième: elle concerne l'exigence d'une diversité de situations telle que puisse être rencontré, pour chaque unité, l'ensemble des formes de plages cartographiques sous lesquelles elle est susceptible d'apparaître.

3) L'utilisation des règles de voisinage au sein d'un outil simulant le retour à la parcelle est l'occasion d'amorcer une réflexion sur la stratégie d'implantation des sondages, question jusqu'alors jamais traitée à fond par les pédologues auteurs de cartes. Des résultats prometteurs (- 40% de sondages à erreur égale) sont obtenus en utilisant, pour définir une telle stratégie, la carte du risque d'erreur fourni par le modèle avec chaque prédiction. Ceci permet de conforter, en y apportant un début d'explication, l'hypothèse formulée en début de mémoire selon laquelle la cartographie du "retour à la parcelle" se trouverait "allégée" par la mobilisation du savoir faire cartographique acquis sur le secteur de référence. L'effort de recherche concernant la formalisation des stratégies de prospection doit cependant être poursuivi dans les voies indiquées en conclusion du sous-chapitre 3.

L'analyse détaillée des erreurs de prédiction a bien montré que, en général, celles ci étaient plus le fait d'insuffisances de l'outil construit dans ce travail que d'un réel défaut de représentativité du secteur de référence. Des perspectives d'amélioration de cet outil existent. Le débat autour des améliorations possibles se situe, pour un part, sur un plan beaucoup plus général que le cadre pédologique initial. Globalement trois grandes voies d'amélioration peuvent être identifiées.

La première concerne la prise en compte des limites pédologiques et de la connaissance s'y attachant. Il s'agit vraisemblablement de la voie la plus porteuse d'espérances. Les erreurs de délimitation représentent en effet une composante majoritaire de l'erreur de prédiction

---

15 WILDING et DREES (1983) différencient deux composantes de la variabilité spatiale des sols:
- "systematic variability": variabilité spatiale qui peut être reliée à une cause connue (relief, pédogénèse,...)
- "random variability": composante non reliable à une cause connue,

La plus ou moins grande "aptitude d'un milieu à la cartographie" pourrait donc s'exprimer sous la forme du rapport entre ces deux composantes (systematic/random)

observée. Actuellement, l'absence de procédures permettant de recueillir une description du tracé des limites entre unités de sol d'une part, la taille trop grande des pixels utilisés d'autre part interdit toute amélioration dans ce sens. Pour espérer lever cette contrainte, une réflexion préalable doit s'engager pour arriver à obtenir une représentation des objets géographiques (sondages, limites, plages cartographiques) non soumise à un découpage rigide en pixels, tout en conservant les facilités de manipulation que ceux-ci procurent. Par ailleurs l'incertitude et la continuité pédologique sous-jacente à une limite pédologique doivent être également représentée. A cet égard, l'introduction de la théorie des ensembles flous apparaît une solution séduisante (BURROUGH, 1992).

La deuxième voie d'amélioration possible est la prise en considération de la diversité interne des "milieux pédologiques" et des unités pédologiques dans la construction du modèle. Il a été maintes fois souligné que les performances du modèle n'étaient pas homogènes sur l'ensemble des secteurs de validation. Une explication possible est qu'il n'est pas pris assez en considération, lors de sa construction, la diversité de détail des milieux et des unités pédologiques. Concrètement, cette diversité pourrait, par exemple, être intégrée au niveau du choix du motif d'organisation des unités de sols permettant de définir la variable "sens de variation". Pour l'instant, la toposéquence représente le motif unique pour tout le secteur. Un choix différent de motif pour le milieu molassique de la rive gauche (correspondant en fait à une unité pédopaysagère de la carte au 1/250.000) permettrait peut être d'améliorer les performances de l'outil de prédiction sur le secteur de Montmau.

Enfin, la troisième voie d'amélioration consisterait à introduire de nouvelles méthodes de combinaison de sources d'informations incertaines. En effet, lorsque les prédictions issues des relations de voisinage d'une part, et des relations sols-paysage d'autre part sont combinées, la synergie attendue ne se produit pas, peut être à cause de l'indigence de la méthode de combinaison. Des travaux existent sur ce sujet qu'il conviendrait d'explorer plus avant.

CONCLUSION GENERALE

Le travail de recherche entrepris s'inscrit dans la perspective de généraliser une carte des sols réalisée sur un **secteur de référence** à l'ensemble de la **petite région naturelle** qu'il est censé représenter. Plus particulièrement, l'accent a été mis sur l'utilisation des **lois de distribution des unités de sol** dégagées au niveau de la carte du secteur de référence. L'hypothèse retenue dans ce travail est que ces lois sont **stables** sur l'ensemble de la petite région naturelle. Elles pourraient donc être utilisées pour **prédire** la présence des unités de sol du secteur de référence à l'extérieur de celui-ci, sur de nouveaux périmètres d'étude situés dans la même petite région naturelle. Deux types de lois ont été identifiées:

- les lois fondées sur les **relations sols-paysage** ; elles doivent permettre de prédire les unités de sols à partir de la connaissance d'autres éléments du milieu naturel d'accès plus aisé (relief, géologie,...);
- les lois fondées sur les **relations de voisinages entre unités de sol** ; elles doivent permettre de prédire les unités de sols à partir de quelques points connus où ces unités seraient identifiées avec certitude au moyen de sondages à la tarière.

L'objectif du travail était donc de **formaliser** ces lois de distributions de sol puis de tester leur aptitude à prédire les sols à l'extérieur du secteur de référence. Ceci s'est traduit par la réalisation puis l'utilisation d'un **outil informatique** automatisant l'opération de **retour à la parcelle** (cartographie d'un nouveau périmètre utilisant la connaissance acquise sur le secteur de référence).

La méthode retenue pour construire cet outil comporte trois étapes:

- la formalisation du retour à la parcelle;
- l'extraction, à partir de la carte du secteur de référence, des lois de distribution des unités de sol selon le formalisme défini à l'étape précédente;
- la recherche d'algorithmes permettant de combiner en un point les prédictions de plusieurs lois de distribution.

Au cours de la première étape, l'opération de retour à la parcelle est d'abord analysée en détail dans le but d'élaborer un premier **schéma de fonctionnement** de cette opération. Ensuite, sur la base de ce schéma, une **formalisation mathématique** du retour à la parcelle est proposée. Enfin, on réalise son **automatisation** au moyen d'outils informatiques. A l'issue de cette étape, des choix fondamentaux pour la suite du travail sont effectués.

1) Le retour à la parcelle est considéré comme un **processus à étapes**. La première étape concerne la prédiction des unités de sol grâce aux lois sols-paysage. Les étapes suivantes correspondent chacune à l'introduction d'un nouveau sondage. Celui-ci suscite la mobilisation de lois de voisinage qui modifient ou précisent les prédictions précédemment établies.
2) Les prédictions d'unités de sol prennent la forme, en tout point, d'une série de **probabilités** de présence des différentes unités du secteur de référence. Ces probabilités sont destinées à évoluer au fil des étapes, sous l'action de **règles "si (prémisse) alors (conclusion)"** correspondant chacune à une loi de distribution d'unités de sol. Compte tenu du cadre

probabiliste retenu, les conclusions des règles utilisées doivent être également sous la forme d'une série de probabilités. Au terme du processus, l'unité de sol obtenant, en un point donné, la plus forte probabilité est finalement prédite.

3) Les prédictions s'appliquent sur une **grille de points** couvrant la zone à cartographier. Chaque point représente un pixel de 50m de côté. Un **Système d'Information Géographique** permet de gérer l'information nécessaire, en chaque point, pour élaborer les prédictions (ex: position géographique, altitude,...). Des programmes informatiques spécifiques simulent la succession des différentes étapes citées plus haut. En fin de processus, le Système d'Information Géographique est de nouveau utilisé pour afficher les résultats sous forme de **cartes de prédiction** d'unités de sol du secteur de référence.

Le formalisme général et le cadre d'utilisation des lois de distribution des unités de sol étant dès lors fixé, **la deuxième étape** consiste à rechercher des algorithmes permettant d'**extraire automatiquement** ces lois à partir de la carte d'un secteur de référence. **Le secteur de référence de la Moyenne Vallée de l'Hérault** est choisi comme exemple. Pour chacun des deux types de lois cités plus haut, des algorithmes distincts sont recherchés.

1) Les lois fondées sur les relations sols-paysage sont extraites au moyen d'une méthode d'analyse de données existante, **la segmentation**. Cette méthode utilise comme source de données un ensemble de points du secteur de référence pour lesquels sont connus, d'une part, l'unité de sol les concernant et, d'autre part, des descripteurs du milieu naturel d'accès aisé, extraits des cartes géologiques et topographiques. Dans ce cas, l'apport du présent travail est de proposer et mettre en oeuvre une méthode d'interprétation des résultats de la segmentation susceptible de tenir compte de la nature, du niveau et de la propagation des erreurs sur les variables utilisées dans l'analyse.
2) L'extraction des lois fondées sur le voisinage des unités de sol fait au contraire l'objet d'un algorithme original: les points situé à l'intérieur du secteur de référence sont, tour à tour, considérés comme des lieux de sondage. Pour chaque "sondage" appartenant à une unité A donnée, les autres points de la carte sont affectés chacun à des **"zones d'isoprédiction"** au vu de critères définissant leur position relative vis à vis du sondage supposé. Les points ainsi traités sont ensuite dénombrés par unité de sol et par zone d'isoprédiction. En répétant ce processus à partir de tous les points de l'unité A, il est ainsi possible d'obtenir les probabilités d'apparition d'unités de sol sur un point sachant, d'une part, que l'unité A est reconnue sur le sondage et, connaissant, d'autre part, la zone d'isoprédiction dans laquelle le point en question se trouve. Le découpage de l'espace autour du sondage en zones d'isoprédiction tient évidemment compte de la distance par rapport à ce sondage, les zones d'isoprédiction se répartissant autour en couronnes concentriques. Elle tient compte également du **sens de voisinage**, nouvelle variable permettant de redécouper ces couronnes afin de tenir compte de l'existence de **motifs d'organisation des unités de sols** (ici la toposéquence), la couverture pédologique n'étant que rarement isotrope.

La **troisième étape** de la méthode mise en oeuvre dans ce travail aborde le problème de **la combinaison de plusieurs prédictions** s'appliquant en un même point. Selon la formalisation du retour à la parcelle adoptée, ce problème comporte deux aspects bien distincts:

- la combinaison, en un point, de deux sources d'information différentes ("sols-paysage" et "voisinage");
- la combinaison des différentes prédictions issues de la source "voisinage"; il y a en effet autant de prédictions différentes sur un point donné que de sondages réalisés.

Ces deux types de combinaisons, bien différentes l'une de l'autre, font chacune parallèlement l'objet d'une étude détaillée aboutissant à deux formules de calcul spécifiques. Cette voie a été en effet préférée à l'utilisation d'une formule générale (approche Bayesienne par exemple) ne permettant pas de tenir compte du contexte géographique propre à ce travail.

L'outil informatique de prédiction des unités de sol ainsi construit est appliqué sur 3 **secteurs de validation** ayant fait l'objet d' une cartographie conventionnelle. Leur superficie totale est limitée (130 ha) mais ils englobent la totalité des unités de sol importantes du secteur de référence.

A partir de l'analyse du fonctionnement de l'outil et des erreurs de prédictions commises, il est possible de dégager les constats suivants.

1) Les règles sols-paysage dégagées ne permettent pas de prédire seules les unités du secteur de référence. Les erreurs de prédiction obtenues sont en effet élevées (60%). La cause principale de cet échec est l'imprécision des critères de paysage utilisés. En effet, si les ambitions en terme de précision des prédictions sont accordées avec la précision des critères de paysage (regroupement d'unités de sols en unités de carte à petite échelle), les erreurs de prédiction retrouvent un niveau plus satisfaisant (26%). Ce dernier résultat permet d'envisager la production de cartes à petite échelle d'une petite région naturelle à partir de son secteur de référence.
2) Les règles de voisinage permettent au contraire de prédire les unités de sol avec des ratios nombre de sondages/ erreur de prédiction généralement satisfaisants quelles que soient les options expérimentées. Les erreurs obtenues sur l'ensemble des secteurs de validation varient de 25 à 35%. De plus, ces erreurs sont liées dans plus de 50% des cas à des défauts de tracé de limite inférieurs à 35m ("erreurs de délimitation"). Dans le détail, les résultats varient en fonction de plusieurs facteurs.
   a) **Le milieu pédologique abordé**. Le secteur de validation sur molasse où l'organisation de la couverture pédologique est particulièrement complexe présente des erreurs de prédiction systématiquement plus élevées que les deux autres secteurs. Il existerait donc des milieux moins propices que d'autres à la prédiction des sols, sans doute les mêmes que ceux réputés "difficiles à cartographier".
   b) **La stratégie d'implantation des sondages**. Une prospection dans laquelle l'implantation des sondages est décidée en fonction d'un critère mesurant son utilité permet d'économiser, à erreur de prédiction égale, jusqu'à 40% des sondages par rapport à un quadrillage systématique. La recherche de stratégies optimales constitue donc une source potentielle de progrès importante.
   c) **La densité de sondages**. Comme prévu, l'erreur de prédiction diminue lorsque le nombre de sondages augmente. Cependant cette diminution est faible (10% au maximum) comparé à l'effet des deux facteurs précédents. La multiplication des

sondages sans remise en cause de la stratégie de leur implantation (et éventuellement des lois de distribution) n'est donc pas forcément une solution efficace pour améliorer la qualité des prédictions.

3) L'association des deux types de règles ne permet pas d'améliorer la qualité des prédictions. Apparemment, sur le milieu expérimental choisi, il existe une trop grande différence de qualité entre les deux sources d'information pour qu'une éventuelle synergie s'exprime.

Au terme de ce travail de recherche, il convient maintenant de revenir sur les deux questions qui ont motivé sa mise en oeuvre:

- les lois de distribution des unités de sol sont elles stables au sein d'une petite région naturelle?
- Si oui, peut-on envisager la réalisation de cartes automatiques suite à une étude de secteur de référence?

Au vu des résultats, l'hypothèse de stabilité des lois de distribution des unités de sol n'a pas été fortement mise en défaut par l'expérimentation entreprise. Cette dernière ne permet pas pour autant de conclure définitivement sur la réalité de l'hypothèse testée. Pour celà, il faudrait multiplier le nombre de secteurs de validation en constituant un échantillon couvrant la totalité de la petite région naturelle. Nul doute alors que les résultats des prédictions se dégraderaient, les secteurs de validation actuels ayant été choisis à proximité du secteur de référence. Par ailleurs, de nouvelles petites régions naturelles exemples doivent être abordées.

Cependant, le travail réalisé à permis de déplacer la question posée. Il est en effet maintenant démontré que des lois de distribution des unités de sols établies à l'occasion d'une cartographie de secteur de référence peuvent être **localement stables** sur des zones situées à l'extérieur de ce secteur. Désormais, deux nouvelles questions se posent:

- les limites géographiques au delà desquelles la stabilité des lois de distribution des unités de sol n'est plus vérifiée correspondent-elles bien aux limites de la petite région naturelle ?
- si non, la modification ou l'extension du périmètre du secteur de référence, selon des éléments de choix qui restent à définir, permet-elle d'espérer une réponse positive à la première question?

L'outil informatique de prédiction des sols élaboré au cours de ce travail pourrait être utilisé dans les futurs protocoles expérimentaux susceptibles d'apporter des éléments de réponse à ces questions.

La réalisation de cartes de sol automatiques à partir d'un secteur de référence ne pourrait être envisagée un jour que si l'outil actuel était complété et amélioré. L'effort doit être poursuivi selon trois voies.

La première consiste à remettre en cause certaines des solutions mises en oeuvre dans l'outil actuel, à la fois pour extraire et pour utiliser les lois de distribution des unités de sol. Certaines des améliorations font déjà l'objet de discussions et propositions (voir conclusions de 2ème et 3ème partie). Elles intéressent souvent des domaines scientifiques débordant largement la pédologie (analyse de donnée, informatique, traitement de la connaissance,...)

La seconde vise à compléter l'outil actuel en intégrant les autres lois de cartographie ignorées ou sous-utilisées au cours du présent travail. Trois aspects se dégagent.

1) **L'identification des unités de sols à partir de sondages pédologiques.** Dans un travail d'informatisation parallèle à celui-ci (LEDREUX 1992), les lois correspondantes ont été programmées et intégrées à l'outil de prédiction mais la lourdeur de mise en oeuvre de ces lois n'a pas permis de les valider. Une approche numérique du problème, sur la base des travaux de GIRARD (1983), pourrait également être envisagée.
2) **Le tracé des limites.** L'analyse des erreurs de prédiction montre qu'une grande marge de progrès existe par cette voie. Un éventuel travail dans ce sens devrait s'accompagner d'une révision conceptuelle de la notion de limite (théorie des "ensembles flous"),
3) **La stratégie d'implantation des sondages.** Insuffisamment approfondi dans le présent travail, cet aspect doit être privilégié en raison de son intérêt potentiel en terme de réduction d'erreur. De plus, il permet de formaliser et d'évaluer expérimentalement la pertinence des raisonnements mis en oeuvre au cours de la prospection pédologique pour lesquels il n'existe encore aucune étude dans la bibliographie consultée.

La troisième voie accompagne naturellement les deux premières. Il faut en effet repenser les solutions informatiques mises en oeuvre pour automatiser le retour à la parcelle. C'est d'abord un impératif scientifique dans la mesure où l'outil actuel ne permet pas d'explorer les nouvelles voies d'amélioration proposées. C'est ensuite un impératif technique si l'objectif est de produire à terme un outil d'extrapolation des secteurs de référence utilisable par des tiers. Ceci dépasse bien entendu les seules compétences d'un pédologue. Un dialogue doit donc être amorcé avec des spécialistes de traitement informatique des données géographiques et d'intelligence artificielle.

Outre les efforts à réaliser en matière d'amélioration de l'outil, il conviendrait également de prévoir dans quelles conditions pratiques un tel type d'outil pourrait être intéressant et viable. Dans cette perspective, s'impose en particulier une étude détaillée sur la qualité, le mode et le coût d'acquisition des données d'entrées nécessaires (données sur le milieu naturel, données ponctuelles sur la couverture pédologique,...).

Enfin, au delà des deux questions ayant motivé le travail de recherche, il est possible que des démarches intellectuelles proches de celle qui a été formalisée dans ce travail soient également pratiquées dans d'autre domaines scientifiques (géologie, phyto-sociologie,...) Si c'était le cas, les efforts de formalisation entrepris trouveraient leur prolongement et leur débouché dans la conception de nouveaux systèmes d'information géographiques capables de stocker du raisonnement humain et d'utiliser ce raisonnement dans un but d'auto-enrichissement. Ainsi se trouveraient démultipliées, pour des domaines scientifiques variés, les possibilités d'inventaire des ressources et d'études de fonctionnement indispensables pour gérer notre milieu naturel.

BIBLIOGRAPHIE

ASCENSIO E.,1984 - Aspects climatologiques des départements de la région Languedoc-Roussillon. Paris, Ed. Centre technique du matériel de la direction de la météorologie : pp 37-53

ASPINALL R.,1992 - An inductive modelling procedure based on Bayes'theorem for analysisof pattern in spatial data. Int. J. of Geographical Information System, 6 (2) : 105-122.

ASTER., 1990 - Base de données sol-drainage, schéma d'organisation général et defonctionnement. Document INRA Montpellier - CEMAGREF Antony (division drainage), 11pp.

ASTLE W.L., WEBSTER R., LAWRANCE C.J., 1969 - Land classification for management planning in the Luangwa valley of Zambia. Journal of applied Ecology, 6, 143-169.

BACCINI A., 1975 - Aspect synthétique de la segmentation et traitement de variables qualitatives à modalités ordonnées. Thèse université P. Sabatier. Toulouse.

BAILLE M., BOURRELLY L., LAGACHERIE P., 1988. : Modélisation de la connaissance d'un pédologue: application à la reconnaissance d'unités de sol. actes du colloque "Application de l'Intelligence Artificielle à l'agriculture, l'agrochimie et aux industries agro- alimentaires. pp 54 - 73.

BAIZE D., 1986 - Couvertures pédologiques, cartographie et taxonomie. Science du sol, 24 (3) : 227-243.

BECKETT P.H.T., WEBSTER, 1971 - Soil variability : a review. Soils fert, 34 : 1-15.

BENICHOU P., Le BRETON O., 1987 - Prise en compte de la topographie des champs pluviométriques statistiques : la méthode Aurelhy. Agrométéorologie des zones de moyenne montagne. INRA, les colloques de l'INRA, 39 : 51-69.

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

BOLSTAD P.V., GESSLER P., LILLESAND T.M., 1990 - Positional uncertainty in manually digitized map data. Int. J. Geographical information systems, 4 (4) : 399-412.

BONFILS P., 1992 - Carte pédologique de la France au 1/100000 - Feuille de Lodève (notice + carte). INRA SESCPF.

BORNAND M., 1972 - Sols des vallées et des plaines alluviales. Sols, paysages, aménagement SES INRA Montpellier, 37-53.

BORNAND M., BARTHES J.P., BONFILS P., 1992 - Carte des pédopaysages du Languedoc Roussillon à l'échelle du 1/250000. Carte + légende.

BOULET R., CHAUVEL A., HUMBEL F.X., LUCAS Y., 1982 - Analyse structurale et cartographie en pédologie. Les cahiers de l'ORSTOM, série pédologie, 19 : 309-351.

BOUROCHE J.M., TENEHAUS M., 1970 - Quelques méthodes de segmentation. R.I.R.O., 2 : 29-42.

BRABANT P., 1989 - La cartographie des sols dans les régions tropicales : Une procédure à 5 niveaux coordonnés. Science du sol, 27 (4) : 369-394.

BREIMAN L., FRIEDMAN J.H., OLSHEN R.A., STONE C.J., 1984 - Classification and regression trees. Wadsworth & Brooks Wadsworth statistics/probability series.

BRGM, 1981 - Carte géologique de la France au 1/50.000, feuille de Pézenas.

BUISSON L., 1989 - Bases de connaissance et information géographique. SIGEO, 1989.

BURROUGH P.A., 1986 - Principles of geographical information systems for land ressources assessment. Oxford science publication monographs on soil and ressources surveyn°12.

BURROUGH P.A., 1992 - Development of intelligent geographical information systems. Int. J. of Geographical Information Systems, 6 (1) : 1-12.

CAILLEMER A., 1955 - Topographie-Photogrammétrie. Ecole Nationale du pétrole. Document de cours n°306.

CALLOT G., 1977 - Logique de la distribution des sols et formations superficielles sur plate forme calcaire . Notion de système géopédologique régional. Exemple de la région Nord Aquitaine. Science du sol, n°4 : 189-205.

CESTRE T., DUVAL O., FAVROT J.C., 1971 - Liste des organismes, bureaux d'études et pédologues intervenant en cartographie des sols. Document SESCPF Orléans - GEPPA.

CHOSSAT M., 1977 - Aide mémoire de mathématique de l'ingénieur. Dunod.

CHRISMAN N.R., 1987 - The accuracy of map overlay : a reassessment. Lanscape and urban planning, 14 : 427-439.

CHRISMAN N.R., 1989 - Modelling errors in overlaid categorical map. Accuracy of spatial database, Goodchild and Gopal, ed Taylor and Francis, 21-35.

CHRISMAN N.R., 1991 - Quality control of data in GIS. Bulletin de la Société Française de la Photogrammétrie et de la Télédétection, 122 : 18-23.

CLARK L.A., PREGIBON D., 1991 - Tree-based models, Statistical models in S. Wadsworth & Brooks, Wadsworth statistics/probability series, 377-419.

CLARKE A.L. , GRUEN A., LOON J.C, 1982 - The application of contour data for generating high fidelity grid digital elevation models. Auto carte 5.

CLOCHARD P., LEENHARDT D. et LEFAY O., 1991 - Carte des sols du secteur de référence d'Adissan-Montagnac (Moyenne vallée de l'Hérault). Carte 1/10000.

CONACHER A.J., DARLYMPE J.B., 1977 - The nine unit land surface model : an approach to pedogeomorphic research. Geoderma, 18 : 1-154.

CPCS, 1967 - Classification des sols . Multigraph 87 pp.

DENT D., YOUNG A., 1981 - Soil survey and land evaluation. Allen & Unwin.

DEPRAETERE C., 1990 - OROLOG, logiciel de calcul de modèles numériques de terrain à partir de courbes de niveau. ORSTOM, Notice OVNIH du Laboratoire d'hydrologie n°6.

DEPRAETERE C., 1990 - Support de cours à la session de formation sur le logiciel LAMONT. ORSTOM.

DEPRAETERE C., 1991 - LAMONT, Logiciel d'application des modèles numériques de terrain. ORSTOM Notice OVNIh du Laboratoire d'hydrologie n°4.

DUNN R., HARRISON A.R., WHITE J.C., 1990 - Positional accuracy and measurement error in digital databases of land use : an empirical study. International Journal of Geographical Information Systems, 4 (4) : 385-398.

DUPERET A., 1989 - Contribution des MNT à la géomorphométrie. DEA en sciences de l'information géographique, ENSG-IGN, IMAGEO-CNRS.

ESRI, 1989 - ARC/INFO référence manual.

FAVROT J.C., 1981 - Pour une approche raisonnée du drainage agricole en France : La méthode des secteurs de référence. C.R. académie d'agriculture de France, séance du 6 mai 1981, : pp 716-723.

FAVROT J.C., 1989 - Une stratégie d'inventaire cartographique à grande échelle : la méthode des secteurs de référence. Science du sol, 27 (4) : 351-368.

FAVROT J.C., BOUZIGUES R., HERVE J.J., CESTRE T., 1981 - Recommandations pour la réalisation des études de sols préalables au drainage dans le cadre des "secteurs de référence" et des projets à la parcelle. Bulletin d'information du CEMAGREF, n° 283 : 41-55.

FEENY V., 1988 - Méthode d'analyse du contenant des unités cartographiques. DEA fédéral de pédologie, INAPG 57 pp.

FISHER P.F., BALACHANDRAN C.S., 1989 - STAX : a Turbo Prolog rule based system for soil taxonomy. Computers and geosciences, 15 (7) : 295-324.

FRIDLAND V. M., 1972 - Pattern of the soil cover. Israel programm for scientific translations Jerusalem 1976 (ed. PrYaalon).

FROLOV Y.S., MALING D.H., 1969 - The accuracy of area measurement by point counting technique. Cartogr. J. 6 : 21-35.

GAULTIER J.P., 1990 - Projet DONESOL, étude détaillée, 4 fasicules. Dpt Science du Sol INRA.

GENEST C., ZIDEK J.V., 1986 - Combining probability distributions : A critique and an annotated bibliography. Statistical science, 1 (1) : 114-148.

GEPPA, 1967 - Modèle de marché pour une étude de sols. Ministère de l'agriculture - direction des aménagements ruraux, service de l'hydraulique.

GIRARD M.C., 1983 - Recherche d'une modélisation en vue d'une représentation spatiale de la couverture pédologique. Application à une région des plateaux jurassiques de Bourgogne. Thèse d'état. INAPG sols n°13.

GOODCHILD M., GOPAL S., 1989 - Accuracy of spatial databases. Taylor & Francis London-New York-Philadelphia.

GRZEBYK M., 1991 - Etude quantitative de séquences pédologiques. DEA Centre de géostatistique de Fontainebleau, 43 pp.

HATON J.P., 1991 - Le raisonnement en intelligence artificielle. collection iia, ed InterEditions.

HOLE F.D., 1978 - An approach to landscape analysis with enphasis on soils. Geoderma, 21 : 1-23.

HUGGETT R.J., 1975 - Soil lanscape system : a model of soil genesis. Geoderma, 13 : 1-22.

IGN, 1983 - Carte topographique au 1 : 25 000. Feuille de Pézenas (2644 Est).

INRA, 1992 - Etude des flux d'eau et de polluants en milieu méditerranéen viticole: le programme ALLEGRO-Roujan. Note interne AIP "eau", 6pp

JAMAGNE M., 1967 - Bases et techniques d'une cartographie des sols. Annales agronomiques, n° hors série vol. 18, 142 pp.

JAMET O., 1991 - Mesure de la qualité de l'information d'occupation du sol dans un SIG. Bulletin de la Société Française de la Photogrammétrie et de la Télédétection, 122 : 29-34.

JENNY H., 1941 - Factors of soil formation. New York, Mac Graw hill C°, 281 pp.

KING D., 1986 - Modélisation cartographique du comportement des sols basée sur la mise en valeur du "Marais de Rochefort". Thèse de docteur ingénieur INAPG 173 pp + annexes.

KOLLIAS V.J., 1988 - Logic programming in soil classification. Soil survey and land evaluation, 8 (2) : 94-99.

LAGACHERIE P., 1986 - Elaboration de blocs diagrammes a partir de modèles numériques de terrain de l'IGN : Exemple du secteur de référence de la Bresse Jurassienne. B.T.I Ministère de l'agriculture, 406 : 3-10.

LAGACHERIE P., 1987 - Synthèse générale sur les études de secteur de référence drainage. CNABRL, INRA Montpellier SES n°591.

LAGACHERIE P., LEDREUX C., 1991 - Essai de modélisation du raisonnement cartographique d'un pédologue. Le projet SAPRISTI. actes du colloque SIG-GIS ,ed Hermès.

LAGACHERIE P., LEDREUX C., 1992 - Spécifications fonctionnelles pour un projet de création d'un outil informatique d'aide à la cartographie. Note interne INRA Montpellier, 7pp

LEENHARDT D., 1991 - Spatialisation du bilan hydrique. Propagation des erreurs d'estimation des caractéristiques du sol au travers des modèles de bilan hydrique. Cas du blé dur d'hiver. These ENSA Montpellier, 123pp + annexes.

LEFORT G., 1967 - Mathématiques pour les sciences biologiques et agronomiques. édition Armand Colin, collection U.

LEGROS J.P., 1975 - Occurrence des podzols dans l'Est du Massif Central. Science du sol, 1 : 37-49.

LOVELAND, 1992 - A spatially-distributed soil, agroclimatic and soil-hydrological model to predict the effects of climate change on land use within the European Community. EEC Project, proposal N° PL910069, 7 pp.

LOWELL K., 1991 - Utilizing discriminant function analysis with a geographical Information system to model ecological succession spatially. Int. J. of Geographical Information System, 5 (2) : 175-192.

MACDOUGALL E. B., 1975 - The accuracy of map overlays. Landscape and planning, 2 : 23-30.

MAFFINI G., ARNO M., BITTERLICH W., 1989 - Observations and comments on the generation and treatment of error in digital GIS data. Accuracy of spatial databases Taylor and Francis, 55-67.

MAIGNIEN R., 1969 - Manuel de prospection pédologique. ORSTOM initiations - documentations.

MAILING D.H., 1989 - Measurement from maps, principle and methods of cartometry. Pergamon press.

MARTIN CLOUAIRE R., 1992 - Combinaison de plusieurs sources d'information. Note interne INRA Laboratoire d'intelligence artificielle, 12 pp.

MIDDELKOOP H., MILTENBURG J. and MULDER N.J., 1989 - Knowledge engineering for image interpretation and classification : a trial run. ITC Journal, 1 : 27-32.

MOLLET J.M., 1991.- Aptitude à la diversification en grandes cultures. Etude agro-pédo-climatique et modélisation du bilan hydrique dans la Moyenne Vallée de l'Hérault. D.A.A. ENSA Montpellier 72pp.

ODEH I.O.A., CHITTLEBOROUGH D.J., Mc BRATNEY A.B., 1991 - Elucidation of soil-landform interrelationships by canonical ordination analysis. Geoderma, 49 : 1-32.

PAVAT J.L., 1986 - Contribution à l'étude de la ressemblance entre types de sol, apport de l'informatique. Application aux secteurs de référence. DEA USTL - ENSAM 77 pp.

PEDRO G., 1989 - L'approche spatiale en pédologie . Fondements de la connaissance des sols dans le milieu naturel. Réflexions liminaires. Science du sol, 27 (4) : 287-300.

PLANCHON O., 1991 - Etude spatialisée des écoulements sur les versants et conséquences sur l'hydrologie et l'érosion. Exemple en Savane humide. ORSTOM.

POURGATON H., 1977 - Introduction à l'étude statistique de la notion de province pédologique. DAA "science du sol et aménagement" ENSA de Montpellier, 27 pp et annexes.

RUELLAN A., DOSSO M., FRITSCH E., 1989 - L'analyse structurale de la couverture pédologique. Science du sol, 27 (4) : 319-334.

SCHELLING J., 1970 - Soil genesis, soil classification and soil survey. Geoderma, 4 : 165-193.

SERVANT J., 1972 - Méthodologie de la carte pédologique. Sols, paysages, aménagment SES INRA Montpellier, 5-13.

SHAFER G., 1976 - A Mathematical theory of evidence. Princeton N.J.

SHOVIC H.F. MONTAGNE C., 1985 - Application of a statistical soil-landscape model to an order III Wildland soil survey. Soil science society of America Journal, 49 : 961-967.

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. D.A.A. mention milieu physique, option sciences du sol et du bioclimat. 61pp.

SKIDMORE A.K., RYAN P.J., DAWES W., SHORT D., O'LOUGHLIN E., 1991 - Use of expert system to map forest soil from geographical information system. Int.J. Geographical Information Systems, 5 (4) : 431-445.

SMECK N.E., RUNGE E.C.A., MACKINTOSH E.E., 1983 - Dynamics and Genetic modelling of soil systems. in Pedogenesis and soil taxonomy, Concepts and interactions. Elsevier Amsterdam.

SOIL SURVEY STAFF, 1951 - Soil survey manual. USDA agricultural handbook.

SONQUIST J.N. and MORGAN J.N., 1964 - The detection of interaction effects. Monograph 35, survey research center, institute for Social research. University of Michigan.

TWERY M.J., ELMES G.A., YUILL C.B., 1991 - Scientific exploration with an intelligent GIS : Predicting species composition from topography. AI Applications, 5 (2) : 45-53.

VEREGIN H., 1989 - A review of error models for vector to raster conversion. The operational cartographer, 7 (11) : 11-15.

VEREGIN H., 1989 - Error modelling for the map overlay operation. Accuracy of spatial database. Goodchild and Gopal Taylor and Francis.

VINK A.P.A., 1963 - Aspects de pédologie appliquée. La Baconnière Neuchatel.

VOGEL C., 1988 - Génie cognitif. Ed. Masson.

VOLTZ M., 1991 - Apports de l'analyse de l'organisation des sols a l'estimation spatiale des paramètres de transfert. Séminaire du département de science du sol INRA, 17 pp.

WALKER P.A., MOORE D.M., 1988 - SIMPLE. An inductive modelling and mapping tool for spatially-oriented data. International Journal of Geographical Information Systems Taylor and Francis, 2 : 347-363.

WALSH S.J., LIGHTFOOT D.R., BUTLER D.R., 1987 - Recognition and assessment of error in Geographical Information Systems. Photogrammetric Engineering and Remote Sensing, 53 (10) : 1423-1430.

WEBSTER C., 1990 - Rule-based spatial research. Int. J. of Geographical Information System, 4 (3) : 241-260.

WEBSTER R., OLIVER M.A., 1990 - Statistical methods in soil and land resource survey. Spatial Information System.

WILDING L.P., DREES L.R., 1983 - Spatial variability and pedology in Pedogenesis and soil taxonomy. 1 - Concepts and interactions. Elsevier Amsterdam.

YOELI P., 1986 - Computer executed production of a regular grid of height points fromdigital contours. The american cartographer, 13 (3) : 219-229.

ZADEH L., 1978 - Fuzzy sets as a basis for a theory of possibility. Fuzzy sets and systems, 1 (1) : 3-28.

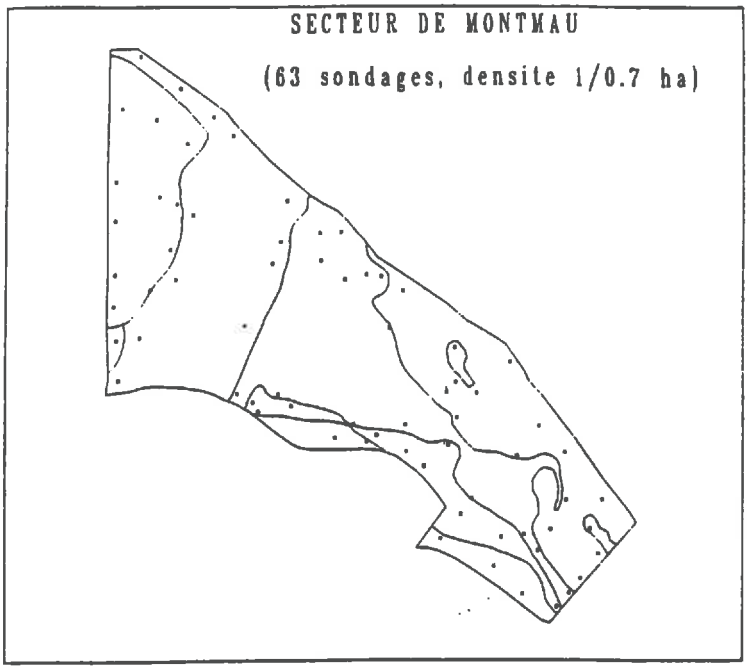
TABLE DES MATIERES

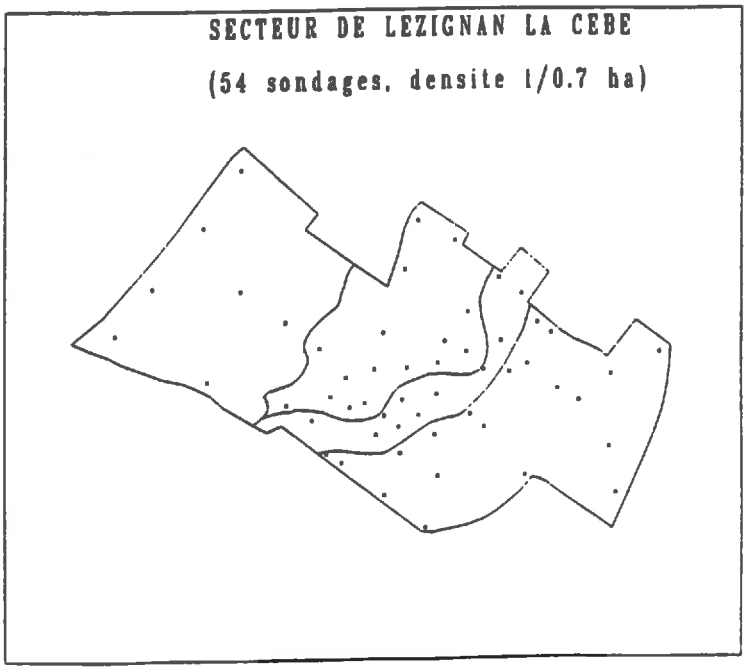
LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES PLANCHES

LISTE DES ANNEXES





Annexe 2 : Plans de sondages des cartographies de secteurs de validation

Annexe 3 : Méthode de production du MNT par interpolation des courbes de niveau (D'après DEPRAETERE 1990)

La méthode d'interpolation utilisée dans le logiciel OROLOG est inspirée de Yoeli (1986). Cette méthode consiste à calculer la valeur moyenne pondérée de l'altitude interpolée à l'aide de fonctions splines cubiques le long de quatre axes.

2.1./Propriétés des fonctions splines cubiques appliquées à un transect:

A l'origine, le terme "spline" s'appliquait à une languette métallique flexible utilisée en architecture navale et en aéronautique afin de faciliter le tracé d'une courbe régulière passant par une série de points donnés , les propriétés élastiques de l'instrument garantissant la continuité de la pente et de la courbure aux points imposés (Delhomme, 1976).

Par analogie, les courbes splines correspondent à des fonctions polynomiales ajustées par morceau sur un transect. Ce type de courbe est particulièrement adapté à la représentation de phénomènes continus.Il existe plusieurs familles, chacune étant caractérisée par le degré de la fonction polynomiale qui sert d'élément de base à la construction de la fonction spline. La famille la plus utilisée dans les problèmes d'interpolation est celle des splines cubiques dont la fonction polynomiale élémentaire est de la forme:

$$y = f(x) = a\,x^3 + b\,x^2 + c\,x + d.$$

Chaque élément de courbe spline cubique est donc caractérisé par quatre paramètres. La courbe spline cubique est un interpolateur exact ce qui signifie qu'elle passe par les points de mesure. Elle obéit de surcroît à une minimisation du carré de la norme de la dérivée seconde sur l'intervalle d'interpolation. Il est possible de démontrer que cela revient à minimiser l'énergie de flexion, c'est-à-dire à obtenir la courbe la plus lisse possible:

$$\int_C f''(x)^2\,dx \text{ minimum}$$

Le calage des éléments ("morceaux") de la courbe se fait donc en considérant deux points consécutifs et en imposant la continuité des dérivées secondes en chacun de ces deux points. Une courbe spline cubique est donc une succession de polynômes du troisième degré calés sur chaque intervalle défini par deux points de mesure consécutifs et se raccordant les uns aux autres par une condition de continuité des courbures.

Dubrulle (1981) a démontré que la résolution des (n-1) systèmes associés au calage des (n-1) polynômes cubiques se ramenait en fait à la résolution d'un système unique de type:

$$\begin{bmatrix}\Psi\\ \alpha\end{bmatrix} = \begin{bmatrix}K & E\\ E^t & O\end{bmatrix}^{-1} \begin{bmatrix}Z\\ O\end{bmatrix}$$

où K est une matrice (n,n) construite à l'aide de la fonctionnelle $K(x_i, x_j) = a\,|x_i x_j|^3$ pour l'ensemble des couples (i,j) de mesure, et E est une matrice (n,2):

$$E = \begin{bmatrix}1 & x_1\\ \ldots & \ldots\\ 1 & x_n\end{bmatrix}$$

Z est le vecteur colonne des n valeurs observées (z_i) aux points x_i, et (Ψ,α) est le vecteur des coefficients de la fonction spline cubique exprimée sous la forme:

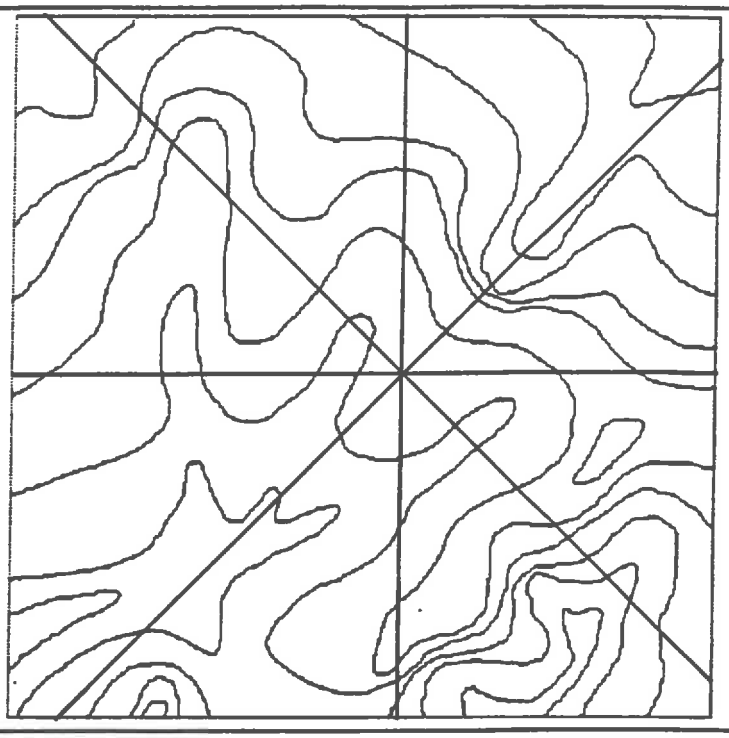
$$\sigma(x) = \alpha + \beta x + \sum_{i=1}^{n} \Psi_i (x_i - x)^3$$

On évite ainsi un calage de proche en proche. La généralisation à un espace à deux dimensions (fonction spline de type "plaque mince") utilise la fonctionnelle $h^2 \text{ Log } |h|$, le calcul des coefficients Ψ et α résulte de la résolution d'un système analogue à celui ci-dessus.

2.2./le système d'axes d'interpolation des courbes splines cubiques:

Pour chaque point du MNT, l'interpolation se fait le long de quatre axes (figure 1) :

figure 1 : Les quatre axes d'interpolation



- axe Ouest Est, chaque axe correspond à un numéro de point par profil du MNT.
- axe Sud Nord, chaque axe correspond à un profil du MNT.
- axe Nord-Ouest Sud-Est, chaque axe correspond à une diagonale descendante.
- axe Sud-Ouest - Nord-Est, chaque axe correspond à une diagonale ascendante.

Le nombre d'axes d'interpolation pourrait être plus élevé, cela multipliant d'autant le nombre de calculs. Dans le cas de quatre axes, on fait l'hypothèse qu'il suffit de rechercher l'information dans 8 directions pour obtenir une valeur locale de l'altitude cohérente au regard des courbes de niveau environnantes. Pour reprendre l'image de l'oeil de poisson, cela revient à admettre que cet oeil ne comporte que 8 facettes.

Le pas du MNT permet de définir l'espacement entre les axes: égal au pas pour les axes Sud Nord et Ouest Est, l'espacement doit être divisé par $2^{1/2}$ pour les axes diagonaux.

Une fois défini le pas du MNT, la première étape du calcul consiste à rechercher toutes les intersections entre les courbes de niveau et le système d'axes. Ces intersections serviront de points de calage pour le calcul des courbes splines le long des axes.

2.3./ Méthode de calcul des altitudes à partir des valeurs interpolées sur les 4 axes:

L'altitude de chaque point du MNT sera obtenue en faisant la moyenne pondérée des altitudes calculées sur les axes.

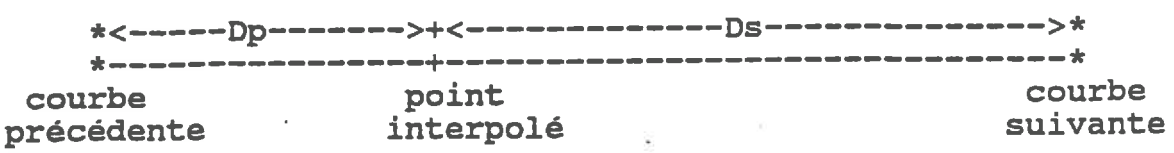
Si l'on admet que les points de calage sur les axes sont exacts, la validité de l'interpolation en un point devrait dépendre essentiellement des caractéristiques suivantes :

- la pente ainsi que la distance entre les deux courbes de niveau encadrant le point. Plus la pente est forte, plus proches sont les courbes et plus précise sera la valeur interpolée ;

- la proximité du point par rapport à l'une des deux courbes de niveau encadrantes. Plus le point est proche de celui d'une courbe, plus leurs valeurs seront voisines et meilleure sera la précision de l'interpolation.

La pondération utilisée par Yoeli consiste à faire la somme des inverses des distances aux courbes de niveau. Soient Dp la distance à la courbe précédente et Ds la distance à la courbe suivante, la pondération W sera:

$$W = 1 / Dp + 1 / Ds$$



Cette pondération permet de prendre en compte à la fois la distance entre les courbes et la proximité du point par rapport à l'une des deux courbes.

La valeur Z la plus probable de l'altitude en un point est égale à la moyenne pondérée des valeurs interpolées Z_1, Z_2, Z_3, Z_4 sur les axes:

$$Z = \text{Somme}_{i=1 \text{ à } 4} (W_i Z_i) / \text{Somme}_{i=1 \text{ à } 4} (W_i)$$

Il est possible de calculer l'erreur quadratique moyenne entre la valeur de la moyenne pondérée et les quatre valeurs interpolées de départ.

Soit $D_i = Z - Z_i$:

L'erreur quadratique moyenne E de la moyenne pondérée sera:

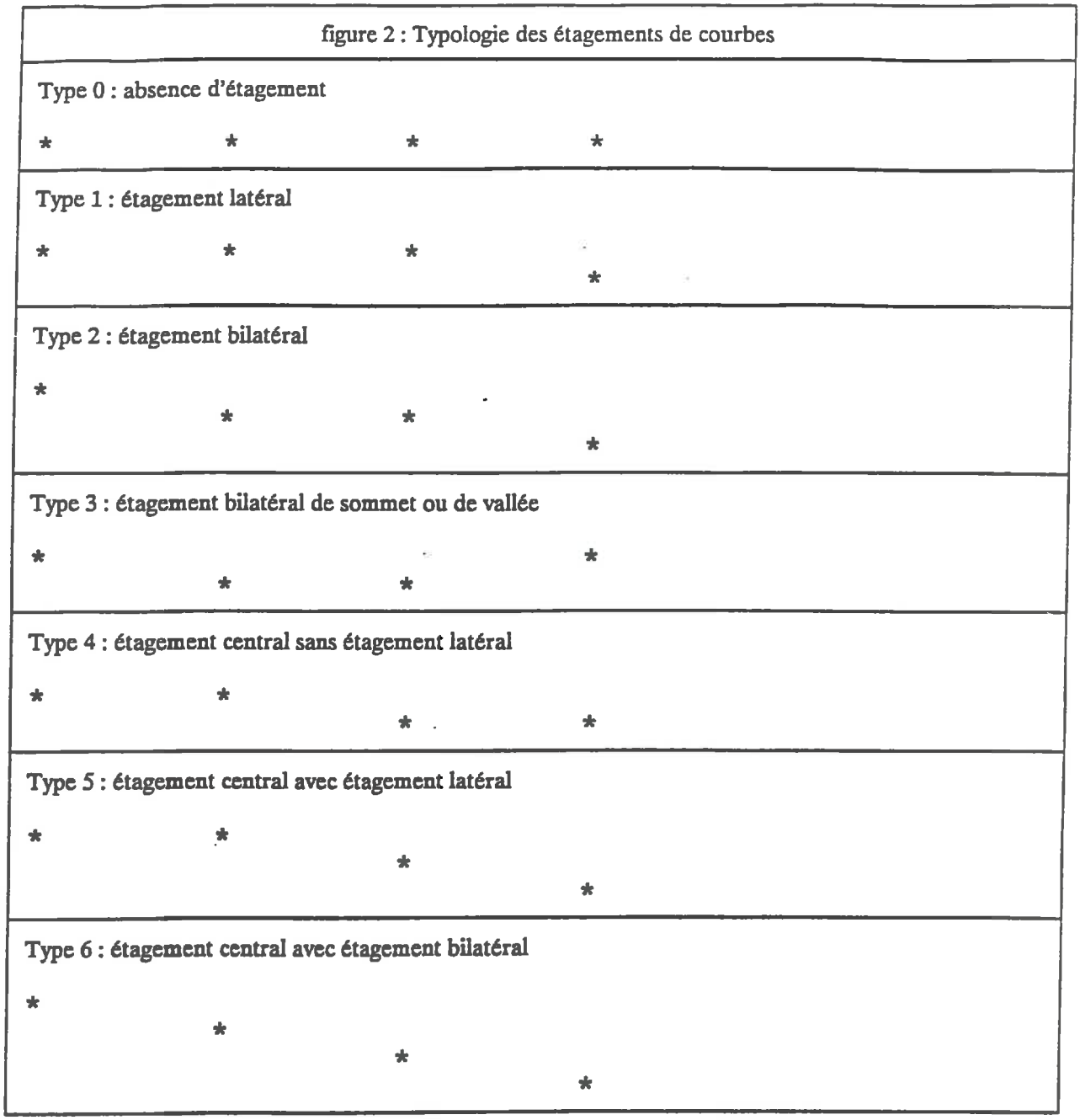
$$E = (\text{Somme}_{i=1 \text{ à } 4}(D_i^2 W_i) / (3 \times \text{Somme}_{i=1 \text{ à } 4}(W_i)))^{1/2}$$

2.4./ Surpondération en fonction de la morphologie locale des courbes de niveau:

Le critère de pondération de Yoeli ne prend pas en compte la différence d'altitude entre les courbes de niveau encadrantes. L'interpolation spline cubique peut être responsable de fortes oscillations de la courbe qui sont hors de proportion par rapport au phénomène étudié. Ces oscillations sont moins fréquentes lorsque les points de calage présentent une tendance simple. Au niveau topographique, un ensemble de courbes de niveau d'altitudes décroissantes (ou croissantes) le long d'un axe représente un étagement simple. En revanche, une structure de points de calage présentant des changements de tendance (structure que l'on peut qualifier de chaotique) sera potentiellement propice à l'apparition d'oscillations. Comme nous l'avons vu précédemment, les fonctions cubiques sont interpolées sur des "morceaux" comportant quatre courbes de niveau. Une typologie de l'étagement de ces quatre points de calage le long de l'axe permet de définir le contexte morphologique dans lequel l'interpolation du point a été faite.

Six types de structure sont distingués en fonction de l'étagement des 4 courbes de niveau (figure 2).

figure 2 : Typologie des étagements de courbes



Cette typologie est établie pour chaque point sur les quatre axes. Elle permet de définir un critère de surpondération morphologique SPM calculé comme suit en fonction du type de site morphologique SM:

$$SPM = C_{sm}^{SM}$$

Le coefficient de surpondération C_{sm} doit être supérieur à 1.

La moyenne des valeurs d'altitudes interpolées sur les quatre axes pondérée à la fois avec le critère de la somme des inverses de distance aux courbes encadrantes et le critère de surpondération morphologique sera égale à :

$$Z = \text{Somme}_{i=1 \text{ à } 4} (SPM_i W_i Z_i) / \text{Somme}_{i=1 \text{ à } 4} (SPM_i W_i)$$

Cette surpondération morphologique privilégie donc les directions où l'étagement des courbes locales est le plus marqué. Elle semblerait correspondre à la méthode intuitivement suivie par des "experts humains" travaillant à partir de cartes. Cette méthode de surpondération morphologique nous parait être une façon pertinente de mettre des oeillères sur notre oeil de poisson à 8 facettes.

Annexe 4: production des fichiers dérivés du MNT

3 variables dérivées du MNT sont utilisées pour la formalisation des lois sols-paysage. Il s'agit de la pente, de la courbure moyenne et de l'encaissement. L'acquisition des deux premières utilise une méthodologie mis au point par DUPERET (1989). Cette méthodologie sera présentée dans une première partie de cette annexe. La dernière a été proposée par DEPRAETERE (199O). Son calcul fera l'objet d'une deuxième partie. La rédaction de l'annexe 4 s'inspire largement des travaux des deux auteurs cités.

1. Calcul des variables "pente" et "courbure moyenne"

Les variables "pente" et "courbure moyenne" représentent les dérivées premières et secondes du champ des altitudes matérialisé par le MNT. La démarche consiste donc à:

- établir, à partir du MNT, l'équation du plan représentant, en tout point, ce champ des altitudes,
- dériver l'expression obtenue pour obtenir les variables désirées.

1.1. Représentation mathématique du champ des altitudes

Aucune représentation mathématique ne permet de rendre compte parfaitement du relief d'une surface quelconque. L'utilisation de fonctions polynomiales par morceaux permet néanmoins de fournir une bonne approximation du relief à représenter. Dans la méthodologie utilisée par DUPERET (1989), ces fonctions polynomiales sont constituées par des développements de Taylor à l'ordre 2 dont les coefficients sont déterminés par ajustement sur une fenêtre 3X3 centrée sur un point (x_0,y_0) d'altitude z_0 (figure 1).

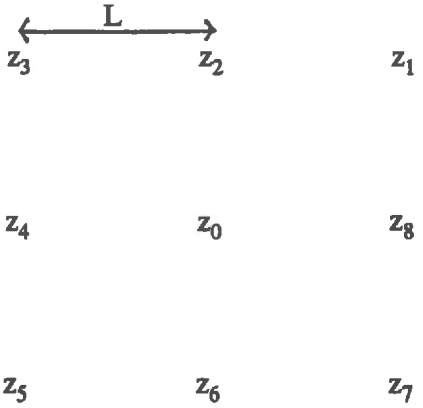


Figure 1: fenêtre 3X3 sur laquelle est réalisé l'ajustement du développement de Taylor

L'équation générale du plan représentant au mieux le champ d'altitude au voisinage du point central de la fenêtre est donc:

$$z = H(x,y) = a + bx + cy + 1/2(dx^2 + 2exy + fy^2) + \epsilon \qquad [1]$$

avec x = latitude y = longitude z = altitude

Remarque: x et y sont exprimées en unité de pas (= L) et selon un repère ayant pour origine le point (x_0,y_0) (en (x_0,y_0), $x=y=0$).

Les coefficients du polynôme, calculés par moindre carrés, s'expriment de la manière suivante (DUPERET 1989):

$a= 5/9\ z_0 + 2/9(z_2 + z_4 + z_6 + z_8) - 1/9(z_1 + z_3 + z_5 + z_7)$
$b= 1/6(z_1 - z_3 - z_4 + z_5 + z_7 + z_8)$
$c= 1/6(z_1 + z_2 + z_3 - z_5 - z_6 - z_7)$
$d= -2/3\ z_0 + 1/3(z_1 + z_3 + z_5 + z_7) + 1/3(z_8 + z_4) - 2/3(z_2 + z_6)$
$e= 1/4(z_1 - z_3 + z_5 - z_7)$
$f= -2/3z_0 + 1/3(z_1 + z_3 + z_5 + z_7) -2/3(z_8 + z_4) + 1/3(z_2 + z_6)$

Si l'on veut que la formule [1] s'exprime en unités normalisée (x, y et z en mètres), il faut, de plus, multiplier les coefficients b et c par 1/L et les coefficients d, e, et f par $1/L^2$.

1.2. Calcul des variables dérivées

1.2.1. La pente

Le vecteur pente a pour expression générale (DUPERET 1989):

$$Pt = Grad\ (H) \qquad [2]$$

Ses coordonnées, obtenues en dérivant l'expression de H, sont:

$$\delta H/\delta x = b + dx + ey$$
$$\delta H/\delta y = c + ex + fy \qquad [3]$$

La valeur de la pente au point (x_0,y_0) situé au centre de la fenêtre sont donc $(x=y=0)$:

$$p = \sqrt{(b^2+c^2)} \qquad [4]$$

ou, en exprimant P en unités normalisées

$$p = 1/L\sqrt{(b^2+c^2)} \qquad [5]$$

1.2.2. La courbure moyenne

La courbure moyenne se définit en (x_0,y_0) comme le laplacien de z. Ce qui donne en dérivant l'expression du polynome de la formule [1]:

$$cm = d + f \qquad [6]$$

cm s'exprime en % par unité de pas

2. Calcul de l'encaissement

La variable "encaissement" représente une mesure du dénivelé entre le point central et ses 8 voisins (cf figure 1). Elle correspond à la somme de chacun des dénivelés:

$$ec = \Sigma_{(i=1...8)}\ z_0 - z_i \qquad [7]$$

ANNEXE 5

Présentation de la démarche permettant d'estimer l'écart type d'erreur sur l'indice d'impureté $i(X_m)$:

L'estimation de l'erreur sur $i(X_m)$ consiste à remonter la filière aboutissant à son calcul. Elle dépasse largement le cadre étroit de la segmentation puisqu'il convient de prendre en compte les erreurs sur les données de l'ensemble d'apprentissage. Ces données étant issues de SIG, c'est dans ce domaine d'activité scientifique que seront recherchées les références nécessaires.

Avec la prise de conscience générale de l'importance du problème des erreurs des cartes et de leur propagation au sein des Systèmes d'Information Géographiques (BURROUGH 1986,GOODCHILD et GOPAL 1989, ...) une très abondante bibliographie existe maintenant sur le sujet. Mais, ne sont pas encore apparus, semble-t-il, de formulations complètes sur lesquelles pourrait s'appuyer un calcul rigoureux de $Se[i(X_m)]$, tenant compte de toutes les sources d'erreur émaillant la filière d'obtention de $i(X_m)$.

Dès lors, la démarche mise en oeuvre pour estimer $Se[i(X_m)]$ vise à utiliser et associer des connaissances éparses concernant ces sources et des modèles rendant compte de leur propagation sur certains segments de la filière. En conséquence, il ne saurait s'agir d'une méthode idéale et fixée de façon définitive. Au contraire, compte tenu de l'effort actuel de la communauté scientifique sur le sujet, des améliorations sont à prévoir.

En l'état actuel des outils et connaissances disponibles, les estimations de $Se[i(X_m)]$ devront donc être utilisées avec prudence. En effet de nombreuses lacunes existent dans la connaissance des différents termes de l'erreur, certains d'entre eux n'ayant jamais fait l'objet d'études expérimentales pour estimer leur ordre de grandeur.

Par ailleurs, ces termes se combinent suivant des modes de propagation d'erreur variés compte tenu de la diversité des opérations intervenant depuis la fabrication des cartes jusqu'à l'obtention de $i(X_m)$. Les modèles de propagation d'erreurs associés à ces opérations sont souvent mal connus, insuffisants et peu homogènes pour ce qui concernent leurs entrées et sorties. Ceci impose parfois des hypothèses simplificatrices hasardeuses sur les termes de l'erreur (additivité, indépendance, normalité). Elles s'avèrent en effet indispensables pour définir les calculs d'erreurs, les simplifier et homogénéiser la signification des différents termes d'erreurs combinés.

L'estimation de $Se[i(X_m)]$ repose sur l'examen du processus qui a permis, in fine, de calculer la valeur de $i(X_m)$. Trois étapes se dégagent et seront envisagées dans la suite de ce chapitre:

- la production en tout point des valeurs des variables topo-géologiques (z(x), dz(x), g(x),...) et de l'unité de sol; ceci implique de rechercher, de sélectionner et d'évaluer les sources d'erreur prépondérantes intervenant dans le processus de fabrication et de manipulation des cartes utilisées et susceptibles d'influencer cette production;
- la sélection et le dénombrement des points inclus dans chaque noeud permettant l'estimation de N_m et N_{mj}. Cette opération relève de la propagation d'erreur au sein d'un Système d'Information Géographique,
- le calcul de $i(X_m)$ à partir de N_m et N_{mj}. Cette étape représente un simple calcul arithmétique pour lequel il est possible d'évaluer l'erreur résultante grâce à la formule générale de propagation d'erreur dans une opération arithmétique .

1. ESTIMATION DE L'ERREUR INTERVENANT DANS LE PROCESSUS DE FABRICATION ET DE MANIPULATION DES DOCUMENTS CARTOGRAPHIQUES UTILISES

Il s'agit, dans cette première étape, de répertorier, évaluer et combiner les sources d'erreurs liées à l'acquisition des différentes variables utilisées pour formaliser les lois sols-paysage. Deux sous chapitres seront distingués:

- le premier présentera et justifiera les principes de formalisation et de calcul de l'erreur à ce niveau;
- le second présentera la démarche permettant, pour chaque variable, de fournir une valeur d'écart type d'erreur.

1.1. Principes de calcul de l'erreur sur les variables topo-géologique et sol

Quelles que soient ces variables, dont la spécificité sera évoquée par la suite, il est possible de décomposer l'erreur les affectant suivant les différentes étapes du processus:

- erreurs commises par l'auteur de la carte au cours de la rédaction de la minute,
- erreurs dans le processus de fabrication de la carte (dessin, impression),
- erreurs au cours de la numérisation manuelle des documents
- pour les variables issues du MNT, erreurs liées aux opérations d'interpolation et de dérivation à partir de la couverture des courbes de niveau .

La connaissance de chacun des termes est très inégale et leur association dans un calcul d'erreur résultant est très rare: prenant le cas de la carte topographique, MALING (1989) évoque l'association des deux premiers en supposant leur additivité et leur indépendance. C'est également cette double hypothèse qui est retenue par DUPERET (1991) dans le cas de l'association du premier et du dernier. En conséquence, elle sera également utilisée pour calculer les erreurs sur les variables utilisées. Ainsi par exemple, si $Se_a[o_h(x_i)]$, $Se_b[o_h(x_i)]$, $Se_c[o_h(x_i)]$ et $Se_d[o_h(x_i)]$ désignent les écarts types d'erreur sur la variable $o_h(x_i)$ caractérisant les 4 étapes évoquées ci-dessus, L'utilisation de la formule de propagation d'erreur dans le cas d'une addition de termes indépendants permet de déduire l'écart type d'erreur $Se[o_h(x_i)]$ résultant:

$$Se(v) = \sum_{i=1}^{i=n} \left(\frac{\partial v}{\partial v_i} \right)^2 \times Se^2(v_i) + \sum_{i \neq j} \frac{\partial v}{\partial v_i} \times \frac{\partial v}{\partial v_j} \times Se(v_i) \times Se(v_j) \times r_{ij} \qquad [1]$$

Si le principe de décomposition de l'erreur exposé ci-dessus reste valable quelle que soit la variable utilisée, il convient par ailleurs de distinguer le cas des variables qualitatives et le cas des variables quantitatives.

1.1.1. Cas des variables qualitatives

Il s'agit de variables extraites de la couverture des sols (u(x)), de la couverture géologique (g(x)) et d'une couverture dérivée la carte topographique (rv(x)). les erreurs commises sur ces

variables se matérialisent par l'affectation erronée d'un point à une unité cartographique. Elles auront une répercussion directe sur le dénombrement des noeuds.

Cette affectation à une unité de la carte est réalisée si le point se situe à l'intérieur d'une des plages cartographiques de cette unité. Pour cela, le SIG opère un recouvrement entre la grille de points et la couverture de polygones correspondant à la carte en plage considérée.

La position du point étant supposée connue avec précision puisque imposée par l'utilisateur au travers de la définition de la grille, l'origine de l'erreur d'affectation doit être recherchée au niveau des cartes utilisées.

Depuis quelques années, une abondante littérature existe sur la nature des erreurs issues de ces cartes. La plupart des auteurs (BURROUGH, 1986; WALSH et Al, 1987; CHRISMAN, 1989) s'accordent pour distinguer les erreurs "géométriques" ("positional error" ou "cartographic errors") des erreurs "sémantiques" ("attribute errors"). Les premières sont liées à une erreur dans le tracé des limites de la carte; les secondes font référence à une affectation erronée d'une plage à une classe. Certains auteurs ajoutent à ces deux catégories les "erreurs d'exhaustivité" ou "d'omission" (CHRISMAN, 1990; JAMET, 1990). Comme leurs noms l'indiquent, ces dernières traduisent l'omission de plages cartographiques d'une unité déjà répertoriée par ailleurs sur la carte. La figure 1 illustre ces différents types d'erreur avec leurs conséquences sur la variable qualitative traduisant l'affectation du point à l'une des unités de la carte.

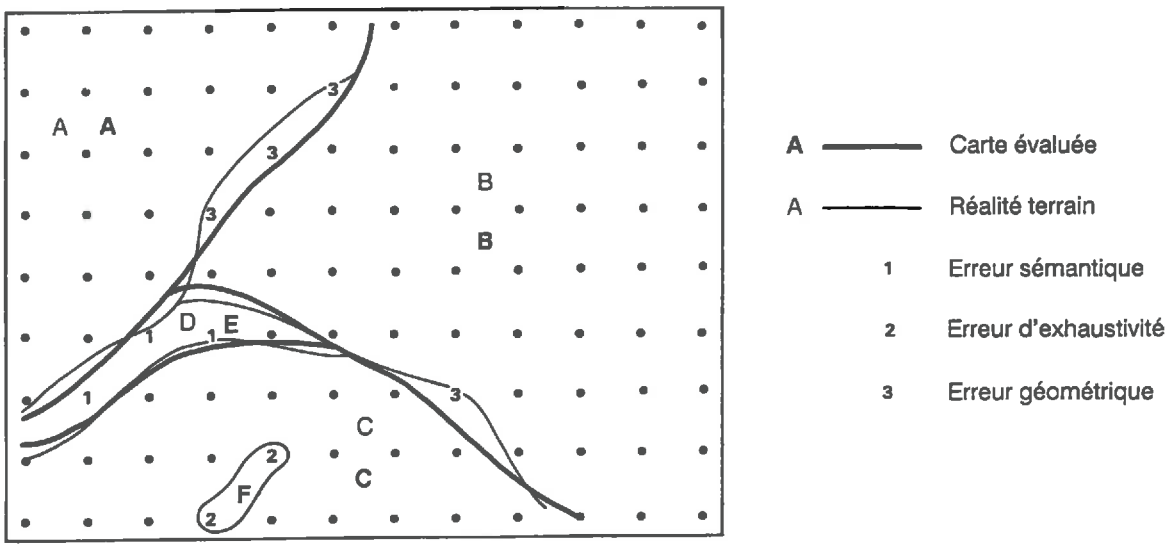


Figure 1: les différentes formes d'erreurs (d'après JAMET 1990)

Compte tenu du contexte du travail il est possible de formuler les trois hypothèses simplificatrices suivantes.

1) On suppose que le risque de mal affecter une plage cartographique à une unité de sol (erreur sémantique) est négligeable. Il n'y a en effet sur une carte de secteur de

référence qu'un nombre limité de plages ce qui autorise des vérifications systématiques

2) Les densités d'observations effectuées au cours des cartographies concernées sont suffisantes pour considérer les erreurs d'omission comme volontaires. Dans le cas (uniquement vrai pour certains terrains) où des inclusions d'une unité cartographique apparaissent au sein d'une autre sans qu'elles soient délimitées, on considèrera que l'auteur de la carte en a rendu compte dans la description du contenu de l'unité. Ce cas de figure est d'ailleurs prévu dans la structure des bases de données pédologiques (JP GAULTIER 1990) au travers du concept "d'unité complexe".

3) Les risques d'éventuelles omissions de plages cartographiques, susceptibles d'apparaître dans la suite du processus, sont quasi nuls. Comme pour la première hypothèse, cette hypothèse se justifie également par le faible nombre de polygones traités.

A la différence des autres types d'erreurs négligées, les erreurs géométriques apparaissent de façon systématique au cours de toutes les étapes de la filière cartographique. Leur importance et leur propagation au calcul de $i(X_m)$ doivent donc être prises en considération.

Même si l'ensemble des documents cartographiques sont concernés, le problème des erreurs cartographiques est particulièrement important à traiter lorsque il s'agit de cartes "naturalistes" telles que les cartes géologiques ou pédologiques. Si l'on considère, par exemple, le cas des cartes pédologiques, la littérature disponible fournit de nombreuses considérations sur le mode de tracé de leur limites (SOIL SURVEY STAFF, 1953; MAIGNIEN, 1969; VINK, 1963; JAMAGNE, 1967; SERVANT, 1972; DENT et YOUNG, 1981). Il en ressort le fait que les limites pédologiques peuvent être abruptes ou graduelles mais que, de toute façon, les moyens de les localiser (sondages, critères "de surfaces") ne sont ni assez nombreux, ni assez précis pour que la limite tracée par le pédologue puisse être considérée comme dépourvue d'incertitude.

La nature "floue" des limites de ce type de carte est également bien identifiée dans la bibliographie traitant de SIG (p.ex. BURROUGH, 1986; MAFFINI et al, 1989; CHRISMAN 1987,...). De nouvelles sources d'incertitudes sont évoquées, en particulier les erreurs dues à la digitalisation manuelle des documents (BOLSTADT et al, 1990; DUNN et al, 1989). La représentation de cette incertitude insiste sur la nature statistique des limites. Celle-ci peut être représentée par la figure 2 (MAFFINI et al, 1989). Elle suppose que soit définie une loi de distribution de la probabilité de position de la limite. La plupart des auteurs considèrent implicitement ou explicitement que cette loi prend la forme d'une courbe "en cloche" centrée sur la limite tracée. Les définitions explicites de cette courbe utilise une loi normale (MAFFINI et al, 1989) ou une loi binomiale (DUNN et al, 1989). Compte tenu de l'absence, dans le domaine étudié, de références bibliographiques sur la détermination expérimentale des paramètres de ces lois, la loi normale sera préférée car elle s'avère la plus aisée à manipuler. Il est en effet possible de définir la distribution des probabilités uniquement en fixant l'écart type Se(l) (figure 2).

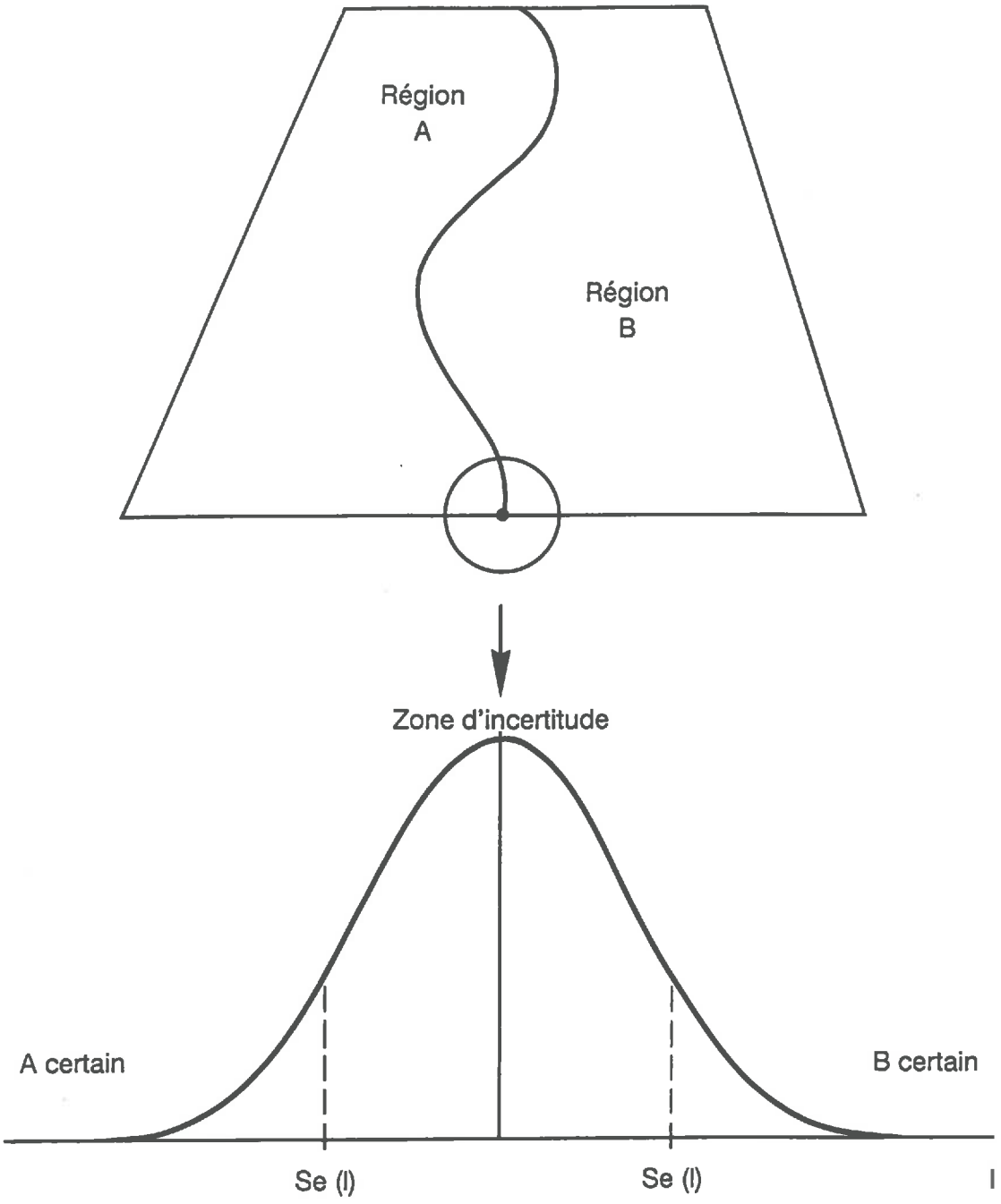


Figure 2: représentation statistique d'une limite de carte (d'après MAFFINI et al 1989)

1.1.2 Cas des variables quantitatives:

Ces variables proviennent exclusivement de couvertures dérivées de la carte topographique. La formalisation des erreurs susceptibles d'influencer les valeurs N_m et N_{mj} est beaucoup plus immédiate que précédemment. elle se traduit par la définition d'un écart type d'erreur associé à toute variable quantitative. Deux types de variables peuvent être distingués.

1) Pour DI, issue d'une mesure de distance à un cours d'eau, Se(di(x)) correspondra à l'écart type d'erreur sur la position de ce cours d'eau. Dans ce cas, l'erreur due à l'algorithme du SIG mesurant les distances sera considérée sinon nulle, tout au moins négligeable.
2) Pour les variables de relief, l'estimation de l'écart type d'erreur tiendra compte de l'erreur altimétrique de la carte et des erreurs d'interpolation entre courbes de niveau. De plus, pour les variables dérivées (pt(x),cm(x),ec(x)), un calcul de propagation d'erreur au travers de l'expression mathématique les définissant sera effectué.

1.2. Mise en oeuvre des principes retenus pour estimer, sur chaque variable, son écart type d'erreur

Chaque étape du processus sera successivement anaysée pour déterminer, par document cartographique, sa contribution à l'erreur totale. Puis, le calcul de l'écart type d'erreur total sera mis en oeuvre, pour chaque variable, suivant la formule [1] définie au précédent chapitre.

1.2.1 Erreurs liées à la réalisation de la minute de terrain:

1.2.1.1 Carte pédologique

L'erreur sur le contenu de l'unité n'étant pas considérée dans le cadre de ce travail, sera prise en compte uniquement l'erreur de tracé des limites entre unités de sol adjascentes. Cette erreur est reconnue comme systématique compte tenu de la mauvaise lisibilité des critères de différenciation de deux unités. L'objectif de l'estimation est d'évaluer l'écart type d'erreur sur la position de la limite.

Il existe peu de bibliographie sur le sujet. Seuls, deux standards de qualité quantifiés ont été trouvés. Le premier émane du soil survey staff de l'USDA (1953): "on detailed soil maps to be published at about 2 inches to the mile (1:31,680), boundaries should be accurate within at least 100 feet". Ainsi, l'échelle de travail correspondante étant de 1:20 000 (Soil survey staff 1953), la précision d'une limite devrait être au minimum de 15 mm sur la minute. Le second standard de qualité émane du GEPPA (1967): pour des études "détaillées" (entre 1 et 2 sondages / hectares) " à 35 - 50 mètres près en moyenne, elle (la précision des limites) sera suffisante". .

Les deux standarts ne s'accordent apparamment pas. Le standart du GEPPA apparait plus proche et plus réaliste pour ce qui concerne la carte de sol étudiée. La valeur de 50 mètres sera donc considérée comme l'erreur maximum autorisée.

Les exigences du calcul d'erreur nécessitent que l'erreur soit fournie sous la forme d'un écart type d'erreur. Si le choix est fait de représenter la limite sous forme d'une loi de probabilité

normale (cf figure 2), il est possible de donner une valeur à cet écart type à partir de l'erreur maximum grâce à la table de répartition de la loi normale. L'erreur maximum peut, en effet, être considérée comme l'expression d'un intervalle de confiance à un niveau α très faible. Ainsi:

$$Se\ (l) = Emax\ /\ t(\alpha) \qquad [2]$$

avec: Se(l): écart type d'erreur sur la position de la limite
Emax: erreur maximum

Suivant la valeur de α, il en résulte les valeurs suivantes de Se(l):

$\alpha = 0.01$: Se(l) = 21 m
$\alpha = 0.001$: Se(l) = 16 m
$\alpha = 0.0001$: Se(l) = 13 m

1.2.1.2. Carte géologique

Il n'existe aucune référence écrite concernant la précision des limites sur les cartes géologiques. En l'absence de ces références, seront appliqués les standarts GEPPA relatifs aux études semi-détaillées. Ils correspondent à une erreur maximum de 200m soit, en appliquant le même raisonnement que précédemment, à la gamme d'écart type d'erreur suivante:

$\alpha = 0.01$ Se(l) = 86 m
$\alpha = 0.001$ Se(l) = 65 m
$\alpha = 0.0001$ Se(l) = 54 m

1.2.1.3. Carte topographique

Deux types d'erreurs doivent être prises en compte en fonction des variables dérivées de cette carte.

L' erreur planimétrique. Elle se propage aux variables relatives à une distance (DI). Il existe des références sur ce types d'erreur exprimée en unité "papier": CAILLEMER (1955) évalue "l'erreur moyenne" d'un levé topographique à 0.2 mm. Plus récemment, MAILING (1989) propose, pour estimer l'écart type d'erreur au cours du levé, une fourchette de 0.33 à 0.54 mm.

L'erreur altimétrique. Elle est relative aux points de la courbe de niveau et se propage aux variables du MNT. CAILLEMER donne comme estimation une "erreur moyenne" de 1 mètre. MAILING cite le résultat de l'ajustement de la formule de KOPPE pour une carte topographique au 1/25 000:

$$Se_a\ (z(x)) = 1 + \tan \alpha \qquad [3]$$

avec: Se_a (z(x)): écart type d'erreur sur l'altitude attaché à l'étape de création de la carte
α : pente du terrain.

Sachant que la prise en compte de la pente dans le calcul d'erreur est impossible eut égard à la complexité des calculs qu'elle entrainerait, Se_a (z(x)) sera considérée comme constante et égal au résultat de la formule ci dessus, tan α prenant la valeur de la moyenne des tanα du secteur étudié. Ceci revient à évaluer Se_a (z(x)) à 1,2 m. Cette valeur sera retenue pour la suite des calculs.

1.2.2. erreurs de dessin, et d'édition des cartes

En l'absence de références spécifiques pour chaque carte, leurs processus de fabrication seront considérés comme identiques. La propagation d'erreur au travers de ce processus a été étudiée en détail par MAILING (1989) dans le cas d'une carte topographique. Les résultats de cette étude sont consignés dans le tableau suivant:

Opération	Ecart type d'erreur
Dessin au propre	entre 0.06 et 0.18 mm
Reproduction (typons)	entre 0.10 et 0.20 mm
Impression	entre 0.17 et 0.30 mm

***Tableau 1:** évaluation des erreurs dues à la fabrication de la carte (d'apres MALING 1989)*

1.2.3. Erreurs de digitalisation manuelle

Les erreurs de digitalisation manuelle correspondent aux erreurs commises au cours du le calage du document sur la table à digitaliser (erreurs sur les repères) et aux erreurs de suivi des limites avec le curseur. Très récemment, des références expérimentales ont été élaborées. A partir des écarts types d'erreur obtenus par BOLSTAD et al (1990) à la suite de répétitions de digitalisation manuelle, il est possible de déduire, en prenant pour chaque source d'erreur envisagée les résultats extrèmes, un intervalle de valeurs situé entre 0.05 mm et 0.08 mm. Par ailleurs, une étude empirique effectuée sur des cartes au 1/25.000 malencontreusement digitalisées deux fois (DUNN et al, 1990) conduit à la détermination d'une "erreur moyenne" (correspondant en fait à l'écart moyen absolu) située entre entre 1.6 et 3.00 m soit entre 0.064 et 0.12 mm. Appliquant une hypothèse de normalité des erreurs, ceci conduit à un écart type variant entre 0.08 et 0.15 mm .

Ainsi, compte tenu de ces deux expériences, l'intervalle de valeurs d'écarts type d'erreur retenu sera entre 0.5 mm et 0.15 mm .

1.2.4. Erreurs de conversion des couvertures pour obtenir les variables du fichier

Ce type d'erreur concerne uniquement les variables dérivées du MNT. il correspond à l'erreur affectant le processus d'interpolation des courbes de niveau permettant d'obtenir z(x). Cette erreur se propage ensuite dans les opérations arithmétiques calculant les autres variables caractérisant le relief.

1.2.4.1. Estimation des erreurs altimétriques d'interpolation du MNT

CLARKE et Al (1982) ont évalué les erreurs d'interpolations générées par différents algorithmes utilisés à cette époque. Les résultats d'écarts type d'erreur, exprimés par un pourcentage de l'intervalle entre courbes de niveau interpolées, oscillent entre 5% et 27 %. Malheureusement, l'algorithme utilisé pour produire le MNT est postérieur à cette étude. Les essais de YOELI (1986) sur son propre algorithme conduisent à une valeur de 0.52 m pour des courbes de niveau d'intervalle 10 mètres. Ceci place l'algorithme testé dans le bas de la fourchette d'erreur précédemment citée ce qui n'est pas surprenant puisque l'algorithme de YOELI va dans le sens des améliorations proposées par CLARKE et Al à l'issue de leurs études (prise en compte de plusieurs directions de l'espace).

Compte tenu de ces références, et sachant en plus que les courbes de niveau sont dessinées tous les 5m sur la carte topographique utilisée, il sera retenu la valeur arrondie de 0.30 m comme écart type d'erreur plausible sur l'altitude z(x) estimée en tout point du MNT .

1.2.4.2. Propagations des erreurs aux variables issues du MNT

4 variables sont concernées (dz(x), pt(x), cm(x), ec(x)). Elles seront envisagées séparément. Ces variables étant déduites de z(x) par des opérations arithmétiques, les calculs d'erreurs utiliseront tous la formule générale de calcul d'erreur. L'hypothèse simplificatrice d'indépendance des termes sera systématiquement utilisé en dépit de son caractère hasardeux dans ce contexte de calcul.

1.2.4.2.1. <u>variable dz(x)</u>

La variable dz(x) est déduite par la soustraction:

$dz(x) = z(x) - z_0(x)$ (cf formule [9], chapitre 4)

La valeur $z_0(x)$ correspond à la valeur en chaque point d'un plan fictif corrigeant la dérive d'altitude créée par la pente générale du lit du fleuve. Elle ne comporte pas d'erreur dans la mesure où elle est déduite d'une fonction linéaire de la latitude, celle-ci étant imposée par l'utilisateur lors de la fabrication,au sein du SIG, de la grille de points supportant le modèle. Donc l'écart type d'erreur estimé pour z(x) sera également l'écart type d'erreur sur dz(x) (1.2m).

1.2.4.2.2. <u>autres variables (pt(x), cm(x), ec(x))</u>

Ces variables découlent directement de calculs à partir de l'altitude. L'application de la formule de propagation d'erreur dans une formule arithmétique est mise en oeuvre. Les calculs détaillés font l'objet de l'annexe 5bis.

1.2.5. Combinaison des différentes sources d'erreurs

Comme exposé au paragraphe 1.1., la combinaison des différents termes d'erreurs répertoriés dans les chapitres précédents utilise les hypothèses d'additivité et d'indépendance. Ces hypothèses permettent d'appliquer la formule de propagation d'erreur définie (formule [1]).

Les variables pour lesquelles les références sont disponibles sous forme de fourchettes d'erreur, voient leurs écarts type d'erreur cumulés également fournis sous forme de fourchettes. Les références exprimées en unités "papier" sont converties en unités "terrain" selon l'échelle de représentation du document étudié. Compte tenu de l'imprécision des références collectées, tous les calculs sont arrondis au mètre près pour ce qui concerne les erreurs planimétriques et au dm près pour ce qui concerne les erreurs altimétriques.

Les résultats sont présentés dans le tableau suivant:

Code de la variable	Nom de la variable	Ecart type d'erreur sur variable (var. quantitative)	Ecart type d'erreur sur position limite (var. qualitative)
u (x)	unité de sol	-	entre 13 et 21 m
g (x)	unité géologique	-	entre 55 et 89 m
rv (x)	rive de l'Hérault	-	entre 9 et 17 m
z (x)	altitude	1.3m	-
dz (x)	dénivelé/Hérault	1.3m	-
di (x)	distance au cours d'eau	entre 9 et 17 m	-
pt (x)	pente	0.6°	-
cm (x)	courbure moyenne	1.5° par 50m	-
ec (x)	encaissement	11m	-

***tableau 2:** estimation des erreurs sur les variables utilisées*

En ce qui concerne les cartes géologiques et pédologiques, ils révèlent la contribution prépondérante de l'étape 1 dans l'erreur totale. Par contre, les étapes de fabrication et de digitalisation des documents entrent pour une part non négligeable dans l'erreur totale sur les variables issues de cartes topographiques.

2. ESTIMATION DE L'ECART TYPE D'ERREUR SUR LES DENOMBREMENTS DES POINTS DES NOEUDS DE L'ARBRE DE CLASSIFICATION ET DES POPULATIONS CORRESPONDANTES POUR CHAQUE UNITE DE SOL

La valeur de N_m est directement conditionnée par la sélection des points appartenant à X_m au moyen des variables topo-géologiques. La valeur de N_{mj} suit la même logique. La différence est qu'elle est, en plus, conditionnée par la variable sol

Estimer l'erreur sur N_m et N_{mj} revient en fait à évaluer la propagation des erreurs affectant les variables topo géologiques et pédologique au travers d'une opération de sélection utilisant divers opérateurs ("et", "ou", "<", ">" et "=") et mettant en jeu plusieurs plans d'information (les "couvertures" utilisées) fournisseurs de variables. Ceci correspond en fait à l'une des fonctionnalités principales caractérisant un SIG. A ce titre, ce type de propagation d'erreurs fait l'objet, depuis quelques années, d'une littérature importante. C'est en particulier le cas de l'opérateur "et" directement associé à la fonction "overlay" disponible dans tous les SIG modernes Cependant, dans le cas étudié, cette littérature ne peut apporter de références directement opérationnelles. En effet:

- l'objectif des modèles présentés est d'évaluer l'erreur globale sur l'ensemble de la carte et non, comme l'exige cette application, de l'évaluer spécifiquement pour des domaines identifiés de l'espace couvert par la carte;
- la propagation de l'erreur au travers de l'opérateur "et" varie considérablement dans son expression théorique (VEREGIN 1989) et dans ses résultats expérimentaux (WALSH et Al 1987) en fonction du degré de corrélation d'erreurs de chaque plan d'information; l'hypothèse simplificatrice de l'indépendance des erreurs (MAC DOUGALL 1976) est, dans ce cas, peu réaliste dans la mesure où certains plans d'information sont utilisés dans la réalisation d'un autre (cas p.ex. de la carte pédologique dont certaines unités sont manifestement limitée par une route et délimitées comme telles);
- le mélange de conditions élémentaires portant indifféremment sur des variables qualitatives et sur des variables quantitatives n'est pas étudié de façon explicite dans la littérature consultée.

Ainsi, en l'absence de modèles de propagation susceptibles de fournir, a **priori**, une estimation des écarts type d'erreur sur N_m et N_{mj}, ils seront estimés a **posteriori** sur chaque noeud produit par l'arbre et chaque sous ensemble de membres d'une unité donnée(noté X_{mj}). Pour cela, seront identifiés, au sein de ces ensembles, des sous-ensembles de points "marginaux" appelés **domaine d'incertitude** d'un noeud X_m et notés $E(X_m)$ ou $E(X_{jm})$ selon les cas. Les dénombrements de ces points marginaux seront utilisés comme un estimateur de l'écart type d'erreur:

$$Se(N_m) = card(E(X_m)$$
$$Se(N_{mj}) = card\ (E(X_{mj})) \qquad [4]$$

Les figures 3a et 3b (pages suivantes) illustrent la méthode de définition des domaines d'incertitude .

Les $E(X_m)$ correspondent à l'union logique, dans l'ensemble X_m, des domaines d'incertitude élémentaires (notés ed_m) créés par l'application de chaque dichotomie (notée d_m).

Les $E(X_{mj})$ sont, eux, définis par l'union logique, dans chaque sous ensemble X_{mj}, de $E(X_m)$ et d'un domaine d'incertitude élémentaire e_{sol} généré par la sélection sur l'unité de sol.

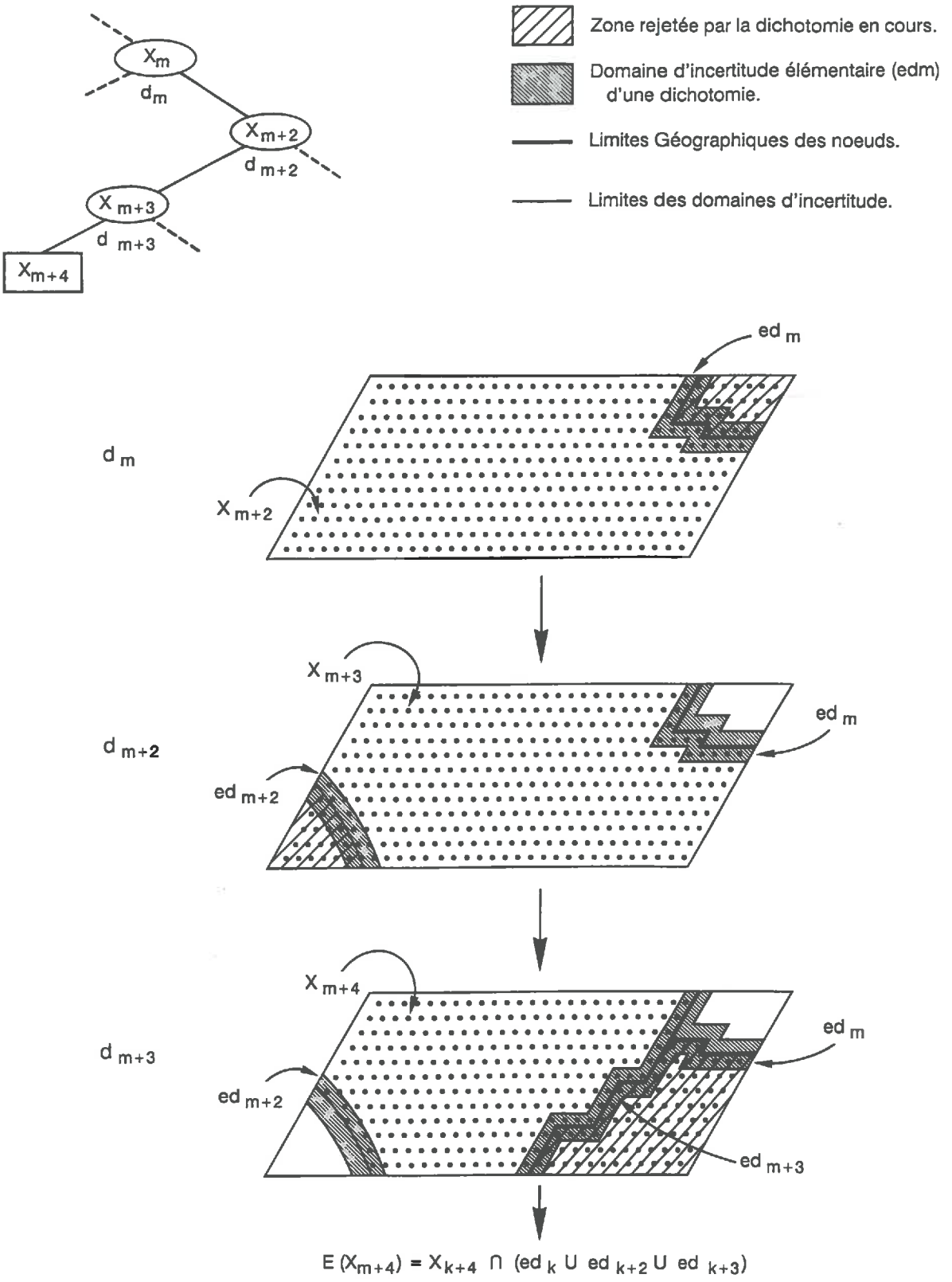


Figure 3a: *détermination du domaine d'incertitude d'un noeud* X_m

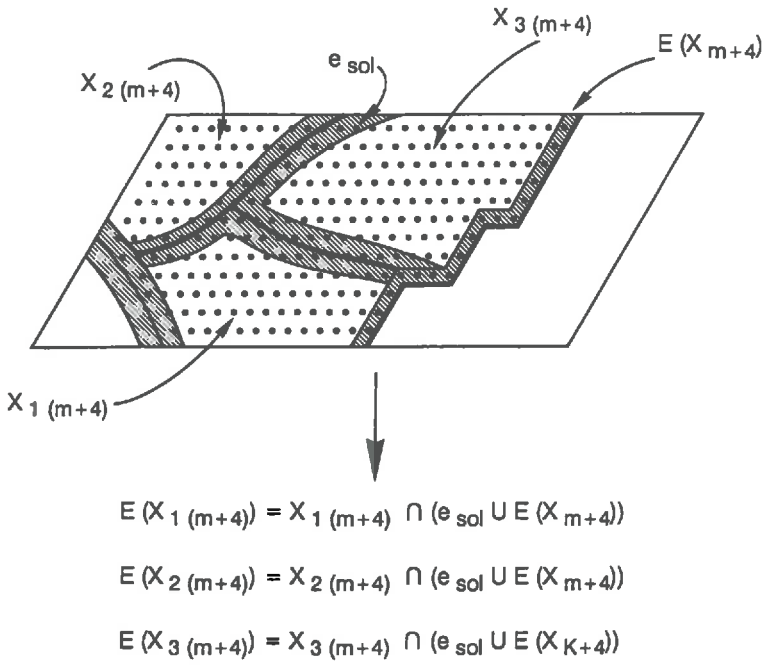


Figure 3b: détermination du domaine d'incertitude de X_{mj}

L'existence des domaines d'incertitude élémentaires est liée aux incertitudes sur les variables topo-géologiques et pédologiques étudiées dans le chapitre précédent. Leur définition et l'identification de leur membres prend une fois de plus des formes distinctes en fonction de la nature qualitative ou quantitative des variables en cause.

1) Dans le cas simple d' une condition portant sur une variable quantitative o_h, le domaine d'incertitude ed_m lié à une dichotomie se définira comme suit:

$$ed_m = \{o(x_i) / (o_h^* - Se[o_h(x_i)] < o_h(x_i) < o_h^* + Se[o_h(x_i)]\} \quad [6]$$

avec $o_h(x_i)$: variable explicative en x_i
o_h^*: valeur seuil permettant de définir les noeuds produits par la dichotomie

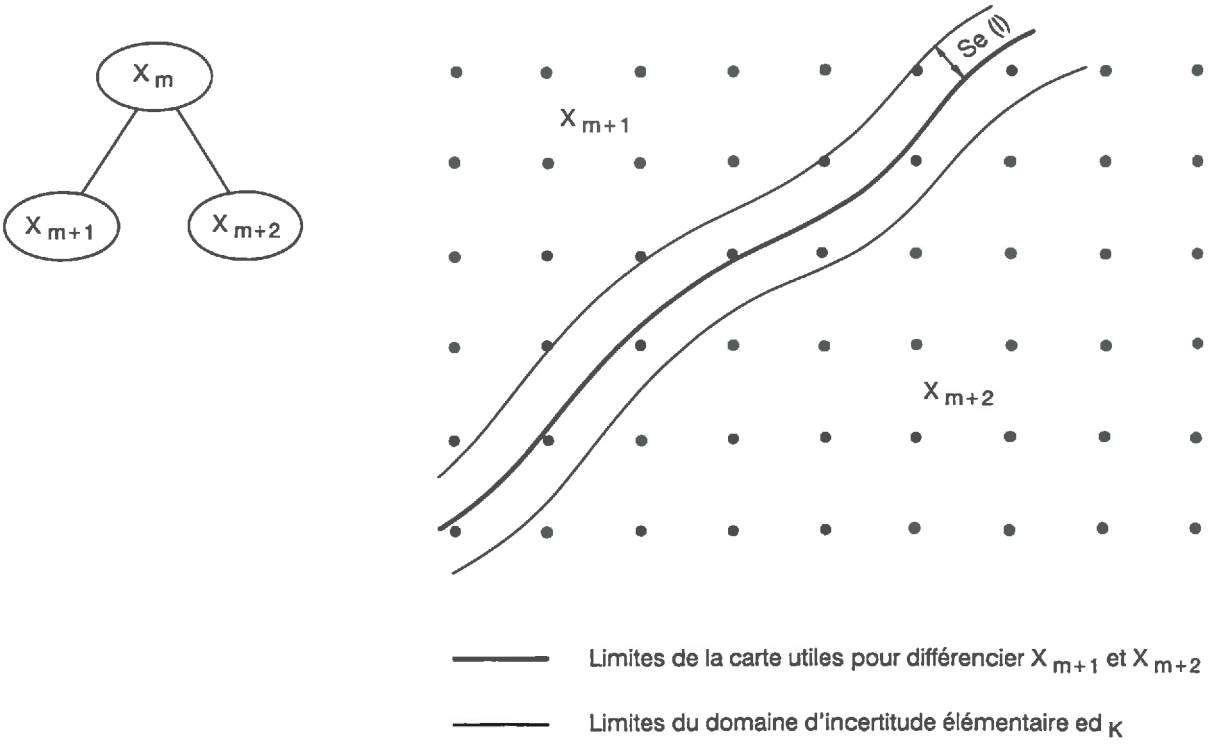


***Figure 4:** Détermination du domaine d'incertitude élémentaire d'unne dichotomie utilisant une variable qualitative*

2) Dans le cas où la condition porte sur une variable qualitative, il sera fait appel au concept de "bande epsilon" (figure 4) défini en premier par CHRISMAN (1982) (cité par CHRISMAN, 1989). Suivant ce concept, il est possible de définir géométriquement un domaine d'incertitude d'une limite de carte par une bande centrée autour de cette limite Si cette bande présente une demi-largeur égale à l'écart type d'erreur de position de la limite (Se(l)), l'aire incluse dans chaque ensemble représente une "forme d'erreur moyenne sur cette aire" ("some form of mean error in area" (CHRISMAN, 1989)). En conséquence, la population de points tombant dans cette aire sera considérée également comme un estimateur du domaine d'incertitude élémentaire correspondant. Seront donc considérés membres du domaine d'incertitude élémentaire ed_m les points situés à une distance inférieure à Se(l) de part et d'autre des limites de la carte utilisées pour définir la dichotomie.

Un programme en FORTRAN a été mis au point afin d'automatiser, pour chaque X_m ou X_{mj}, la détermination des $E(X_m)$ et $E(X_{mj})$ et leur dénombrement qui permet d'estimer l' écart type d'erreur sur N_m et N_{mj}. Il figure en annexe 6.

3. ESTIMATION DE L'ECART TYPE D'ERREUR SUR $I(X_M)$

La valeur de tout $Se(N_m)$ et $Se(N_{mj})$ pouvant être estimée par la démarche exposée ci-dessus, il est désormais possible d'aborder la troisième étape de la démarche. Elle consiste à déduire des estimations précédentes une estimation de $Se[i(X_m)]$ puisque $i(X_m)$ représente une expression arithmétique dans laquelle interviennent les termes N_m et N_{mj}. Le calcul d'erreur sur l'indice d'impureté $i(X_m)$ se déroule en 2 étapes successives:

- calcul d'erreur sur les pr_{mj} qui entrent dans l'expression de $i(X_m)$;
- calcul d'erreur sur $i(X_m)$ à partir des écarts type d'erreur sur les pr_{mj} ($Se(pr_{mj})$ obtenus dans la première étape.

Ces calculs sont fondés sur la formule générale de propagation d'erreurs dans une opération arithmétique quelconque:

Soit $v = f(v_1,\ldots,v_i,\ldots,v_n)$

$$Se\left(o_h(x_1)\right) = \sqrt{\left[Se_a^2\left(o_h(x_1)\right)\quad Se_b^2\left(o_h(x_1)\right)\quad Se_c^2\left(o_h(x_1)\right)\quad Se_d^2\left(o_h(x_1)\right)\right]} \qquad [7]$$

Avec: $Se(v)$: écart type d'erreur sur u
$Se(v_i)$: écart type d'erreur sur v_i
$Se(v_j)$: écart type d'erreur sur v_j
r_{ij} : coefficient de corrélation entre les termes v_i et v_j

3.1. Calculs d'erreur sur pr_{mj}

On sait que:

$$pr_{mj} = N_{mj}/N_m \qquad [8]$$

Il en résulte, en appliquant la formule [7]:

$$Se^2(pr_{mj}) = \left[\frac{\partial pr_{mj}}{\partial N_m}\right]^2 x Se^2(N_m) + \left[\frac{\partial pr_{mj}}{\partial N_{mj}}\right]^2 x Se^2(N_{mj}) + \frac{\partial pr_{mj}}{\partial N_m} x \frac{\partial pr_{mj}}{\partial N_{mj}} x Se(N_m) x Se(N_{mj}) x r_{ij}$$

$r(N_m,N_{mj})$, coefficient de corrélation des deux termes, sera estimé peu différent de 1 compte tenu du fait que les processus permettant d'obtenir ces deux termes sont largement confondus. Il vient donc en calculant les dérivées partielles de la formule précédente

$$Se^2(pr_{mj}) = \left[Se^2(N_m) x \frac{N_{mj}^2}{N_{mj}^4}\right] + \left[Se(N_{mj}) x \frac{1}{N_m^2}\right] + \left[\frac{-N_{mj}}{N_m^2} x \frac{1}{N_m} x Se(N_m) x Se(N_{mj})\right]$$

on pose:

$F(N_m) = Se(N_m)/N_m$: erreur relative sur N_m

$F(N_{mj}) = Se(N_{mj})/N_{mj}$: erreur relative sur N_{mj}

La formule précédente devient:

$$Se^2(pr_{mj}) = \left[\frac{N_{mj}^2}{N_m^2} \times \left(F^2(N_m)+F^2(N_{mj})\right)\right] - \left[\frac{N_{mj}}{N_m^2} \times \frac{1}{N_m} \times N_m \times F(N_m) \times N_{mj} \times F(N_{mj})\right]$$

$$Se^2(pr_{mj}) = pr_{mj}^2 \times \left(F^2(N_m) + F^2(N_{mj}) - F(N_m)\,F(N_{mj})\right)$$

Calcul d'erreur sur $i(X_m)$:

On sait que:

$$i(X_m) = 1 - \Sigma_{(j)}\, pr_{mj}^2$$

L'application de la formule [7] donne donc:

$$Se^2\left[i(X_m)\right] = \sum_{j=1}^{j=v}\left(\frac{\partial\, i(X_m)}{\partial\, pr_{mj}}\right)^2 \times Se^2(pr_{mj}) + \sum_{i\neq j}\frac{\partial\, i(X_m)}{\partial\, pr_{mj}} \times \frac{\partial\, i(X_m)}{\partial\, pr_{mi}} \times Se(pr_{mi}) \times Se(pr_{mj}) \times r_{ij}$$

Dans ce cas, il n'est pas possible de donner une estimation quantifiée du coefficient de corrélation. Par défaut, l'hypothèse simplificatrice de l'indépendance des erreurs ($r_{ij} = 0$) sera donc appliquée. La formule précédente devient, en dérivant les termes:

$$Se^2\left[i(X_m)\right] = \sum_{j=1}^{j=v} 4pr_{mj}^2 \times Se^2(pr_{mj})$$

ou,

$$Se\left[i(X_m)\right] = 2\sum_{j=1}^{j=v} pr_{mj} \times Se(pr_{mj})$$

Annexe 5bis : Calculs d'erreurs pour les variables dérivées du MNT

1. Calcul d'erreur pour la variable "pente":

On sait que (cf annexe 4):

$$pt(x) = \frac{\sqrt{b^2 + c^2}}{D}$$

avec : D = 50m (pas du MNT)

$$a = \frac{1}{6}(z_1 - z_3 - z_4 - z_5 + z_7 + z_8)$$

$$b = \frac{1}{6}(z_1 + z_2 + z_3 - z_4 - z_5 - z_6)$$

$z_1,\ldots z_8$: altitude des points autour du point sur lequel la pente est calculée (cf figure 1, annexe4)

Selon la formule générale de propagation d'erreur dans une opération arithmétique on a:

$$Se^2\left(pt(x)\right) = \sum_{i=1}^{i=8}\left[\frac{\partial\, pt(x)}{\partial\, z_i}\right]^2 x Se^2(z_i) + \sum_{i\neq j} \frac{\partial\, pt(x)}{\partial\, z_i} x \frac{\partial\, pt(x)}{\partial\, z_j} x\, Se(z_i) \times Se(z_j) \times r_{ij}$$

Pour simplifier les calculs, on fera l'hypothèse que les altitudes en chaque point ne sont pas corrélées entre elles ($r_{ij} = 0$). Le deuxième terme de l'addition ci-dessus est donc considéré comme nul. Par ailleurs on supposera que l'écart type d'erreur est égal pour tous les points considérés. Il sera noté Se(z). Dès lors, le calcul de l'écart type d'erreur devient:

$$Se^2\left(pt(x)\right) = Se^2(z) \times \left[\frac{\partial\, pt(x)}{\partial\, z_i}\right]^2$$

Se(z) étant connu, il faut déterminer les dérivées partielles de pt(x).

En dérivant pt(x) on obtient :

$$\frac{\partial\, pt(x)}{\partial\, z_i} = \frac{1}{50}\left[\frac{\partial \sqrt{b^2+c^2}}{\partial\, b} \times \frac{\partial\, b}{\partial\, z_i} + \frac{\partial \sqrt{b^2+c^2}}{\partial\, c} \times \frac{\partial\, c}{\partial\, z_i}\right]$$

$$= \frac{1}{50}\left[\frac{b}{\sqrt{b^2+c^2}} \times \frac{\partial\, b}{\partial\, z_i} + \frac{c}{\sqrt{b^2+c^2}} \times \frac{\partial\, c}{\partial\, z_i}\right]$$

En remplaçant chaque $\frac{\partial\, b}{\partial\, z_i}$ et $\frac{\partial\, c}{\partial\, z_i}$ par leurs valeurs on obtient:

$$\frac{\partial\, pt(x)}{\partial\, z_1} = \frac{1}{50}\left[\frac{1}{6} \times \frac{b+c}{\sqrt{b^2+c^2}}\right] \qquad \frac{\partial\, pt(x)}{\partial\, z_5} = \frac{1}{50}\left[\frac{1}{6} \times \frac{-(b+c)}{\sqrt{b^2+c^2}}\right]$$

$$\frac{\partial\, pt(x)}{\partial\, z_2} = \frac{1}{50}\left[\frac{1}{6} \times \frac{c}{\sqrt{b^2+c^2}}\right] \qquad \frac{\partial\, pt(x)}{\partial\, z_6} = \frac{1}{50}\left[\frac{1}{6} \times \frac{-c}{\sqrt{b^2+c^2}}\right]$$

$$\frac{\partial\, pt(x)}{\partial\, z_3} = \frac{1}{50}\left[\frac{1}{6} \times \frac{c-b}{\sqrt{b^2+c^2}}\right] \qquad \frac{\partial\, pt(x)}{\partial\, z_7} = \frac{1}{50}\left[\frac{1}{6} \times \frac{b-c}{\sqrt{b^2+c^2}}\right]$$

$$\frac{\partial\, pt(x)}{\partial\, z_4} = \frac{1}{50}\left[\frac{1}{6} \times \frac{-b}{\sqrt{b^2+c^2}}\right] \qquad \frac{\partial\, pt(x)}{\partial\, z_8} = \frac{1}{50}\left[\frac{1}{6} \times \frac{b}{\sqrt{b^2+c^2}}\right]$$

En remplaçant ces valeurs dans l'expression de $Se\left(pt(x)\right)$, on obtient:

$$Se^2\left(pt(x)\right) = Se^2(z) \times \left[\frac{1}{50^2} \times \frac{1}{36(b^2+c^2)} \times \left(2(b+c)^2+2(b-c)^2+2b^2+2c^2\right)\right]$$

$$= Se^2(z) \times \left[\frac{1}{50^2} \times \frac{1}{36(b^2+c^2)} \times \left(2(b^2+2bc+c^2)+2(b^2-2bc+c^2)+2b^2+2c^2\right)\right]$$

$$= Se^2(z) \times \left[\frac{1}{50^2} \times \frac{1}{36(b^2+c^2)} \times 6(b^2+c^2)\right]$$

Donc :

$$Se\left(pt(x)\right) = 0{,}008165\ Se(z)$$

Pour $Se(z) = 1{,}3$ m , $\boxed{Se\left(pt(x)\right) = 0.00979 \quad \text{soit} \simeq 1\% \quad \text{soit} \simeq 0{,}6°}$

2. Calcul d'erreur sur cm(x)

On sait que (cf annexe 4):

$cm(x) = d + f$

avec: $d = \frac{1}{3}(z_8+z_4) - \frac{2}{3}(z_2+z_6) + \frac{1}{3}(z_1+z_3+z_5+z_7) - \frac{2}{3}z_0$

$f = \frac{1}{3}(z_2+z_6) - \frac{2}{3}(z_8+z_4) + \frac{1}{3}(z_1+z_3+z_5+z_7) - \frac{2}{3}z_0$

En remplaçant d et f par leurs valeurs il vient:

$$cm(x) = \frac{1}{3}\left(2(z_1+z_3+z_5+z_7) - (z_2+z_4+z_6+4z_8)\right)$$

Si l'on fait les mêmes hypothèses simplificatrices que pour le calcul précédent (indépendance des z_i et égalité de tous les $Se(z_i)$), le calcul de l'écart type d'erreur sur cm(x) devient:

$$Se^2\left(cm(x)\right) = \sum_{i=1}^{i=8}\left[\frac{\partial\ cm(x)}{\partial\ z_i}\right]^2 \times Se^2(z_i)$$

Chaque dérivée partielle de cm(x) doit être calculée

$$\frac{\partial\ cm(x)}{\partial\ z_1} = \frac{\partial\ cm(x)}{\partial\ z_3} = \frac{\partial\ cm(x)}{\partial\ z_5} = \frac{\partial\ cm(x)}{\partial\ z_7} = \frac{1}{3}$$

$$\frac{\partial\ cm(x)}{\partial\ z_2} = \frac{\partial\ cm(x)}{\partial\ z_4} = \frac{\partial\ cm(x)}{\partial\ z_6} = \frac{\partial\ cm(x)}{\partial\ z_8} = -\frac{2}{3}$$

$$\frac{\partial\ cm(x)}{\partial\ z_0} = -\left(-\frac{4}{3}\right)$$

Donc

$$Se^2\left(cm(x)\right) = \frac{4 + 16 + 16}{9} \times Se^2(z) = 4\ Se^2(z)$$

Donc

$$Se\left(cm(x)\right) = 2\ Se(z)$$

Avec Se(z) = 1,3 m

$$Se\left(cm(x)\right) = 2{,}6\ \%\ /\ 50\ m$$
$$= 1{,}5°\ /\ 50\ m$$

3. Calcul d'erreur sur ec(x)

On sait que (Annexe 4):

$$cm\ (x) = \sum_{i=1}^{i=8}(z_0 - z_i)$$

Si l'on fait les mêmes hypothèses simplificatrices que pour les calculs

précédents, le calcul de l'écart type d'erreur sur ec(x) devient:

$$Se^2\left(ec(x)\right) = \sum_{i=1}^{i=8}\left[\frac{\partial\ ec(x)}{\partial\ z_i}\right]^2 \times Se^2(z_i)$$

Les dérivées partielles de ec(x) sont les suivantes:

$$\frac{\partial\ ec(x)}{\partial\ z_0} = 8$$

$$\frac{\partial\ ec(x)}{\partial\ z_i} = -1 \quad i = 1,\ldots 8$$

Donc:

$$Se^2\left(ec(x)\right) = Se^2(z) \times 71$$

soit

$$Se\left(ec(x)\right) = Se(z) \times \sqrt{71}$$

Avec Se(z) = 1,3m

$$\boxed{Se\left(ec(x)\right) = 11m}$$

Annexe 6: Programme de calcul de l'écart type d'erreur sur i(X_m)

```
c       Déclaration et initialisation des variables
        INTEGER Z(3802),R(3802),G(3802),S(3802),N(17),NE(17)
        INTEGER G1(3802),G2(3802),RT(3802),GT(3802),GT1(3802),GT2(3802)
        REAL D(3802),P(17),FN(17),FP(17),SEP(17),ES(3802),EG(3802)
        NT=0
        NET=0
        FNT=0
        SEI1=0
        SEI=0
        XP=1
        DO 1 I=1,3802
        RT(I)=1
        GT(I)=1
        GT1(I)=1
        GT2(I)=1
1       CONTINUE
        DO 2 J=1,17
        N(J)=0
        NE(J)=0
        P(J)=0
        FN(J)=0
        FP(J)=0
        SEP(J)=0
2       CONTINUE
c
c       Lecture fichiers entrées
        OPEN(1,file='erseg25.txt')
        DO 3 I=1,3802
        READ(1,10)
        Z(I),D(I),R(I),G(I),S(I),ES(I),EG(I),G1(I),G2(I)
10      FORMAT(I5,1X,F12.3,1X,I5,1X,I3,1X,I2,1X,2(F12.3,1X),I1,1X,I1)
3       CONTINUE
c
c       Définition du noeud étudié
        ZMIN=167
        ZMAX=9999
        DMIN=-9999
        DMAX=9999
        DO 4 I=1,3802
        IF (R(I).NE.2) THEN
        RT(I)=0
        ENDIF
        IF ((G(I).EQ.2).OR.(G(I).GE.5)) THEN
        GT(I)=0
        ENDIF
        IF ((G1(I).EQ.2).OR.(G1(I).GE.5)) THEN
        GT1(I)=0
        ENDIF
        IF ((G2(I).EQ.2).OR.(G2(I).GE.5)) THEN
        GT2(I)=0
        ENDIF
4       CONTINUE
c
c       Comptages de NT ,N(J),NE(J),NET
```

```
      DO 5 I=1,3802
      IF ((Z(I).GT.ZMIN).AND.(Z(I).LE.ZMAX)) THEN
        IF ((D(I).GT.DMIN).AND.(D(I).LE.DMAX)) THEN
      IF (RT(I).EQ.1) THEN
       IF (GT(I).EQ.1) THEN
          NT = NT+1
          N(S(I))=N(S(I))+1
          IF ((EG(I).LE.86).AND.(GT1(I).NE.GT2(I))) THEN
              'NE(S(I))=NE(S(I))+1
            NET = NET+1
         ELSE IF (D(I).LE.17) THEN
              NE(S(I))=NE(S(I))+1
                NET = NET+1
      ELSE IF ((D(I).LE.(DMIN+17)).OR.(D(I).GE.(DMAX-17))) THEN
            NE(S(I))=NE(S(I))+1
            NET = NET+1
      ELSE IF ((Z(I).LE.(ZMIN+13)).OR.(Z(I).GE.(ZMAX-13))) THEN
            NE(S(I))=NE(S(I))+1
            NET = NET+1
      ELSE IF (ES(I).LE.21) THEN
            NE(S(I))=NE(S(I))+1
          ENDIF
        ENDIF
       ENDIF
      ENDIF
       ENDIF
5      CONTINUE
       PRINT *,NT,NET
c
c       Calcul de P(J),FN(J),FNT,SEP(J),FP(J),IP
     FNT = REAL(NET)/REAL(NT)
       DO 6 J=1,17
       P(J) = REAL(N(J))/REAL(NT)
       IF (N(J).NE.0) THEN
        FN(J)= REAL(NE(J))/REAL(N(J))
      FP(J)= SQRT(FNT**2+FN(J)**2-(FNT*FN(J)))
        SEP(J)= P(J)*FP(J)
        XP = XP - P(J)**2
       ENDIF
6     CONTINUE
       DO 15 J=1,17
       PRINT *,J,P(J),FP(J)
15     CONTINUE
c
c       Calcul de SEI
       DO 7 J=1,17
     SEI1 = SEI1 + ((P(J)**2)*(SEP(J)**2))
7      CONTINUE
     SEI = 2*SQRT(SEI1)
c
c       expression des résultats
     PRINT *,'Incertitude sur Nt = ',FNT,' (',NET,')'
       PRINT *,'indice impureté (XP) = ',XP
       PRINT *,'incertitude sur XP = ',SEI
       OPEN (3,file='a:ND65.rs1')
       WRITE (3,12) NT,XP,SEI,(J,P(J),FP(J),J=1,17)
12     FORMAT (I4,1X,F5.3,1X,F6.4,1X,/,(I2,1X,2(F6.4,1X)))
       END
```

Annexe 7: Correspondance entre les unités du secteur de référence et celles des unités au 1/100.000 et 1/250.000.

Unités 1/250.000	Unités 1/100.000	Unités secteur de référence
552 T	60 a 64, 65	17 3, 5, 8, 11, 12 13, 15, 16
309 A	77 79	7 6
309 B	78	6,9
309 V	75	10,14
173 A	82 a 82 b	1, 4 2

Annexe 8: le programme d'extraction des règles de voisinage à partir de la carte du secteur de référence (stratégie "grid survey")

```
c     Déclaration des variables indexées
      INTEGER L(3802),M(3802),NTH(53),NTB(53),NTI(53),NPE(53)
      INTEGER NP(53),NT(53),NH(53,17),NB(53,17),NI(53,17)
        REAL X(3802),Y(3802),DT(53),DM(53),RP(53),GINH(53),GINB(53)
      REAL GINI(53),PH(53,17),PB(53,17),PI(53,17)
c     initialisation des variables
        NU=1
      NBU=0
        ICM=0
c       lecture des données d'entrée
        OPEN (1,file='acmvois.txt')
        OPEN (2,file='acmpot.txt')
        DO 1 I=1,3802
        READ(1,100) X(I),Y(I),L(I),M(I)
100     FORMAT(2(F12.3,1X),I2,1X,I5)
1       CONTINUE
        DO 2 IC=1,53
        READ (2,101) NPE(IC)
101     FORMAT(I3)
2       CONTINUE
c       décompte nombre d'individus dans unité L
        DO 3 I=1,3802
        IF (L(I).EQ.NU) THEN
        NBU = NBU+1
        ENDIF
3       CONTINUE
c
c
        DO 4 IC=1,53
        IF(RP(IC-1).GT.10) THEN
c       initialisation des variables renseignées par comptages
        DT(IC)=0
        NTH(IC)=0
        NTB(IC)=0
        NTI(IC)=0
        GINH(IC)=1
        GINB(IC)=1
        GINI(IC)=1
      DO 5 I=1,17
        NH(IC,I)=0
        NB(IC,I)=0
        NI(IC,I)=0
5       CONTINUE
c
       PRINT * ,IC
c       comptage des points dans les couronnes
          DO 6 I=1,3802
          IF(L(I).EQ.NU) THEN
            DO 7 J=1,3802
            D=SQRT((X(J)-X(I))**2+(Y(J)-Y(I))**2)
            DZ=M(J)-M(I)
            IF(D.GT.(IC-1)*50.AND.D.LE.IC*50) THEN
```

```
            DT(IC)=DT(IC)+D
           IF(DZ.GT.17) THEN
              NTH(IC)=NTH(IC)+1
              NH(IC,L(J))=NH(IC,L(J))+1
           ENDIF
           IF(DZ.LT.-17) THEN
              NTB(IC)=NTB(IC)+1
              NB(IC,L(J))=NB(IC,L(J))+1
           ENDIF
           IF(DZ.GE.-17.AND.DZ.LE.17) THEN
              NTI(IC)=NTI(IC)+1
              NI(IC,L(J))=NI(IC,L(J))+1
           ENDIF
         ENDIF
7        CONTINUE
       ENDIF
6      CONTINUE
c      Calcul des proportions d'unités dans couronnes
       DO 8 I=1,17
         IF(NTH(IC).NE.0) THEN
           PH(IC,I)=100*(REAL(NH(IC,I))/REAL(NTH(IC)))
         ELSE
           PH(IC,I)=-9.99
         ENDIF
         IF(NTB(IC).NE.0) THEN
           PB(IC,I)=100*(REAL(NB(IC,I))/REAL(NTB(IC)))
         ELSE
           PB(IC,I)=-9.99
         ENDIF
         IF(NTI(IC).NE.0) THEN
           PI(IC,I)=100*(REAL(NI(IC,I))/REAL(NTI(IC)))
         ELSE
           PI(IC,I)=-9.99
         ENDIF
8        CONTINUE
c        Calcul des indices d'impuretés
         DO 9 I=1,17
       IF(NTH(IC).NE.0) THEN
          GINH(IC)=GINH(IC)-(PH(IC,I)**2)
        ELSE
        GINH(IC)=9.999
       ENDIF
       IF(NTB(IC).NE.0) THEN
          GINB(IC)=GINB(IC)-(PB(IC,I)**2)
        ELSE
        GINB(IC)=9.999
       ENDIF
       IF(NTI(IC).NE.0) THEN
          GINI(IC)=GINI(IC)-(PI(IC,I)**2)
        ELSE
        GINI(IC)=9.999
       ENDIF
9      CONTINUE
c        Evaluation représentativité comptages
         NT(IC)=NTH(IC)+NTB(IC)+NTI(IC)
         NP(IC)=NBU*NPE(IC)
         RP(IC)=REAL(NT(IC))/REAL(NP(IC))
c
```

```
          DM(IC)=DT(IC)/NT(IC)
          ICM=ICM+1
        ENDIF
4       CONTINUE
c
c     Sorties résultats
        OPEN(3,file='unt1.sur')
        OPEN(4,file='unt1.sou')
        OPEN(5,file='unt1.idm')
        OPEN(6,file='unt1.syn')
        DO 10 IC=1,ICM
        WRITE(3,103) (PH(IC,I),I=1,17)
        WRITE(4,103) (PB(IC,I),I=1,17)
        WRITE(5,103) (PI(IC,I),I=1,17)
103     FORMAT(17(1X,F6.2))
        WRITE(6,104) IC,DM(IC),RP(IC),GINH(IC),GINB(IC),GINI(IC)
104     FORMAT(I2,2X,F7.2,2X,F6.2,2X,3(F4.3,2X))
10      CONTINUE
        STOP
        END
→
```

ANNEXE 9: Le programme de prédiction des unités de sol avec des règles de voisinage (stratégie "grid survey")

```
c
c     Déclaration initialisation des variables
      INTEGER NP(4000),NS(4000),NU(4000),ND(4000)
      INTEGER NUP(4000),NBS(4000),L(4000),L0(4000)
      REAL P1(17,30,17),P2(17,30,17),P3(17,30,17),P(4000,17)
      REAL
PMX(4000),ER(4000),X(4000),Y(4000),X0(4000),Y0(4000),W(4000)
      DO 1 I=1,3916
      PMX(I)=0
      ER(I)=0
      NUP(I)=0
      NBS(I)=0
      W(I)=0
      DO 2 JP=1,17
      P(I,JP)=0
2     CONTINUE
1     CONTINUE
c
c        lecture des fichiers d'entrées
      OPEN(1,file='unt.sur')
      OPEN(2,file='unt.sou')
      OPEN(3,file='unt.idm')
      OPEN(4,file='srpix.txf')
      OPEN(5,file='srson.txf')
      READ (1,101) (((P1(K,IC,J),J=1,17),IC=1,30),K=1,17)
      READ (2,101) (((P2(K,IC,J),J=1,17),IC=1,30),K=1,17)
      READ (3,101) (((P3(K,IC,J),J=1,17),IC=1,30),K=1,17)
101   FORMAT (17(1X,F6.2))
         DO 10 I=1,3916
      READ (4,102) NP(I),X(I),Y(I),L(I)
10    CONTINUE
102   FORMAT (I5,1X,2(F12.3,1X),I5)
      DO 11 IS=1,3916
      READ (5,103) NS(IS),X0(IS),Y0(IS),L0(IS),NU(IS),ND(IS)
11    CONTINUE
103   FORMAT (I5,1X,2(F12.3,1X),I5,1X,I2,1X,I2)
c
c     sélection de la règle à appliquer
      DO 3 IS=1,3916
      IF (NU(IS).NE.0.AND.NU(IS).LE.17) THEN
       IF (ND(IS).GE.2) THEN
        PRINT*,'sondage n°',IS
        DO 4 I=1,3916
           IF (NUP(I).EQ.0) THEN
            IF (NP(I).EQ.NS(IS)) THEN
             NUP(I)=NU(IS)
         ELSE
          D=SQRT((X(I)-X0(IS))**2+(Y(I)-Y0(IS))**2)
           IF (MOD(D,50).EQ.0) THEN
            ICO = D/50
              ELSE
            ICO= 1+AINT(D/50)
```

```
                ENDIF
              WO = 1
c             WO = REAL(1/REAL(ICO**2))
c
c             calculs des probabilités par pixels
                IF (ICO.LE.30) THEN
               IF ((L(I)-L0(IS)).GT.17) THEN
                DO 5 JP=1,17
                IF (P1(NU(IS),ICO,JP).GE.0) THEN
                P(I,JP)=((W(I)*P(I,JP))+(P1(NU(IS),ICO,JP)*WO))
     c               /(W(I)+WO)
                ENDIF
5               CONTINUE
               ELSE IF (L(I)-L0(IS).LT.-17) THEN
                DO 6 JP=1,17
                IF (P2(NU(IS),ICO,JP).GE.0) THEN
                P(I,JP)=((W(I)*P(I,JP))+(P2(NU(IS),ICO,JP)*WO))
     c               /(W(I)+WO)
                ENDIF
6               CONTINUE
               ELSE
                DO 7 JP=1,17
                IF (P3(NU(IS),ICO,JP).GE.0) THEN
                P(I,JP)=((W(I)*P(I,JP))+(P3(NU(IS),ICO,JP)*WO))
     c               /(W(I)+WO)
               ENDIF
7               CONTINUE
                 ENDIF
              NBS(I)=NBS(I)+1
              W(I)=W(I)+WO
                ENDIF
               ENDIF
         ENDIF
4        CONTINUE
        ENDIF
          ENDIF
3     CONTINUE
c
c     Expression des résultats
      DO 8 I=1,3916
      IF (NUP(I).EQ.0) THEN
          NUE=0
       DO 9 JP=1,17
          NUE=NUE+1
       IF (P(I,JP).GT.PMX(I)) THEN
           NUP(I)=NUE
        PMX(I)=P(I,JP)
          ENDIF
9         CONTINUE
       ER(I)=100-PMX(I)
      ELSE
       ER(I)=1
       NBS(I)=0
      ENDIF
8     CONTINUE
c
c     Sorties des résultats
      OPEN (6,file='a:sr2000.prb')
```

```
      OPEN (7,file='a:sr2000.res')
      OPEN (8,file='a:sr2000.txt')
      DO 13 I=1,3916
      WRITE (6,104) NP(I),(P(I,JP),JP=1,17)
      WRITE (8,106) NP(I),NBS(I),NUP(I),ER(I)
      WRITE (7,105) NP(I),NUP(I),ER(I)
13    CONTINUE
104      FORMAT (I5,1X,17(F6.2,1X))
105   FORMAT (I5,',',I2,',',F6.2)
106   FORMAT (I5,',',I5,',',I2,',',F6.2)
      STOP
      END
→
```

Annexe 10: Le programme de prédiction des unités de sol avec les règles de voisinage (stratégie "free survey"

```
c
c     Ce programme réalise une cartographie des sols en appliquant
c       les règles de voisinage apprise sur un SR (fichiers unt.*)
c       les fichiers *pix contiennent les coordonnées (x,y,z) et les n°
c       des points .les fichiers *son contiennent, en plus des précédentes,
c     les n° de l'unité d'après la carte de validation.
c     les sondages sont effectués selon des densité maxi autorisées
c     croissantes, à chaque densité ,seuls les pixels dont les erreur prévues
c       (er(i)) excèdent 50%, sont effectivement sondés.
c
c
c       Déclaration initialisation des variables
        INTEGER NP(180),NS(180),NU(180),ND(180)
        INTEGER NUP(180),NBS(180),L(180),L0(180),NTC(180)
        REAL P1(17,30,17),P2(17,30,17),P3(17,30,17),P(180,17)
        REAL PMX(180),ER(180),X(180),Y(180),X0(180),Y0(180),W(180)
        DO 1 I=1,179
        PMX(I)=0
        ER(I)=0
        NUP(I)=0
        NBS(I)=0
        W(I)=0
        NTC(I)=0
        DO 2 JP=1,17
        P(I,JP)=0
2       CONTINUE
1       CONTINUE
c
c     lecture des fichiers d'entrées
        OPEN(1,file='unt.sur')
        OPEN(2,file='unt.sou')
        OPEN(3,file='unt.idm')
        OPEN(4,file='rbpix.txf')
        OPEN(5,file='rbson.txf')
        READ (1,101) (((P1(K,IC,J),J=1,17),IC=1,30),K=1,17)
        READ (2,101) (((P2(K,IC,J),J=1,17),IC=1,30),K=1,17)
        READ (3,101) (((P3(K,IC,J),J=1,17),IC=1,30),K=1,17)
101     FORMAT (17(1X,F6.2))
      DO 3 I=1,179
        READ (4,102) NP(I),X(I),Y(I),L(I)
3       CONTINUE
        DO 55 IS=1,179
        READ (5,103) NS(IS),X0(IS),Y0(IS),NU(IS),ND(IS),L0(IS)
55      CONTINUE
102     FORMAT (I5,1X,2(F12.3,1X),I5)
103     FORMAT (I5,1X,2(F12.3,1X),I2,1X,I3,1X,I3)
c
c
c       Première densité traitée (1 sondage tous les 400 m)
c
        NSD=0
        ERSD=999
```

```
      DO 51 WHILE (ERSD.GT.50)
c
c     Selection des sondages à effectuer et mise à jour pixels sondés
      ERSD=0
      DO 4 IS=1,179
      IF ((ND(IS).EQ.86.OR.ND(IS).EQ.8).AND.((ER(IS).GT.50
     c  .AND.NTC(IS).EQ.0).OR.ERSD.EQ.0)) THEN
        NSD=NSD+1
        NTC(IS)=1
c      PRINT*,'sondage n°',IS
       DO 5 I=1,179
      IF (NUP(I).EQ.0) THEN
      IF (NTC(I).EQ.1.AND.NP(I).EQ.NS(IS)) THEN
       NUP(I)=NU(IS)
        NUR=0
       DO 6 JP=1,17
        NUR=NUR+1
        IF (NUP(I).EQ.NUR) THEN
         P(I,NUR)= 100
       ELSE
        P(I,NUR)= 0
       ENDIF
6       CONTINUE
c
c      selection de la règle à appliquer et calcul pondération (W)
       ELSE IF (NU(IS).NE.18) THEN
        D=SQRT((X(I)-X0(IS))**2+(Y(I)-Y0(IS))**2)
        IF (MOD(D,50).EQ.0) THEN
         ICO = D/50
       ELSE
         ICO= 1+AINT(D/50)
       ENDIF
c        WO = 1
        WO = REAL(1/REAL(ICO**3))
c
c     calculs des probabilités par pixels et par unités (P(I,JP))
       IF (ICO.LE.30) THEN
          IF ((L(I)-L0(IS)).GT.17) THEN
           DO 7 JP=1,17
           IF (P1(NU(IS),ICO,JP).GE.0) THEN
           P(I,JP)=((W(I)*P(I,JP))+(P1(NU(IS),ICO,JP)*WO))
   c       /(W(I)+WO)
           ENDIF
7          CONTINUE
          ELSE IF (L(I)-L0(IS).LT.-17) THEN
           DO 8 JP=1,17
           IF (P2(NU(IS),ICO,JP).GE.0) THEN
           P(I,JP)=((W(I)*P(I,JP))+(P2(NU(IS),ICO,JP)*WO))
   c       /(W(I)+WO)
           ENDIF
8          CONTINUE
          ELSE
           DO 9 JP=1,17
           IF (P3(NU(IS),ICO,JP).GE.0) THEN
           P(I,JP)=((W(I)*P(I,JP))+(P3(NU(IS),ICO,JP)*WO))
   c       /(W(I)+WO)
          ENDIF
9          CONTINUE
```

```
          ENDIF
            NBS(I)=NBS(I)+1
            W(I)=W(I)+WO
          ENDIF
        ENDIF
      ENDIF
5       CONTINUE
        ENDIF
4       CONTINUE
c
c       Expression des résultats (unité prévue (NUP(I),erreur prévue (ER(I))
        DO 10 I=1,179
        IF (NUP(I).EQ.0) THEN
       NUE=0
        DO 11 JP=1,17
       NUE=NUE+1
        IF (P(I,JP).GT.PMX(I)) THEN
       NUP(I)=NUE
         PMX(I)=P(I,JP)
       ENDIF
11      CONTINUE
        ER(I)=100-PMX(I)
        IF (ER(I).GT.ERSD.AND.(ND(I).EQ.86.OR.ND(I).EQ.8)) THEN
          ERSD = ER(I)
        ENDIF
       ELSE
        ER(I)=1
        NBS(I)=0
       ENDIF
10      CONTINUE
        PRINT*,ERSD
51      CONTINUE
c       Sorties des résultats
c
c       OPEN (7,file='b:rb8sb0.prb')
        OPEN (8,file='b:rb8sb0.res')
c       OPEN (9,file='b:rb8sb0.txt')
        DO 12 I=1,179
c       WRITE (7,104) NP(I),(P(I,JP),JP=1,17)
c       WRITE (9,106) NP(I),NBS(I),NUP(I),ER(I)
        WRITE (8,105) NP(I),NUP(I),ER(I)
12      CONTINUE
104    FORMAT (I5,1X,17(F6.2,1X))
105     FORMAT (I5,',',I2,',',F6.2)
106     FORMAT (I5,',',I5,',',I2,',',F6.2)
        PRINT*, NSD,' sondages réalisés pour la densité 8'
c
c
c
c     Densité suivante (1 sondage tous les 200m)
c
        ERSD=999
        DO 52 WHILE (ERSD.GT.50)
c       Selection sondage à effectuer et mise à jour pixels sondés
c
        DO 13 I=1,179
        IF (NTC(I).EQ.0) THEN
        PMX(I)=0
```

```
      NUP(I)=0
      ENDIF
13    CONTINUE
      ERSD=0
      DO 14 IS=1,179
       IF ((ND(IS).EQ.86.OR.ND(IS).EQ.8.OR.ND(IS).EQ.4)
     c  .AND.ER(IS).GT.50.AND.NTC(IS).EQ.0) THEN
         NSD=NSD+1
         NTC(IS)=1
c       PRINT*,'sondage n°',IS
        DO 15 I=1,179
      IF (NUP(I).EQ.0) THEN
       IF (NTC(I).EQ.1.AND.NP(I).EQ.NS(IS)) THEN
        NUP(I)=NU(IS)
         NUR=0
        DO 16 JP=1,17
         NUR=NUR+1
         IF (NUP(I).EQ.NUR) THEN
          P(I,NUR)= 100
        ELSE
         P(I,NUR)= 0
        ENDIF
16       CONTINUE
c
c     selection de la règle à appliquer et calcul pondération (W)
        ELSE IF (NU(IS).NE.18) THEN
         D=SQRT((X(I)-X0(IS))**2+(Y(I)-Y0(IS))**2)
          IF (MOD(D,50).EQ.0) THEN
           ICO = D/50
         ELSE
           ICO= 1+AINT(D/50)
         ENDIF
c         WO = 1
          WO = REAL(1/REAL(ICO**3))
c
c     calculs des probabilités par pixels
        IF (ICO.LE.30) THEN
           IF ((L(I)-L0(IS)).GT.17) THEN
            DO 17 JP=1,17
            IF (P1(NU(IS),ICO,JP).GE.0) THEN
            P(I,JP)=((W(I)*P(I,JP))+(P1(NU(IS),ICO,JP)*WO))
   c        /(W(I)+WO)
            ENDIF
17          CONTINUE
            ELSE IF ((L(I)-L0(IS)).LT.-17) THEN
            DO 18 JP=1,17
            IF (P2(NU(IS),ICO,JP).GE.0) THEN
            P(I,JP)=((W(I)*P(I,JP))+(P2(NU(IS),ICO,JP)*WO))
   c        /(W(I)+WO)
            ENDIF
18          CONTINUE
           ELSE
            DO 19 JP=1,17
            IF (P3(NU(IS),ICO,JP).GE.0) THEN
            P(I,JP)=((W(I)*P(I,JP))+(P3(NU(IS),ICO,JP)*WO))
   c        /(W(I)+WO)
           ENDIF
19          CONTINUE
```

```
          ENDIF
            NBS(I)=NBS(I)+1
            W(I)=W(I)+WO
          ENDIF
         ENDIF
          ENDIF
15        CONTINUE
       ENDIF
14      CONTINUE
c
c       Expression des résultats
        DO 20 I=1,179
        IF (NUP(I).EQ.0) THEN
       NUE=0
         DO 21 JP=1,17
       NUE=NUE+1
         IF (P(I,JP).GT.PMX(I)) THEN
        NUP(I)=NUE
          PMX(I)=P(I,JP)
       ENDIF
21      CONTINUE
        ER(I)=100-PMX(I)
        IF (ER(I).GT.ERSD.AND.(ND(I).EQ.86.OR.ND(I).EQ.8
   c    .OR.ND(I).EQ.4)) THEN
          ERSD = ER(I)
        ENDIF
       ELSE
        ER(I)=1
        NBS(I)=0
       ENDIF
20      CONTINUE
        PRINT*,ERSD
52      CONTINUE
c
c       Sorties des résultats
c       OPEN (10,file='b:rb4sb0.prb')
        OPEN (11,file='b:rb4sb0.res')
c       OPEN (12,file='b:rb4sb0.txt')
        DO 22 I=1,179
c       WRITE (10,104) NP(I),(P(I,JP),JP=1,17)
c       WRITE (12,106) NP(I),NBS(I),NUP(I),ER(I)
        WRITE (11,105) NP(I),NUP(I),ER(I)
22      CONTINUE
        PRINT*, NSD,' sondages réalisés pour la densité 4'
c
c
c
c     Densité suivante (1 sondage tous les 100m)
c
        ERSD=999
        DO 53 WHILE (ERSD.GT.50)
c
c       selection des sondages à effectuer et mise à jour pixel sondés
        DO 23 I=1,179
        IF (NTC(I).EQ.0) THEN
        PMX(I)=0
        NUP(I)=0
        ENDIF
```

```
23      CONTINUE
        ERSD=0
        DO 24 IS=1,179
         IF ((ND(IS).EQ.86.OR.ND(IS).EQ.8.OR.ND(IS).EQ.4
     c    .OR.ND(IS).EQ.2.OR.ND(IS).EQ.62).AND.
     c    ER(IS).GT.50.AND.NTC(IS).EQ.0) THEN
           NSD=NSD+1
           NTC(IS)=1
c         PRINT*,'sondage n°',IS
          DO 25 I=1,179
        IF (NUP(I).EQ.0) THEN
         IF (NTC(I).EQ.1.AND.NP(I).EQ.NS(IS)) THEN
          NUP(I)=NU(IS)
            NUR=0
          DO 26 JP=1,17
            NUR=NUR+1
            IF (NUP(I).EQ.NUR) THEN
              P(I,NUR)= 100
          ELSE
           P(I,NUR)= 0
          ENDIF
26          CONTINUE
c
c       choix de la règle à appliquer et calcul pondération (W)
          ELSE IF(NU(IS).NE.18) THEN
           D=SQRT((X(I)-X0(IS))**2+(Y(I)-Y0(IS))**2)
            IF (MOD(D,50).EQ.0) THEN
             ICO = D/50
          ELSE
             ICO= 1+AINT(D/50)
          ENDIF
c           WO = 1
            WO = REAL(1/REAL(ICO**3))
c
c       calculs des probabilités par pixels
          IF (ICO.LE.30) THEN
             IF ((L(I)-L0(IS)).GT.17) THEN
              DO 27 JP=1,17
              IF (P1(NU(IS),ICO,JP).GE.0) THEN
              P(I,JP)=((W(I)*P(I,JP))+(P1(NU(IS),ICO,JP)*WO))
    c         /(W(I)+WO)
              ENDIF
27            CONTINUE
             ELSE IF (L(I)-L0(IS).LT.-17) THEN
              DO 28 JP=1,17
              IF (P2(NU(IS),ICO,JP).GE.0) THEN
              P(I,JP)=((W(I)*P(I,JP))+(P2(NU(IS),ICO,JP)*WO))
    c         /(W(I)+WO)
              ENDIF
28            CONTINUE
             ELSE
              DO 29 JP=1,17
              IF (P3(NU(IS),ICO,JP).GE.0) THEN
              P(I,JP)=((W(I)*P(I,JP))+(P3(NU(IS),ICO,JP)*WO))
    c         /(W(I)+WO)
             ENDIF
29            CONTINUE
          ENDIF
```

```
            NBS(I)=NBS(I)+1
            W(I)=W(I)+WO
          ENDIF
         ENDIF
          ENDIF
25        CONTINUE
       ENDIF
24      CONTINUE
c
c       Expression des résultats
        DO 30 I=1,179
        IF (NUP(I).EQ.0) THEN
       NUE=0
         DO 31 JP=1,17
       NUE=NUE+1
         IF (P(I,JP).GT.PMX(I)) THEN
        NUP(I)=NUE
          PMX(I)=P(I,JP)
       ENDIF
31      CONTINUE
        ER(I)=100-PMX(I)
        IF (ER(I).GT.ERSD.AND.(ND(I).EQ.86.OR.ND(I).EQ.8
    c   .OR.ND(I).EQ.4.OR.ND(I).EQ.2.OR.ND(I).EQ.62)) THEN
          ERSD = ER(I)
        ENDIF
       ELSE
        ER(I)=1
        NBS(I)=0
       ENDIF
30      CONTINUE
        PRINT*,ERSD
53      CONTINUE
c
c       Sorties des résultats
c       OPEN (13,file='b:rb2sb0.prb')
        OPEN (14,file='b:rb2sb0.res')
c       OPEN (15,file='b:rb2sb0.txt')
        DO 32 I=1,179
c       WRITE (13,104) NP(I),(P(I,JP),JP=1,17)
c       WRITE (15,106) NP(I),NBS(I),NUP(I),ER(I)
        WRITE (14,105) NP(I),NUP(I),ER(I)
32      CONTINUE
        PRINT*, NSD,' sondages réalisés pour la densité 2'
c
c
c     Densité suivante (densité 1)
c
        ERSD=999
        DO 54 WHILE (ERSD.GT.50)
c
c       selection pixels à effectuer et mise à jour pixels sondés
        DO 33 I=1,179
        IF (NTC(I).EQ.0) THEN
        PMX(I)=0
        NUP(I)=0
        ENDIF
33      CONTINUE
        ERSD=0
```

```
      DO 34 IS=1,179
     IF (ER(IS).GT.50.AND.NTC(IS).EQ.0) THEN
        NSD=NSD+1
        NTC(IS)=1
c      PRINT*,'sondage n°',IS
       DO 35 I=1,179
     IF (NUP(I).EQ.0) THEN
      IF (NTC(I).EQ.1.AND.NP(I).EQ.NS(IS)) THEN
       NUP(I)=NU(IS)
         NUR=0
       DO 36 JP=1,17
         NUR=NUR+1
         IF (NUP(I).EQ.NUR) THEN
          P(I,NUR)= 100
       ELSE
        P(I,NUR)= 0
       ENDIF
36       CONTINUE
c
c     choix de la règle à appliquer et calcul pondération (W)
        ELSE IF (NU(IS).NE.18) THEN
         D=SQRT((X(I)-X0(IS))**2+(Y(I)-Y0(IS))**2)
          IF (MOD(D,50).EQ.0) THEN
           ICO = D/50
        ELSE
           ICO= 1+AINT(D/50)
        ENDIF
c         WO = 1
          WO = REAL(1/REAL(ICO**3))
c
c     calculs des probabilités par pixels
        IF (ICO.LE.30) THEN
           IF ((L(I)-L0(IS)).GT.17) THEN
            DO 37 JP=1,17
            IF (P1(NU(IS),ICO,JP).GE.0) THEN
            P(I,JP)=((W(I)*P(I,JP))+(P1(NU(IS),ICO,JP)*WO))
  c         /(W(I)+WO)
            ENDIF
37          CONTINUE
           ELSE IF (L(I)-L0(IS).LT.-17) THEN
            DO 38 JP=1,17
            IF (P2(NU(IS),ICO,JP).GE.0) THEN
            P(I,JP)=((W(I)*P(I,JP))+(P2(NU(IS),ICO,JP)*WO))
  c         /(W(I)+WO)
            ENDIF
38          CONTINUE
           ELSE
            DO 39 JP=1,17
            IF (P3(NU(IS),ICO,JP).GE.0) THEN
            P(I,JP)=((W(I)*P(I,JP))+(P3(NU(IS),ICO,JP)*WO))
  c         /(W(I)+WO)
           ENDIF
39          CONTINUE
        ENDIF
          NBS(I)=NBS(I)+1
          W(I)=W(I)+WO
       ENDIF
      ENDIF
```

```
          ENDIF
35        CONTINUE
        ENDIF
34      CONTINUE
c
c       Expression des résultats
        DO 40 I=1,179
        IF (NUP(I).EQ.0) THEN
       NUE=0
        DO 41 JP=1,17
       NUE=NUE+1
        IF (P(I,JP).GT.PMX(I)) THEN
       NUP(I)=NUE
        PMX(I)=P(I,JP)
       ENDIF
41      CONTINUE
        ER(I)=100-PMX(I)
        IF (ER(I).GT.ERSD) THEN
          ERSD = ER(I)
        ENDIF
       ELSE
        ER(I)=1
        NBS(I)=0
       ENDIF
40      CONTINUE
        PRINT*,ERSD
54      CONTINUE
c       Sorties des résultats
c       OPEN (16,file='b:rb1sb0.prb')
        OPEN (17,file='b:rb1sb0.res')
c       OPEN (18,file='b:rb1sb0.txt')
        DO 42 I=1,179
c       WRITE (16,104) NP(I),(P(I,JP),JP=1,17)
c       WRITE (18,106) NP(I),NBS(I),NUP(I),ER(I)
        WRITE (17,105) NP(I),NUP(I),ER(I)
42      CONTINUE
        PRINT*, NSD,' sondages réalisés pour la densité 1'
        STOP
        END
→
```