MODELISATION STATISTIQUE DE LA STRATIFICATIOND'UN ESPACE REGIONAL POUR L'ESTIMATION DE LA SUPERFICIE D'UN THEME AU SOL

Haja ANDRIANASOLO

I PRESENTATION

Vouloir connaître les réalités d'une région, même restreintes aux seules superficies de thèmes quelconques (cultures, fôrets,...) a toujours posé des problèmes tels que le recours aux démarches de l'estimation statistique est nécessaire. Démarches dont l'utilisation demande la définition, de population ou univers, d'individus élémentaires ou unités de sondage, de base de sondage, d'échantillon, de stratification, de grappe, etc... Or dans les pays en voie de développement il est très rare que l'incomplétude des éléments disponibles, permette effectivement la constitution de ces divers éléments. Si en pays riches il est envisageable de mettre en oeuvre de façon rigoureuse de telles méthodes, c'est parce que sont disponibles des listes d'exploitations agricoles dans le cas d'enquêtes d'exploitants, et qu'existent des cartes à grandes échelles ou des moyens de prises de photographies aériennes dans le cas de sondages aréolaires.

Ces capacités rendent possibles le travail sur une représentation la moins "fausse" possible de la réalité. En effet par exemple dans le cas d'enquêtes d'exploitants agricoles, le modèle "mental" que l'on a est que toute culture où qu'elle soit et quoiqu'elle soit, dans la région est le fait d'un exploitant. Et qu'en échantillonnant des exploitations on échantillonne les cultures.

Pour l'affinement des précisions et des extrapolations, viennent ensuite les notions de découpage de l'univers en strates. Malheureusement dans le contexte des pays sous développés, le concept d'exploitant agricole est assez difficile à utiliser, puisque la majeure partie de la population cultive toujours quelque part, ou fait cultiver, sans parler des imbrications familiales, sociales et économiques. La base de sondage serait la population entière, dont on n'a aucune connaissance exhaustive, loin s'en faut. De même pour les méthodes aréolaires l'inexistence de cartes, cadastres à jour ou photographies aériennes est un facteur très limitatif sinon rédhibitoire. Il manque donc l'essentiel: une représentation de la réalité.

Or les données images satellitaires, de par leurs nature et caractéristiques, constituent aujourd'hui un modèle de choix représentant les réalités d'une région. En effet elles couvrent exhaustivement de grandes superficies, sont prises périodiquement, et rendent compte des phénomènes présents, en subdivisant l'aire couverte en éléments élémentaires: les pixels, qui suivants l'état de ceux-ci prennent une valeur déterminée (la valeur prise par un pixel est caractérisée par une intégrale des réflectances de l'ensemble des phénomènes couverts). Ainsi suivant les longueurs d'onde, les résolutions spatiales et spectrales, peut-on avoir une certaine représentation des états et types de phénomènes, de leurs forme, géométrie, extension, et localisation. Et même par induction sur la base de modèles adéquats, et d'études de l'agencement des objets représentés est-il possible d'obtenir des éléments de compréhension et de définition du mode d'organisation de l'espace par les hommes: capacité à appréhender la relation entre système naturel et système social.

Profitant de ce raccourci technologique, les pays en voie de développement, ont alors la possibilité d'avoir une idée de leurs réalités, du moment que des modèles basés sur les images satellitaires sont produits.

Notre propos est de fournir une modélisation possible de la stratification pour l'estimation de surface à partir de ce type de représentation.

II. IDENTIFICATION DU MODELE GÉNÉRAL.

Ainsi que vu ci-dessous, les modèles d'estimations basés sur les notions telles que celles d'exploitants agricoles, rencontrent difficilement les spécificités des pays pauvres. Aussi nous orienterons-nous vers celui de l'estimation aréolaire, dont les concepts sont par ailleurs simples: partition de la superficie totale à étudier en N surfaces élémentaires, tirage aléatoire d'un échantillon de m d'entre eux, enquêtes des éléments de cet échantillon, et estimation pour la population des N éléments par expansion directe en multipliant le résultat de l'échantillon par N/m.

II.1 DESCRIPTION DU MODELE.

Soient (voir figures a et b) :

- une aire S
- une culture quelconque i

L'objectif est d'estimer l'extension de i dans l'aire S.

Définition: on appelle **segment** une portion de territoire, aux limites bien définies (E. Houseman, 1979).

Dans les méthodes d'échantillonnage d'aires, la surface totale S est subdivisée en segments. Un segment constitue l'unité élémentaire d'échantillonnage.

Soient:

- N le nombre total de segments de l'univers (N=S/s, où s: aire d'un segment)
- E un échantillon de m segments (m<N)
- j un segment quelconque

Par enquête on obtient $x_{i,j}$ la surface de i dans j faisant partie de E

Ce qui permet l'obtention de la superficie moyenne de i par segment:

$$\overrightarrow{x_i} = \frac{1}{m} \sum_{j=i}^{m} x_{i,j}$$

Un estimateur de l'extension totale de i dans S est par expansion directe:

$$\hat{x}_i = N * \overrightarrow{x_i}$$

Dont la variance est:

$$\text{var}(\overrightarrow{x_i}) = \frac{N-m}{N-1} * \frac{1}{m(m-1)} \sum_{j=i}^{m} (x_{i,j} - \overrightarrow{x_i})^2$$

D'autre part étant démontré que tout découpage de l'univers en strates est source de gain, dans le cas de ce modèle il suffit de considérer chaque strate séparément et de sommer les résultats sur toutes les strates.

Pour une stratification en H strates, on aurait si h représente une strate:

$$\overrightarrow{x_{i,h}} = \sum_{j=i}^{mh} \frac{x_{i,h,j,}}{m_h}$$

où :

m_h nombre de segments de h

$x_{i,h,j}$ extension de i dans le segment j de h

$\overrightarrow{x_{i,h}}$ extension moyenne de i par segment de h

L'estimateur est : $\hat{x}_i = \sum_{h=1}^{H} N_h * \overrightarrow{x_{i,h}}$

où :

$\overrightarrow{x_{i,h}}$ extension moyenne de i par segment dans h

N_h nombre totalde segment de h

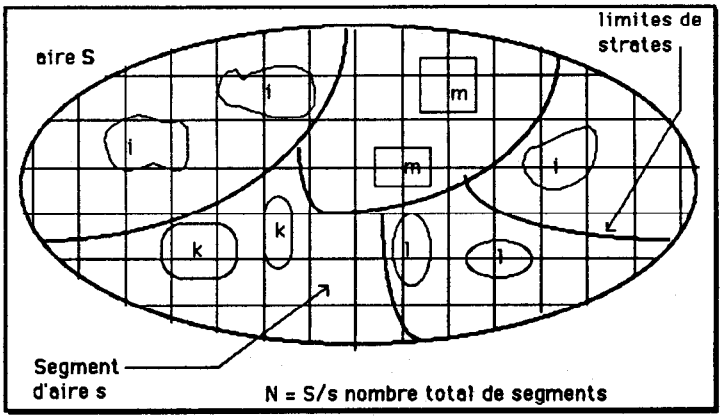


Figure a-

Identification du modèle d'estimation statistique (sondage aréolaire)

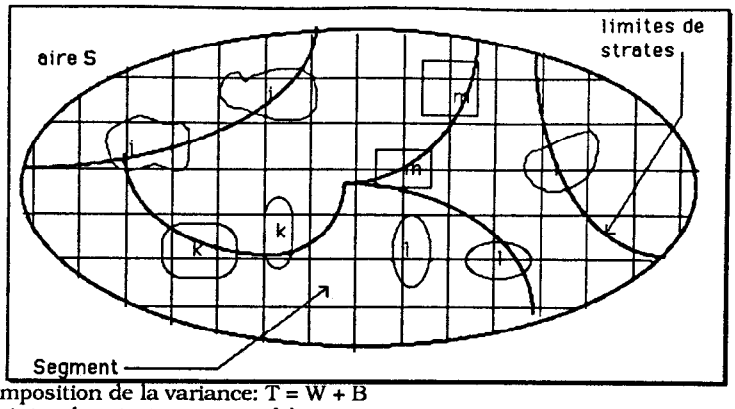


Figure b-Cas d'une mauvaise stratification

Dans la relation de décomposition de la variance: T = W + B
W est grand (la variance intra des strates est grande)
B est petit (la variance inter des strates est petite)
Chaque phénomène ciblé se retrouve dans plusieurs types de strate.
Une strate est un magma (mélange) de phénomènes.
Les phénomènes sont à cheval sur plusieurs strates.

Cependant malgré son apparente simplicité, la mise en oeuvre pratique et efficace de ce modèle d'estimation exige que certaines conditions soient remplies. En particulier que la variation entre segments soit la plus petite possible (que chaque segment possède tous à peu prés les mêmes proportions des phénomènes existant), et qu'ils soient tous à peu prés de la même taille. La contrainte de la minimisation des variations inter-segments, exige formellement une partition en strates efficace de l'aire étudiée. En effet plus les strates sont "pures" du point de vue des phénomènes ciblés, plus les variations entre les segments dans chaque type de strate diminuent. A la limite dans le cas ou une strate correspond à un phénomène (culture), il suffirait de ne faire les enquêtes que dans cette strate, pour estimer ce phénomène dans l'univers.

Un point crucial est donc l'élaboration d'un modèle apte à générer des strates tel que leur variance interne soit minimisée, sur le critère mesuré sur les segments.

III. IDENTIFICATION DU MODELE DE STRATIFICATION.

III.1 RÉALITÉS: PROTOTYPE ET ENVIRONNEMENT.

Ainsi que précédemment vu, les données image satellitaires étant une traduction synthétique des réalités d'une aire géographique quelconque, ont un énorme potentiel d'information. Très rapidement ont été rappelés que peuvent y être approchés les phénomènes présents par leur état qua-

litatif, leur forme, leur agencement, leur organisation etc...Elles constituent donc une réalité complexe. Aussi selon l'approche en modélisation nous restreindrons nous à certains aspects de cette réalité, qui nous seraient utiles.

III.1.a Prototype.

L'objectif est la capacité à l'obtention de partitions de l'espace tel que la variance des phénomènes ciblés y soit la plus petite possible. Le problème est donc la découverte (dé-couvrir) des limites de ces parties de l'espace, sous la contrainte qu'à la limite ce phénomène, ne se retrouve que dans l'une d'entre elles. La découverte de ces limites constitue le sujet de la modélisation.

Etant donnée la richesse des données satellitaires, la recherche de ces limites pourrait être conduite du point de vue des différents états qualitatifs présents, des différentes formes présentes, des agencements et organisations de celles-ci, ou de tout autre élément participant de cette réalité. Le *prototype* que nous tenterons d'approcher est la *stratification par les états qualitatifs*. Pourquoi?

On sait que les données satellitaires sont un modèle de représentation du monde, dans lequel les objets prennent une valeur précise suivant leur réflectance, dans les différentes bandes de longueur d'onde de prélèvement de l'information. Il y a relation directe entre réflectance et objets sources de ces réfléctance. Ces objets étant eux-mêmes définis par une infinité d'attributs possibles, mais dont l'un des plus importants au niveau de la télédétection est justement l'état qualitatif (la qualité). En particulier au niveau de cultures (phénomènes végétatif), la nature de l'objet intervient (arbre, fleurs,...), mais plus encore l'état de cette nature est primordial. La télédétection fait une différentiation entre, par exemple, arbre vert (un état) et arbre dont les feuilles jaunissent (un autre état). Si donc on se dote d'un modèle faisant ressortir les limites de ces différents états qualitatifs, l'objectif serait atteint. En effet supposons que nous ayons une culture cible, dans un état particulier, le fait d'arriver à découvrir une/des strates la circoncrivant tout spécialement, nous rapproche sensiblement de l'objectif, puisque sa dispersion dans les strates tend alors à décroître. Et ici un autre attribut des données image satellitaires prend toute son importance: la date de la prise de vue. Car s'il y a des dates auxquelles la majeure partie des thèmes présents sont dans des états qualitatifs semblables, il y en a d'autres où elles ne le sont pas. Mais cet aspect dépasse le cadre de cette modélisation (abordé par l'auteur dans sa thèse, voir bibliographie). De ce qui précède il ressort que le modèle s'appuiera essentiellement sur l'aspect spectral des données satellitaires.

III.2 IDENTIFICATION.

III.2.a Le modèle.

Suivant notre approche, le modèle doit fournir des classes-strates tel que la variance intra-strate soit la plus petite possible, ce qui tendrait à regrouper les individus pixels selon leur état qualitatif représenté par la valeur radiométrique. La minimisation de cette variance intra constitue alors le critère de classification.

Sachant que lorsqu'une population est divisée en classes, sa variance totale T se décompose en la somme T = W + B de:

- la variance à l'intérieur des classes W,

- la variance entre les classes B,

le critère veut que la classification recherchée soit telle que W soit minimisée (respectivement que B soit maximisée, puisque T est une donnée constante).

Or de tels modèles de classification statistique existent, et sont reconnus sous le nom de "classification par réallocations et optimisation (du critère ci-dessous)". A titre d'exemple on peut mentionner ceux de Forgy (1965), Jancey (1966), MacQueen (1967), etc...

III.2.b Les variables.

L'identification du modèle en soi n'est pas nécessairement suffisant. Suivant les variables décrivant les individus pixels, les partitions se modifieront. Cela veut dire qu'au niveau de l'optimisation du critère, il faut inclure une identification de celles-ci. Identification qui à priori semble devoir être empirique.

En effet au départ on ne dispose que des valeurs des pixels, dans les différentes bandes de prélèvement des capteurs ("canaux"). Ainsi les données "MSS Landsat" disposent de quatre bandes, "Spot" de trois, "Thematic Mapper" de sept, etc... Théoriquement on pourrait donc s'en contenter. Cependant, de façon tout aussi empirique, il peut être fait usage d'autres variables communément admises en télédétection: les indices, et les axes factoriels.

Les indices sont des rapports de combinaisons linéaires de canaux, et sont répertoriés. Ainsi peuvent être mentionnés pour les données "MSS Landsat":

- l'indice de végétation verte $\frac{K7-K5}{K7+K5}$.

- l'indice de brillance $\frac{K6-K4}{K6+K4}$.

- l'indice des sols $\frac{K4-K5}{K4+K5}$.

- etc...(Ki représentant le canal i).

Quant aux axes factoriels, ils sont classiquement le résultat de la transformation des variables, en combinaisons linéaires successives maximisant la variance résiduelle, tout en exigeant l'indépendance de ces combinaisons. Les facteurs sont tels qu'ils expliquent la totalité de la variance des individus.

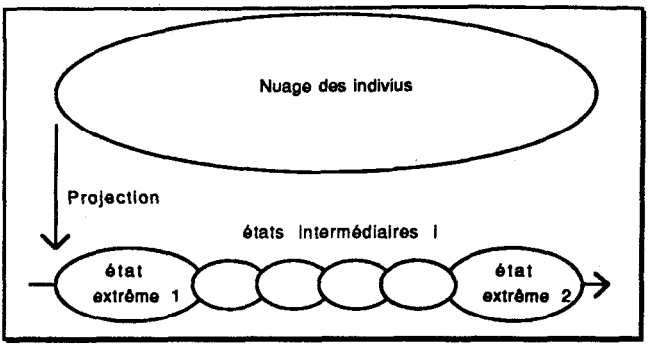
Dans la phase d'estimation, ces variables seront testées suivant un plan d'expérience, qui permettra la détermination du meilleur ensemble. Par itération, ultérieurement c'est cet ensemble qui devrait être utilisé.

Cependant suivant la problématique et la logique de l'approche du modèle adopté, une identification conceptuelle peut conduire à un ensemble de variables supplémentaires.

La variable répondant le plus aux critères du modèle devrait être telle qu'à travers elle, les différences entre les divers états présents dans les données, soient exacerbées. Que les variances entre ces classes d'état y soient maximisées.

Sachant que ces classes existent structurellement dans les données, et sachant qu'elles traduisent les réalités régionales (classes de sols nus, sols inondés, sols partiellement recouverts de végétation,, complètement recouverts, tel que la végétation puisse être de très chlorophylienne (verte), à peu (scénéscente), etc...), le fait de projeter et de maximiser la dispersion entre ces états sur une telle variable, équivaudrait à les ordonner d'une certaine manière sur celle-ci. Ordonnancement qui devrait s'organiser suivant une échelle organisant les données entre les deux structures les plus fortes qu'elles contiennent (voir figure 1.)

Figure 1



Cette variable doit donc rendre compte de façon optimale des différences entre états. Supposant que les classes sont connues, elle doit avoir une variance maximale entre les classes, et minimale en leur intérieur. Le rapport (variance inter/variance intra) doit être maximum. Ce modèle est celui d'un axe d'une fonction discriminante sur des classes d'une partition.

Mais sur quelles classes calculer cet axe?.

Au sens des structures existant dans les données, il y en a deux qui caractérisent les oppositions les plus fortes. Travaillant sur des représentations qui sont des états sur le terrain, les deux oppositions les plus fortes pourraient, par exemple être entre "eau profonde" et "sols nus en plein soleil". Entre ces deux états extrêmes, viennent s'organiser les autres états "intermédiaires". L'objectif est donc de construire l'axe de la fonction discriminante sur l'ensemble des individus pixels organisés en deux classes. Mais deux classes dont les caractéristiques doivent être telles que les critères présidant à leur formation s'inscrivent dans la logique du modèle, et de l'approche. Car l'efficacité d'un axe discriminant dépend directement de l'éloignement des classes (variance inter (B) grande) et de leur compacité (variance intra (W) petite). L'organisation en deux classes nécessaire est donc tel que, dans la relation d'analyse de la variance:

$T = W + B$

- W soit minimisée

- B soit maximisée

C'est là le critère du modèle général de stratification (classification) que nous avons retenu précédemment. Nous l'utiliserons donc aussi dans la phase relative aux variables.

(Le propos étant une stratification, et non une reconnaissance de cultures, l'application présentée dans la partie "estimation" montre que le calcul sur une partition en deux classes satisfait les contraintes du modèle).

IV. ESTIMATION.

Les données sont sur une région du nord-ouest de Madagascar, qui constitue une partie d'image de 1000 pixels sur 920. Les données satellitaires sont des images "Landsat MSS", de l'année 1981 du mois d'avril, voir images 1 à 13 (pour comparaison une image du mois d'août a aussi été utilisée. Les calculs illustratifs sont sur le mois d'avril).

Se situant dans le cadre de l'estimation de superficies de cultures, dont celle concernée est ici le riz.

Les estimations rigoureuses des paramètres pour le calcul des variables, doivent théoriquement être menées suivant les règles usuelles de l'échantillonnage. A savoir que pour des estimations de la moyenne de population, pour des niveaux de confiance donnés, il est possible d'obtenir la taille de l'échantillon nécessaire. Ainsi peut-on rappeler que si N est une taille d'échantillon:

$$N = \frac{r^2 \cdot var'(x)}{\frac{I^2}{4}}$$ où :

- r: valeur associée au niveau de précision désirée. (r = 1.96 pour un niveau de confiance de 95%)

- var'(x): estimation de la variance sur x

- I: amplitude de l'intervalle de l'estimation de la moyenne, autour de celle de la population.

Cependant dans le cas présent, les dimensions de la partie d'image utilisée ont permis un calcul direct sur la totalité de celle-ci. L'univers est donc restreint à cette partie d'image

Les indices suivants ont été calculés directement (Ki désignant le canal i)

$\frac{K7-K5}{K7+K5}$ "végétation verte", $\frac{K6-K4}{K6+K4}$ "végétation jaune", $\frac{K4-K5}{K4+K5}$ "sols"

Images de la région test
d'après les canaux originaux

Image 1 - Canal 4

Image 2 - Canal5

Image 3 - Canal6

Image 4 Canal7

Images de région test
d'après les axes factoriels

Image 5 - Axe 1

Image 6 - Axe 2

Image 7 - Axe 3

Image 8 - Axe 4

Images de la région test
d'après les indices et l'axe unique modélisé

Image 9 - "IVG"

(végétation verte)

Image 10 - "IVJ"

(végétation jaune)

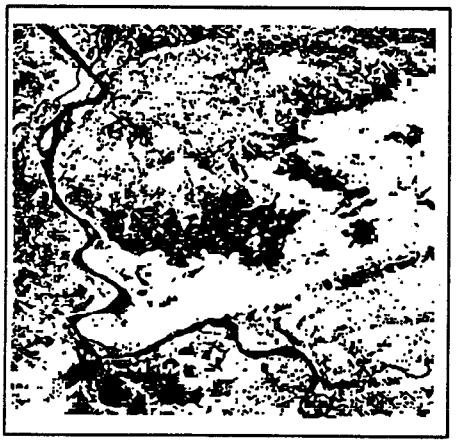


Image 11 - "SOL"

Image 12 - Axe "UNIQUE

Les coefficients des combinaisons pour les axes factoriels:

axe 1: 0.566 * K4 + 0.586 * K5 - 0.306 * K6 - 0.493 * K7

axe 2: 0.418 * K4 + 0.361 * K5 + 0.673 * K6 + 0.227 * K7

axe 3: 0.699 * K4 - 0.624 * K5 - 0.266 * K6 + 0.227 * K7

axe 4: 0.125 * K4 - 0.371 * K5 + 0.619 * K6 - 0.681 * K7

Calcul de l'axe fonction discriminante:

a) caractéristiques des deux classes les plus fortes:

- classe 1:

canal 4: moyenne 37.0, variance 9.1

canal 5: moyenne 41.1, variance 31.3

canal 6: moyenne 45.4, variance 34.7

canal 7: moyenne 31.5, variance 62.2

- classe 2:

canal 4: moyenne 36.3, variance 10.0

canal 5: moyenne 38.7, variance 47.0

canal 6: moyenne 68.5, variance 105.4

canal 7: moyenne 66.7, variance 196.1

b) coefficients des canaux pour l'axe de la fonction discriminante:

axe: -0.012092 * K5 + 0.079628 *K7

Remarquons que dans tous les cas, les variables utilisées ont été centrées réduites.

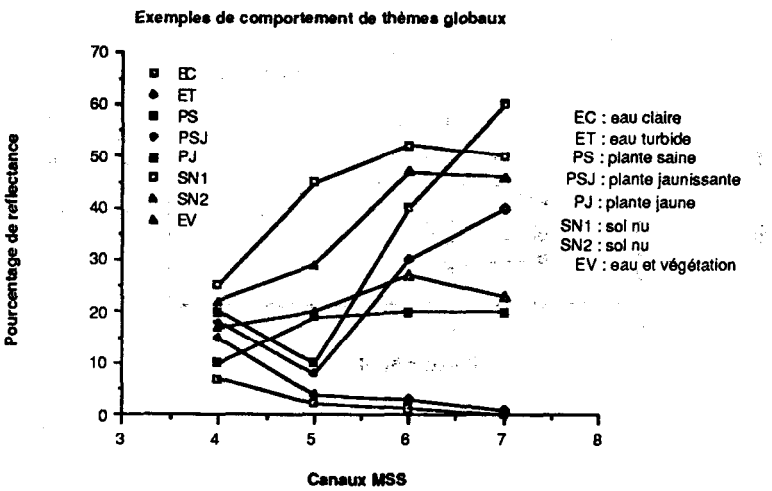
IV.1 PARTITION DE L'IMAGE.

En règle générale, pour une variable de contrôle donnée, le gain d'une stratification est sensible avec six à huit classes, et même si apparemment l'augmentation du nombre de strates pourraient apporter plus, le gain n'est plus aussi déterminant par rapport aux autres éléments de la mise en oeuvre: coûts, complexité, etc... Les données satellitaires du fait de leur

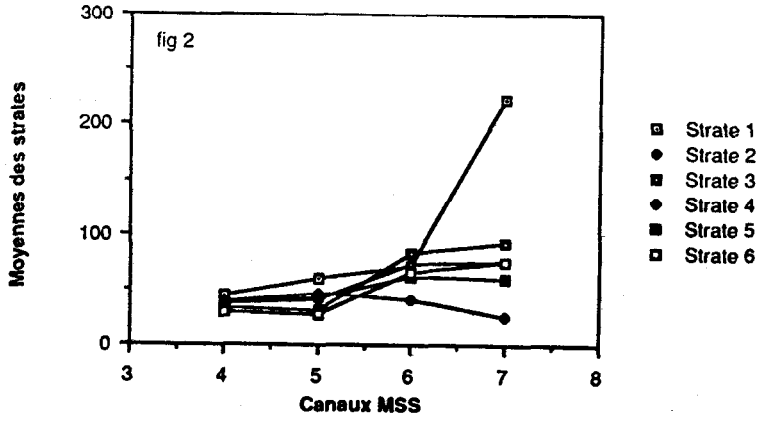
aspect synthétique, pourraient facilement permettre de générer un grand nombre de strates. Ainsi que l'on vient de le mentionner, ceux sont les éléments d'ordre pratique (compromis coûts et précision) de mise en oeuvre qui sera déterminant dans la limitation du nombre de strates (cependant dans le cadre de l'évolution de la présente modélisation, les problèmes de détermination du nombre de classes d'une partitions seront ultérieurement abordés). Dans la cas actuel, une partition en six classes est utilisée.

IV.1.a Description des partitions.

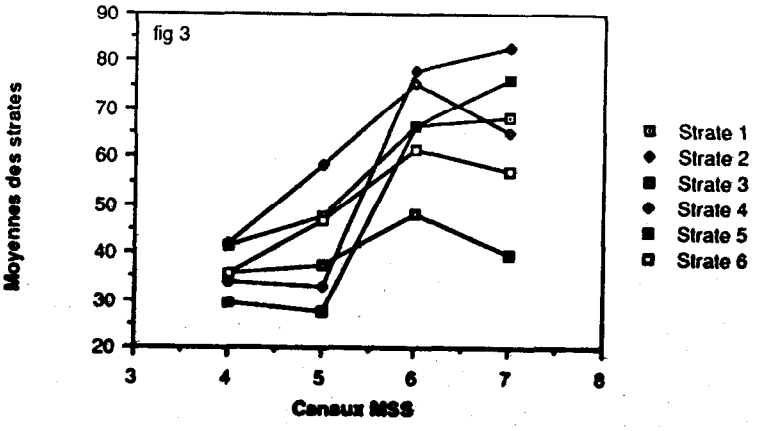
Bien qu'une analyse précise et détaillée des classes obtenues puisse être faites d'un point de vue strictement télédétection, nous n'en effectuerons les descriptions qu'en rapportant les comportements des classes obtenues sur les canaux originaux. Des exemples de comportements globaux de thèmes sont donnés dans la figure suivante, au moyen desquels il est possible de situer thématiquement les résultats des partitions. Ainsi peut-on constater que les classes générées n'ont pas du tout de comportements identiques, voir figures 2 à 7 toutes relatives au mois d'avril, et les images 14 à 19.

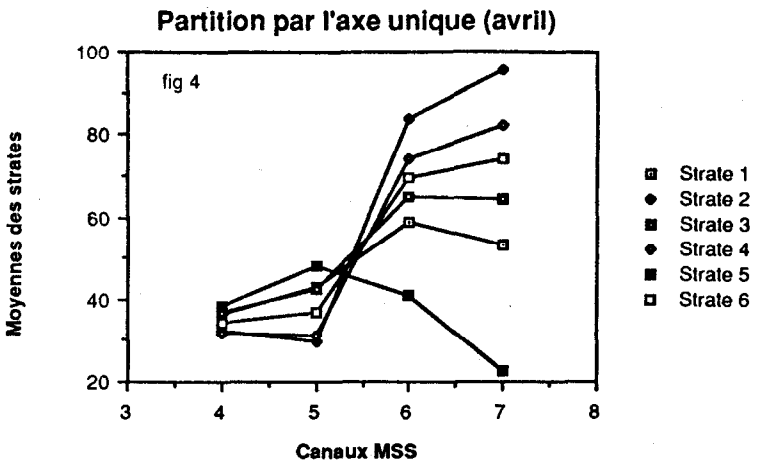


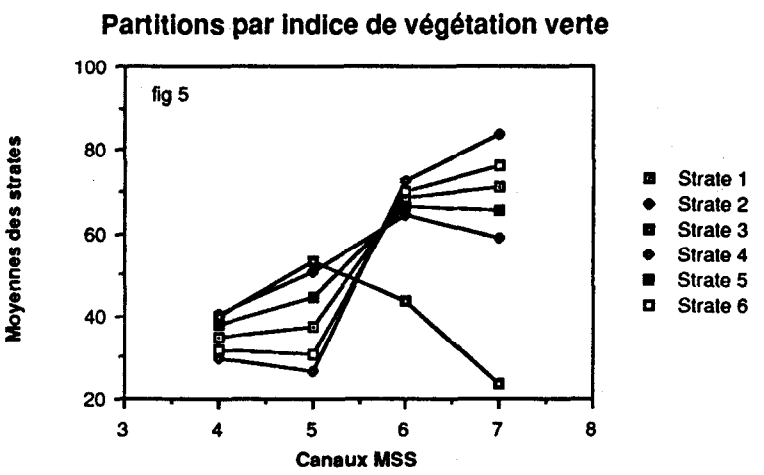
Partitions par les canaux originaux (avril)



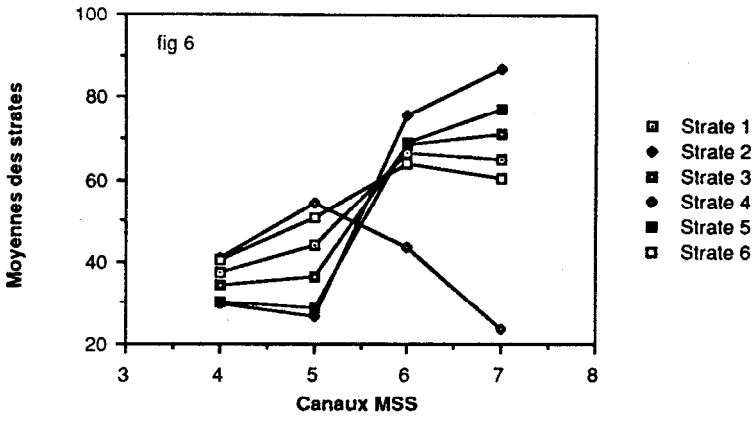
Partition par les axes factoriels (avril)



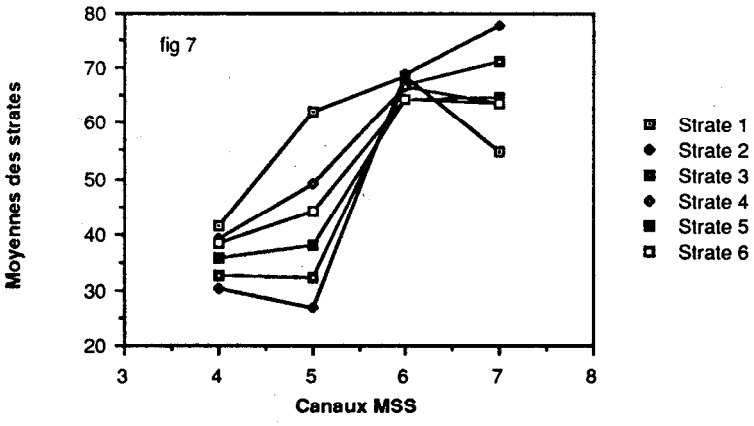

Partition par l'axe unique (avril)
fig 4
Moyennes des strates
100
80
60
40
20
3
4
5
6
7
8
Canaux MSS
Strate 1
Strate 2
Strate 3
Strate 4
Strate 5
Strate 6


Partitions par indice de végétation verte
fig 5
Moyennes des strates
100
80
60
40
20
3
4
5
6
7
8
Canaux MSS
Strate 1
Strate 2
Strate 3
Strate 4
Strate 5
Strate 6

Partition par indice végétation jaune



Partition par indice des sols (avril)



V. VALIDATION.

Une stratification efficace conduit à minimiser la variance interne W, et par la même occasion maximiser $\frac{B}{T}$ qui est le rapport de la variance entre les strates à la variance totale: rapport de corrélation Plus ce rapport tend vers 1 meilleure est la stratification. En effet cela veut alors dire que, W tendant à zéro, tous les individus de chaque strate prise une par une, tendent tous vers une même valeur moyenne. Inversement si il tend vers 0, toutes les strates ont la même valeur moyenne, et donc théoriquement il n'est pas besoin de stratification.

Dans l'utilisation des données satellitaires qu'en fait ce modèle, les strates obtenues représentent une partition suivant des états qualitatifs. En particulier si on ne s'intéresse qu'à une culture, celle-ci ne sera pas seule dans la ou les strates où elle se trouverait. En effet tout autre phénomène dans le même état qualitatif qu'elle, se retrouvera à ses côtés. Ce qui entraîne que dans les calculs des paramètres de W et B, la culture cible se retrouve noyée. Ainsi si on code en 0/1 l'absence/présence de la culture, la moyenne qu'on en obtiendra tendra vers des valeurs très petites, puisqu'en terme d'effectif, le sien est très inférieur à celle de la strate. Ce qui a une incidence directe sur la valeur du rapport de corrélation En effet on se retrouve avec des strates qui apparemment tendent toutes vers une moyenne de zéro, puisque la majorité des individus y sont codée à zéro (seule le thème cible est à 1). Ce qui fait aussi tendre B vers de petites valeurs, inférieures à W. Une conclusion trop rapide serait que la stratification est inadéquate, de même que le modèle.

En fait le point fondamental, dont une validation doit être réalisée, est que le *thème (culture, etc...) auquel on s'interesse se retrouve localisé dans un minimum de strates, et que dans la relation décomposant sa variance, W soit nettement inférieur à B et que le rapport* $\frac{B}{T}$ *tende vers 1. Ceci car en fait, la population réellement étudiée est justement composée des individus pixels représentant le thème ciblé* (population qui peut être étendue à un ensemble de phénomènes se trouvant dans un même état qualitatif). Le cas idéal étant qu'il soit dans une seule et unique strate. L'étape de validation se décompose alors en deux:

- Au niveau des ensembles de variables vérifier lequel est le plus efficace, et au niveau des applications ultérieures l'utiliser, puisqu'il aura été identifié et validé. Ce que nous ferons au moyen de la comparaison des différentes valeurs de W au niveau des strates, suivant les variables utilisées. Ce qui constituera aussi une validation globale au niveau de tous les thèmes présents, car on peut considérer -étant donné le modèle de représenta-

tion qu'est la télédétection- que chaque strate tend à ne comporter que des phénomènes à des états qualitatifs identiques. Pour le vérifier il suffit de considérer qu'on peut, dans le cas présent, admettre que le riz constitue la totalité des phénomènes pouvant se trouver dans le même état qualitatif que lui. Ainsi une strate où il se trouverait ne serait composer que de riz. Le rapport de corrélation tendrait alors vers 1, puisque les strates sont d'opposition maximale (0/1 soit présence/absence), et que les individus les composant sont soit à zéro, soit à un (inexistence de mélange).

- Au niveau du modèle lui même, élaborer: identifier et estimer, un paramètre qui puisse effectivement le valider. C'est un paramètre qui devra traduire la "détermination" du thème par la stratification, et le montrer clairement si tel est le cas. Ce niveau de validation doit tenir compte du nombre de strates comportant le thème. Plus il y en a, moins la stratification est efficace. L'objectif est d'assurer un coefficient de détermination élevé et remplir les conditions de minimisation de W sur le thème (T= W + B).

Ces deux conditions impliquent l'existence d'une relation d'ordre entre partitions. Relation qui les ordonne suivant la dispersion (dispersion spatiale) du thème dans les classes. Et le coefficient rendant compte de la correlation thème/stratification doit en tenir compte.

Ces deux niveaux de validation peuvent paraître redondants. Cependant nous les effectuerons toutes deux, au moins durant cette modélisation, car si le premier ne peut traduire une détermination élevée -ainsi que nous l'avons précédemment vu- vis à vis d'un thème, nous attendons du second qu'il le fasse.

Le premier niveau concerne plus les variable: lesquelles sont les plus efficaces?; et le second le modèle: les strates déterminent elles vraiment le thème ciblé?.

Images de la région test
d'après les differentes partitions

Image 14
par les canaux originaux

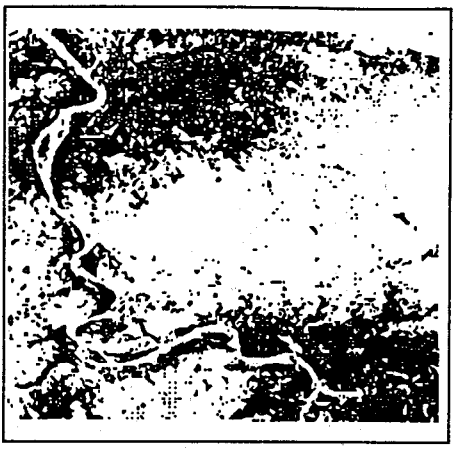


Image 15
par les axes factoriels

Image 16 - par "IVG"

Image 17 - par "IVJ"

Images de la région test
d'après les differentes partitions

Image 18 - par "SOL"

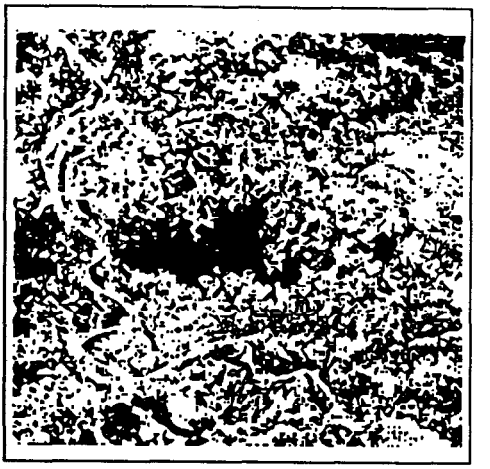


Image 19 - par l'axe "UNIQUE"

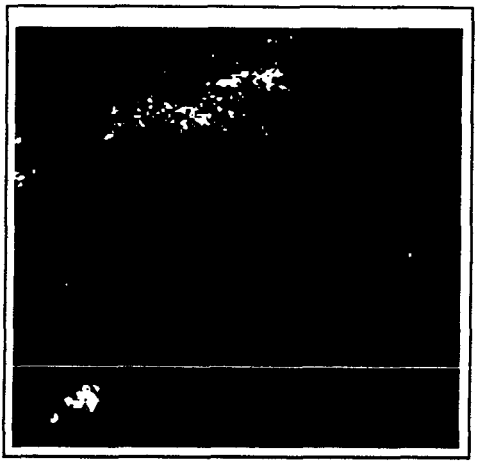


Image 13 - La "Vérité terrain riz"

V.1 PREMIER NIVEAU DE VALIDATION: LES VARIABLES.

Dans les figures 8, 9, 10 et le tableau 1 sont reportées les différentes valeurs obtenues pour W, B, et le rapport de corrélation $\frac{B}{T}$.

La comparaison porte sur trois groupes de variables:

- fin du mois d'avril 1981, fin de saison des pluies (rizières inondées). Variables utilisées: les canaux originaux, les axes factoriels, les indices de végétation verte, jaune et de sol, l'axe unique ci-dessous modélisé.

- fin du mois d'août, en saison sèche (riz à maturité, peu avant la récolte). Variables utilisées: canaux originaux, axes factoriels, axe unique modélisé.

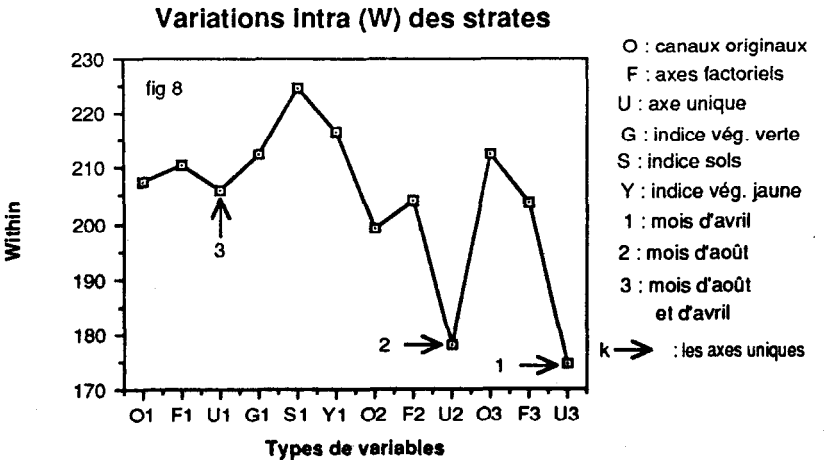
- avril et août simultanément, sur les variables: canaux originaux, axes factoriels, et axes uniques modélisés.

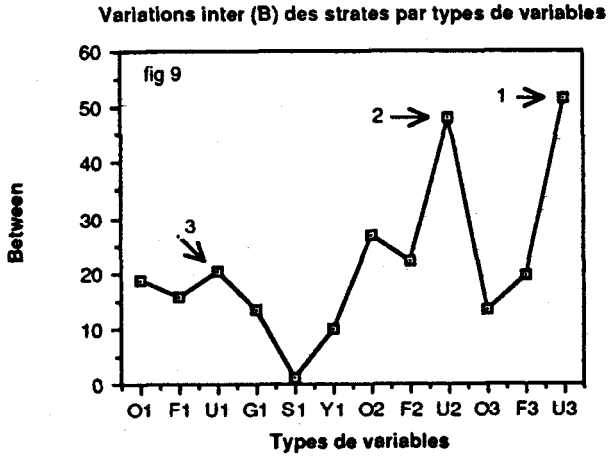
TP	Within	Between	Neta2
O1	0,01556967	0,00139391	0,0822
F1	0,01579522	0,00116703	0,0688
U1	0,01543219	0,00151823	0,0896
G1	0,01594395	0,00099359	0,0587
S1	0,01684339	0,00009253	0,0055
Y1	0,01623219	0,00073421	0,0433
O2	0,01493044	0,00201515	0,1189
F2	0,01529999	0,00165379	0,0975
U2	0,01335295	0,00360181	0,2124
O3	0,01593457	0,00101447	0,0599
F3	0,01526979	0,00146480	0,0875
U3	0,01309894	0,00386572	0,2279

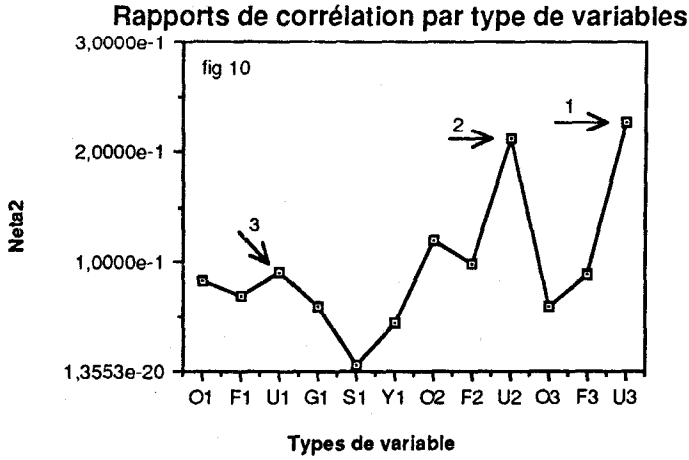
Tableau 1-Légende du tableau 1 (valable pour toute la suite):
- O : canaux originaux - F : axes factoriels
- U : axe unique modélisé
- G : indice végétation verte
- S : indice des sols
- Y : indice végétation jaune

- 1 : mois d'avril - 2 : mois d'août
- 3 : mois d'avril et d'août

Par exemple: "U3": "axes uniques modélisés au mois d'avril et d'août" "O2":"les canaux originaux du mois d'août"


Variations intra (W) des strates
fig 8
Within
230
220
210
200
190
180
170
3
2
1
O1 F1 U1 G1 S1 Y1 O2 F2 U2 O3 F3 U3
Types de variables
O : canaux originaux
F : axes factoriels
U : axe unique
G : indice vég. verte
S : indice sols
Y : indice vég. jaune
1 : mois d'avril
2 : mois d'août
3 : mois d'août et d'avril
k : les axes uniques


Variations inter (B) des strates par types de variables
fig 9
Between
60
50
40
30
20
10
0
3
2
1
O1 F1 U1 G1 S1 Y1 O2 F2 U2 O3 F3 U3
Types de variables



A travers ces trois groupes seront analysés l'efficacité relative des dates et des variables en combinaison.

D'emblée constatons que dans tous les cas les variances dans les strates sont très inférieures à celles entre les strates. Il est donc impossible à ce niveau de conclure sur le degré de signification de la détermination du thème par les stratifications. Le coefficient associé est en effet très bas, en fait il tend systématiquement vers zéro (voir la figure 10).

Ne pouvant avoir des certitudes sur l'hypothèse de dépendances ou d'indépendance des strates et du thème, on se restreint ici à l'analyse des variables.

a) Dans tous les cas de dates, l'axe unique modélisé s'avère être le plus efficace. En particulier avec ce type de variable, l'utilisation de deux dates semble apporter un certain gain. Bien qu'il faille remarquer que les résultats du mois d'août (saison sèche) s'en rapprochent: 48.02 de part de la variation totale de la population, contre 51.54 pour les deux dates.

b) Dans le cas d'utilisation d'une seule date, les canaux originaux s'avèrent toujours supérieurs aux axes factoriels. C'est uniquement dans le cas de l'utilisation des deux dates en même temps que les axes factoriels sont plus efficaces que les canaux originaux.

c) Les différents indices, testés uniquement sur le mois d'avril, sont peu efficaces. Pour celui des sols cela est explicable assez facilement: en effet ils ne sont guère visible à cette date, ou assez peu.

d) Mis à part le cas de la variable unique modélisée, l'utilisation des deux dates n'apporte en rien le gain, qu'on aurait été en droit d'espérer. En ef-

fet l'axe unique modélisé en avril, les canaux originaux, les axes factoriels et l'axe unique modélisé en août, le surclassent nettement.

Ainsi sans parler de ses avantages supplémentaires: facilité d'interprétation des classes, volume de données plus faible, etc..., l'axe unique modélisé s'avère toujours supérieur à tout autre ensemble de variables. Du moins est-ce le cas avec celles qui sont comparées ici, et qui regroupent la majorité de celles usuellement utilisées en télédétection jusqu'à présent.

(Rappelons qu'il s'agit ici de stratification).

V.2 SECOND NIVEAU DE VALIDATION: LE MODELE.

Dans ce second niveau, il nous faudra montrer si de façon significative ou non, le thème visé est déterminé par la mise en oeuvre du modèle. Il faut donc tester l'hypothèse H_0 de l'indépendance de la stratification et de la localisation/dispersion spatiale du thème. Pour cela, il a été vu que les partitions doivent être classées.

Le modèle que nous nous proposons de mettre en oeuvre ici, est tel qu'une partition en un nombre k de classes, est d'autant meilleure que -le thème étant dans un état qualitatif donné-, celui-ci se trouve dans un minimum de strates. Ce qui veut dire que la stratification sert effectivement à quelque chose. Car si en effet le thème se retrouve éparpillé sur un grand nombre de strates, la variance B s'en trouve réduite, de telle façon qu'un modèle stratifié ne se justifie pas. Aussi les classes d'une partition seront-elles classées de deux manière: l'une par rapport à la vérité terrain de localisation du thème, et l'autre par rapport à la classification proposée par le modèle. Ces deux classements feront ensuite l'objet d'une comparaison qui devra conclure sur la validité ou non du modèle.

Supposons que parmi les k classes d'une partition, l contiennent le thème. Alors ces l sont classées de 1 à l sur la partie modèle. Ceci car plus il y a de strates contenant le thème, moins celui-ci est déterminé par le modèle qui est d'autant moins efficace. Dans le classement dit de "vérité terrain", toutes ces strates sont classées à 1: la vérité terrain est par définition la référence.

Ainsi:

- Si seule une strate contient le thème -le meilleur cas- l'accord sera total entre les deux classements, puisque cette unique strate est classée à 1 dans les deux cas.

- Si deux strates contiennent le thème, l'accord diminue, car si sur l'une les deux classements sont respectivement à 1, sur l'autre elle est de 2 sur la partie modèle et de 1 sur la partie "vérité terrain".

Ainsi de suite jusqu'au cas où l strates contiennent le thème.

Les strates qui ne contiennent pas le thème sont dans les deux classements mis identiquement à (l+1), pour un accord total sur elles.

Ainsi on reconnait un modèle où deux relations d'ordres partielles classent des individus, qui dans notre cas sont des classes-strates. La comparaison et le test de l'indépendance de telles relations, rencontrent les spécificités de la correlation de rang de Spearman ou du "taux" de Kendall, dont l'efficacité par rapport au coefficient de Pearson, dans le cas où celui-ci s'applique, est de 91%.

V.2.a Estimation.

Suivant les variables utilisées dans le cadre du modèle global, nous allons relever le nombre de strates contenant le thème, réaliser les classements, estimer le r_S de Spearman, et tester l'hypothèse H_0 d'indépendance des deux classements.

Les partitions sont en six classes. Le nombre de strates comprenant le thème va donc de 1 -meilleur cas- à 6, en fait ici de 1 à 5. Les résultats (tableau 2) sont visualisés sur la figure 11.

V.2.b Etude des résultats.

Avec l'axe unique modélisé, aux mois d'avril, d'août, d'avril et d'août simultanément, et avec les axes factoriels du mois d'août, la probabilité de rejeter H_0 alors qu'elle est bonne est de 0.01. On peut donc dire avec un niveau de confiance de 99% que la stratification détermine efficacement le thème. Le modèle est le plus efficace avec l'axe unique utilisé en août, et en avril et août simultanément, avec r_S à 1. Viennent ensuite l'axe unique en avril et les axes factoriels en août, avec r_S à 0.9714286.

Avec un niveau de confiance de 95% les canaux originaux aux mois d'août, d'avril et d'août simultanément, déterminent aussi efficacement le thème. La valeur de r_S est alors de 0.8571429.

Les autres variables s'avèrent inefficaces au niveau de la détermination du thème. Ainsi:

- les axes factoriels en avril

- les canaux originaux en avril

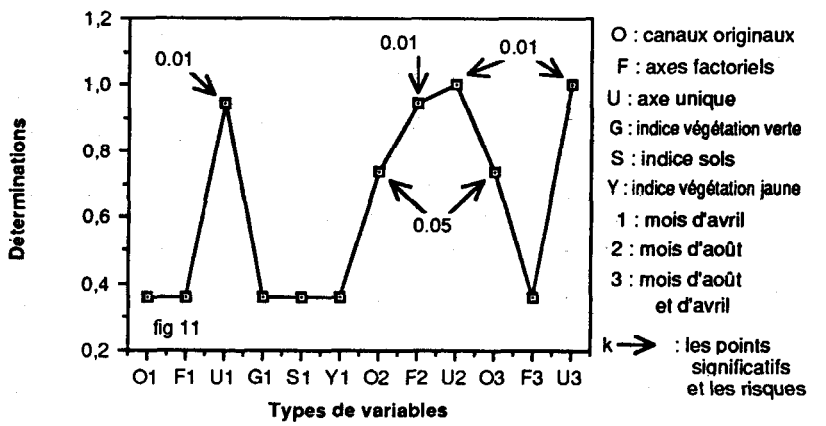
- les indices de végétation verte et jaune en avril
- l'axe des sols en avril
- les axes factoriels en avril et août simultanément

sont dotés d'un r_s de 0.6.

TP	CO	RS
O1	--%	0,6000
F1	--%	0,6000
U1	99%	0,9714
G1	--%	0,6000
S1	--%	0,1428
Y1	--%	0,6000
O2	95%	0,8571
F2	99%	0,9714
U2	99%	1,0000
O3	95%	0,8571
F3	--%	0,1428
U3	99%	1,0000

Tableau 2 - TP : types de variable
- CO : niveaux de confiance
- RS : coefficients de Spearman

Détermination du thème par les stratifications



VI. APPLICATIONS.

Mise à part cette application ayant servie de support à la démonstration du modèle, celui-ci, en particulier la partie relative à la construction de classes par décomposition de la variance totale, a été utilisée pour une étude de l'éco-système en zone semi-aride (application sur l'Egypte) d'une part, et d'autre part pour l'étude et l'analyse de l'évolution d'un paysage sahélien (application sur la mare d'Oursi Burkina-Faso). Dans ces deux cas la stratégie d'utilisation et d'analyse de l'espace suivant un nombre croissant de classes, développé et mis au point par l'auteur dans sa thèse (voir bibliographie), pour en faire ressortir les hiérarchisations successives, basé sur ce modèle d'élaboration de partitions, a permis l'obtention de résultats dont l'exploitation par les thématiciens concernés s'est révélé correspondre aux besoins et attentes.

VII. CONCLUSION.

La conclusion qui s'impose après l'étape de validation, est qu'effectivement le modèle global de stratification auparavant identifié, est efficace d'une part; et d'autre part qu'il l'est systématiquement le plus avec l'axe unique lui aussi identifié et modélisé. Nous proposons donc ce modèle: partition par décomposition de la variance totale par minimisation de la variance intra W et utilisation de l'axe unique modélisé, qui est celui d'une fonction discriminante sur deux classes, pour la construction de strates ou de grandes classes de phénomènes sur une région quelconque. En effet il est basé directement sur la représentation (spectrale) que fait la télédétection des réalités au sol, et des différences maximales entre ces réalités prises comme un tout.

VIII. BIBLIOGRAPHIE

VIII.1 TELEDETECTION.

ANDRIANASOLO H. Paris - 1987.Analyse statistique des données de télédétection Statistiques agricoles Application sur Madagascar

Manual of remote sensing American Society of Photogrammetry1983.

BARRETT E. C., CURTIS F. Introduction to environmental remote sensing Chapman and Hall, London - 1976.

BAUER M., HIXSON M., BIEHL E.,DAUGHTY C., ROBINSON B., STONER E.

Agricultural scene understanding Jonhson Space Center - 1978.

VIII.2 CLASSIFICATION.

COOLEY W., PAUL R., COHNES Multivariate data analysis John Wiley and Sons, New-York - 1971.

DUDA HART Pattern classification and scene analysis John Wiley and Sons, New-York - 1973.

ANDERBERG Cluster analysis for applications Academic Press - 1973.

HARTIGAN Clustering algorithm

John Wiley and Sons, New-York - 1975.

GENDRE FRANCIS Analyse statistique multivariée Librairie Droz - 1976.

VIII.3 STATISTIQUE.

HOUSEMAN E. Area frame sampling in agriculture Department of Agriculture Washington DC. - 1979.

WARREN GILCHRIST Statistical modelling John Wiley and Sons - 1984.

COCHRAN W. Sampling techniques John Wiley and Sons, New-York - 1977.

DESABIE J. Théorie et pratique des sondages Dunod - 1966.

SIEGEL Non parametric statistics Mac Graw Hill - 1956.