Statistique et analyse des données

P. Monestiez

Méthode de classification automatique sous contraintes spatiales

Statistique et analyse des données, tome 2, n° 3 (1977), p. 75-84

<http://www.numdam.org/item?id=SAD_1977__2_3_75_0>

Statistique et Analyse des Données
3 - 1977

méthode de classification automatique sous contraintes spatiales

APPLICATION A L'ETUDE D'UN "ESSAI A BLANC" EN AGRONOMIE.

M O N E S T I E Z P.

Centre National de Recherches Forestières - CHAMPENOUX - 54280 SEICHAMPS

RESUME

Les individus que l'on cherche à classer sont des points, des parcelles ou des zones d'un champ spatial, décrits par plusieurs variables. En introduisant une contrainte de contiguïté, on cherchera à optimiser les partitions géographiques formées de classes connexes. A travers un exemple, on essaiera d'exhiber les principaux avantages de cette méthode.

ABSTRACT

This paper considers the problem of classifying areas or plots issued from a spatial field. The introduction of a constraint of contiguity in a hierarchical clustering allows to find classes in a single piece. Through an example, we show that this methode is more suitable to get geographical map and to reveal spatial structure.

PRESENTATION

Le but de cette étude est de décrire un phénomène spatial à l'aide d'une classification ascendante hiérarchique. Les individus étant des points ou des parcelles d'un espace géographique, la méthode habituelle consiste simplement à cartographier les partitions retenues dans l'arbre de la C.A.H.

Trois remarques sont cependant à faire quant à l'efficacité d'une telle méthode par rapport à nos objectifs :

- on ne tient pas compte de la position géographique des individus
- ce qui nous intéresse est autant la partition retenue, optimale dans l'espace des variables, que la partition géographique plus fine formée des composantes connexes des classes précédentes. Cette partition contiendra souvent un très grand nombre de classes et engendrera une carte peu lisible.
- une composante connexe peut avoir des caractéristiques assez différentes de la classe dont elle fait partie, par exemple une variance plus forte, ce qui empêchera toute analyse locale de la carte et risquera de nous tromper sur l'homogénéité de certaines zones.

Ces trois considérations nous ont amenés à introduire une contrainte dans l'algorithme ; des individus ou des classes ne seront agrégés que si ils sont contigus. Les noeuds ainsi formés seront connexes, et c'est la partition géographique qui sera optimisée à chaque agrégation.

En plus du résultat pratique obtenu, c'est-à-dire une carte simple et optimisée, la méthode pourra être utilisée afin de décrire les liaisons spatiales, les tendances, aussi bien dans le cas univariable que multivariable.

DISCUSSION ET PERFORMANCES

L'algorithme avec contraintes spatiales devient plus adapté que la méthode sans contrainte lorsque le nuage d'individus ne présente pas de classes très nettement différenciées. De nombreuses partitions peuvent être obtenues pour des indices équivalents. Si l'une d'elle vérifie la contrainte, il devient préférable de la choisir pour sa représentation simple, d'autant plus qu'elle sera tout aussi significative puisque d'indice à peine supérieur.

Par contre, lorsque les points seront indépendants spatialement, le résultat sera médiocre, mais très proche du découpage géographique obtenu par la C.A.H. habituelle. Dans ce cas, il se produira des phénomènes d'inversion, l'indice du noeud devenant inférieur à celui du noeud précédent. Ces inversions ont un sens d'un point de vue géographique ; elles sont

cependant gênantes car la partition correspondant au noeud précédent n'est plus significative et il serait illusoire de vouloir fixer un nombre de classes connexes a priori si on rencontre un grand nombre d'inversions.

D'un point de vue pratique, l'algorithme présente un avantage important. Pour chacun des 5 modes d'agrégation cités dans l'exemple, on a traité les 360 individus en 0.80 mn sur IRIS 80, la moitié de ce temps ayant servi à effectuer la cartographie de 5 partitions. Dans la version du programme la plus générale, il suffira de fournir les données et la description des voisins sous forme d'un fichier-disque, pour pouvoir classer 3000 individus en une vingtaine de minutes sur IRIS 80 (le nombre de variables influant très peu sur le temps et n'influant pas sur la place mémoire qui est un tableau 50 x 3000). Pour l'ultramétrique sous dominante, un algorithme du type de PRIM accroîtrerait encore notablement ces performances.

PARTICULARITES DE CERTAINES STRATEGIES D'AGREGATION

1) Ultramétrique sous dominante

La contrainte de contiguïté est insuffisante pour ce mode. La distance entre deux classes voisines va être obtenue en recherchant le couple de distance minimum. Ce couple sera la plupart du temps formé de points non contigus et souvent distants géographiquement. Les très nombreuses inversions dues à ce phénomène posent des problèmes d'interprétation identiques à ceux de l'algorithme sans contraintes.

Pour ce mode, on ajoutera une contrainte supplémentaire : le minimum sera recherché parmi les couples d'individus contigus. On oblige ainsi l'arbre de longueur minimum à suivre le graphe des contiguïtés.

La possibilité de représenter cet arbre sur la carte va faciliter l'interprétation et empêcher toute inversion.

2) La variance des classes

La variance donnera une bonne représentation des structures spatiales.

Si on rencontre un arbre de forme classique sans inversions, cela voudra dire que l'on est capable d'isoler des zones spatiales assez homogènes de caractères différentes. Il y a donc dépendance entre l'emplacement et les caractéristiques, et entre la taille et la variance. Cela montre simplement

des "tendances", au sens le plus large, à l'intérieur du champ étudié. Si, au contraire, un très grand nombre d'inversions se produisent pour les indices supérieurs on pourra conclure inversement à la "stationnarité". Pour un indice inférieur, l'effectif moyen des classes sera un bon indicateur des liaisons spatiales, un effectif très faible correspondant à l'indépendance.

PRESENTATION DE L'EXEMPLE

L'étude a été faite sur un champ de 30 m sur 37,5 m destiné à des expériences agricoles qui, entre chaque "essai", est mis en repos pendant 2 ans par une culture d'homogénéisation pour "effacer" les effets précédents.

Nous avons utilisé une de ces cultures, en l'occurence de la luzerne comme s'il s'agissait d'un plan d'expérience ; c'est donc un "essai à blanc" qui a été effectué sur 12 lignes de 30 parcelles de 1,25 m x 2,50 m et sur 2 coupes consécutives de luzerne.

Les 2 variables étudiées ont été le rendement en matière verte pour chaque coupe, afin de déceler les éventuelles anomalies dues au terrain et au passé cultural de ce champ, anomalies qui pourraient-être utiles à connaitre pour mettre en place les expériences l'année suivante.

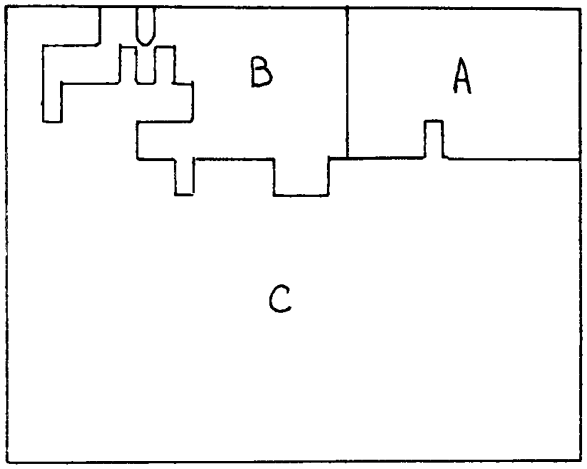
Le volume total de l'étude étant assez important pour 360 individus nous avons essayé de montrer les résultats les plus expressifs agrémentés de quelques commentaires.

1 - MINIMISATION DE LA VARIANCE DES CLASSES

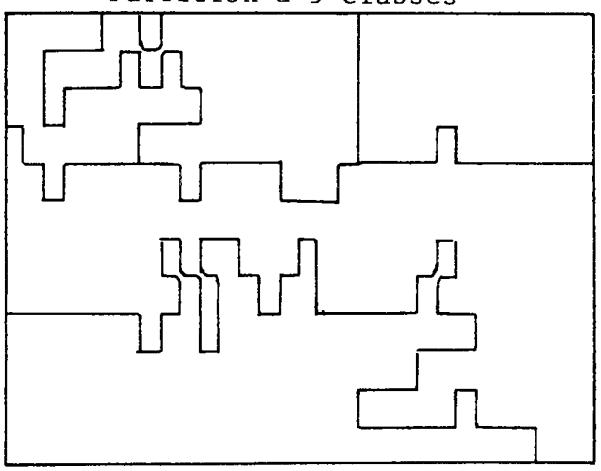
8 voisins sont retenus

parcelle et son voisinage

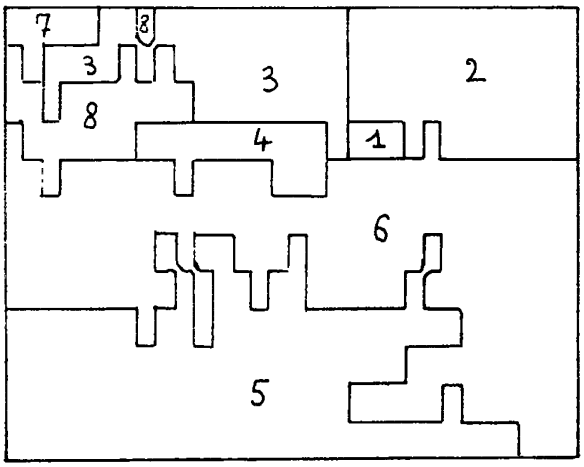
Partition à 3 classes



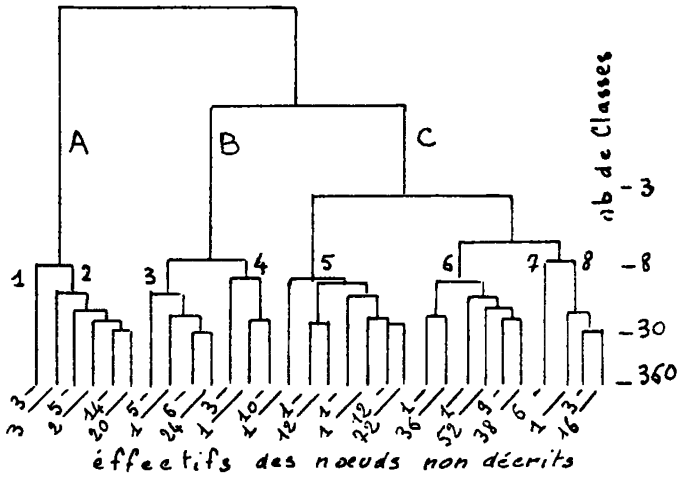
Partition à 5 classes



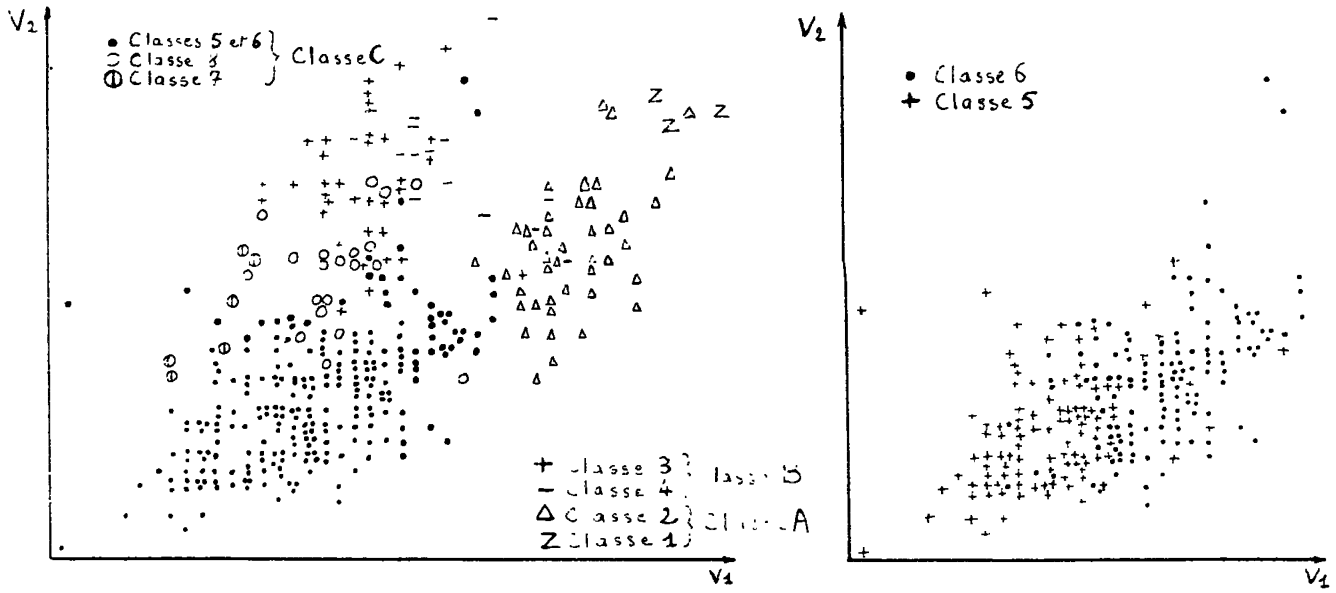
Partition à 8 classes



Arbre supérieur à 30 classes



Nuage des individus (partition à 8 classes)

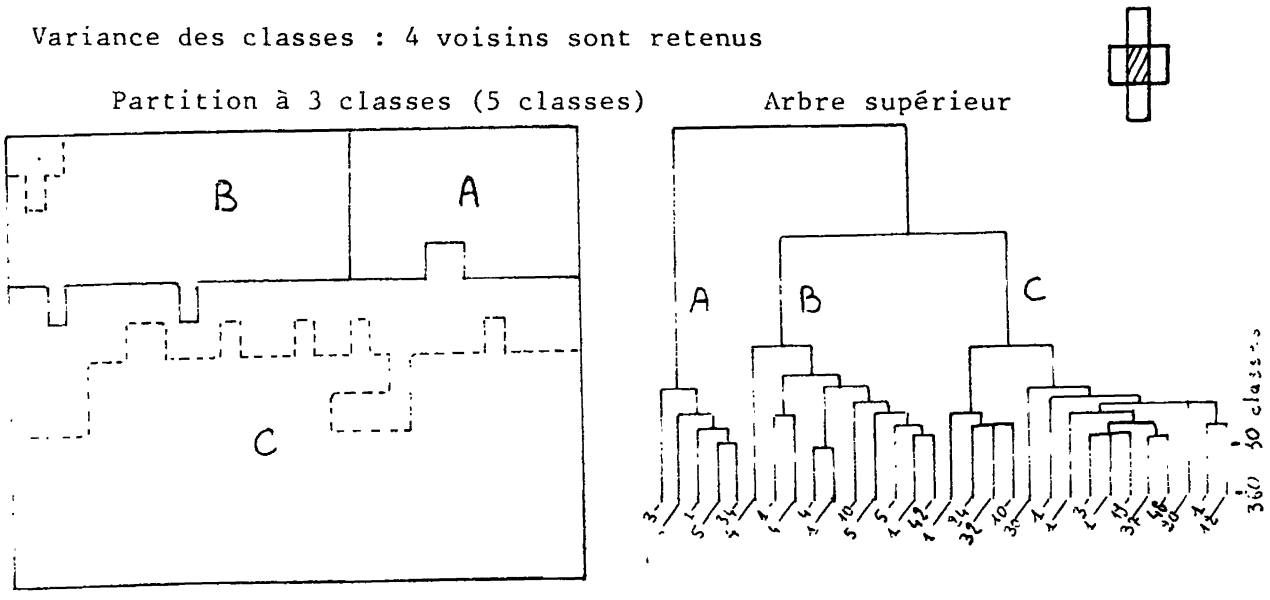


La variance nous montre distinctement deux zones (A et B) assez différentes du reste du champ. Ces deux zones correspondent à deux populations bien visibles sur le nuage de points. Il est évident que l'algorithme sans contrainte aurait permis de les isoler facilement. Les résultats auraient été similaires à une trentaine de points près.

L'intérêt de l'algorithme avec contraintes n'apparait qu'à travers une analyse un peu plus fine. Il est capable de distinguer dans la classe C plusieurs zones importantes de caractéristiques différentes. C'est uniquement avec l'aide de la contrainte de contiguité qu'il est possible de mettre en évidence les classes 5 et 6 alors qu'en l'absence de contrainte, toute partition aurait engendré une multitude de petites composantes connexes amenant à la conclusion que la zone C pouvait être considérée comme "homogène" et de variance forte. On voit ici que ce n'est pas du tout le cas et qu'il y a réellement une tendance.

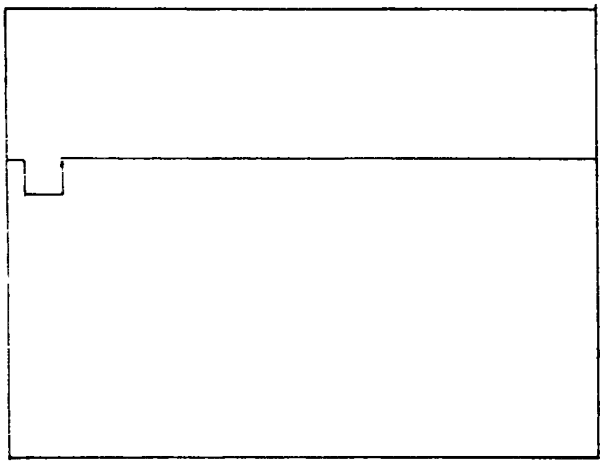
Afin d'améliorer le contour, il est possible de renforcer la contrainte en n'admettant plus que 4 voisins au lieu de 8. On remarquera l'importance du choix de la contrainte sur les résultats.

Variance des classes : 4 voisins sont retenus

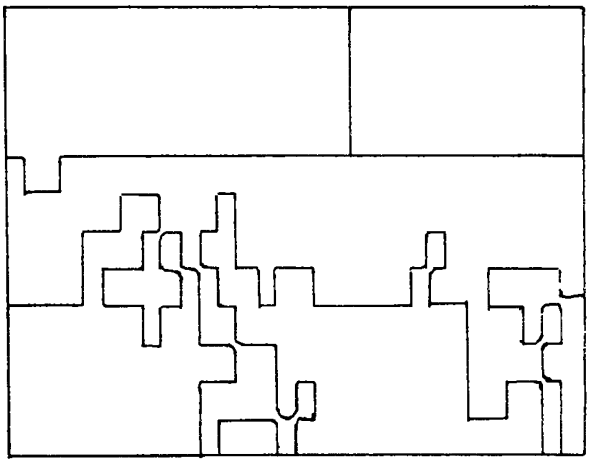


2 - MINIMISATION DU MOMENT CENTRE D'ORDRE 2 DES CLASSES (8 VOISINS)

Partition à 2 classes

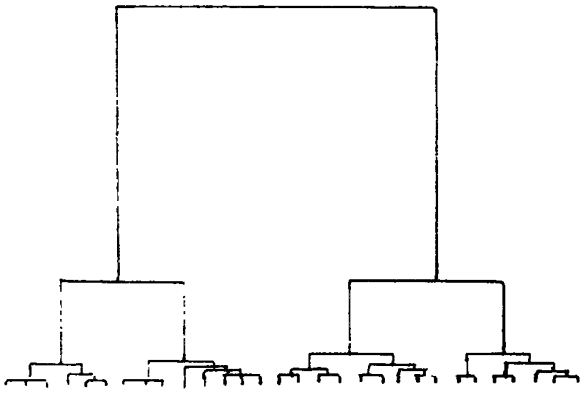


Partition à 4 classes

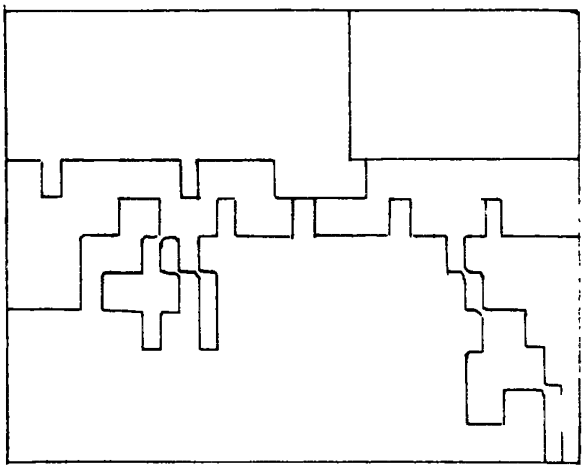


MOMENT D'ORDRE 2 DE LA PARTITION

Arbre supérieur à 30 classes



Partition à 5 classes



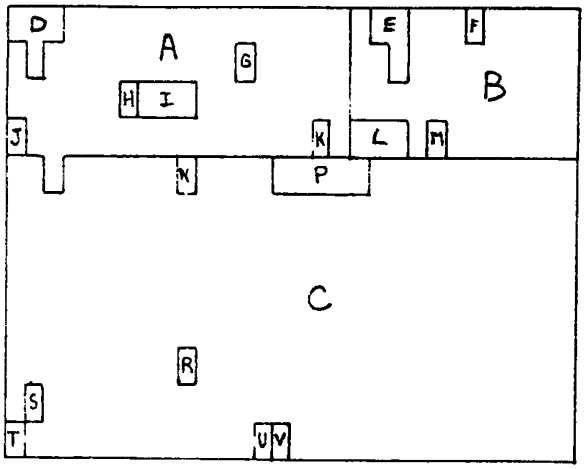
Les critères d'agrégation inertiels, minimisation du moment d'ordre 2 centré des classes et maximisation du moment d'ordre 2 de la partition, apparaissent assez performants pour isoler des zones importantes.

Ils favorisent l'agrégation des classes de faible effectif et permettent ainsi d'obtenir des partitions plus pratiques pour l'utilisateur qui préférera toujours travailler sur des zones de tailles comparables.

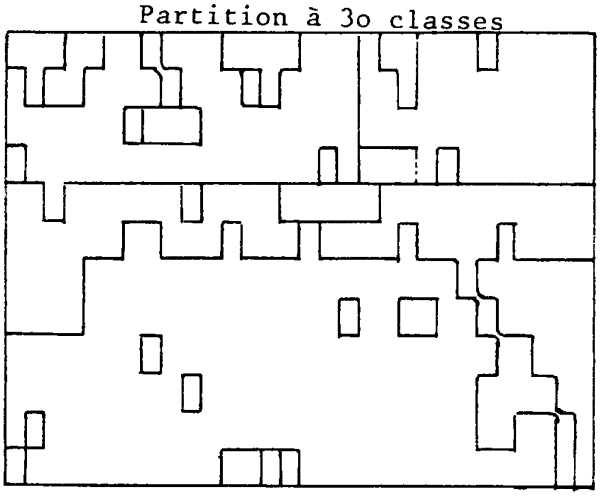
On remarquera que le moment d'ordre 2 des classes met en évidence trois zones de forme rectangulaire quasiment parfaite pour lesquelles on pourra sûrement trouver l'explication dans le passé culturaì.

3 - BARYCENTRE

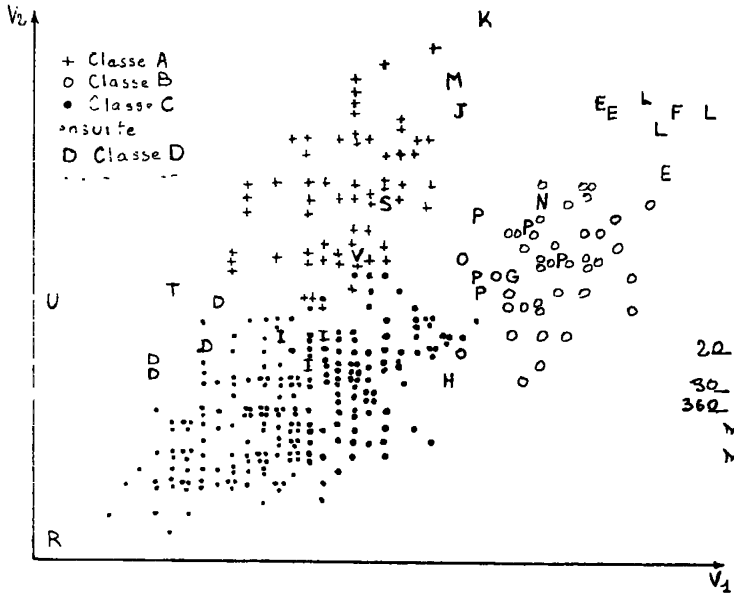
Partition à 20 classes



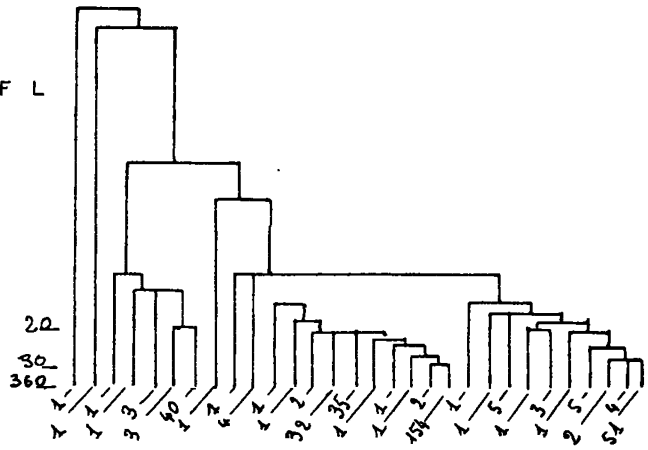
Partition à 3o classes



Nuage des individus (20 classes)



Arbre supérieur à 30 classes

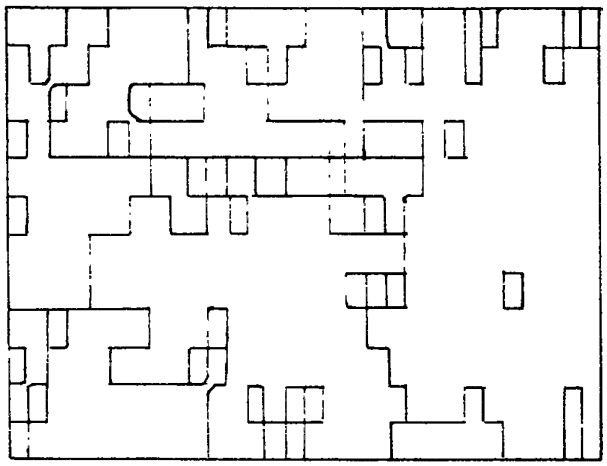


Le critère du barycentre met en évidence les points excentrés du nuage. Ici cela se fera par rapport au nuage des classes voisines à cause de la contrainte. Il faut donc remonter assez loin dans l'arbre pour distinguer des zones plus importantes.

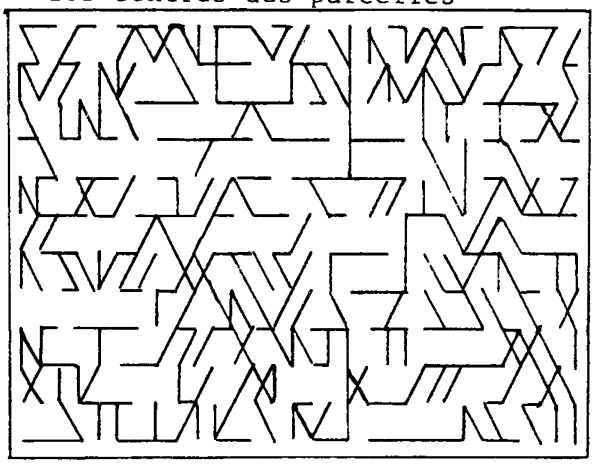
Le critère est particulièrement bien adapté dans notre cas, car il explique par les points localement abérants, les formes plus ou moins complexes des partitions obtenues pour les critères précédents. Il apporte donc un très bon complément d'information.

4 - ULTRAMETRIQUE SOUS DOMINANTE

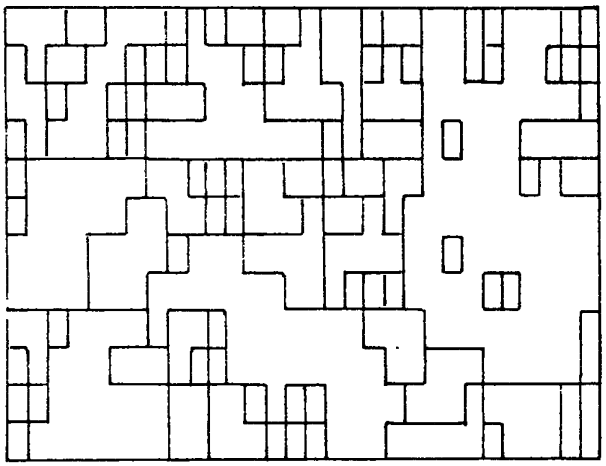
- 8 voisins : partition à 57 classes



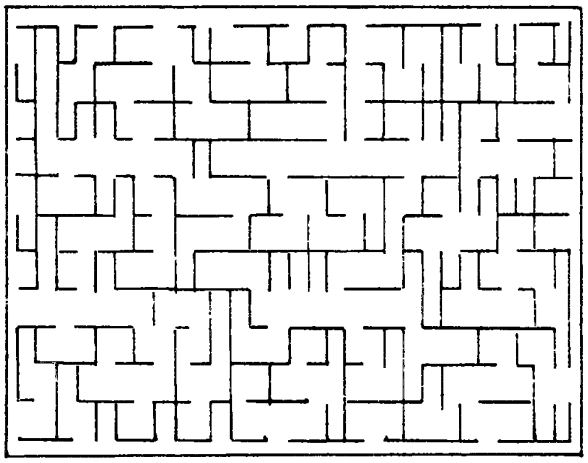
Arbre de longueur minimum liant les centres des parcelles



- 4 voisins : partition à 100 classes



Arbre de longueur minimum



(eux partitions correspondent au même indice)

L'ultramétrique sous dominante se comporte un peu de la même manière que le barycentre. Il faut donc remonter très loin dans l'arbre pour éviter d'avoir une classe importante unique et des points isolés. Ce critère est particulièrement adapté à l'étude des phénomènes locaux. L'analyse de partitions assez fines permet de détailler les liaisons et d'expliquer certains découpages car l'U.S.D. décrit ce que l'on pourrait appeler des "failles" en terme géographique. L'étude de la représentation de l'arbre de longueur minimum pourra dégager les structures spatiales particulières (liaison selon les lignes par exemple) en jouant sur le nombre de voisins. Dans notre cas, on peut uniquement conclure qu'il n'y a pas d'effet directionnel. L'importance des liens en diagonale et le fait qu'ils se croisent

souvent laisse supposer qu'il y a indépendance locale entre un point et son voisin immédiat (1 côté commun), les liaisons observées n'étant dues qu'à une structure sur une plus grande échelle.

CONCLUSION

Il est bien sûr impossible d'analyser un tel exemple en quelques pages. Nous avons essayé de montrer comment la méthode permettait de décrire une structure spatiale. Ici, nous n'avons fait que constater certains phénomènes qu'il faudrait analyser plus en détail afin de les discuter et de les expliquer à l'aide du passé cultural, des hypothèses sur les tendances, et de la connaissance du terrain.

En effet nous touchons au domaine de l'agronomie et nous ne le développerons pas ici. Il faut cependant remarquer que la méthode permet de s'intéresser à des problèmes locaux ; en n'étudiant complétement qu'une branche de l'arbre, on peut décrire un phénomène local en "oubliant" le reste. On pourra ainsi traiter des problèmes généralement difficiles tels que les effets de bordure, les points localement abérants.

La méthode est donc un outil assez performant, par ses possibilités techniques peu contraignantes (3000 individus), par son vaste champ d'application (toutes les recherches de partition géographique), et par les nombreuses variantes qui permettent de traiter des objectifs assez divers (modes et contiguités à adapter au problème).

REFERENCES BIBLIOGRAPHIQUES

J.P. MARBEAU, Géostatistique forestière, thèse de Doc.Ing., (Ecoles des Mines de Fontainebleau), 1976

M. JAMBU, Techniques de classification automatique, thèse de 3e cycle, (L.S.M., I.S.U.P.), 1972