Statistique et analyse des données

Brigitte Escofier
Habib Benali
Kaddour Bachar

# Comment introduire la contiguïté en analyse des correspondances ? Application en segmentation d'image

*Statistique et analyse des données*, tome 15, n° 3 (1990), p. 61-92
<http://www.numdam.org/item?id=SAD_1990__15_3_61_0>

Statistique et Analyse des Données

1990 - Vol. 15 n° 3 - p. 61 - 92

# Comment introduire la contiguïté en analyse des correspondances ?

## Application en segmentation d'image

Brigitte ESCOFIER [1] [2] Habib BENALI [3] Kaddour BACHAR [1]

**Résumé**

*Nous proposons ici des méthodes qui permettent de tenir compte dans une analyse des correspondances multiples d'une structure de proximité sur l'ensemble des individus et d'étudier les relations entre chacune des variables et la structure de proximité.*

*Dans une première approche la structure de proximité est définie par un graphe pondéré. Deux méthodes sont proposées. La première analyse les tendances générales en éliminant l'influence des fluctuations locales par un "lissage" des données. La seconde analyse ces différences locales. Lorsque ce graphe vérifie une certaine propriété (facile à respecter dans toutes les situations courantes) il suffit d'appliquer un programme classique d'analyse factorielle des correspondances (AFC). Dans une autre approche la structure de proximité est décrite par un ensemble de variables numériques données a priori ou déduite d'un graphe. Une analyse factorielle multiple (AFM) [Esc 90] est appliquée aux deux groupes de variables formés, d'une part par les variables qualitatives et d'autre part, par ces variables numériques.*

*Ces trois méthodes sont illustrées par un exemple d'application à la segmentation d'image. Nous abordons ensuite le cas de structure de proximité définies sur les modalités des variables et le cas de l'analyse des correspondances simple.*

[1]IRISA Campus de Beaulieu 35042 Rennes cedex
[2]IUT B.P.1104 56008 Vannes Cedex
[3]U287 INSERM Institut Gustave Roussy 94805 Villejeuif

Manuscrit reçu le 27 avril 1990, révisé le 18 décembre 1990

**Abstract**

*We propose here methods which allow to take into account a neighbouring relation in Multiple Correspondence Analysis and to study the relations between the variables and the neighbouring properties.*

*First, the neighbouring relationships are represented by means of a weighted graph defined on the set of individuals.Two methods are studied. The first one eliminates local variations in order to analyse more easily the general trends and the second one analyses these local variations. We define an easily satisfied condition for the graph in such a way that the application of the two methods needs only a classical Correspondence Analysis program.*

*In an other method, neighbouring relationships are represented by means of numericals variables. Multiple Factor Analysis which allows to analyse together groups of numerical variables and groups of qualitative variables is utilized.*

*All these methods are illustrated on an application-example (image segmentation).*

*Problem of simple Correspondance Analysis is also studied.*

## 1.INTRODUCTION

Nous traitons ici essentiellement du cas de l'analyse des correspondances multiples (ACM). En effet, la notion de contiguïté ou, plus généralement, de proximité, s'y introduit plus naturellement et plus simplement que dans l'AFC de tableau de fréquence quelconque. L'ACM traite des tableaux croisant des individus et des variables qualitatives. Il est fréquent que l'on dispose, en plus des variables analysées, d'une structure de proximité sur l'ensemble des individus. C'est le cas pour des données géographiques ou temporelles. Nous illustrons le problème et les techniques étudiées par un exemple où les individus représentent des zones d'une image bidimensionnelle caractérisées par des paramètres de texture.

On souhaite souvent faire intervenir cette proximité. Soit pour diminuer l'importance des variations locales qui peuvent avoir un caractère partiellement aléatoire et dégager ainsi des tendances générales sur lesquelles les zones auront tendance à se regrouper par région. Soit, inversement, pour analyser les variations locales et regrouper des zones qui diffèrent de leur voisinage de manière analogue. Soit encore pour étudier les relations entre cette proximité et

les variables qualitatives. Il est clair que dans tous les cas, la proximité doit jouer un rôle propre, à côté de celui des variables.

Au niveau des objectifs, il n'y a pas de différence entre le cas des variables numériques (et donc de l'analyse en composantes principales) et celui des variables qualitatives (et donc de l'ACM). Pour les premières, nous avons proposé et testé deux méthodes simples : l'analyse lissée et l'analyse des différences locales [Ben 88] et [Ben 89] dans lesquelles la contiguïté est introduite par un graphe. Le principe de la première est d'analyser un nuage d'individus dérivé du nuage initial en remplaçant les individus actifs de l'analyse par le barycentre de leur voisinage ; celui de l'analyse des différences locales est de représenter un élément par la différence avec ce barycentre. Nous commençons par adapter ces méthodes au cas des variables qualitatives, ce qui pose quelques problèmes techniques et amène à introduire des graphes particuliers.

## 2. GRAPHE DE CONTIGUITE OU DE PROXIMITE

On peut caractériser la contiguïté ou la proximité par un graphe défini sur l'ensemble des n individus noté I. Précisément, on note G la matrice du graphe qui est carrée et de dimension n. Son terme général, noté $g_{ii'}$, est positif ou nul ; il est d'autant plus grand que l'individu i' est considéré comme proche de l'individu i. Si $g_{ii'}$ est différent de 0, on dit que i' appartient au voisinage de i et que $g_{ii'}$ est le poids de i' dans ce voisinage. On note $g_{i.} = \sum_{i'} g_{ii'}$ la somme des poids des voisins et $g_{.i} = \sum_{i'} g_{i'i}$ la somme des poids de i sur ses voisins. On note N la matrice diagonale des poids $g_{i.}$ .

Dans le cas d'une contiguïté, le graphe n'est pas pondéré : seules les zones i' contigues à i (y compris i elle-même) ont un poids $g_{ii'}$ non nul et égal à 1. La pondération permet de nuancer et d'introduire une notion plus souple de plus ou moins grande proximité.

### Graphe bistochastique

A priori nous n'imposons aucune contrainte sur ce graphe. Cependant, nous verrons que l'analyse lissée et l'analyse des différences locales sont beaucoup plus simples et donnent des résultats bien plus satisfaisants lorsque $g_{i.} = g_{.i'} =$ constante.

Cette condition n'est pas toujours satisfaite. Elle ne l'est pas dans un graphe de contiguïté où le nombre de voisins n'est pas constant. En effet, la somme $g_{i.}$ des poids du voisinage d'une zone ayant peu de voisins est plus faible que celle d'une zone ayant un voisinage important. Pour rendre constante $g_{i.}$ , on peut diviser tous les $g_{ii'}$ par $g_{i.}$ mais le graphe n'est plus symétrique et la condition $g_{i.} = g_{.i'} =$ constante n'est toujours pas vérifiée.

Lorsque la somme des lignes et des colonnes est constante, nous pouvons supposer sans perte de généralité, que cette somme vaut 1. La matrice du graphe est alors une matrice bistochastique (la somme des lignes et des colonnes vaut 1) ; nous parlerons dans ce cas de graphe bistochastique.

Cette condition est facile à obtenir pour un graphe symétrique en jouant sur le poids d'un individu dans son propre voisinage (qui apparait dans la diagonale de G), comme nous le montrons dans les deux exemples ci-dessous.

Le premier exemple est celui d'une image plane où l'on veut introduire comme voisinage d'une zone les zones qui lui sont contigües ; le nombre de voisins varie : 3 pour un coin, 5 pour un bord et 8 pour les autres zones.Fixons, par exemple, à 0.1 la valeur du poids de zones contigües, le poids d'une zone dans son propre voisinage sera soit 0.7 (coin), soit 0.5 (bord), soit 0.2 (autre). Ce choix est logique, tous les individus interviennent avec le même poids dans le lissage de leurs voisins ; un individu qui a peu de voisins garde un poids important dans son lissage alors qu'un individu au voisinage riche, pour lequel on connait beaucoup plus de choses sur son environnement, a un poids plus faible.

Ce graphe pondéré présente sur celui d'un poids constant pour les voisins et pour l'individu lui-même, l'avantage de ne pas augmenter l'importance (dans le lissage et donc dans les résultats finaux des analyses) des voisins d'un individu dont le voisinage est pauvre et de ne pas diminuer celle de ces individus peu entourés. En effet un individu i intervient avec le poids $g_{i'i}/g_{i'.}$ sur chaque individu i' et donc sur l'ensemble des points avec le poids $\sum_{i'} g_{i'i}/g_{i'.}$ Du fait que $g_{i.} = g_{.i'} =$ constante, la "perte" de l'information sur i impliquée par son lissage est exactement compensée par le lissage de ses voisins. Ceci n'est pas le cas pour

un graphe de contiguïté non pondéré. Par exemple, pour le coin d'une image qui a trois voisins, $g_{i.}$ vaut 4, $g_{i'i}/g_{i'.}$ vaut donc 1/4 pour lui-même, 1/6 pour 2 de ses voisins et 1/9 pour le dernier; ce coin intervient donc dans l'analyse avec un poids de 25/36. Alors que son voisin le plus central, qui a 9 voisins, intervient dans le lissage avec un poids pratiquement double (49/36). 1/4 pour le coin et 1/6 pour les 4 bords et 1/9 pour les autres)

De même, dans le cas d'un graphe temporel où chacun des individus représente un instant t, et pour un nombre de voisins égal à 4, on peut donner aux instants t-1 et t+1 le poids 0.2, aux instants t-2 et t+2 le poids 0.1 et fixer le poids de t dans son propre voisinage pour que la somme des poids soit égale à 1. Ce poids vaudra donc 0.4 partout, sauf au début et à la fin de la série chronologique où il vaudra 0.7 pour les deux points extrêmes et 0.5 pour leur voisin immédiat.

Un autre type de graphe qui satisfait à ces propriétés est celui qui traduit une situation extrême à laquelle nous nous référerons. C'est celui d'un graphe défini par une partition où les individus de la même classe qu'un individu i ont un poids $g_{ii'}$ égal à l'inverse de l'effectif de cette classe et les individus des autres classes ont un poids nul. Les deux paragraphes suivants concernent le cas de ce graphe de partition ; ce sont des rappels qui permettent d'introduire facilement les généralisations à des graphes quelconques.

## 3. GRAPHE DE PARTITION : ANALYSE INTER

Dans le cas d'un graphe de partition, l'analyse lissée consiste à remplacer le point représentant l'élément i dans l'ACM par le barycentre de sa classe. Le poids affecté à ce barycentre dans l'analyse du nuage est proportionnel à l'effectif de la classe. Cette analyse est une analyse inter, car la dispersion analysée est exactement la dispersion interclasse.

Techniquement, il suffit de construire le tableau qui croise les classes de la partition et les modalités des autres variables et de lui appliquer une analyse factorielle des correspondances (AFC) classique. Ce tableau s'obtient en sommant les lignes du tableau disjonctif complet (TDC) qui correspondent aux individus d'une même classe Dans cette opération, la marge définie sur l'ensemble J des modalités est conservée, ce qui assure que la métrique définie sur $R^J$ est la même que celle induite par le tableau disjonctif complet. D'autre part, en AFC,

la somme de plusieurs lignes est toujours située au barycentre de ces lignes, et le poids qui lui est affecté est la somme des poids de ces lignes.

Dans cette analyse, il est intéressant de mettre le TDC en lignes supplémentaires. On obtient alors la projection des individus initiaux sur les axes et donc une représentation du nuage exact sur les axes d'inertie inter.

Cette analyse est équivalente à celle d'un tableau de même dimension que le TDC dans lequel chaque ligne est la moyenne des lignes de la même classe. Notons que les marges sur I de ce tableau et du TDC étant identiques, les colonnes des deux tableaux sont situées dans le même espace euclidien. Les profils des colonnes du nouveau tableau sont les projections des profils du TDC sur le sous espace engendré par les indicatrices des classes de la partition. Il est donc inutile de mettre en supplémentaire les colonnes du TDC qui se projettent aux mêmes points que les colonnes actives. Mais, par contre, il peut être intéressant d'introduire d'autres variables qualitatives illustratives pour visualiser les positions des barycentres de classes définies par les modalités de variables extérieures.

Notons que le cas des variables qualitatives est beaucoup plus riche que celui des variables numériques. En effet, pour l'illustrer, prenons l'exemple d'une variable décrivant une teinte variant du blanc au noir. La moyenne de cette variable numérique sur une classe d'individus comprenant une moitié de noir et une moitié de blanc est un simple gris et donc identique à celle d'une classe composée d'éléments gris ; tandis que la moyenne de la même variable codée en qualitative traduit la répartition dans les différents niveaux de gris et sépare très bien les deux situations.

## 4. GRAPHE DE PARTITION : ACM CONDITIONNELLE

Dans le cas d'un graphe de partition, l'analyse des différences locales consiste à représenter chaque individu i par un point dont les coordonnées sont les différences entre les coordonnées de cet individu dans l'ACM et celles du barycentre de sa classe. Nous avons étudié cette transformation et introduit une méthode appellée l'analyse des correspondances multiples conditionnelle [Esc 87 et Esc 90b] Pour obtenir les résultats, il suffit d'appliquer un programme d'AFC à un tableau qui dérive soit du TDC, soit du tableau de Burt. Cette facilité technique ne doit pas cacher le fait qu'il s'agit, comme l'ACM d'une méthode à part entière. La partition définit une nouvelle variable qualitative qui sert à conditionner les autres

variables. L'ACM conditionnelle dérive de l'ACM en travaillant sur les variables ainsi conditionnées. Elle peut s'introduire et se justifier théoriquement suivant toutes les présentations classiques de l'ACM (AFC d'un TDC, d'un tableau de Burt, analyse multicanonique..). On obtient une typologie des individus, caractérisés par les variables ainsi conditionnées, liée par des formules de dualité à une typologie des indicatrices conditionnées elles aussi. Le tableau qui dérive du tableau de Burt traduit, comme ce dernier, toutes les liaisons binaires, mais conditionnées par la variable extérieure.

## 5. ACM LISSEE POUR UN GRAPHE BISTOCHASTIQUE

L'ACM lissée est une généralisation immédiate de "l'analyse inter" présentée comme une AFC du tableau de même dimension que le TDC : nous proposons de traiter par l'AFC un tableau dans lequel chaque ligne i est la moyenne pondérée (par les poids $g_{ii'}$) des lignes du TDC qui appartiennent au voisinage de i [Esc 89]. Notons L ce tableau, $l_{ij}$ son terme général, K désigne le tableau disjonctif complet dont le terme général $k_{ij}$ vaut 0 ou 1.On a :

$$L = N^{-1}GK \quad \text{avec } l_{ij} = \sum_{i'} g_{ii'}\, k_{i'j}/g_{i.}$$

### 5.1. Le tableau lissé

Ce tableau n'est pas un tableau disjonctif complet mais il en garde certaines propriétés : la somme sur les modalités d'une même variable des éléments d'une ligne quelconque i vaut 1 et la somme totale d'une ligne i est égale au nombre total de variables. Ce tableau correspond à ce que l'on appelle souvent un "codage flou". Il a pour marges :

$$l_{i.} = \sum_j l_{ij} = \sum_j \sum_{i'} (g_{ii'}\, k_{i'j}/g_{i.}) = k_{i.}$$

$$l_{.j} = \sum_i l_{ij} = \sum_i \sum_{i'} (g_{ii'}\, k_{i'j}/g_{i.}) = \sum_{i'} (g_{.i'}/g_{i.})\, k_{i'j}$$

La marge sur I du tableau lissé L est donc égale à celle du TDC. Quand $g_{i.}$ est égal à $g_{.i'}$ pour tout i et i' (cas du graphe bistochastique) la marge sur J du tableau analysé est égale à celle du TDC. Dans la suite de ce paragraphe nous supposons que cette condition est vérifiée par le graphe. On a alors :

$$l_{i.} = k_{i.} \qquad l_{.j} = k_{.j} \qquad L = GK$$

Le cas d'un graphe quelconque est étudié dans le paragraphe 7.

### 5.2. Les individus

Chaque indidividu est caractérisé par la *répartition de son voisinage dans les différentes modalités des variables*. On retrouve encore, bien entendu, la richesse qui dérive de la nature qualitative des variables.

La condition imposée sur le graphe implique que la marge sur J est égale à celle du TDC, la métrique de $R^J$ définie dans l'AFC de L est donc la même que celle de l'ACM. La ligne i de L étant la moyenne pondérée de son voisinage dans le TDC, alors, dans l'espace vectoriel $R^J$, son profil est situé au barycentre des profils des lignes de son voisinage. Son poids est le même que dans le TDC : tous les individus ont donc le même poids. Le nuage d'individus se déduit donc du nuage défini dans l'ACM en remplaçant chaque point par le barycentre de son voisinage.

Comme dans l'analyse inter, il est utile de mettre le TDC en lignes supplémentaires pour avoir une représentation du nuage initial. Cette représentation qui est une projection sur des axes moins dépendants des variations locales que ceux de l'ACM peut être le résultat essentiel de l'analyse. On peut ensuite utiliser ces projections sur les premiers facteurs dans une classification.

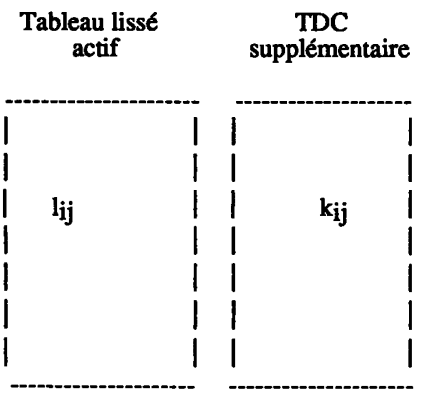
Ceci permet aussi, en comparant les positions des individus initiaux et lissés, de repérer les individus stables (accordés à leur voisinage) et instables (qui diffèrent de leur voisinage) et de juger suivant l'importance des variations des individus, l'influence du lissage sur leur nuage. Cependant, en ACM, les individus sont souvent trop nombreux pour qu'il soit possible de s'intéresser à chacun d'entre eux. On étudiera plutôt la stabilité des classes d'individus définies par les variables de l'ACM comme nous le précisons dans le paragraphe suivant.

### 5.3. Les modalités

La marge sur I de L étant la même que celle du TDC les modalités sont représentées dans le même espace euclidien que les indicatrices du TDC. Ce ne sont pas des variables indicatrices comme dans le TDC, mais des "indicatrices floues" qui prennent sur chaque individu une valeur comprise entre 0 et 1. Leur projections sur les facteurs s'interprétent sans

difficulté. Comme en ACM, le barycentre des modalités d'une même variable est confondu avec le barycentre de l'ensemble des modalités et les facteurs opposent donc entre elles les modalités de chaque variable.

Il est intéressant de mettre le TDC aussi en colonnes supplémentaires (cf. schéma ci-dessous). On obtient ainsi les projections des indicatrices sur des axes moins dépendants des variations locales que dans l'ACM, ce qui a un intérêt en lui-même, mais on obtient aussi (à un coefficient près) les barycentres des classes d'individus définies par les variables qualitatives. L'étude de ces projections et leur comparaison avec les "indicatrices floues" permet d'étudier l'influence du lissage sur chaque indicatrice et par conséquent les relations entre les variables analysées et le graphe de proximité : lorsqu'une "indicatrice floue" est proche de l'indicatrice initiale, la classe définie par cette indicatrice est composée d'éléments proches du point de vue du graphe. Inversement, une grande distance entre les deux indicatrices traduit une classe très dispersée par le graphe.




### 5.4. Liaison entre les modalités et les individus

Les formules de transition s'interprètent aussi très facilement en référence à cette notion "d'indicatrice floue". On note Fs (resp. Gs) le facteur d'ordre s défini sur les individus (resp. les variables). Nous avons les formules de transitions suivantes :

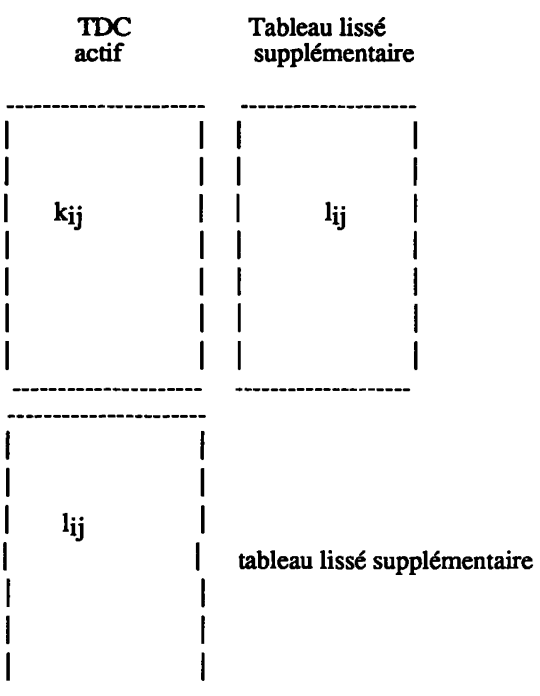
$$F_s(i) = 1/\sqrt{\lambda_s} \ \sum_j l_{ij}/k_{.j} \ G_s(j)$$

$$G_s(j) = 1/\sqrt{\lambda_s} \ \sum_i l_{ij}/k_{i.} \ F_s(i)$$

Ces relations impliquent qu'un individu est situé du côté des modalités que son voisinage possède fréquemment . Inversement, une modalité est attirée par les individus dans le voisinage desquels elle apparait souvent.

### 5.5. Tableau lissé en supplémentaire dans l'ACM

Pour étudier les relations entre les indicatrices et le graphe, au lieu de mettre le TDC en éléments supplémentaires dans l'AFC du tableau lissé, on peut inverser les rôles des deux tableaux en introduisant en éléments supplémentaires dans l'ACM le tableau lissé en colonnes. On peut l'introduire aussi en lignes supplémentaires pour étudier la stabilité de chaque individu.



## 6.ACM DES DIFFERENCES LOCALES

Dans le paragraphe précédent, nous avons diminué l'influence des variations locales, nous cherchons maintenant, au contraire, à analyser ces variations.

Nous supposons encore ici que le graphe de proximité est bistochastique (les autres cas sont évoqués au paragraphe 8). Les tableaux K et L ont alors les mêmes marges sur I et J, marges qui servent comme dans l'ACM à définir les métriques pour analyser les profils des lignes et des colonnes du tableau K - L de terme général $d_{ij} = k_{ij} - l_{ij}$

Posons $G_I$ la matrice ligne de terme général $k_{i.}$ et $G'_J$ la matrice colonne de terme général $k_{.j}$. Nous proposons d'appliquer un *programme d'AFC* au tableau R de terme général $r_{ij}$ :

$$R = K - L + G_I.G_J/k..$$

avec $r_{ij} = d_{ij} + k_{i.}\, k_{.j}/k..$ et $k.. = \sum_j k_{.j} = \sum_i k_{i.}$

### 6.1. AFC du tableau R

Le tableau R n'est autre que la différence entre K et le tableau lissé L à laquelle on ajoute le produit des marges sur I et sur J (communes aux deux tableaux). Ce tableau peut comporter des termes négatifs. Il ne peut pas être considéré comme un tableau de contingence et l'application d'un programme d'AFC n'est qu'une technique pratique pour analyser, par une généralisation de l'analyse des correspondances [Esc 83 et 85], l'écart entre un tableau de données (ici le TDC) et un tableau "modèle" (ici le tableau lissé) de mêmes dimensions et de mêmes marges que le tableau de données. Les deux "marges" du tableau résidu sont égales à celles des deux tableaux. Dans le nuage centré des lignes, chaque point i représente la différence entre le profil de la ligne i dans le tableau de données et son profil dans le tableau modèle. Il en est de même dans le nuage des colonnes. Les métriques des espaces euclidiens sont celles qui sont définies dans les AFC du tableau de données et du tableau modèle. Les relations de transition impliquent que, sur un facteur, les lignes sont situées du côté des colonnes auxquelles elles s'associent plus dans les données que dans le modèle et à l'opposé de celles auxquelles elles s'associent moins dans les données que dans le modèle (et inversement).

### 6.2. Le nuage des individus

Notons que, comme dans les deux autres analyses tous les individus ont des poids égaux. Dans l'AFC de R, un individu est représenté par la différence entre son profil dans le TDC et son profil dans le tableau lissé ; c'est-à-dire par la différence entre ses caractéristiques et la moyenne ou, plus exactement puisqu'il s'agit de variables qualitatives, la répartition des caractéristiques de ses voisins.

Les individus "isolés" (dont l'ensemble des caractéristiques diffère de celles de leur voisins) sont donc éloignés de l'origine et mis en évidence par cette analyse tandis que ceux qui ressemblent beaucoup à leur voisinage sont proches de l'origine.

Les individus dont les différences avec leur voisinage sont analogues seront proches. Mais le cas est différent de celui des variables numériques où deux individus ayant des valeurs initiales très différentes peuvent être très proches dans l'analyse des différences locales. En effet, ici, pour des variables qualitatives, la coordonnée d'un individu i, sur l'axe j, dans l'AFC de R vaut :

$$r_{ij}/r_{i.} = k_{ij}/k_{i.} - l_{ij}/l_{i.} = (k_{ij} - l_{ij})/k_{i.}$$

Cette coordonnée est positive (ou nulle) si i possède la modalité j ( $k_{ij}=1$ et $l_{ij}\leq 1$) et négative (ou nulle) si i ne possède pas la modalité j ( $k_{ij}=0$ et $l_{ij}\geq 0$). Pour que deux individus éloignés de l'origine soient proches, il est donc nécessaire que, d'une part, *ils aient beaucoup de modalités communes et que, d'autre part, la répartition des modalités de leurs voisins soit analogue*. Cette technique permet donc d'analyser les points avec leur contexte.

### 6.3. Le nuage des modalités

Une modalité est représentée par la différence entre son profil dans le TDC et son profil dans le tableau lissé. Dans cette analyse, les modalités d'une même variable ont encore leur barycentre à l'origine.

Les modalités loin de l'origine sont celles qui regroupent des points qui ne sont pas voisins les uns des autres, alors que les modalités concernant des individus contigüs (définissant des zones très connexes), qui sont donc des modalités très liées au graphe, sont proches de l'origine (comme en ACM un effectif faible augmente la distance à l'origine).

L'analyse met en évidence, aux extrémités des axes, les modalités peu liées au graphe et permet ainsi *d'étudier la liaison entre l'ensemble des variables et la structure de proximité*. Dans cette optique, les résultats sont plus simples à dépouiller que dans les analyses précédentes (AFC de L avec le TDC en supplémentaire et inversement) où l'on devait comparer les deux projections de chaque modalité. Dans ces dernières, les modalités bien représentées et expliquées par l'analyse étaient celles qui sont liées au graphe, alors qu'ici, inversement, ce sont celles qui sont peu liées.

Deux modalités loin de l'origine sont proches lorsque ce sont les mêmes individus qui, soit possèdent ces modalités sans que leur voisins généralement la possède, soit ne la possèdent pas alors que leurs voisins la possèdent généralement.

### 6.4. Liaison entre les modalités et les individus

Avec les notations du paragraphe 5.4, les formules de transiton s'écrivent :

$$F_S(i) = 1/\sqrt{\lambda_S} \sum_j (k_{ij} - l_{ij}) / k_{.j} \; G_S(j)$$

$$G_S(j) = 1/\sqrt{\lambda_S} \sum_i (k_{ij} - l_{ij}) / k_{i.} \; F_S(i)$$

Les relations de transition impliquent qu'un individu est situé du côté des modalités qu'il posséde et que ses voisins ne possèdent guère et à l'opposé des modalités que non seulement il ne possède pas, mais que beaucoup de ses voisins possèdent. Réciproquement, une modalité est située du côté des individus qui la possèdent alors que ses voisins ne la possèdent pas.

### 6.5. TDC en supplémentaire

Le TDC peut être mis en lignes supplémentaires dans cette analyse. On obtient ainsi la projection du nuage d'individus défini dans l'ACM. Mais cette projection présente rarement de l'intérêt car la structure sur laquelle les axes sont définis est fondamentalement différente de la structure initiale du TDC. La projection du TDC en colonnes supplémentaires donne, à un coefficient près, les barycentres des classes d'individus de cette analyse.

## 7. ACM LISSEE POUR UN GRAPHE QUELCONQUE

Dans le cas d'un graphe de proximité quelconque, une propriété très importante est perdue : le tableau lissé n'a plus la même marge sur J que le TDC. Ceci implique que :

1) la métrique de l'espace $R^J$ définie dans l'AFC du tableau lissé n'est pas égale à celle qui est utilisée dans l'ACM. Le nuage d'individus analysé (ou projeté en élément supplémentaire) dans cette AFC est donc déformé par le changement de métrique.

2) le poids des modalités dans cette AFC n'est pas le même que dans l'ACM.

La première conséquence est théoriquement gênante. En réalité la plupart du temps les deux marges sont assez peu différentes et cette déformation est si faible qu'il est possible de ne pas en tenir compte.

Ce problème n'apparait pas lorsque le tableau lissé est introduit en éléments supplémentaires dans l'ACM (cf § 5.4).

## 8.ACM DES DIFFERENCES LOCALES, GRAPHE QUELCONQUE

L'analyse des différences entre les profils des individus dans le TDC et dans le tableau lissé est encore possible, mais beaucoup moins satisfaisante .

Pour procéder à cette analyse, les deux marges des deux tableaux n'étant pas identiques, il faut appliquer un programme spécifique [Esc 85].Dans cette étude il faut préciser la métrique de l'espace de représentation des différences des profils; on peut choisir l'une ou l'autre des métriques induite par $k_{.j}/k_{..}$ ou $l_{.j}/k_{..}$ . La différence entre ces deux marges implique que dans l'analyse, le nuage n'est pas centré. Il existe une représentation duale du nuage des colonnes mais il ne représente pas exactement les différences des profils des colonnes des deux tableaux. Les formules de transition sont aussi perturbées.

Si les deux marges $k_{.j}$ et $l_{.j}$ ne sont pas très différentes, on pourra négliger ces écarts dans l'interprétation des résultats. Si ces deux marges sont très différentes, l'intérêt de cette analyse est très limité. Dans le cas de marges très proches, il est aussi possible d'appliquer un programme d'AFC au tableau $r_{ij}$ défini ci-dessus même si, théoriquement, toutes les bonnes propriétés de cette analyse sont perdues.

## 9.ANALYSE FACTORIELLE MULTIPLE AVEC UN GROUPE DE COORDONNEES

Dans la technique que nous proposons maintenant, la notion de proximité ou de contiguïté n'est pas utilisée sous forme de graphe mais d'un ensemble de variables numériques. Dans le cas de données géographiques ou de zones d'une image, ces variables sont les deux coordonnées dans le plan. Plus généralement, lorsque la proximité est induite par une position dans un espace vectoriel, on pourra prendre un système de coordonnées de cet espace. Si la structure est définie directement par un graphe, on peut se ramener à une situation analogue en appliquant une analyse des correspondances à la matrice du graphe. En effet, d'après les travaux de Lebart [Leb84], les premiers facteurs de cette analyse forment un système de coordonnées spatiales qui traduisent bien les contiguïtés.

L'Analyse Factorielle Multiple [Esc 90a] est une méthode d'analyse factorielle qui permet notamment de traiter simultanément des groupes de variables numériques et des groupes de variables qualitatives en équilibrant leur influence respective. Le groupe de variables qualitatives et la proximité jouent alors un rôle symétrique. Ce qui n'était pas du tout le cas dans les analyses précédentes. Comme dans l'analyse lissée, les facteurs auront tendance à regrouper des zones assez connexes.

L'Analyse Factorielle Multiple (AFM) permet aussi de comparer différents groupes de variables en comparant les structures induites sur le même ensemble des individus. Des indices permettent notamment de repérer l'existence de directions de dispersion semblable (appellés facteurs communs), de mesurer l'importance relative de ces directions pour chaque groupe de variables. On mettra ainsi en évidence l'existence et l'importance relative dans l'ensemble des variables qualitatives étudiées de structures liées à la proximité ou au contraire indépendantes de celle-ci. Ceci permet d'étudier les relations entre les variables qualitatives et la structure de proximité d'une manière plus globale que par les deux méthodes précédentes.

L'AFM permet aussi de superposer les graphiques de ces directions communes afin de repérer des éléments dont le comportement n'est pas cohérent, ici des individus qui, contrairement à l'ensemble, ont des valeurs qui ne s'accordent pas à leur voisinage.

## 10.EXEMPLE : SEGMENTATION D'IMAGE

### 10.1 Les données image (figures en fin de texte)

Les méthodes que nous proposons sont appliquées sur deux images. L'une l'image "VOITURE" (Fig. 2) qui provient du CCETT est une image entièrement naturelle où apparaissent différents types de texture : une grille, un fond de broussailles, des zones uniformes comme la calandre ou les roues ....L'autre, venant de l'album de BRODATZ [Bro 66] (Fig. 1), est constituée par un assemblage de cinq textures réelles. La première est composée de 512 fois 512 pixels, l'autre de 256 fois 256 pixels.

La texture microscopique en chaque point est caractérisée par un vecteur d'attributs. Ces attributs sont calculés sur une fenêtre de taille et de forme modulables extraite autour de chaque point d'analyse. Celle-ci est déterminée par la méthode de BARBA-RONSIN [Bar83] et sa forme traduit la texture locale. Schématiquement, on part du point d'analyse dans les 16 directions principales et dans chaque direction, on s'arrête d'autant plus vite que les variations de luminance sont importantes. Les points d'arrêt déterminent un polygone alongé dans les directions les plus homogènes.

Sur cette fenêtre, on calcule des attributs qui décrivent :
- la forme et l'orientation de la fenêtre
- la distribution statistique de la luminance sur la fenêtre
- la répartition des valeurs de la luminance dans différentes orientations.

Mais parmi ces attributs, il y en a qui apportent la même information texturale. Afin de tester les redondances et chercher les plus significatifs, de façon objective et indépendamment de l'image et de ses propriétés texturales, une AFM a été effectuée [Dan 89]. L'interprétation des résultats de celle-ci a permis, en effet, de dégager un ensemble réduit de paramètres non redondants.

Les paramètres retenus, à l'issue de cette AFM sont :
- deux paramètres structuraux : *orientation et excentricité,*
- trois paramètres de luminance : *moyenne, skewness et kurtosis,*
- quatre paramètres de co-occurence (calculés pour deux orientations différentes) :

*homogéneité, énergie et deux moments dipolaires* (pour caractériser des textures microscopiques).

L'ensemble ainsi défini constitue la base de paramètres adaptés aux caractéristiques des images que nous avons analysées.

Nous avons pris, arbitrairement, 2601 points de mesures dans chaque image. Ces points (51 par 51) ont été obtenus par un échantillonage régulier. La surface moyenne des fenêtres est du même ordre de grandeur que le pas d'échantillonage.

### 10.2 Méthodologie

Les paramètres sont codés en variables qualitatives, définies par division en cinq classes d'effectifs égaux, sur les intervalles de valeurs calculées initialement.

Sur l'image VOITURE, nous appliquerons les techniques suivantes.

- Une ACM sans aucune notion de proximité (le résultat servira de référence à des comparaisons),
- une ACM sur un codage qualitatif des données numériques lissées
- une ACM lissée avec deux pondérations différentes,
- une ACM des différences locales
- une AFM avec un groupe de coordonnées de position dans le plan.

Sur l'image BRODATZ nous appliquons une ACM classique, une AFM avec groupe de coordonnées, puis une ACM lissée.

Une classification ascendante hiérarchique (critère de Ward, méthodes des voisins réciproques) est appliquée sur les facteurs des différentes analyses. Elle est suivie d'une partition.

A partir des résultats de la partition, on réalise la segmentation de l'image. Pour visualiser le résultat de la segmentation on construit une image où chaque bloc, centré sur un point de la grille, est rempli avec une "fausse couleur" (ici des niveaux de gris) qui correspond à la classe du point analysé. Notons que l'échantillonage qui réduit considérablement les calculs introduit un défaut dans la segmentation et surtout dans sa représentation. La classe de texture d'une fenêtre de forme et de surface variable entourant un point d'échantillonnage est affectée dans sa représentation à un bloc carré.

**Le choix de la pondération dans les lissages**

Le nombre de voisins d'un point de l'image (les plus proches voisins) varie de huit points dans une zone centrale, à trois points pour une zone située dans le coin, en passant par cinq points dans les bords de l'image. Nous testerons deux pondérations différentes qui correspondent toutes deux à un graphe bistochastique.

Dans l'une, "uniforme", chaque voisin a un poids de 1/10 et on affecte le poids 2/10 pour un point central, le poids 7/10 pour les quatres coins et le poids 5/10 pour un point du bord de l'image. La somme des poids est égale à 1 quelque soit le voisinage. Le point courant, pour une zone qui n'est pas un bord, a ici moins d'importance que l'ensemble de ses voisins.

Dans l'autre "contexte" le poids du contexte, défini par la somme des poids des voisins, égale le poids du point de mesure ; avec la contrainte d'une somme totale égale à 1, on donne un poids de 1/2 pour un point courant et les poids 1/16, 1/6, 1/10 pour les voisins respectivement pour un point central, un point du coin et un point du bord de l'image.

### 10.3 Résultats de l'image "voiture"

Dans toutes nos analyses factorielles, nous avons retenu 25 axes correspondant à plus de 80% d'inertie expliquée. Nous avons vérifié que le nombre de facteurs retenus dans la classification n'influence pratiquement pas les partitions obtenues.

La partition en dix classes que nous avons choisie se justifie, d'une part par rapport aux différents types cellulaires apparaissant dans l'image, d'autre part par la forme de "l'histogramme" des pertes d'inertie. (La présence d'un palier dans l'histogramme des pertes d'inertie indique généralement, une coupure significative en un nombre réduit de classes).

#### 10.3.1 ACM classique (Fig. 3)

Le résultat de la segmentation obtenue après l'ACM sans introduction de la proximité, présente un grand éparpillement des zones et donc peu de classes homogènes et peu de connexité. La description par les paramètres seuls, permet néanmoins de reconnaître, à l'oeil, certaines entités sémantiques: la grille, les roues, la toiture...

### 10.3.2. AFM avec un groupe de coordonnées (Fig. 4 et 5 )

L'AFM, avec les deux coordonnées du point d'échantillonage dans le plan de l'image, apporte une amélioration sensible. Les zones appartenant à un même voisinage sont regroupées dans une même classe, ce qui réduit l'éparpillement et augmente la connexité. Mais les zones éloignées, dans le plan de l'image, appartiennent à des classes différentes même si elles se ressemblent d'un point de vue texture. Par exemple, dans l'arrière plan de l'image où le fond de broussaille forme à droite et à gauche deux classes différentes.

Ceci s'explique très bien puisque la technique utilisée donne autant d'importance aux coordonnées qu'à la texture. Il est possible d'effectuer un nouveau regroupement des classes obtenues par AFM en considérant seulement les caractéristiques de texture; techniquement cela consiste à croiser les classes avec les modalités des variables qualitatives et à effectuer une classification sur les facteurs issus de l'AFC du croisement. Le nombre de classes finales est alors nécessairement inférieur à celui obtenu par AFM seulement. Dans notre exemple on passe de dix classes à cinq classes significatives (au sens de la perte d'inertie), et la qualité de la segmentation est meilleure (Fig. 5).

### 10.3.3 ACM à partir de données numériques lissées (Fig. 6)

Pour comparer les effets d'un lissage de variables numériques et de variables qualitatives, nous avons appliqué une ACM sur les paramètres numériques lissés avant transformation en variables qualitatives (pondération "uniforme" ). Une luminance moyenne lissée, par exemple, sera codée par une valeur "gris" si la zone considérée est noire et entourée de zones blanches ; aussi, deux zones aux caractéristiques différentes peuvent se regrouper.

Les classes obtenues sont nettement moins dispersées que dans l'ACM sans contiguïté. La grille -barreaux et interstices- ne forme plus qu'une entité ; cette classe se retrouve dans d'autres parties de l'image de texture très différente. La segmentation parait moins claire que celles qui sont issues de l'AFM.

### 10.3.4. ACM lissée (Fig. 7 et 8)

L'ACM lissée ou AFC du tableau lissé, pour la même pondération "uniforme" (fig. 8), aboutit à des classes beaucoup plus connexes que le lissage préalable des variables

numériques : ce lissage qui consiste à considérer la répartition du voisinage d'un point dans les différentes modalités des variables est à la fois plus riche et plus stable qu'un lissage numérique. Les classes obtenues sont très cohérentes avec l'analyse visuelle de l'image initiale. La pondération utilisée, qui donne peu d'importance au point d'analyse, écrase beaucoup les disparités locales. Dans la grille, par exemple, barreaux et interstices ne forment qu'une même classe. La pondération "contexte" (fig.7) qui donne plus d'importance au point courant aboutit à des classes beaucoup plus éparpillées et sépare ces deux entités.

### 10.3.5. ACM des différences locales (Fig. 9)

Dans les analyses précédentes nous avons diminué l'influence des variations locales. L'ACM des résidus du lissage, i.e. des différences locales, (cf. paragraphe 6), nous permet d'analyser ces variations locales.

Le résultat de la segmentation, issue de cette analyse, avec la pondération "uniforme", met en relief, de façon nette, l'entité "grille", où les barreaux et leurs interstices forment deux classes typiques. Les autres classes sont très éparpillées. Ceci s'explique car deux zones se regroupent si, à la fois leurs caractéristiques et celles de leurs voisinages se ressemblent. Cette double condition ne permet de mettre en relief que des zones très "typées" (cas de la "grille", par exemple).

### 10.3.6 Conclusion

Cet exemple illustre les possibilités et les qualités des différentes méthodes proposées, mais il est difficile sur cette image d'apporter un jugement précis sur la qualité des segmentations. En effet, dans ce cas, celle-ci dépend du but poursuivi. Pour pouvoir en juger et même la mesurer, il faut connaitre une segmentation donnée a priori. C'est le cas, un peu artificiel, de la deuxième image que nous étudions. Dans de nombreux cas réels une segmentation optimale peut ainsi être définie a priori.

### 10.4. IMAGE BRODATZ

La nature de cette image (Fig.1) (assemblage de cinq textures naturelles distinctes) rend plus aisé le jugement sur la qualité de toute segmentation.En effet, nous disposons d'une connaissance a priori simple: cinq classes différentes. Aussi nous effectuerons une partition en cinq classes qui se justifie aussi au vu de l'histogramme des pertes d'inertie.

### 10.4.1. ACM classique (Fig.10)

Le résultat montre un assez grand éparpillement des 5 classes de texture. Cependant, on différencie relativement bien la texture située au centre et les deux textures situées sur la droite. Par contre, les deux textures situées sur la gauche sont complètement mêlées entre elles.

### 10.4.2. AFM avec groupe de coordonnées (Fig. 11)

La définition par l'introduction de coordonnées, dans le plan de l'image, dans l'AFM, doit fournir, en particulier, une caractérisation de classes extrêmes correspondant aux quatre quadrans du plan et à la région centrale du même plan. Les résultats expriment parfaitement cette caractérisation. La segmentation est grandement améliorée, les 5 textures se différencient globalement, mais comme pour l'image VOITURE, un relatif éparpillement des classes existe encore.

### 10.4.3. ACM lissée (Fig.12)

La connexité et l'homogénéïté des cinq classes sont très nettement améliorées. Les résultats sont presque parfaits. Le "petit" mélange observé sur la figure s'explique localement ; en effet, il existe au niveau microscopique, des textures trés semblables, constatées sur l'image initiale.

Cette analyse est donc tout à fait adaptée à ce type de texture. En lissant le tableau disjonctif complet, on prend en compte la répartition de la texture locale autour d'un bloc, ce qui revient à une définition de la texture à un double niveau.

C'est le meilleur résultat obtenu sur cette image. Il est encourageant mais la diversité des textures naturelles nous interdit de penser qu'il s'agit d'un outil universel. De plus, le choix de paramètres (dans le calcul de la fenêtre et dans la pondération rend délicate son application).

# 11. STRUCTURE DE PROXIMITE SUR LES MODALITES

Il n'est guère envisageable d'introduire n'importe quelle structure de proximité sur l'ensemble des modalités des variables qualitatives

Il peut y avoir une structure de proximité sur les modalités d'une même variable, notamment lorsqu'il s'agit d'une variable numérique codée par classe. Des codages "flous" ont été proposés [Gal 82] par exemple pour tenir compte de la position d'un point proche d'une borne de classes. Mais la problématique est un peu différente.

Le cas d'une variable répétée dans le temps est plus proche de nos préocupations. Le temps introduit une notion de proximité sur cette suite de variables et donc sur la suite de ses modalités. La matrice du graphe associé se décompose en blocs ; les seuls blocs non nuls sont ceux qui sont définis par une même modalité aux différentes époques. Le graphe peut être pondéré pour tenir compte avec souplesse de l'ensemble des périodes précédentes et suivantes. Les équivalents de l'analyse lissée et de l'analyse des différences locales sont envisageables mais il n'y pas d'équivalent de l'AFM dans cette approche.

## 11.1. Analyse lissée

Dans ce cas comme dans celui des individus, on peut souhaiter diminuer les variations temporelles. On peut tenter d'adapter l'analyse lissée.

Prenons, par exemple un tableau dans lequel toutes les variables sont indicées par le temps. On notera $K_{IJT}$ le tableau disjonctif complet associé dans lequel une colonne a un double indiçage, jt , j désignant la modalité d'une variable et t désignant le temps.

Un lissage temporel consisterait techniquement à remplacer dans le tableau disjonctif complet $K_{IJT}$ chaque partie de ligne $k_{ijt}$ correspondant aux modalités d'une variable à un instant donné, par le barycentre $l_{ijt}$ de ses homologues aux différents moments. On note $L_{IJT}$ ce tableau. Le graphe est défini sur le produit JT, mais comme on suppose qu'il est indépendant de j, on le note $g_{tt'}$.

$$l_{ijt} = \sum_{t'} g_{tt'} k_{ijt'}$$

Etudions les propriétés de ce tableau lissé. On vérifie facilement que, dans chaque ligne, la somme des éléments d'une même variable ( à un temps donné ) vaut 1. Il correspond donc à un "codage flou" et sa marge sur I est constante et égale à celle du tableau disjonctif

complet. Il est donc tout à fait logique de lui appliquer une AFC. Dans cette AFC, tous les individus ont le même poids. Ceux qui, dans une étape de leur évolution, sont passés accidentellement dans d'autres classes de l'une ou l'autre des variables ne sont pas aussi éloignés de l'origine que dans l'ACM. Les colonnes sont des indicatrices floues, les modalités qui pouvaient avoir une position extrême en ACM du fait de phénomènes ponctuels sont un peu neutralisées, ce qui permet de mieux percevoir l'évolution temporelle générale. C'est en ce double sens que l'influence des variations temporelle est diminué. L'analyse lissée répond donc bien à la problématique. La marge sur J de ce tableau n'est pas la même que celle du TDC, même dans le cas d'un graphe bistochastique. Ceci ne pose pas de problème dans cette analyse, qu'il s'agisse du point de vue individus ou du point de vue modalités.

La mise en éléments supplémentaires des modalités d'origines peut être une aide intéressante à l'interprétation des résultats. En effet, si une modalité varie dans le temps de façon accidentelle, elle apparaitra dans cette analyse assez éloignée de son homologue lissée. Ceci permet de détecter à la fois les modalités présentant une fluctuation importante, et les périodes correspondant à des variations importantes.

### 11.2. Analyse des résidus

L'analyse des différences ponctuelles pose des problèmes car la marge sur J n'est pas celle du TDC. La situation est tout à fait analogue à celle du graphe quelconque pour les individus : l'analyse des différences entre les profils des lignes des deux tableaux est possible avec un programme spécial mais elle ne permet pas d'étudier les différences exactes entre les profils des colonnes. Elle éloigne de l'origine les individus au parcours très accidenté mais n'a de sens que si les deux marges sont assez peu différentes.

## 12.CAS DE L'ANALYSE DES CORRESPONDANCES SIMPLE

En analyse des correspondances simple, l'ensemble des lignes et des colonnes joue exactement le même rôle; mais il existe cependant de nombreux cas où une structure de proximité existe sur l'un des deux ensembles. Citons par exemple le cas d'un tableau de fréquence de mortalité par cause et par cantons. Les distributions sont définies par des effectifs trop faibles pour être très fiables et une AFC du tableau fait ressortir des accidents locaux et non des tendances générales liées à la structure spatiale. Il existe une structure de

contiguïté évidente sur l'ensemble des cantons et les raisons de la faire intervenir sont analogues à celles que nous avons évoquées pour des individus caractérisés par des variables qualitatives (cf. paragraphe 1). Nous allons tenter d'adapter l'analyse lissée à un tableau de fréquence quelconque en rappelant d'abord le cas d'un graphe de partition.

### 12.1. Graphe de partition : Analyse inter

Dans le cas d'un graphe de partition, l'analyse lissée consiste à remplacer le profil de la ligne i par le barycentre des profils des éléments de sa classe. Le poids affecté à ce barycentre dans l'analyse du nuage est la somme des poids des individus de cette classe. L'analyse lissée est alors exactement l'analyse interclasse.

Techniquement, la situation pour un tableau de fréquence quelconque reste analogue à celle d'un TDC, il suffit de construire le tableau cumulant les lignes du tableau de fréquence qui concernent les individus d'une même classe et de lui appliquer une AFC. Le profil de la somme de plusieurs lignes est toujours au barycentre des profils de ces lignes, si le poids de chacune d'entre elles est celui qui est défini en AFC. La marge sur J et de ce tableau cumulé est la même que celle induite par le tableau de contingence, donc la métrique induite sur RJ est la même.

On peut aussi considérer le tableau L de même dimension que K dont la ligne i est le produit du profil de la classe de i par la valeur de la marge $k_{i.}$. Les deux marges de L sont égales à celles de K et il est clair que l'AFC de ce tableau aboutit aux mêmes résultats.

### 12.2. Graphe de partition : Analyse Intra

Toujours dans le cas d'un graphe de partition, l'analyse des différences locales consiste à remplacer chaque point représentant un individu i par les différences entre les coordonnées de cet individu dans l'AFC et celles du barycentre de sa classe. Là, encore, la situation reste simple. Le tableau L ayant les mêmes marges que K, il suffit d'appliquer une AFC au tableau $R = K - L + G_I.G_J/k$. (cf. paragraphe 6).

Dans le cas d'un graphe de partition cette analyse décompose exactement l'inertie intraclasse et a déja été étudiée dans [Dro 83] et [Benz 83]. Récement une généralisation au cas d'une double partition sur l'ensemble des individus et des variables a été proposée par Cazes et al. dans [Caz 88].

### 12.3. Graphe bistochastique

Considérons le tableau lissé L défini, comme au paragrahe 6 pour l'ACM, par un graphe bistochastique.

$$l_{ij} = \sum_{i'} g_{ii'}\, k_{i'j}$$

Avec un tel graphe, l'influence de chaque ligne sur l'ensemble des données reste exactement celle qu'elle a dans le tableau initial K puisque $\sum_{i'} g_{ii'} = 1$ et la marge sur J de L est égale à celle de K :

$$l_{.j} = \sum_i l_{ij} = \sum_i \sum_{i'} (g_{ii'}\, k_{i'j}/g_{i.}) = \sum_{i'} (g_{.i'}/g_{i.})\, k_{i'j} = k_{.j}$$

Dans le cas général ce tableau L n'a pas la même marge sur I que le tableau K. Cette intéressante propriété est vérifiée pour des tableaux qui, comme les TDC ou par exemple les tableaux de notes dédoublées, ont une marge sur I constante. Pour un tableau quelconque, la valeur de $l_{i.}$ qui est la moyenne pondérée des valeurs de son voisinage est différente de $k_{i.}$ :

$$l_{i.} = \sum_{i'} g_{ii'}\, k_{i'.} \neq k_{i.}$$

Le poids affecté à i dans l'AFC de L est donc différent de celui qui lui est affecté dans l'AFC de K.

Le profil de la ligne i du tableau L s'écrit donc :

$$l_{ij} \,/\, l_{i.} = \sum_{i'} g_{ii'}\, k_{i'j} \,/\, \sum_{i'} g_{ii'}\, k_{i'.} = (\sum_{i'} g_{ii'}\, k_{i'.}\, (k_{i'j} \,/\, k_{i'.}) \,/\, (\sum_{i'} g_{ii'}\, k_{i'.})$$

C'est donc la moyenne pondérée des profils des lignes de son voisinage pour les poids $g_{ii'}\, k_{i'.}$ ; on retouve la logique de l'AFC qui affecte à chaque point une importance proportionelle à son effectif. Notons que ceci ne représente le profil de la région entourant i que si les poids $g_{ii'}$ sont égaux, condition difficile à respecter avec un graphe bistochastique.

Il est possible d'obtenir un tableau L' dont les profils des lignes sont identiques à ceux des lignes de L et qui a la même marge sur I. Il se déduit de L en multipliant chaque ligne i par le produit ki./li.. Mais ce tableau L' ne se déduit pas de K pas un lissage bistochastique et il n'y a aucune raison que sa marge sur J soit celle de K. Cette inégalité induit certaines conséquences faisant perdre à l'analyse des propriétés intéressantes (cf. paragraphe 7). En effet la métrique définie dans l'AFC du tableau lissée n'est pas égale à celle induite par le tableau initial. Le nuage ainsi analysé est alors déformé par cette métrique.

L'application de l'analyse des différences locales se heurte à un problème que nous avons déja rencontré aux paragraphes 8 et 11.2 : une des marges du tableau lissé est généralement différente de celle du tableau initial K.

L'analyse des différences entre le profils des individus dans le tableau de contingence et dans le tableau lissée n'est possible que si les deux marges des deux tableaux sont presque identiques. Dans le cas contraire, cette analyse présente peu d'intérêt (cf. paragraphe 8).

## 13.COMPARAISON AVEC D'AUTRES TECHNIQUES

Les techniques les plus proches sont celles qui utilisent la notion de graphe de contiguïté. L'analyse locale introduite par Lebart et Aluja [Leb 84] et [Leb 88] et implantée dans le logiciel SpadN [Spa 89] permet d'étudier les liaisons locales et a été utilisé notamment pour traduire des contiguïtés géographiques. La comparaison avec l'analyse des différences locales est étudiée dans [Ben 88] dans le cas des variables numériques, le cas des variables qualitatives est analogue. Rappelons simplement ici que le principe de l'analyse locale est de calculer des "inerties locales" en supprimant de l'inertie globale les couples qui ne sont pas reliés par le graphe, ce qui revient à faire une analyse d'un nuage d'arêtes du graphe au lieu du nuage d'individus. Reprise par Le Foll [LeF 82] et généralisée à un graphe pondéré, elle est aussi présentée par Carlier [Car 85] comme une analyse factorielle du tableau initial, l'espace où est situé le nuage des variables étant muni d'une semi-métrique. La différence essentielle entre ces deux analyses dont certains objectifs sont semblables est que dans l'analyse des différences locales, contrairement à l'analyse locale, les nuages d'individus et de variables s'interprêtent comme dans une analyse classique et se déduisent des nuages initiaux par des transformations simples.

On peut aussi rapprocher les techniques proposées de celles qui consistent à analyser des projections du nuage des variables sur des sous-espaces déterminés par un autre tableau pour trouver la part de variabilité qui dépend (ou ne dépend pas) de l'autre tableau (on en trouve un exemple dans [Dol 87]). Une étude systématique en est faite dans la thèse de Sabatier [Sab 87] qui reprend la terminologie de Rao [Rao 64] "analyse sur variables instrumentales".Ici, "l'autre tableau" est le graphe, mais les deux techniques proposées ne se traduisent par une projection que dans le cas d'un graphe de partition.

On peut encore rapprocher ces techniques de l'analyse canonique des correspondanes (ACC) introduite par Ter Braak [Ter 86] et [Ter 87] reprise et appliquée par Lebreton et al. à des variables qualitatives [Lebr 88 a et b]. Dans l'ACC, "l'autre tableau" est un tableau de fréquence (Dans l'exemple de [Lebr 88], c'est un tableau croisant des sites et des espèces tandis que les variables qualitatives sont définies sur les sites). Parmi ses multiples propriétés, l'ACC est une analyse des barycentres définis par le tableau de fréquence (dans l'exemple les espèces pondérées par leur effectif dans les sites). Cet aspect la rapproche de l'analyse lissée, et plus particulièrement de l'analyse statistique spatiale [Ben89] utilisant une métrique particulière qui maximise la variance de ces barycentres et en fait une généralisation de l'analyse discriminante [Che 87].

## 14.CONCLUSION

Une partie importante des données que nous traitons nécessite, en plus d'un tableau, la prise en compte d'information complémentaire. C'est le cas de l'analyse de la texture d'une image, de l'étude des causes de mortalité dans les différents cantons, où il existe une structure de proximité entre les éléments. L'introduction de cette information par un graphe pondéré, ou par un ensemble de variables numériques, nous permet de traduire la plupart des situations que nous avons rencontrées.

Nous avons montré dans l'exemple illustratif de segmentation d'image que les trois méthodes proposées étaient très simples à mettre en oeuvre en ACM. La contrainte sur le graphe, utile pour l'application de l'analyse de l'analyse lissée et de l'analyse des différences locales, n' a posé aucun problème et a permis au contraire de traduire de manière tout à fait cohérente les proximités. La mise en oeuvre de ces deux techniques est très simple puisqu'elles ne nécessittent qu'une transformation du tableau disjonctif complet et une AFC classique. L'interprétation et la richesse des résultats sont ceux d'une AFC. La prise en

compte de la proximité est très nette et dans le cas de l'image de Brodatz le résultat est tout à fait satisfaisant. Nous avons aussi montré l'importance du choix des poids du graphe qui déterminent les résultats. Dans ce choix, il parait difficile d'imposer, ou même de suggérer une règle, car l'importance à accorder au voisinage dépend de l'objectif poursuivi. Nous avons montré aussi que le lissage (et les différences locales) traduit sur des variables qualitatives était une notion beaucoup plus riche que sur des variables numériques puisqu'il traduit la répartition du voisinage et non seulement la moyenne. L'introduction de la proximité par un groupe de variables numériques dans une AFM est aussi très simple. Le problème de pondération du voisinage ne s'y pose pas puisque le rôle des groupes est automatiquement équilibré. L'analyse lissée et l'AFM rapprochent toutes deux les zones connexes mais donnent des résultats différents dont la valeur dépend de l'objectif.

Des classifications sur les facteurs issus de ces analyses ont permis d'introduire aussi la notion de proximité dans une classification sur variables qualitatives.

Pour l'ACM, nous disposons donc d'une série de techniques qui paraissent résoudre la plupart des problèmes à condition de bien choisir les paramètres, éventuellement après quelques tâtonnements.

Les deux techniques utilisant directement la notion de graphe peuvent se replacer dans un cadre géométrique très général : elles généralisent à une contrainte non linéaire la notion de variables instrumentales introduites par Rao [Rao 64]. Les nombreux liens que nous avons vus (paragraphe 13) avec d'autres méthodes soulignent encore leur intérêt méthodologique.

L'introduction de la proximité entre modalités en ACM et en AFC simple est possible mais les techniques sont plus lourdes à mettre en oeuvre et théoriquement moins satisfaisantes .

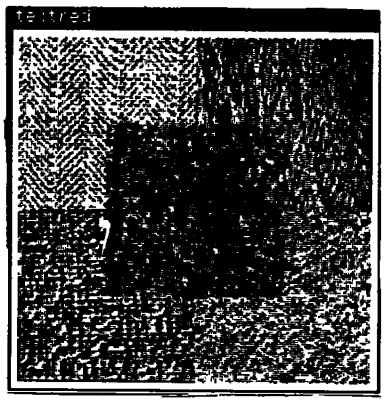


fig.1 Image BRODATZ

fig.2 Image VOITURE

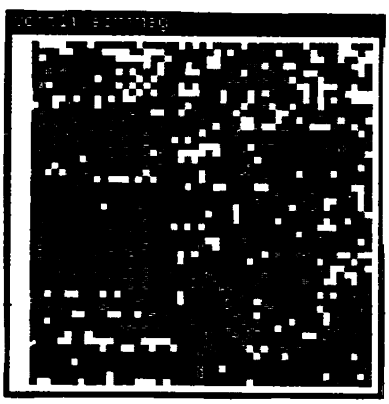


fig.3 ACM classique

fig.4 AFM

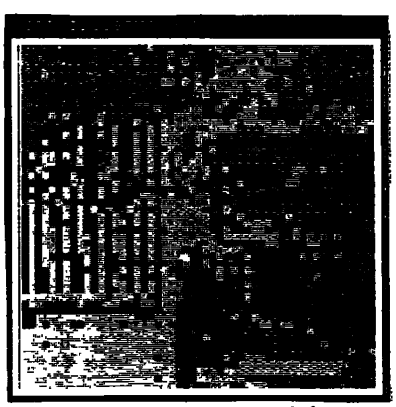


fig.5 AFM classes réduites

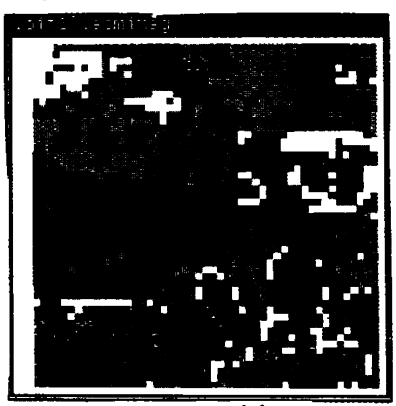


fig.6 lissage numérique

fig.7 lissage qualitatif

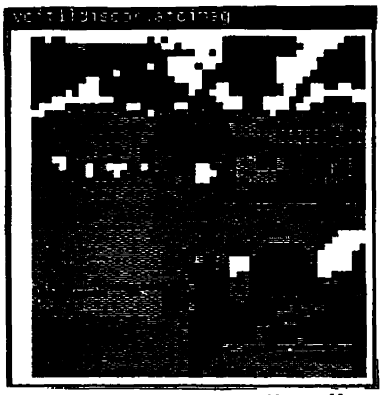


fig.8 lissage qualitatif uniforme

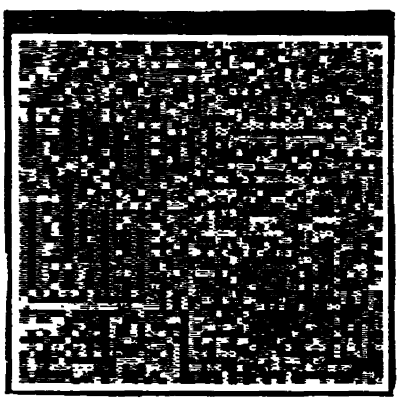


fig.9 AFC des résidus

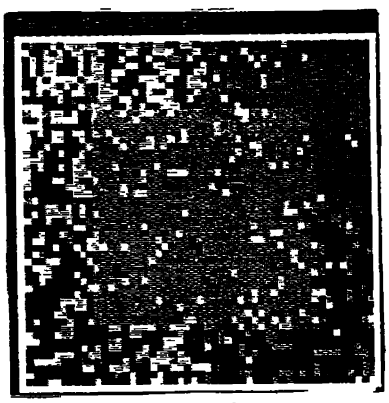


fig.10 ACM classique

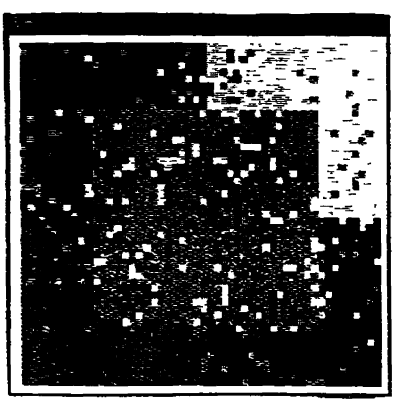


fig.11 AFM

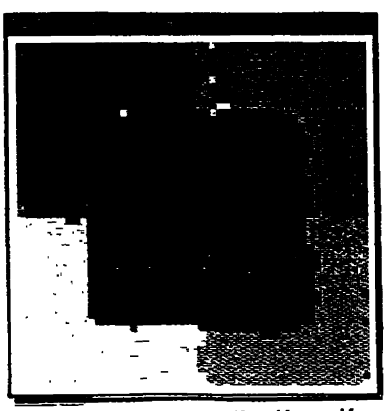


fig.12 lissage qualitatif uniforme

**Références**

[Aluj 83] **Aluja T. et Lebart L.** : Factorial Analysis upon a graph. Bulletin technique du centre de statistique et informatique appliqué. Vol. 3 pp 4-34,1983.

[Barb 83]**D.Barba J.Ronsin** "New method in texture analysis in the context of image segmentation ". In Eupsico 83, Erlangen ,1983.

[Ben 89] **Benali H.** : Analyse statistique spatiale. Application à la mortalité par cancer chez l'homme en France pour la période 1978-1984. Rapport Interne du CIRC A paraitre.

[Ben 89] **Benali H. et Escofier B.** : Smooth factorial analysis and factorial of local differences. Proc. Inter . Conf. on multiway matrices, North Holland Amsterdam, 1989

[Ben 90] **Benali H. et Escofier B.** Analyse factorielle lissée et analyse des différences locales. Revue de Statistique Appliquée XXXVIII (2) pp 55-76, 1990

[Benz 83] **Benzecri J.P.** : Analyse de l'inertie intraclasse par l'analyse d'un tableau de correspondance. C.A.D. VIII n°3 p.351-358, 1983

[Bro 66] **Brodatz P.** Textures : a photographic album for artists and designers. Dover Publications, Inc, New-York, 1966

[Car 85] **Carlier A.** : Applications de l'analyse factorielle de l'évolution et de l'analyse intrapériode. Statistique et analyse des données Vol. 10 pp 27-53, 1985

[Caz 88] **Cazes P. Chessel D. Doledec S.** : L'analyse des correspondances interne d'un tableau partitionné : son usage en hydrobiologie Revue de Statistique Appliquée XXXXVI n° 1 pp 39-54, 1988

[Che 87] **Chessel D.,Lebreton J.D.,Yoccoz N.**, Propriétés de l'analyse canonique des correspondances : une illustration en hydrobiologie Revue de Statistique Appliquée XXXV n°4 pp 55-77, 1987

[Dan 89] **Daniel T.** "Traitement numérique d'images appliqué à l'analyse texturale de roches déformées." Thèse Université Rennes1, 1989.

[Dol 87] **Doledec S. et Chessel D.** : Description d'un plan d'observation complet par projection des variables. Acta oecologica Oecol. Gen. Vol.8 n°3 pp.403-426, 1987

[Dro 83] **Drouet D. Escofier B.** Comparaison de plusieurs tableaux de fréquence. Cahiers de l'analyse des données. Vol. VIII n°4, 1983

[Esc 90 a] **Escofier B. Pages J.** : Analyse factorielles simples et multiples. Objectif, méthodes et interprétation Dunod Paris, deuxième édition, 1990

[Esc 90b] **Escofier B.** :Analyse des correspondances multiples conditionnelle.Revue de Modulad, n° 5 pp.1-12, 1990

[Esc 89] **Escofier B.** : Multiple Correspondence Analysis and neighbouring relation. Data Analysis. Learning symbolic and numeric knowledge. Nova Science Publisher pp. 55-64, 1989

[Esc 85] **Escofier B.** : Analyse factorielle en référence à un modèle. Application au traitement des tableaux d'échanges. Revue de statistique appliquée n°2, 1985

[Esc 87] **Escofier B.** : Analyse des correspondances multiples conditionnelle. Proc. Inter. Symp. Data Analysis and Informatics Versailles. pp.13-22 North Holland Amsterdam, 1987

[Esc 83] **Escofier B.** : Analyse de la différence entre deux mesures définies sur le produit de deux ensembles. Cahiers de l'analyse des données. Vol. VIII n°3 pp.325 -329, 1983

[Gal 82] **Gallego** Codage flou en analyse des correspondances. Cahiers de l'analyse des données. Vol. VII n° 4 pp.413-430, 1982

[Leb 84] **Lebart L.** : Correspondance analysis of graph structure. Bulletin technique du centre de statistique et informatique appliqué. Vol. 2 pp. 5-19, 1984

[Lebr 88a] **Lebreton J.D. Chessel D. Prodon R. Yoccoz N.** : L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. Variables de milieu quantitatives. Acta oecologica Oecol. Gen. Vol.9 n°1pp.53-67, 1988

[Lebr 88b] **Lebreton J.D.Chessel D.Richardot-Coulet M.,Yoccoz N.** : L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. Variables de milieu quantitatives Acta oecologica Oecol. Gen. Vol.9 n°2 pp.137-151, 1988

[LeF 82] **Le Foll Y.** : Pondérations des distances en analyse factorielle.Statistique et analyse des données n°7,1, pp.13-31, 1982

[Rao 64] **Rao C.R.** : The use and interpretation of Principal Component Analysis in applied research. Sankhya A 26,4 p.329-358, 1964

[Sab 87] **Sabatier R.** : Méthodes factorielles en analyse des données. Approximation et prise en compte de variables concomitantes. Thèse. Université de Montpellier, 1987

[Spa 89] **Spadn** : Système Portable pour l'Analyse des Données. Manuel de référence Cisia, 1989

[Ter 86] **Ter Braak C.J.F.** : Canonical correspondence analysis : a new eigenvector technique for multivariate direct gradient analysis. Ecology, 67(5), pp.1167-1179, 1986

[Ter 87] **Ter Braak C.J.F.** : The analysis of vegetation-environnment relationships by canonical correspondence analysis. Vegetatio, 69, pp.69-77, 1987