Ce texte est, après diverses mises à jour et corrections, celui paru sous le même titre dans : *La revue de Modulad*, n°27, juin 2002, p 1- 22, (INRIA). Cf. également : LEBART, L. (2000): Contiguity Analysis and Classification, In: W. Gaul, O. Opitz and M. Schader (Eds): *Data Analysis*. Springer,Berlin, 233--244.

Séminaire SAMOS. Le 13 décembre 2002

CLASSIFICATION ET ANALYSE DE CONTIGUÏTE

Ludovic Lebart

CNRS - Ecole Nationale Supérieure des Télécommunications

46 rue Barrault, 75013, Paris, France.

lebart@enst.fr

Lorsque des observations statistiques (multivariées) sont associées à un graphe (séries temporelles, données géographiques), les variances et covariances locales permettent de prendre en compte la dépendance des observations vis-à-vis de la structure de graphe. L'analyse de contiguïté permet alors de confronter structures locales et globales. Si le graphe est associé à une partition, on retrouve l'analyse linéaire discriminante de Fisher. On étudie ci-dessous le cas où le graphe n'est pas une donnée externe, mais est construit à partir des données elles-mêmes, à partir des plus proches voisins de chaque observation. Cette étude permet de mettre en évidence des zones d'inégales densité, et des structures intermédiaires entre celles que détectent les méthodes factorielles, et celles mises en évidence par les méthodes de classification. L'idée de trouver une métrique susceptible de mettre en évidence des classes remonte aux travaux de Art *et al.* [ART82] et Gnanadesikan *et al*.[GNA82]. Nous présentons ici la contribution de l'analyse de contiguïté à de telles approches.

Considérons n sommets d'un graphe symétrique $\boldsymbol{G}$ dont la matrice associée est $\mathbf{M}$ ($m_{ii'} = 1$ si les sommets i et i' sont joints par une arête $m_{ii'} = 0$ sinon). Ces sommets sont simultanément décrits par p variables, (x_{ij}) est la valeur de la variable j pour le sommet i). Une telle situation se présente quand les sommets représentent des instants, des zones géographiques. L'analyse de contiguïté utilise simultanément une matrice des covariances locales $\mathbf{C}$ et une matrice des covariances globale $\mathbf{V}$.

La minimisation du quotient: $\mathbf{u'Cu/u'Vu}$ ($\mathbf{u}$ étant un p-vecteur) engendre un outil de visualisation permettant de déplier certaines structures non linéaires, et généralisant l'analyse discriminante linéaire de Fisher dans le cas de classes empiétantes.

Après quelques résultats préliminaires relatifs à la visualisation de quelques structures de graphe par l'analyse des correspondances, l'analyse de contiguïté est définie.

La seconde partie est consacrée à la situation dans laquelle la structure de graphe n'est pas exogène, mais établie à partir de la matrice de données $\mathbf{X}$ elle-même, sous la forme de la série

des graphes des plus proches voisins. Quelques possibilités d'exploration de données sont esquissées.
La troisième partie évoque, essentiellement à propos d'un exemple numérique, l'apport possible de méthodes telles que l'analyse de contiguïté au cas de la visualisation de résultats de classification, en complément ou en concurrence avec les *Self Organizing Map* de Kohonen.

1. Variance et covariance locales, graphes de contiguïté

Cette section considère le cas d'un ensemble d'observations, (*n* objets décrits par *p* variables), conduisant à une *(n,p)* matrice **X**, ayant une structure de graphe *a priori*. Les *n* observations sont les sommets d'un graphe symétrique ***G*** dont la matrice *(n, n)* associée est **M** ($m_{ii'} = 1$ si les sommets *i* et *i'* sont joints par une arête, $m_{ii'} = 0$ sinon).

1.1 Variance locale v*(y) d'une variable y

y étant une variable aléatoire prenant ses valeurs sur chaque sommet *i* d'un graphe symétrique ***G*** , avec *m/2* arêtes, une première définition de la variance locale $v^c(y)$ est:

$$v^c(y) = (1/2m)\sum^{(c)} (y_i - y_{i'})^2$$

Le symbole $\Sigma^{(c)}$ signifie: somme pour tout *i* et *i'* tels que les sommets *i* et *i'* sont joints par une arête.
Une écriture équivalente, utilisant la matrice binaire **M** = ($m_{ii'}$) associée au graphe ***G***, est :

$$v^c(y) = (1/2m)\sum m_{ii'}(y_i - y_{i'})^2$$

Notons que si ***G*** est un graphe complet (toutes les paires *(i,i')* sont jointes par une arête), $v^c(y)$ n'est rien d'autre que *v(y)*, la variance empirique classique. Quand les observations sont distribuées aléatoirement sur le graphe, $v^c(y)$ et *v(y)* estiment tous deux la variance de *y*.
Le coefficient de contiguïté *c(y)* de Geary [GEA54] , généralisant le ratio de Von Neumann [VON41], s'écrit : $c(y) = v^c(y) / v(y)$.
Une valeur du coefficient de contiguïté *c(y)* significativement inférieure à 1 indique une autocorrélation spatiale positive pour la variable *y*. Beaucoup de coefficients ont été proposés dans la même veine; cf. par exemple [RIP81]; [CLI81]; [ANS95].

Une modification va être faite sur la définition du coefficient *c(y)* (cf. [MOM88], [ESC89]) pour rendre la variance locale compatible avec la variance "intra" ("*within*") quand le graphe décrit une partition des observations (i.e. une série de cliques [sous-graphes complets] disconnectées).
On note par **N** la *(n,n)* matrice diagonale ayant le degré de chaque sommet *i* comme élément diagonal n_i (n_i dénote ici n_{ii}). **y** est le vecteur dont la $i^{ème}$ composante est y_i.

On a $n_i = \Sigma_k m_{ik}$.

La variance locale sera redéfinie comme:

$$v^*(y) = (1/n)\sum (y_i - m_i^*)^2$$

Dans cette dernière formule, la *moyenne locale* est définie comme :

$$m_i^* = (1/n_i)\sum_k m_{ik} y_k$$

C'est la moyenne des valeurs adjacentes au sommet *i*.

Notons que si le graphe ***G*** est régulier (i.e. si n_i est constant) : $v^*(y) = v^c(y)$.

1.2 Bornes pour c(y)

On rappelle dans cette section que les vecteurs propres calculés à partir de l'analyse des correspondances (AC) d'une matrice **M** associée au graphe ***G*** ont des propriétés optimales vis-à-vis du coefficient de contiguïté (cf. par exemple : [LMP98]).

Pour une variable centrée réduite *y*, le coefficient *c(y)* s'écrit (**I** désignant la matrice unité, et **N** la matrice diagonale définie plus haut):

$$c(y) = \mathbf{y}'(\mathbf{I} - \mathbf{N}^{-1}\mathbf{M})'(\mathbf{I} - \mathbf{N}^{-1}\mathbf{M})\ \mathbf{y} \ /\ \mathbf{y}'\mathbf{y}$$

Donc le minimum de *c(y)*, que l'on notera μ, est la plus petite racine de:

$$(\mathbf{I} - \mathbf{N}^{-1}\mathbf{M})'(\mathbf{I} - \mathbf{N}^{-1}\mathbf{M})\ \psi = \mu\ \psi \qquad (1)$$

Si le graphe est régulier (si tous les sommets sont adjacents à un même nombre *a* d'arêtes) la matrice **N** s'écrit : **N** = a**I**, donc $\mathbf{N}^{-1}\mathbf{M}$ est symétrique, et l'équation (1) peut s'écrire:

$$(\mathbf{I} - \mathbf{N}^{-1}\mathbf{M})^2\ \psi = \mu\ \psi$$

ce qui implique :

$$(\mathbf{I} - \mathbf{N}^{-1}\mathbf{M})\ \psi = \sqrt{\mu}\ \psi$$

$$\mathbf{N}^{-1}\mathbf{M}\ \psi = (1 - \sqrt{\mu})\ \psi$$

Notons que les *formules de transition* correspondant à l'AC de la matrice **M** s'écrivent, pour le premier facteur: $\mathbf{N}^{-1}\mathbf{M}\ \phi = \varepsilon\sqrt{\lambda}\ \phi$

si ε = +1, le facteur est dit direct, alors que si ε = -1, le facteur est dit inverse [BEN73]).

Un facteur inverse correspond en fait, dans ce cas, à une valeur propre négative de la matrice de données symétrique initiale **M** (matrice associée au graphe).

Puisque *c(y)* est positif, la valeur minimale μ correspond à la valeur maximum de λ notée λ_{max} pour un facteur direct (ε = +1). Donc, la borne inférieure de *c(y)* est:

$$Min\ [\ c(y)\ ] = (1 - \sqrt{\lambda_{max}})^2$$

Ce minimum est atteint quand ψ est le premier facteur ϕ obtenu à partir de l'analyse des correspondances de la matrice **M**. Alors, la séquence des premiers facteurs ϕ_r correspond à une séquence de variables **N**-orthogonales ayant la propriété de contiguïté extrémale.

Cette propriété explique la bonne qualité de la description des graphes par l'Analyse des Correspondances de leur matrice associée)(cf. [BEN73]; [LSB98]).

1.3 Visualisation de graphes

La figure 1 représente, à titre d'exemple, un graphe à 25 sommets associé à un damier.

La matrice de contiguité **M** associée à ce graphe peut être stockée sous forme condensée à partir des adresses des cases non vides [matrice (5, 25) ici au lieu de (25, 25) , 5 étant un majorant du degré maximum du graphe].

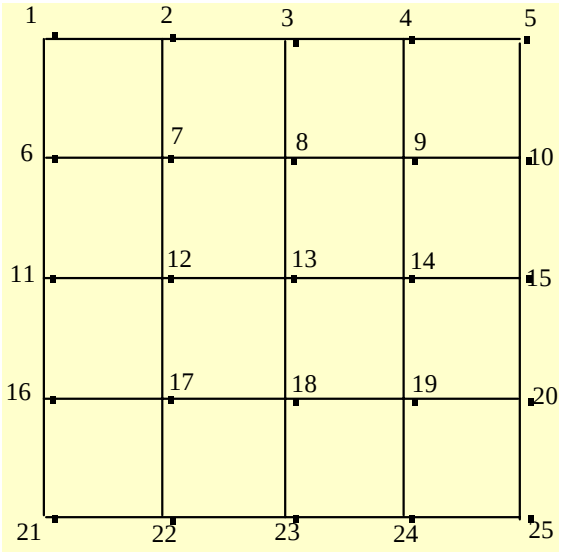


Figure 1 : Graphe associé à un damier (5 x 5)

Les algorithmes pourront utiliser ce codage réduit. Ce détail aura son importance lorsqu'il s'agira de travailler sur les *graphes des plus proches voisins* de milliers d'observations.

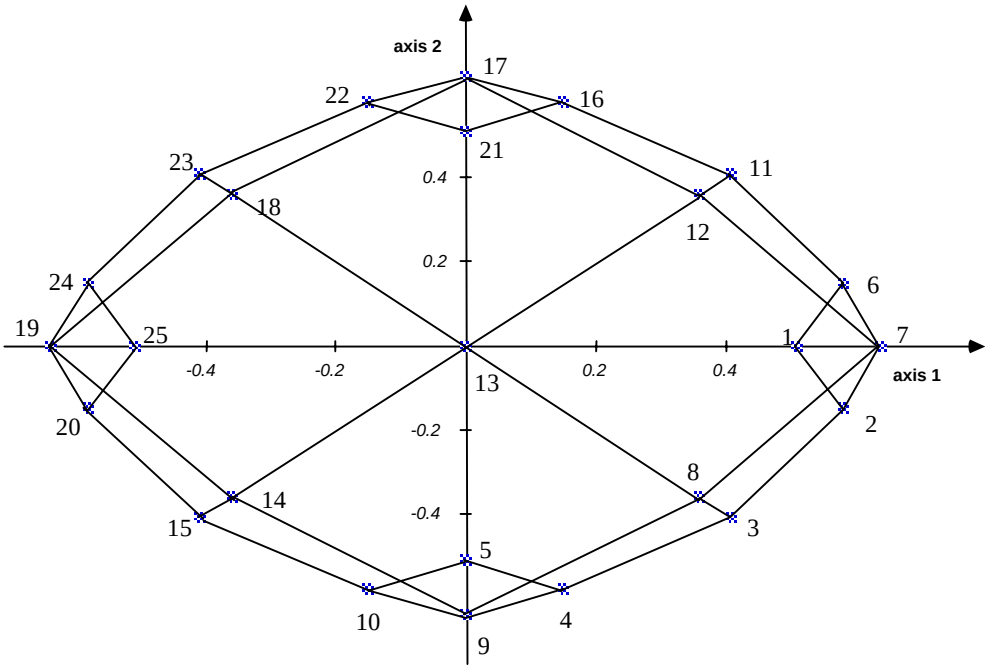


Figure 2 : Représentation du damier dans le plan principal d'une analyse en composantes principales de la matrice **M**.

La figure 2 nous montre que l'analyse en composantes principales de la matrice **M**, qui reconstitue une grande partie du damier, et notamment sa symétrie par rapport au sommet 13. La plus grande valeur propre est double ($\lambda_1 = \lambda_2 = 3.98$) et la plan de la figure 2 rend compte de 31.8 % de la variance.

Toutefois, la figure 2 ne fournit pas une reconstitution d'aussi bonne qualité que la figure 3, provenant d'une analyse des correspondances de la même matrice **M**. Cette analyse fournit aussi une première valeur propre double ($\lambda_1 = \lambda_2 = 0.814$), alors que le plan rend compte de 32.24 % de la variance.

Ce type de représentation par analyse des correspondances nous permettra d'obtenir des visualisations des graphes des plus proches voisins qui se construisent facilement à partir du tableau de données initial.

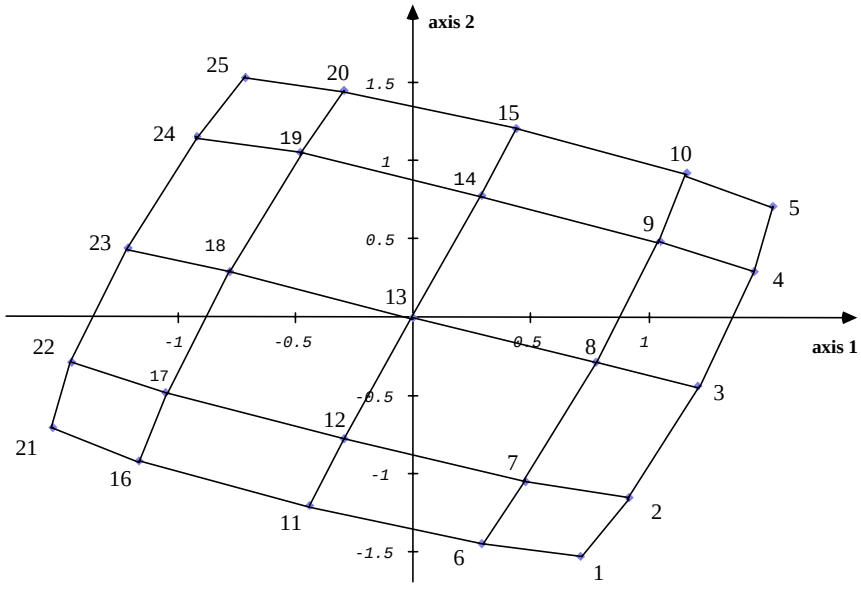


Figure 3 : Représentation du damier dans le plan principal d'une analyse des correspondances de la matrice **M**.

1.4 Analyse en composantes principales locales

Le coefficient de contiguïté peut être généralisé :

1) à de plus grandes distances entre sommets sur le graphe;
2) à des observations multivariées (ces deux généralisations sont traitées dans [LEB69]).

On peut en effet définir comme distance entre deux sommets d'un graphe le plus court chemin reliant ces deux sommets en empruntant des arêtes du graphe. Le graphe correspondant à la distance k est associé à la matrice **M**(k) – **M**(k-1), où **M**(k) désigne la $k^{ème}$ puissance booléenne de la matrice (**I** + **M**) (**I** étant la matrice unité). Ceci permet un calcul de variance locale en fonction de k, qui donne une variante, dans le cas discret et isotropique, du *variogramme* introduit par Matheron [MAT63] et largement utilisé en géostatistique.

Cette section est consacrée à la seconde généralisation : l'analyse d'ensembles d'observations multivariées ayant une structure de graphe *a priori*.
La matrice des covariances locales généralise maintenant la variance locale.
Si **X** désigne la *(n,p)* matrice donnant les valeurs des *p* variables pour chacun des *n* sommets du graphe, décrite par sa matrice associée **M**, la matrice des covariances locales peut s'écrire :

$$\mathbf{V}^* = (1/n)\, \mathbf{X}'(\mathbf{I} - \mathbf{N}^{-1}\mathbf{M})'(\mathbf{I} - \mathbf{N}^{-1}\mathbf{M})\mathbf{X}$$

La diagonalisation de la matrice des corrélations correspondante (Analyse en composantes principales locales) fournit une description des corrélations locales, qui peut être comparée aux résultats d'une analyse en composantes principales classique ne prenant pas en compte la structure de graphe. Les comparaisons entre matrices des covariances locales et globales peuvent se faire par analyses "Procustéennes" (Tucker [TUC58]; Schönemann [SCH68]; Gower [GOW84], Lafosse [LAF85]).
Si le graphe est fait de *k* sous-graphes disjoints complets, **V*** n'est autre que la classique *matrice des covariances "within" ou "intra"* utilisée en analyse discriminante linéaire.
En fait, **V*** coïncide avec la matrice *intra* dans ce cas particulier parce que nous avons modifié la définition de la variance locale (cf. section 1.1). Le Foll (cf. [LEF82]) a généralisé la formule de la covariance locale dans le cas d'observations pondérées; voir aussi : [ALU84]. Une revue et une synthèse de nombreuses approches se trouve dans [MEO93].

2. Analyse de contiguïté et graphes des plus proches voisins

2.1 Analyse de contiguïté

Soit **u** un vecteur définissant une combinaison linéaire *u(i)* des *p* variables pour le sommet *i*:

$$u(i) = \Sigma_j\, u_j\, y_{ij} = u'\, y_i$$

Avec les notations précédentes, la variance locale de la variable artificielle *u(i)* vaut :

$$v^*(u) = \mathbf{u}'\ \mathbf{V}^*\ \mathbf{u}$$

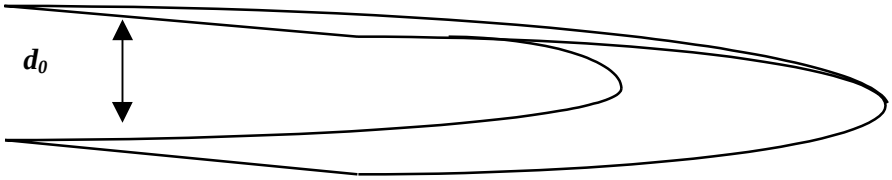
Le coefficient de contiguïté de cette combinaison linéaire s'écrit:

$$c(u) = \mathbf{u}'\ \mathbf{V}^*\ \mathbf{u}\ /\ \mathbf{u}'\ \mathbf{V}\ \mathbf{u}$$

où **V** est la matrice des covariances classique du vecteur **y**. La recherche de **u** qui minimise *c(u)* donne des fonctions de contiguïté minimale: Ces fonctions sont, en un sens, les

combinaisons linéaires de variables distribuées de façon la plus continue possible sur le graphe. Au lieu d'allouer une observation à une classe spécifique, (comme cela est fait en analyse discriminante) ces fonctions alloue cette observation à une zone particulière du graphe. Donc, elles peuvent être utilisée pour discriminer entre des classes empiétantes, pourvu que la relation entre observations soient décrite par un graphe. Ainsi, Faraj [FAR93] l'utilise pour discriminer simultanément plusieurs variables nominales, alors que Chateau [CHA99] suggère de l'utiliser quand l'ensemble de classe possède une structure *a priori*.
Les résultats précédents peuvent être enrichis de plusieurs manières. Il est facile de construire un graphe de contiguïté à partir de la matrice de données elle-même: n'importe quel seuil appliqué à l'ensemble des *n(n-1)* distances ou similarités entre observations permet de définir une relation binaire, et donc un graphe symétrique..
Si le nuage de *n* points décrits par *p* variables est concentré dans un espace *p*-dimensionnel le long d'une hypersurface repliée comme le montre la Figure 4, un graphe ***G*** peut être construit, avec la matrice associée **M** telle que $m_{ii'} = 1$ si les observations (sommets du graphe) *i* et *i'* sont à une distance inférieure à d_0, $m_{ii'} = 0$ sinon.

La section 1.2 suggère que l'AC d'une telle matrice **M** va déplier le diagramme puisqu'il n'y a pas d'arête du graphe joignant les deux branches du "fer à cheval".



Le seuil de distance ***d*** *doit être inférieur à* $\boldsymbol{d_0}$ *pour que le nuage soit déplié par analyse de contiguïté*

Figure 4 : Dépliage par Analyse de contiguïté

L'analyse de contiguïté réalise un "dépliage" similaire, puisque les observations éloignées sont ignorées lors du calcul de la matrice des covariances locales. Nous avons ici un cas particulier de "*projection pursuit algorithm*" [BUR91].
On considérera ci-dessous les matrices construites à partir des *k* plus proches voisins (notés: PPV) de chaque observation.
Ces matrices ont l'avantage d'être « non-paramétriques » puisqu'elles ne demandent pas de choix de seuils de distances, et la série des matrices est naturellement incrémentée en fonction du nombre *k* de plus proches voisins, sans nécessiter la définition d'un « pas ».

De plus, les graphes obtenus sont, en général, plus souvent connexes.

2.2 Sélection du meilleur graphe de contiguïté

Des analyses de contiguïté ont été réalisées sur les classiques "IRIS" de Fisher, utilisant différents graphes selon le nombre de PPV (Figure 5). Les données "IRIS" contiennent 150 individus correspondant à 3 espèces de fleurs, chaque espèce comprenant 50 observations.

Commentaires sur la figure 5

L'axe horizontal représente le nombre *k* de plus proches voisins retenus pour construire un graphe de contiguïté (ce nombre varie de 3 à 149). Quatre courbes sont représentées.

- La courbe "losanges noirs", proche de la diagonale principale du cadre rectangulaire, donne le nombre de sommets du graphe (en fait le pourcentage de sommets du graphe par rapport à un graphe complet ayant *n(n-1)* sommets) : plus on retient de plus proches voisins, plus ce pourcentage augmente. Cette proportion est une fonction approximativement linéaire du nombre de plus proches voisins.

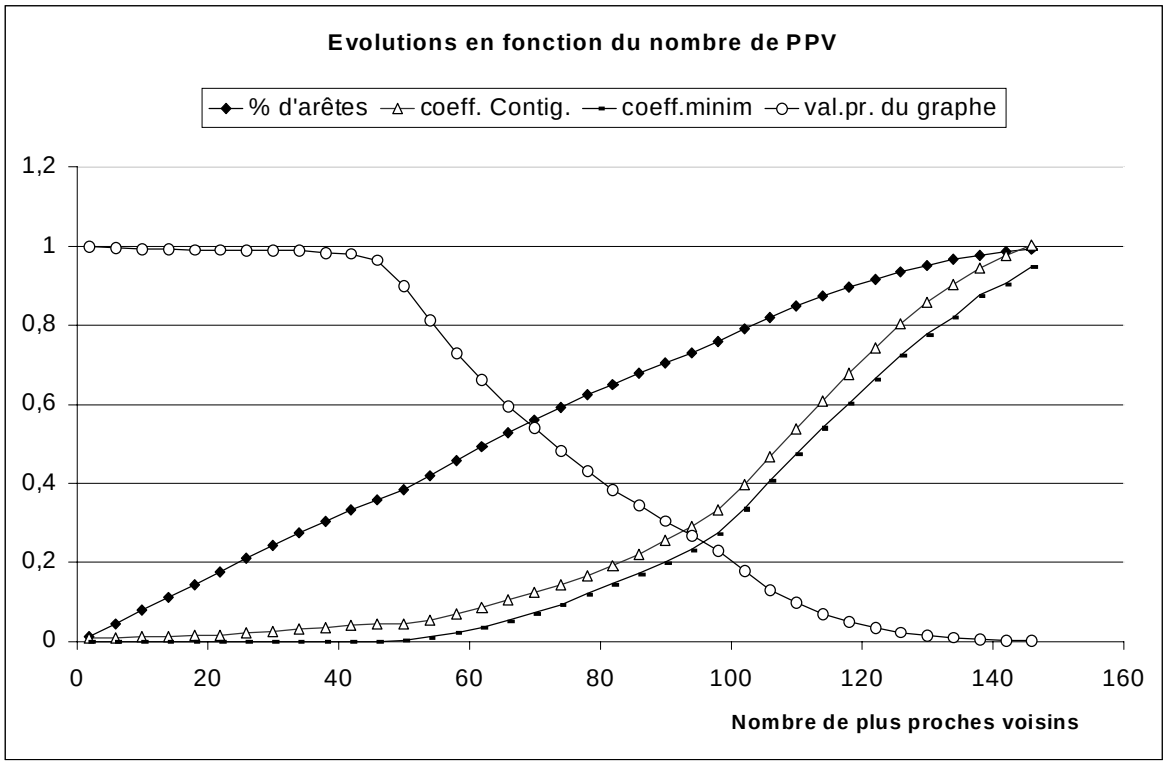


Figure 5 : Graphes et ratios de contiguïté en fonction du nombre de PPV.

- La courbe "triangles blancs", située sous la précédente, représente la plus petite valeur propre de **u'Cu/u'Vu**, c'est-à-dire le plus petit coefficient de contiguïté d'une combinaison linéaire des quatre variables originales. Un changement de pente dans la trajectoire est discernable pour environ 50 plus proches voisins, et également autour de 100 plus proches voisins. On sait que pour ce jeu de données "Iris", un groupe de 50 observations est très éloigné des autres. Ceci explique l'allure de la courbe observée précédemment.
- La courbe sans symbole immédiatement sous la précédente décrit la trajectoire du coefficient de contiguïté minimum *Min [c(y)]*, que nous donnent les AC des matrices **M** associées aux différents graphes des k plus proches voisins, pour les valeurs successives de k.
- Finalement, l'unique courbe décroissante de ce graphique représente la première valeur propre λ_{max} de l'analyse des correspondances de **M** (chaque point correspond à une diagonalisation d'une matrice *(150, 150)*). Cette information est équivalente à celle donnée par la courbe *Min [c(y)]* , puisque *Min [c(y)]* = $(1- \sqrt{\lambda_{max}})^2$, mais l'isolement d'un groupe d'observation est encore plus lisible ici : la valeur propre 1 apparaît en effet en analyse des correspondances quand le graphe dégénère en composantes connexes disjointes (cf., par exemple, [CAZ86], [LEB93]). Ainsi, la décroissance marquée de la courbe après 50 voisins traduit bien l'existence d'un groupe isolé.

2.3 Un critère utilisant l'information *a priori* sur les groupes

La Figure 6 montre la trajectoire du critère W/T (variance intra ou *within* W, divisée par variance totale T) en fonction du nombre de plus proches voisins.

Contrairement aux quantités étudiées jusqu'ici, ce critère prend en compte l'appartenance aux classes connues *a priori* (les trois espèces de fleurs). Ce critère a été calculé pour le premier axe issu de chaque analyse de contiguïté. Il est présenté ici comme une fonction du nombre de plus proches voisins gardés.

La ligne droite horizontale en tirets correspond à la valeur de ce même critère (0.030) pour une analyse linéaire discriminante réalisée sur le même jeu de données. Evidemment, le premier axe principal d'une analyse de contiguïté qui ignore totalement l'appartenance *a priori* des observations aux classes ne peut concurrencer la première fonction discriminante qui utilise cette information pour, précisément, rendre minimum le critère W/T.

La droite horizontale en trait gras correspond à la valeur du critère (0.063) fournie par le premier facteur d'une ACP du jeu de données (150 individus, 4 variables).

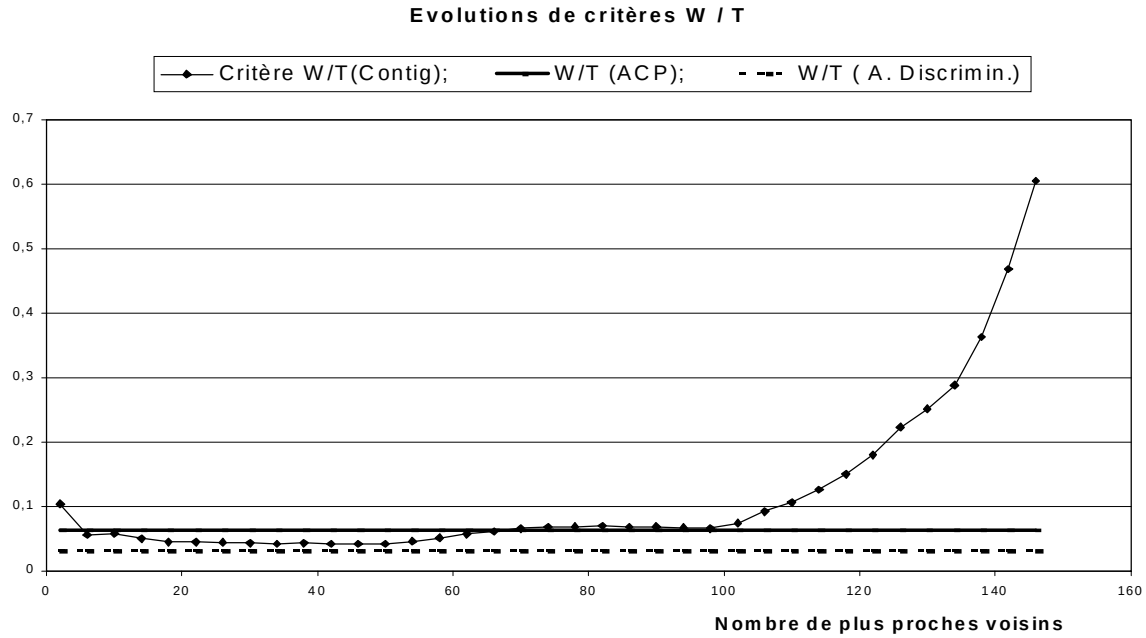


Figure 6 : Critère W/T (premier axe) en fonction du nombre de PPV.

La trajectoire de notre critère est située sous cette ligne en pointillé pour des nombres de plus proches voisins compris entre 4 et 70. Si l'on garde moins de 40 plus proches voisins (cette borne supérieure est donnée par le premier changement de pente du graphique précédent), la trajectoire du critère reste très proche de la valeur théorique minimale donnée par la ligne horizontale continue. La valeur minimale du critère est de 0.0365, très inférieure à celle du premier facteur de l'analyse en composantes principales. En fait, le minimum théorique de 0.030, fourni par l'analyse discriminante, est une estimation par resubstitution (calculée sur l'échantillon d'apprentissage, et non sur un échantillon-test), qui donne une idée optimiste de la qualité de la discrimination.

2.4 Discussion

On ne doit pas oublier que chaque point des graphiques des figures 5 et 6 correspond à une représentation des observations dans un espace à deux ou trois dimensions (ou plus si nécessaire) . Ces représentations ne sont pas publiées ici, faute de place.

Elles mettent en évidence à la fois l'existence de classes et les possibilités de discriminer entre ces classes. Lorsque les classes ont une existence réelle (ce qui est le cas pour les iris de Fisher), elles sont redécouvertes, avec, si l'on peut dire, l'analyse discriminante en prime. L'ensemble du processus est "non supervisé", puisque le nombre optimal de plus proches voisins est donné par des études de trajectoires analogues à celles de la figure 5.

Pourquoi utiliser des graphes issus de plus proches voisins, et non des graphes calculés à partir de simples seuils de distances ? Plusieurs expériences nous ont montré que ces derniers graphes sont souvent non connexes (sauf pour de grandes valeurs du seuil, mais il est alors

difficile de détecter des petites classes). De plus, les valeurs de critères externes comme W/T sont moins favorables. Ainsi, pour l'exemple particulier des iris, le minimum de W/T pour un graphe construit à partir de seuil de distance est de 0.044, sensiblement supérieur au minimum atteint par les PPV.

En conclusion de cette section, l'approche non-paramétrique mettant en œuvre des matrices des covariances locales (calculées à partir de la série des graphes des plus proches voisins) permet :

- De détecter des classes potentielles, après sélection du nombre de plus proches voisins;
- D'obtenir simultanément une visualisation des observations et de ces classes, avec d'éventuels *dépliages* permis par le caractère non-linéaire de l'opération ;

Elle permet par ailleurs mettre en œuvre une analyse discriminante classique lorsque le graphe de contiguïté est externe (graphe associé avec une partition *a priori*, ou une partition construite par une méthode de classification automatique).

3. Visualisation de partitions.

Les cartes de Kohonen [KOH89] répondent à une demande certaine des utilisateurs : obtenir à la fois une partition d'un ensemble d'objets, d 'individus ou d'observations, et une visualisation simple et pratique des proximités entre les classes de cette partition.

Le prix à payer pour cette représentation ergonomique est l'absence de transparence et de critère explicite pour la définition des classes, et l'absence de procédures de validation.

Dans le cas de méthodes de partitionnement directe comme les *k-means* ou les « nuées dynamiques », on tend à optimiser des critères explicites, bien qu'on doive se contenter le plus souvent d'optima locaux, sauf cas particuliers favorables.

Lorsqu'on désire avoir une vue d'ensemble des classes obtenues par ces méthodes « transparentes » et une appréciation de leurs proximités respectives, on procède parfois à une projection des centres de classes sur le plan factoriel principal provenant d'une analyse (Analyse en composantes principales – ACP–, Analyse des correspondances simple –AC–, ou multiple –ACM–, selon les cas) avec pour *variables actives* les variables ayant servi à construire la partition.

De telles projections, sans dépliage ni déploiement, donnent une image souvent trop complexe d'une réalité elle-même complexe. On peut dire que l'on reste trop près des données, et trop loin de l'utilisateur, alors que les cartes de Kohonen se situent (ergonomiquement parlant) près de l'utilisateur, et, dans certains cas, plus loin des données. On peut penser en effet, pour les cartes de Kohonen, que la contrainte exprimée par la grille ne peut que pénaliser la

minimisation du critère W/T (variance interne sur variance totale) par rapport à ce que ferait un algorithme sans contrainte. Mais nous avons affaire dans les deux cas (Kohonen et k-means sans contrainte) à des optima locaux, et il est donc difficile d'énoncer des résultats ayant une portée générale sur la comparaison des algorithmes.

3.1 L'exemple numérique illustratif :

L'exemple choisi concerne une enquête au cours de laquelle les individus ont répondu au questionnaire sémiométrique [STE92]. Ce questionnaire consiste à noter une suite de 210 mots de la langue française (le questionnaire et les enquêtes correspondantes existent en plusieurs langues et pour plusieurs pays) selon une échelle de notes de 1 à 7, la note minimale 1 correspondant aux mots produisant chez la personne interrogée une sensation très désagréable, la note 7 à une sensation très agréable. On travaillera ici sur un tiers des mots (un sur trois par ordre alphabétique) pour des raisons d'encombrements graphiques. L'échantillon est de 3360 individus. Les analyses en composantes principales des notes donnent des configurations stables dans l'espace des six premiers axes principaux. Cette stabilité est vérifiable par *bootstrap*, mais aussi dans le temps (enquêtes périodiques entre 1990 et 1998 réalisées par la SOFRES) et dans l'espace (d'un pays à un autre). Cette permanence de structure constitue un cas favorable pour étudier, dans le cadre d'une classification, les positions respectives des classes.

Une analyse en composantes principales (ACP) du tableau (3360 x 70) nous donne les coordonnées factorielles des 70 mots sur les 7 premiers axes principaux.

Dans un premier temps, on procédera à une représentation des associations entre mots utilisant une carte de Kohonen.

Puis on procédera à une classification des mots en 12 classes en utilisant un algorithme assez classique. On représentera les classes par leurs projections dans le plan principal de l'ACP de base. Enfin, on représentera les mêmes classes dans le plan principal de l'analyse de contiguïté, et on discutera des mérites respectifs de chacune des méthodes.

3.2 Carte de Kohonen

La figure 7 représente une « carte auto-organisée » construite selon la variante de l'algorithme proposée par Kleiweg [KLE96]. Cet algorithme comporte aussi la possibilité de tracer des traits d'épaisseurs variables, un trait épais séparant des classes éloignées dans l'espace bien que contiguës sur la carte. Enfin, il permet également un tracé de l'arbre de longueur minimale entre les objets à classer, ce qui constitue, même dans le cas d'école de 70 objets à classer, un enchevêtrement assez inextricable. Ces options n'ont pas été utilisées ici. Des perfectionnements similaires, mais plus élaborés, ont été apportés dans [ROU99] et [COT97] qui suggèrent des enrichissements graphiques permettant de remédier au caractère figé et non-métrique de la grille.

La première qualité de cette représentation est son extrême clarté, due à sa compatibilité avec les formes usuelles de l'édition (rectangles, lignes, cases).

3.3 Classification et projection sur le plan principal de l'ACP

La figure 8 représente les 12 enveloppes convexes de 12 classes obtenues par classification des mots selon la méthode mixte suivante : classification hiérarchique utilisant la distance euclidienne usuelle dans l'espace des axes principaux de l'ACP (critère de Ward pour le recalcul des distances), coupure du dendrogramme au niveau de 12 classes, optimisation de la partition obtenue par réaffectation itérative des objets à classer (mots) (*k-means* à partir de la coupure précédente utilisée comme partition de démarrage).
Ces 12 enveloppes convexes sont tracées dans le premier plan factoriel de l'ACP.

On représente habituellement les centres de gravités des classes (surtout lorsqu'il y a des milliers d'objets classer), mais ici on a représenté les classes par leurs enveloppes convexes de façon à garder visibles les compositions des classes (comme dans le cas de la carte de Kohonen de la figure 7), mais aussi la forme et donc la dispersion des classes, ce qui constitue une information complémentaire. Ces douze classes, établies sans contraintes, ne peuvent être que de meilleure qualité, du point de vue de leurs variances internes, que des classes équivalentes obtenues sous contraintes dans les *self organizing maps* .

VIDE ROMPRE MURAILLE	FUSIL ARMURE	DANGER	INCONNU CHARNEL	SENSUEL NUDITE DESIR	SUBLIME		FECONDER ENFANCE CAMPAGNE ANIMAL
			VITESSE		REVER		FLEUVE
METALLIQUE INTERDIRE				AVENTURIER	EAU LUNE		MONTAGNE
MATERIEL RIGIDE				FEU			ESCALADER
PRODUIRE	COMMANDER	CERTITUDE		ECRIRE			BLEU
HONNEUR PRUDENCE				SCIENCE REFLECHIR INVENTEUR			
OR			CONSOLER			JUSTICE	ABSOLU
HERITER GRATUIT		PURETE PAIX DOUCEUR	CONFIANCE	PATIENCE MODERATION ATTACHEMENT	RAISON	SACRE NOBLE	TRADITION

Figure 7 : Carte de Kohonen décrivant les associations entre mots.

3.4 Classification et projection sur le plan principal de l'Analyse de contiguïté

La figure 9 représente les enveloppes convexes des 12 classes déjà obtenues, mais la projection se fait maintenant dans le plan des deux premiers axes de l'analyse de contiguïté selon le graphe des trois plus proches voisins.

Malgré un changement de signe de l'axe horizontal, on peut noter que la figure 9 est sensiblement plus lisible que la figure 8, il y a moins de superpositions de points. En revanche, les axes initiaux sont légèrement modifiés. Il faudrait évidemment compléter cette représentation par celle obtenue sur le plan des axes 3 et 4, car la compression dans un espace bidimensionnel peut être tout simplement impossible, ou conduire à des distorsions et à des contresens dans l'interprétation.

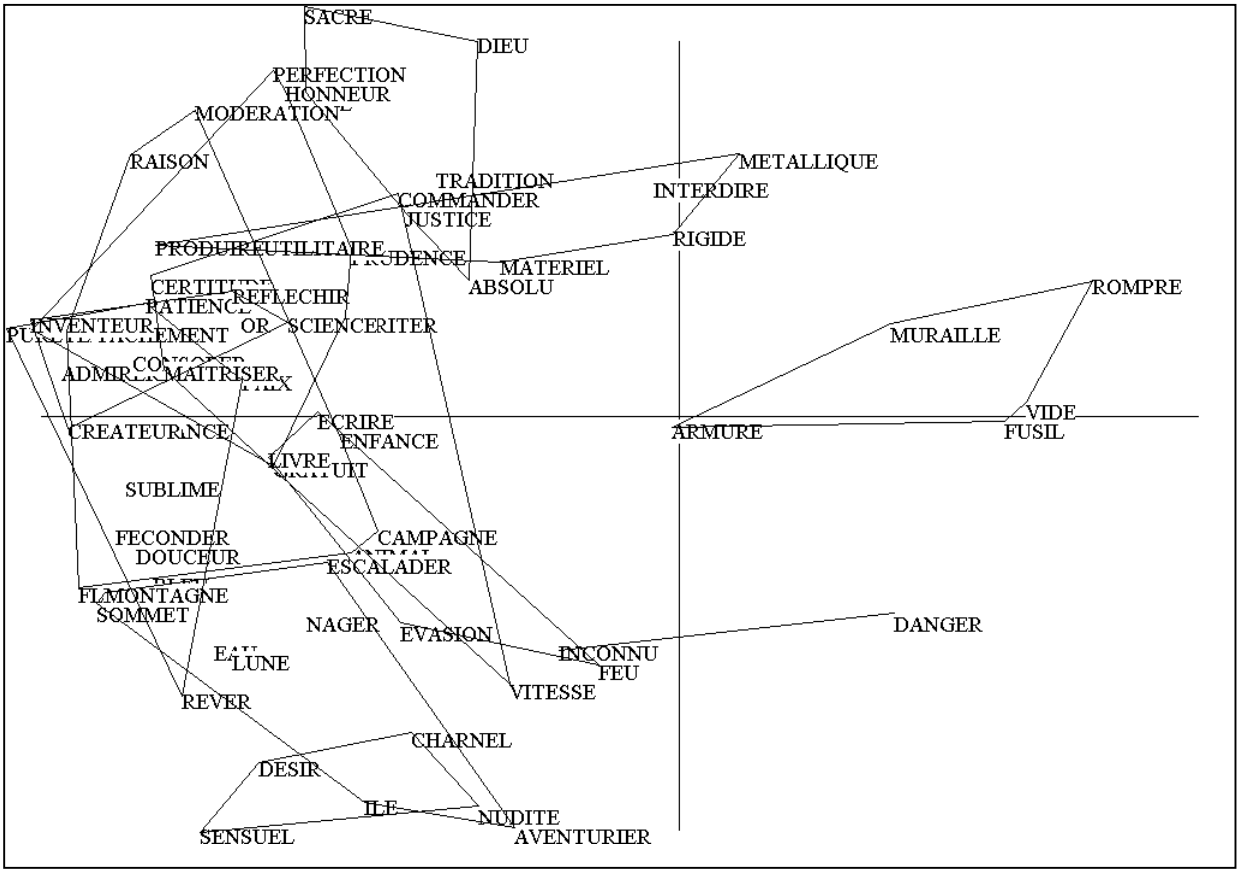


Figure 8 : Enveloppes convexes des 12 classes dans le plan principal de l'ACP.

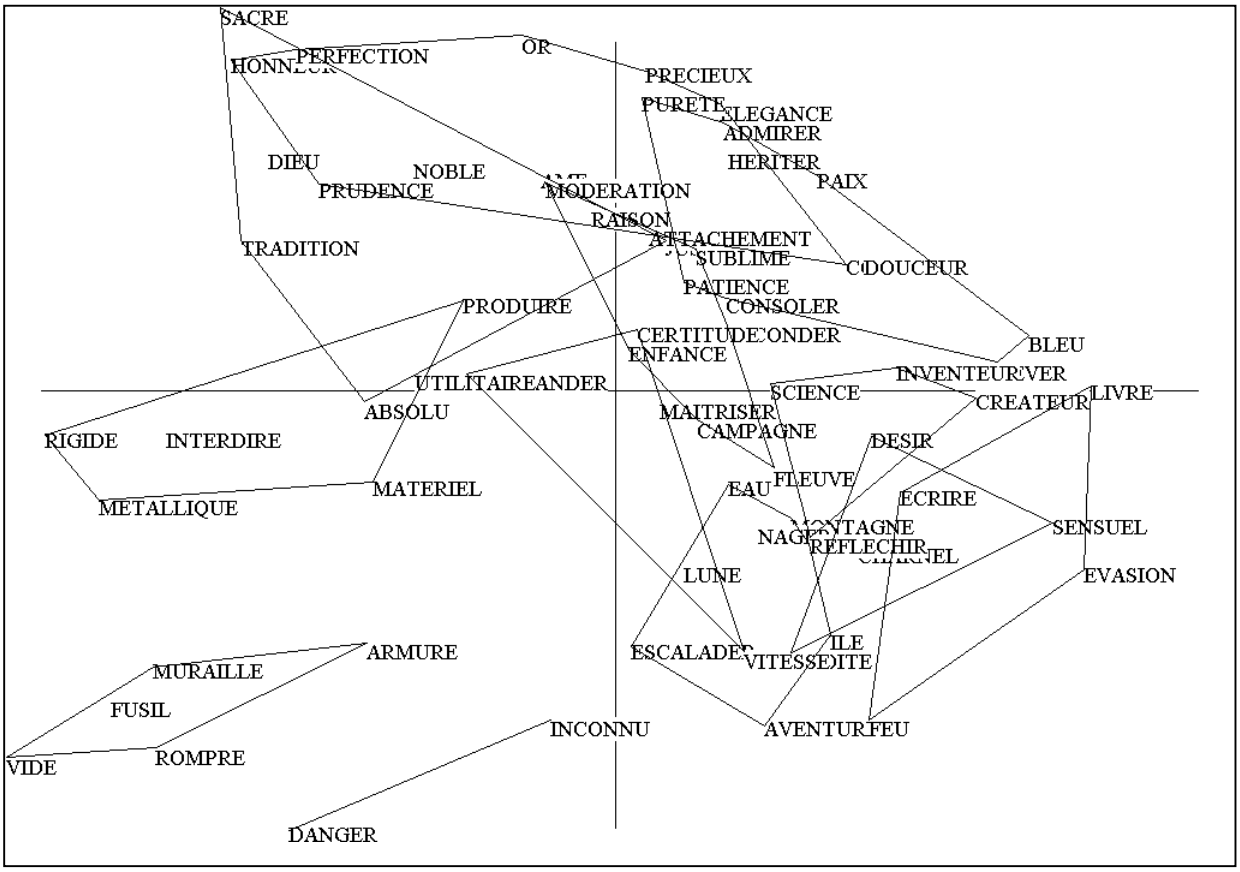


Figure 9 : Enveloppes convexes des mêmes 12 classes dans le plan principal de Contiguïté

3.5 Commentaires et discussions

La constellation « vide, rompre, muraille, fusil, armure » (à droite de la figure 8, à gauche de la figure 9, en haut à gauche de la carte de Kohonen) constitue visiblement un élément stable de la classification, de même que le couple voisin « danger, inconnu ». Il s'agit de regroupements communs aux trois représentations.

En revanche, « sensuel » par exemple, (en bas à gauche de la figure 8, à droite en bas de la figure 9) est fort éloigné des mots « rigide, métallique, rompre » dans la représentation euclidienne (ACP, figure 8) ainsi que dans la représentation non-linéaire (analyse de contiguïté de la figure 9) mais figure sur la première ligne de la carte de la figure 7.

Il semble bien qu'il s'agisse ici d'un défaut de la carte imputable à la contrainte trop forte que constitue la structure de départ sous forme de grille, malgré la souplesse laissée par la possibilité de cases vides.

La confrontation des trois figures permet en fait de découvrir certaines incohérences au niveau des positions relatives des classes. Mener à bien une comparaison systématique demanderait aussi de faire varier les nombres de classes à la fois dans le cas des cartes auto-organisées [dont la taille a été fixée arbitrairement à (8 x 8)] et dans le cas des classifications classiques pour lesquelles le nombre de classe a été fixé, au vu du dendrogramme, à 12.

Deux remarques pratiques : lors des consultations sur écran, la lisibilité des figures 8 et 9 est améliorée par l'utilisation de la couleur (couleurs différentes pour chaque classe). De plus, ces figures sont aussi plus lisibles sur écran si seuls les centres de classes sont représentés, les compositions internes s'obtenant par exemple par un clic de souris.

En conclusion à l'issue de cette exploration empirique :

- La représentation des *classes obtenues sans contrainte* sur le ou les plans factoriels provenant de l'ACP contient le plus d'éléments de validation statistique (chaque point peut être assorti d'une zone de confiance *bootstrap* par exemple).
- Les traits structuraux de la représentation factorielle sous-jacente sont conservés.
- Plusieurs partitions peuvent être représentées sur le même fond.
- Les règles de construction et d'interprétation des proximités sont connues.
- L'amélioration de la lisibilité par analyse de contiguïté (et les éventuels *dépliages* qu'elle permet) réalise un premier compromis entre respect de la structure initiale et lisibilité du graphique plan.
- L'approche par carte auto-organisée relève d'une option radicalement différente : les classes obtenus avec contraintes pourraient être *a priori* de moins bonne qualité, mais s'insèrent dans un cadre lisible. Les proximités locales semblent bien conservées, mais les grandes oppositions et la géométrie du nuage de départ peuvent être profondément altérées.
- Enfin, les représentations sous-jacente à l'analyse de contiguité peuvent faire l'objet de validation statistique (zones de confiance bootstrap par exemple).

BIBLIOGRAPHIE

[ALU84] ALUJA T. and LEBART L. (1984): Local and Partial Principal Component Analysis and Correspondence Analysis, *COMPSTAT Proceedings*. Physica Verlag, Vienna, 113-118.

[ANS95] ANSELIN L. (1995): Local indicators of spatial association . *Geog. Anal.*, 27, 2, 93-115.

[ART82] ART D., GNANADESIKAN R., KETTENRING J.R. (1982): Data Based Metrics for Cluster Analysis, *Utilitas Mathematica*, 21 A, 75-99.

[BEN73] BENZECRI, J.P. (1973): *Analyse des Données: Correspondances*. Dunod, Paris.

[BUR91] BURTSCHY B., LEBART L. (1991): Contiguity analysis and projection pursuit. In*: Appl. Stoch. Mod. and Data Anal*. R. Gutierrez et al., Eds, World Scientific, Singapore, 117-128.

[CAZ86] CAZES P. (1986) Correspondance entre deux ensembles et partition de ces deux ensembles, *Les Cahiers de l'Analyse des Données*, vol.XI, no.3, 335-340.

[CHA99] CHATEAU F. (1999): Structured Discriminant Analysis. *Communic. in Stat.*, 255-256.

[CLI81] CLIFF A.D. et ORD J.K. (1981): *Spatial Processes: Models and Applications*. Pion, London.

[COT97] COTTRELL M., ROUSSET P. (1997): The Kohonen Algorithm: a powerful tool for analysing and representing multidimensional qualitative and quantitative data. In: *Biological and Artificial Computation : From Neuroscience to Technology*. J. Mira, R. Moreno-Diaz, J. Cabestany, (eds), Springer, 861-871.

[ESC89] ESCOFIER B. (1989): Multiple correspondence analysis and neighboring relation. In: *Data Analysis, Learning Symbolic and Numeric Knowledge*. Diday E. (ed.), Nova Science Publishers, New York, 55-62.

[FAR93] FARAJ A. (1993): Analyse de contiguité: une analyse discriminante généralisée à plusieurs variables qualitatives. *Revue Statist. Appl*., 41, (3), 73-84.

[GEA54] GEARY R.C. (1954): The Contiguity Ratio and Statistical Mapping, *The Incorporated Statistician*, 5, 115-145.

[GNA82] GNANADESIKAN R., KETTENRING J.R. et LANDWEHR J.M. (1982): Projection Plots for Displaying Clusters, In: *Statistics et Probability*. G. Kallianpur *et al*., eds, North-Holland.

[GOW84] GOWER J. C. (1984): Procrustes analysis. In: *Handbook of Applicable Mathematics*. 6, Lloyd E.H. (ed.), J. Wiley, Chichester, 397-405.

[KLE96] KLEIWEG P. (1996*): Een inleidende cursus met practica voor de studie Alfa-Informatica*. Master's thesis, Rijksuniversiteit Groningen, 1996.

[KOH89] KOHONEN T.(1989): *Self-Organization and Associative Memory*. Springer-Verlag, Berlin, 3rd edition, 1989.

[LAF85] LAFOSSE R. (1985): *Analyse Procustéenne de deux tableaux*. Thèse, Université de Toulouse.

[LEF82] LE FOLL Y. (1982): Pondération des distances en analyse factorielle. *Statist. et Anal. des Données*. 7, 13-31.

[LEB69] LEBART L. (1969): Analyse Statistique de la Contiguité, *Publ. de l'ISUP*. XVIII, 81-112.

[LEB00] LEBART, L. (2000): Contiguity Analysis and Classification, In: W. Gaul, O. Opitz and M. Schader (Eds): *Data Analysis*. Springer,Berlin, 233--244.

[LEB01]LEBART, L. (2001): Representing words and texts through contiguity analysis. In: *ASMDA 2001, 10th International Symposium on Applied Stochastic Models and Data Analysis*. G. Govaert, J. Janssen, N. Limnios (eds), UTC, Compiègne, 654-659.

[LSB98] LEBART, L., SALEM, A. and BERRY, L. (1998): *Exploring Textual Data*. Kluwer, Dordrecht.

[LEB93] LEBART L., MIRKIN B. (1993): Correspondence Analysis and Classification. In*: Multivariate Analysis: Future Directions 2*. Cuadras C.M. and Rao C.R., (eds), North-Holland, 341-357.

[LMP98] LEBART L., MORINEAU A. PIRON M., L. (1998): *Statistique Exploratoire Multidimensionnelle,*. Dunod, Paris.

[MAT63] MATHERON G. (1963): Principles of geostatistics. *Economic Geology*. 58, 1246-1266.

[MEO93] MEOT A., CHESSEL D. et SABATIER R. (1993): Opérateur de voisinage et analyse des données spatio-temporelles. In *Biométrie et environnement*, Lebreton J.-D., Asselain B., (eds), Masson, Paris, 45-71.

[MOM88] MOM A. (1988): *Methodologie Statistique de la Classification des reseaux de transport*. Thèse, Université des Sciences et Techniques du Languedoc, Montpellier.

[RIP81] RIPLEY B. D. (1981): *Spatial Statistics*. J. Wiley, New York.

[ROU99] ROUSSET P. *Application des algorithmes d'auto-organisation à la classification et à la prévision.* Thèse, Univ. Paris I, UFR Math. et Info.

[SCH68] SCHONEMANN P. H. (1968): On two-sided orthogonal procrustes problems. *Psychometrika*. 33, 19-33.

[STE92] STEINER J.-F. and AULIARD, O. (1992): La sémiometrie: un outil de validation des réponses. In *La Qualité de l'Information dans les Enquêtes* , ASU (eds), Dunod, Paris, 241--274.

[TUC58] TUCKER, L. R. (1958): An inter-battery method of factor analysis. *Psychometrika*. 23, (2).

[VON41] VON NEUMANN, J.(1941): Distribution of the ratio of the mean square successive differences to the variance. *Ann. of Math. Statistics*. 12, 367-395.