REVUE DE STATISTIQUE APPLIQUÉE

MICHEL FORTIN

Sur un algorithme pour l'analyse des données et la reconnaissance des formes

Revue de statistique appliquée, tome 23, n° 2 (1975), p. 37-46

<http://www.numdam.org/item?id=RSA_1975__23_2_37_0>

SUR UN ALGORITHME POUR L'ANALYSE DES DONNÉES ET LA RECONNAISSANCE DES FORMES [1]

Michel FORTIN*

"Les facultés de l'esprit qu'on définit par le terme, analytiques, sont en elles-mêmes fort peu susceptibles d'analyse. Nous ne les connaissons que par leurs résultats". (Edgar Allen Poe).

§ 1 :

Nous allons décrire ici un algorithme capable, dans un sens évidemment restreint, de procéder à une première analyse sur certains types de données et de fournir ainsi à son utilisateur un point de départ valable pour une analyse plus poussée.

De façon précise, nous considérons un ensemble d'objets ou d'individus décrits par un certain nombre de caractéristiques sinon numériques, du moins codifiables. La seule propriété essentielle sera de pouvoir mesurer la similitude (ou inversement la "distance") de deux quelconques de ces objets. Le choix d'une telle mesure est un problème difficile, surtout pour des données non numériques, et nous ne l'aborderons pas ici. Nous nous contenterons, cette mesure étant donnée, d'en tirer un maximum d'information et de fournir à l'analyste une image aussi dégrossie que possible de ses données brutes. Notre démarche sera essentiellement heuristique et nous n'aurons jamais besoin des propriétés strictes d'une métrique en ce qui regarde le choix d'une distance entre les objets. Par contre, pour fixer les idées et justifier intuitivement la méthode proposée, nous parlerons le plus souvent de nos objets comme de points du plan ou de l'espace euclidien $\mathbf{R}^n$.

§ 2 : LES POSSIBILITES DE LA METHODE.

Considérons donc un cas très simple : un ensemble de points du plan que nous voulons décrire à une tierce personne de façon aussi simple et

---

(1) Article remis le 28/11/73, révisé le 22/5/74.

(*) Département de Mathématiques, Faculté des Sciences, Université Laval, Québec, Canada, GlK 7P4.

complète que possible. Par exemple on donnera une information considérable si l'on dit que dans la figure 2.1 les points dessinent la lettre A. Si

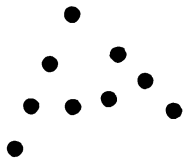

Figure 2.1

nous voulons construire un algorithme capable, même très imparfaitement, de nous décrire un ensemble d'objets, nous voyons donc que nous sommes conduits à simuler un phénomène de perception. Or le sens commun nous apprend que même dans des cas très simples, un observateur pourra donner d'un même ensemble de points des descriptions différentes, basées sur des points de vue différents mais également valables, et fournissant des informations complémentaires et parfois, à première vue, contradictoires.

Par conséquent, on ne saurait attendre d'un algorithme qu'il nous fournisse une image unique et objective. Au contraire, il est souhaitable de pouvoir, selon notre désir, lui faire adopter un point de vue donné.

Voyons maintenant quelles sont les descriptions que nous pouvons espérer obtenir, sachant que la seule information de départ est fournie par les distances entre les points. Nous devons tout d'abord éliminer les conditions de symétrie et d'orientation. En fait, nous ne pourrons conserver que des notions "topologiques", en particulier la connexité.

En fait, notre algorithme pourra décrire deux types de structures :

i) *une classification, des objets* : c'est-à-dire une partition de notre ensemble en sous-ensembles naturels (ou groupes)

ii) *une ordination* : la description d'un chemin "continu" décrivant le passage d'un comportement extrême à un autre par une suite de comportements intermédiaires.

Les deux types de structures pourront être présents dans un cas donné. Par exemple, certains groupes d'une classification pourront être interprétés comme des associations naturelles d'objets semblables alors que d'autres contiendront des cas extrêmes et leur voie de liaison. Il appartiendra à l'utilisateur de distinguer le genre de la structure qui lui est proposée.

D'autre part, l'utilisateur pourra modifier le point de vue de l'algorithme, de la considération des structures locales, jusqu'à la situation globale de l'ensemble tout entier. Mais pour décrire exactement cette possibilité nous devons faire intervenir une définition plus précise de la notion de structure.

§ 3 : STRUCTURE VERSUS UNIFORMITE

Revenons à notre cas simple des points dans le plan. Le seul cas où nous n'aurons rien à décrire est celui d'une *répartition uniforme* de ces points :

il n'y a alors aucune classification naturelle ni de chemin privilégié pour passer d'un comportement à un autre. De même si les points, sans être distribués uniformément, sont un échantillon d'une variable aléatoire uniforme nous pourrons certes y distinguer des structures à un niveau très local mais nous perdrons toute image précise dès que nous nous placerons à un niveau plus global. Au contraire un ensemble de points présentant des zones à forte concentration séparées par des vides, sera aisément séparé en groupes et présentera des structures fortes. Nous avons donc retenu pour la construction de notre algorithme de mesurer cette concentration des points et nous avons cru devoir exiger de cette mesure les propriétés suivantes :

i) la mesure de concentration (ou inversement de dispersion) doit être contrôlable afin de distinguer les variations locales à un extrême et à l'autre de ne conserver que les variations très globales, en lissant les comportements aléatoires locaux.

ii) elle doit être uniforme sur un ensemble uniforme de points, elle doit devenir uniforme au niveau global lorsque les points sont échantillonnés sur une variable uniforme.

Ces exigences sont loin d'assurer l'unicité d'une telle mesure. Nous avons retenu celle décrite ci-dessous qui nous a permis d'obtenir de bons résultats mais des travaux sont en cours pour étudier d'autre possibilités.

Soit donc un point x (un objet de notre ensemble). Nous savons mesurer les distances d(x, y) (resp. les similitudes s(x, y)) lorsque y parcourt notre ensemble d'objets. Conservons donc les m objets ou individus les plus proches (les plus semblables) à x. Nous les noterons $v_i (i = 1, m)$ et nous les appellerons les *points voisins de* x, l'indice i variant selon l'ordre de proximité. Ainsi v_1 sera le plus proche voisin du point x. Nous calculons maintenant

$$d(x) = \sum_{i=1}^{m} d^2(x, v_i) \qquad \text{dans le cas d'une distance} \qquad (3.1)$$

$$c(x) = \sum_{i=1}^{m} s^2(x, v_i) \qquad \text{dans le cas d'une similitude.} \qquad (3.2)$$

Nous dirons que d(x) est la dispersion au point x et que c(x) est la concentration au point x.

Ces deux mesures vérifient bien les propriétés i) et ii), le passage du niveau local en niveau global se faisant en augmentant m.

§ 4 : LE PRINCIPE DE L'ALGORITHME.

Supposons donc la dispersion (ou la concentration) calculée en tous les points de l'ensemble. Dans notre problème modèle du plan nous pourrions en donner une image très simple : il suffirait de rendre plus ou moins foncées les régions selon le degré de concentration. De même on pourrait penser à une maquette topographique tridimensionnelle où la cote (l'altitude) varierait en fonction de la concentration. Il s'agit maintenant de décrire ce que nous voyons.

Comme nous avons rejeté au point de départ toute notion d'orientation ou de symétrie, l'idée la plus naturelle serait de procéder par ordre de concentration : placer d'abord la tache la plus noire ou la montagne la plus haute et procéder ensuite à partir de ce point, par ordre décroissant, en rattachant successivement les points aux structures existantes ou en leur permettant de créer une nouvelle structure. Le rattachement d'un point se fera au moyen de ses points voisins. Un point qui se rattacherait à deux ou plusieurs structures sera appelé un lien ou un pont et nous permet de connaître un chemin entre deux structures. Parmi tous les liens possibles entre deux groupes nous ne conserverons que celui ayant la plus forte concentration. Dans notre modèle topographique, il s'agirait d'un col, dans la vallée séparant deux sommets.

§5 : DESCRIPTION COMPLETE DE L'ALGORITHME.

Soit N le nombre total d'objets.

Etape 1) Considérons un des objets (points) que nous noterons O_i. On calculera d'abord les N−1 distances (ou resp. similitudes) entre ce point O_i et tous les autres points de l'ensemble. On retiendra alors les m plus petites distances afin de calculer la dispersion ou la concentration par la formule (3.1) ou (3.2). Rappelons que m est un paramètre arbitraire : pour m "petit" (par exemple 1, 2, 3, 4 dans le cas de 100 objets) on considère des variations locales de concentration alors que pour m croissant on lisse de plus en plus ces variations locales pour se tenir compte que des structures globales. La dispersion ainsi calculée au point O_i sera conservée dans le tableau d_i ($i = 1, N$). Par ailleurs on conservera dans un tableau $V = (v_{ij})$ où $i = 1, N$ et $j = 1, p$ les indices des p plus proches voisins du point i. On a ici un second paramètre arbitraire p (qui doit être $\leqslant m$) et qui déterminera comme nous le verrons plus loin les niveaux d'exigence pour le rattachement d'un point à une structure existante.

Notons que dans la suite de l'algorithme les distances ne seront plus utilisées et que par conséquent on n'aura pas à les conserver en mémoire.

En résumé à la fin de l'étape 1) nous avons en mémoire pour chaque point O_i une mesure de concentration ou de dispersion d_i et la liste de ses p voisins v_{ij} ($j = 1, p$). Nous dirons que le tableau $V = (v_{ij})$ est le tableau des points voisins.

Etape 2) Nous ordonnons maintenant les points par ordre croissant de dispersion (décroissant de concentration) et nous créons un dictionnaire permettant de connaître pour chaque point O_i son numéro d'ordre par rapport aux dispersions et réciproquement de connaître le point correspondant à un numéro d'ordre donné.

Etape 3) Nous balayons maintenant les points dans l'ordre décroissant des concentrations.

Le premier point forme automatiquement le premier groupe et y entraîne ses p voisins. Nous voyons donc ici que le paramètre p nous permet d'étendre plus ou moins loin le voisinage d'un point. Nous considérons maintenant un

point Q quelconque. Deux cas peuvent se préciser :

A) le point est déjà classé dans un groupe (il y a été attiré par un des points précédents)

B) le point n'est pas déjà classé.

Dans le premier cas, le point Q attire dans son groupe tous ceux parmi ses voisins qui ne sont pas encore classés. On dit alors que Q est l'attracteur de ces points et on conserve cette information dans un tableau qui à chaque point fait correspondre son attracteur (ou O si le point est le centre d'un nouveau groupe). Si un des voisins, P, est déjà classé dans un autre groupe que celui du point Q, on est en présence d'un lien entre deux groupes. Nous convenons alors de dire que celui de Q et de P qui a la plus faible concentration est le point qui crée le lien. Nous vérifions alors si un lien avait déjà été créé entre les groupes en question. Si oui, nous ne conservons que le lien créé par le point ayant la plus forte concentration, (recherche du point col).

Dans le second cas nous devons choisir entre la création d'un nouveau groupe ou le rattachement du point Q à un groupe existant. Le critère sera le suivant : si le plus proche voisin de Q est plus dense (possède une concentration plus forte ou une dispersion plus faible) nous savons que ce point est nécessairement déjà classé et nous rattachons Q à son groupe. Dans le cas contraire, nous créons un nouveau groupe et le traitement des voisins de Q est celui décrit pour le premier cas.

Par conséquent, lorsque l'algorithme est terminé nous avons à notre disposition les informations suivantes.

1) Pour chaque point un numéro de groupe et son attracteur dans le groupe. Cela nous permet de construire une liste des groupes et d'en tracer un graphe (arbre) à partir du tableau des attracteurs. Ce graphe peut fournir une image des liens entre les éléments du groupe. Il pourra servir à décider si nous somme en présence d'un groupe "naturel" ou d'un chemin continu.

2) Nous avons aussi une liste des liens inter-groupes, qui définit à son tour un graphe, duquel on peut tirer un arbre en ne conservant entre deux groupes que le chemin de densité maximum. Cet arbre peut ensuite servir à construire une hiérarchie parmi nos groupes.

Nous travaillons en ce moment sur la partie graphique de la présentation de ces résultats. Mais voyons plutôt sur des exemples les renseignements que l'on peut en tirer.

6 : RESULTATS.

Exemple 1) La figure 6.1 présente, dans le plan, un ensemble de points qui sera fourni à l'algorithme de classification. L'œil humain reconnait facilement la présence de deux groupes liées par un col (un sablier ?). Les points sont disposés uniformément sur une grille et toute méthode de classification basée uniquement sur la distance conduira à des résultats fantaisistes. Cependant pour un nombre de points voisins m assez grand on peut s'attendre à observer un minimum de concentration dans le col. En effet, les points du col doivent

chercher plus loin que les autres leurs voisins (de même que les points périphériques). Les numéros de la figure 6.1 représentant la classification obtenue avec 8 points voisins m = p = 8. On voit donc que notre algorithme se débrouille très bien dans une situation reconnue comme étant des plus difficiles en classification automatique.

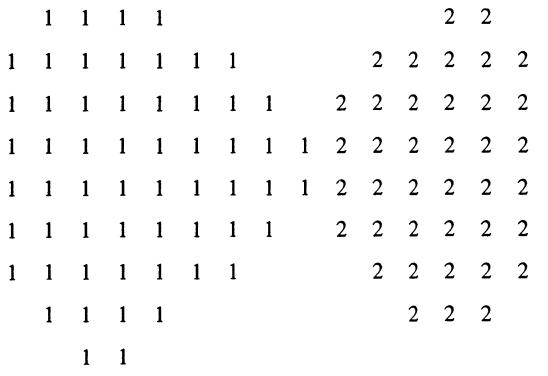

Figure 6.1 – Sablier (points équidistants)

Exemple 2) Les données de Fisher sur les Iris sont classiques en analyse discriminante et en classification automatique. Rappelons qu'il s'agit de 150 individus de trois espèces d'Iris à savoir Setosa, Versicolor et Virginica sur lesquels on a pris 4 mesures : longueur et largeur des pétales ainsi que des sépales. Nous avons donc à classifier un ensemble de 150 individus, chacun des individus étant caractérisé par un vecteur de $\mathbf{R}^4$. Après avoir normalisé les variables i.e. soustrait la moyenne et divisé par l'écart-type, nous avons utilisé notre algorithme en nous servant de la distance euclidienne. Nous avons rapidement constaté que Iris Setosa formait un groupe aisément reconnaissable et nous sommes restreints à traiter Iris Versicolor et Iris Virginica (50 individus respectivement). Ces deux espèces se recoupent et il n'est pas aisé de distinguer une frontière valable.

Nous avons utilisé ici m = 8, p = 5. Ces valeurs ont été choisies au hasard, mais les résultats pour d'autres valeurs montrent qu.elles sont représentatives. Les résultats bruts nous donnent l'existence de 9 groupes ou structures distinctes. Ces groupes sont formés respectivement comme suit :

Groupe 1 : 19 Versicolor, 1 Virginica
Groupe 2 : 19 Versicolor, 1 Virginica
Groupe 3 : 16 Virginica
Groupe 4 : 10 Virginica, 1 Versicolor
Groupe 5 : 10 Virginica, 7 Versicolor
Groupe 6 : 5 Virginica,
Groupe 7 : 6 Virginica,
Groupe 8 : 4 Versicolor,
Groupe 9 : 1 Virginica.

Cette décomposition entre Versicolor et Virginica résulte évidemment d'une connaissance a priori, à savoir que nos individus appartiennent à deux espèces distinctes. Considérons maintenant le graphe des liens entre les groupes.

Dans le graphe présenté les liens sont pondérés par le numéro dans l'ordre des densités du point formant le lien. Il convient aussi de préciser par rapport au même ordre, le numéro du point ayant créé le groupe. Un groupe apparaissant tôt sera d'après nos conventions plus cohérent qu'un groupe apparaissant en queue de liste où on retrouve des points périphériques plus ou moins isolés. Afin de simplifier la lecture du graphe nous n'avons retenu qu'un arbre générateur minimal, (le chemin pour passer d'un nœud à un autre est minimal pour les numéros d'ordre de densité).

Sous nos conventions voici donc le graphe.

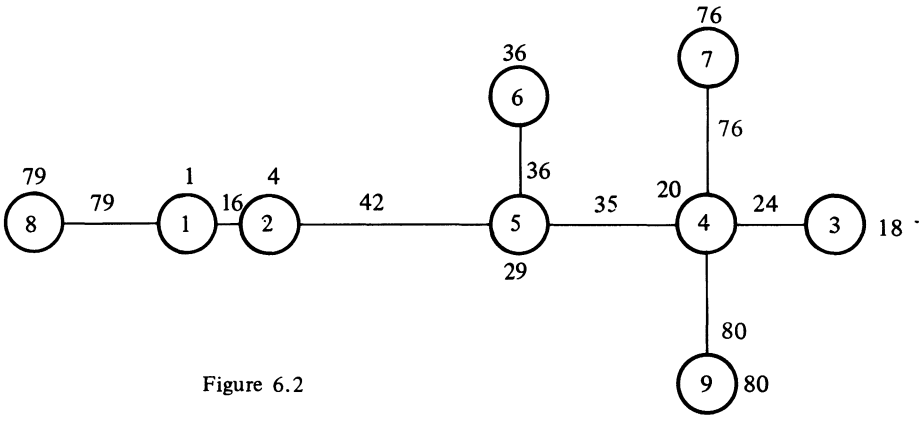

Figure 6.2

Nous pouvons maintenant tirer de cet arbre une hiérarchie selon un procédé classique (cf [4]) en reportant les liens par ordre d'apparition. L'axe de la hiérarchie est ici défini par l'ordre des dispersions et concentrations et non par les distances comme dans une hiérarchie normale. Le résultat est le suivant

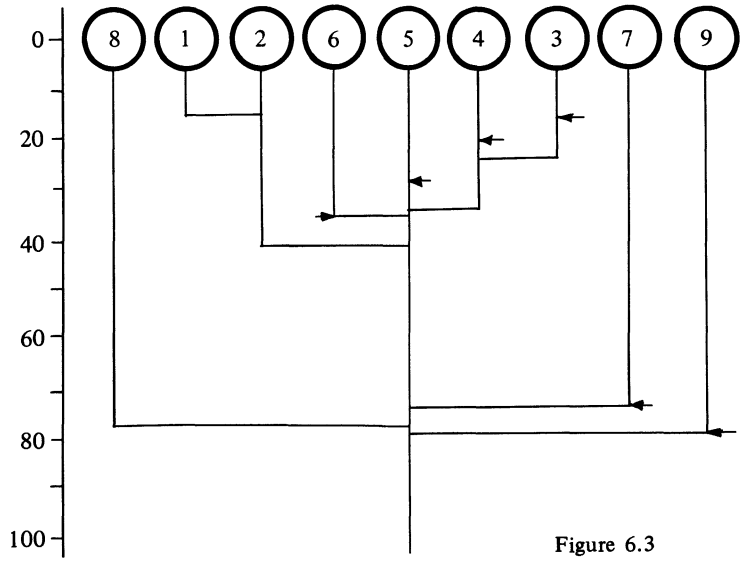

Figure 6.3

Tentons momentanément d'oublier toute connaissance a priori sur l'existence de deux espèces et tentons d'interpréter les figures 6.2 et 6.3.

Tout d'abord on constate que les groupes 6, 7, 8, 9 se rattachent dès leur formation à un groupe déjà existant : ainsi le groupe 7 est créé par le point 76 qui crée aussi le lien de groupe avec le groupe 4. Par conséquent, sauf si l'on s'intéresse aux structures locales on pourra regrouper 7 et 9 avec 4, 6 avec 5 et 8 avec 1. Considérons maintenant les cinq premiers groupes qui sont séparés de leurs voisins par de véritables vallées.

On distingue en premier lieu deux regroupements principaux : les groupes 1 et 2 d'une part et les groupes 3 et 4 auxquels se rattachent ensuite 5 et 6 d'autre part. Par ailleurs la figure 6.2 nous indique que le groupe 5 est une sorte d'intermédiaire entre les groupes 2 et 4.

Nous pouvons donc en déduire la description suivante : Les groupes 1 et 2 d'une part, 3 et 4 d'autre part correspondent à deux structures bien définies. Le groupe 5 se rattache d'abord à 3 et 4 mais il est aussi un intermédiaire entre 2 et 4. Les groupes 6, 7, 8, 9 sont peu différenciées et se rattachent à 5, 4, 1, 4 respectivement. Les groupes 7, 8, 9 sont des structures locales et périphériques (faible concentration, peu de différenciation).

Si on compare avec la description à priori on s'aperçoit effectivement que 1 et 2 correspondent à Versicolor, 3 et 4 à Virginica et que 5 est effectivement composé de représentants de ces deux espèces.

Si on place sur la figure 6.2 la coupure entre 2 et 5 on obtient en tout 10 individus mal classés ce qui est très correct étant donné la difficulté du problème.

Exemple 3) La figure 6.4 représente les données de Ruspini (cf [3]) l'algorithme sépare sans difficulté les 4 groupes principaux. La variation des paramètres permet de discerner avec plus ou moins de netteté les structures locales. Les résultats présentés ont été obtenus avec m = p = 5.

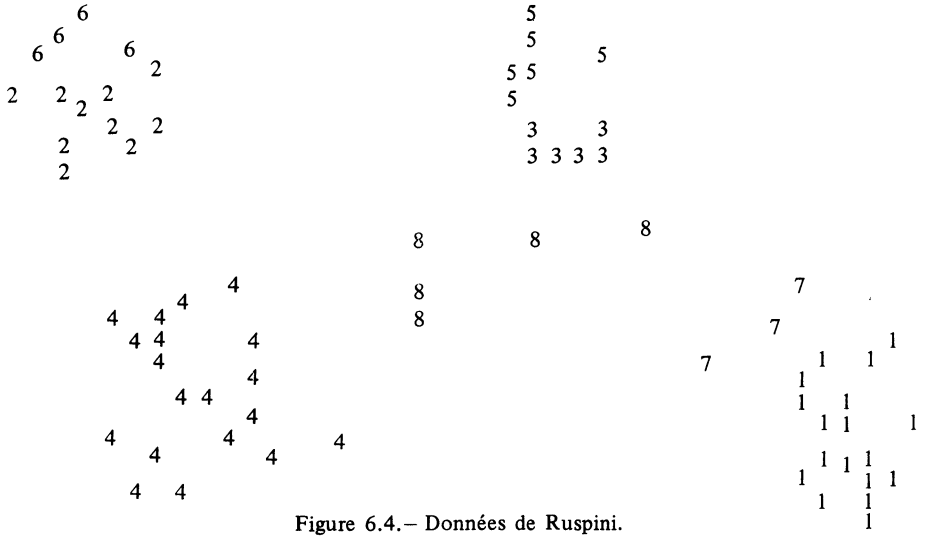

Figure 6.4.– Données de Ruspini.

§ 7 : CONCLUSION

Les approches au problème de la reconnaissance des formes et de la classification automatiques sont multiples et il ne nous est guère possible ici de comparer de façon complète l'algorithme que nous proposons aux méthodes déjà existantes. Nous renverrons aux ouvrages de Benzecri [1] et de Sokal [6] pour une vue générale de la question et à [2], [3], [4] et [7] pour quelques considérations précises ayant inspiré notre travail. Disons simplement que la notion de classification est floue et qu'il serait du plus haut intérêt d'en faire l'étude théorique.

Pour ce qui est de notre algorithme, nous croyons que ses principales qualités sont les suivantes : il n'est ni aléatoire ni itératif mais parfaitement "déterministe" et de plus très rapide par rapport à d'autres méthodes possédant ces mêmes qualités comme les algorithmes de classification hiérarchique. (Par exemple sur IBM 370-155, il nous faut moins de 2 secondes pour effectuer la classification de 100 points de $\mathbf{R}^4$). De plus on peut en tirer plus d'information sur les liens entre les objets que d'une classification hiérarchique normale.

Une autre qualité est l'adaptabilité. Celle-ci suppose cependant une interaction homme-machine constante : l'utilisateur doit être en mesure d'apprécier rapidement la qualité de l'image qui lui est fournie afin de choisir celle qui lui parait la mieux adaptée à ses besoins. Nous espérons adjoindre d'ici peu à notre programme des moyens graphiques permettant une visualisation immédiate des résultats, en particulier des différents graphes fournis par le programme.

Disons pour terminer que nous nous proposons d'étudier aussi de façon plus précise la notion de niveau d'une classification que nous avons introduite ici de façon intuitive.

(On peut obtenir le programme FORTRAN utilisé en écrivant à l'auteur au Département de Mathématiques de l'Université Laval. Pour obtenir une copie sur ruban magnétique il conviendrait de fournir le ruban. L'ordinateur de l'Université Laval est un IBM 370-158, et pour les usagers de calculateurs autres que ceux de IBM, il sera souhaitable de vérifier au préalable la compatibilité des systèmes).

BIBLIOGRAPHIE

[1] BENZECRI J.P. – L'analyse des données. Tomes I – II Dunod, Paris 1973.

[2] DIDAY E. – Une nouvelle méthode en classification automatique et reconnaissance des formes : la méthode des nuées dynamiques. *Revue de Statistique Appliquée* vol XIX, no. 2, 1970.

[3] DIDAY E. – The Dynamic Clusters Method in Non-Hierarchical Clustering. Rapport IRIA-72004 Informatique Numérique. Domaine de Voluceau 78 Rocquencourt France.

[4] GOWER J.C., ROSS G.J.S. – Minimum Spanning Trees and Single Linkage Cluster Analysis. *Appl. Stat.* 18, p. 54-64, 1969.

[5] RUBIN J. – Optimal Classification into Groups : an Approach for Solving the Taxonomy Problem. *J. Theoret. Biol.* 15, p. 103-144, 1967.

[6] SOKAL R.R. – Numerical Taxonomy (nouvelle édition à paraître).

[7] ZAHN C.T. – Graph Theoretical Methods for Detecting and Describing Gestalt Clusters. *IEEE Transactions on Computers,* vol C-20, no 1, pp. 68-86, 1971.