Statistique et analyse des données

Résumés des communications faites à Vannes dans le cadre des premières journées nationales sur la classification

Statistique et analyse des données, tome 2, n° 2 (1977), p. 27-44.

<http://www.numdam.org/item?id=SAD_1977__2_2_27_0>

METHODES COMBINATOIRES ET STATISTIQUES DANS LE TRAITEMENT DES DONNEES DU COMPORTEMENT

par I. C. LERMAN
Département de Mathématiques et Informatique - I.R.I.S.A.
Université de Rennes - B.P. 25 A - 35031 - RENNES CEDEX
N° Tél : 36/48/15

Nous tenterons dans cet exposé qui s'inscrit sous la rubrique "Bilans et Perspectives" de faire le point sur l'ensemble des méthodes et réalisations que nous avons, au bout de longues années, élaboré avec l'aide et la collaboration de nombreux chercheurs en Statistique et Informatique Appliquées. Ces travaux ont un triple caractère : Combinatoire et Statistique pour l'élaboration des méthodes, Informatique pour leur mise en oeuvre et de Participation à la Recherche dans diverses disciplines des Sciences Humaines, Economiques et de la Nature, pour la validité des méthodes, leur progrès et leur promotion.

La plupart des méthodes couramment utilisées en Analyse des Données se réfèrent à une représentation euclidienne. Nos méthodes qui se refèrent à une représentation mathématique finie constituent une véritable stratégie combinatoire et statistique dans le traitement des grands tableaux des données. S'il s'agit de situer ces méthodes par rapport à celles, très pratiquées en France, de J.P. Benzecri, on peut signaler rapidement que ces dernières sont géométriques et procèdent de l'analyse métrique de la dépendance ; alors que les nôtres sont essentiellement combinatoires et procèdent de l'analyse de la corrélation entre structures finies de même type. Cela permet de rester plus près du langage posé par le Spécialiste et de s'adapter avec souplesse à n'importe quel type de tableau de données.

En effet, les variables descriptives, telles qu'elles se manifestent dans les Sciences de l'Homme et de la Nature, sont rarement numériques et il importe que la méthode de synthèse de l'Information respecte la représentation naturelle des données. D'autre part, la structure de condensation la plus adéquate pour le problème posé par le spécialiste a le plus souvent un caractère fini ; partition, chaîne de partitions, ordre,... et le critère de synthèse doit en tenir compte tout en ayant un fondement statistique clair. Enfin, les algorithmes proposés doivent être justifiés du point de vue du critère qu'ils optimisent. Ce sont ces points de vue que nous avons cherché à développer de façon assez systématique dans les méthodes présentées.

ETUDE DE DIFFERENTES FORMES DE "REPRESENTATION" EN CLASSIFICATION AUTOMATIQUE : BILAN, RETOMBEES, PERSPECTIVES.

par E. DIDAY
IRIA et Université de Paris IX
Domaine de Voluceau - Le Chesnay - 78150 - ROCQUENCOURT
N° de Tél : 954/90/20

Après avoir étudié une famille d'algorithmes de partitionnement efficace et apte à traiter des grands tableaux ; nous nous sommes penchés sur le problème de la représentation d'une classe d'objets. Différents types de représentation ont été choisis et ont débouché sur des programmes nouveaux aux retombées diverses : analyse factorielle typologique, lissage typologique, sélection typologique de paramètres, régression typologique, analyse discriminante typologique, la décomposition de mélanges de lois de probabilité, les distances adaptatives.

REMARQUES SUR LES DISTANCES ENTRE PARTITIONS

par : J.P. BARTHELEMY

Ecole Nationale Supérieure de Chronométrie et de Micromécanique

25030 - BESANÇON CEDEX
N° de Tél : 80/78/33 (poste 47)

Une démarche fondamentale, en classification, est de construire une partition qui minimise un certain critère. Lorsqu'il s'agit de construire l'agrégation d'une famille a de partitions il est naturel de choisir comme critère une "fonction d'éloignement" d'une partition à a. C'est ce qu'a fait Régnier, il y a longtemps déjà, en prenant pour fonction d'éloignement : $\Delta(P,a) = \sum_{Q \in a} D(P,Q)$ où $D(P,Q)$ est le cardinal de la différence symétrique des relations d'équivalence associées à P et Q.

D'une manière générale si π est l'ensemble des partitions d'un ensemble fini X et si D est une distance sur π, pour tout $a = (Q_1, \ldots, Q_m) \in \pi^m$, on peut définir l'éloignement de P à a en procédant comme suit : On considère le vecteur $\vec{V}(P,a) = (D(P,Q_1), \ldots, D(P,Q_m))$ de R^m et on pose : $\Delta(P,A) = N(\vec{V}(P,a))$ où N est une norme sur l'espace vectoriel R^m (à noter que l'on peut "récupérer" dans l'expression de la norme une pondération sur les Q_i).

Il est donc essentiel, pour la construction de ces fonctions d'éloignement, de "répertorier" les distances sur , c'est ce que l'on se propose de faire ici :

1°) Il y a les distances induites par une distance sur 2^X, elles sont de la forme $D(P,Q) = \sum_{A \in P,\ B \in P} v(A \cap P)\, d(A,B)$ où v est une mesure strictement positive sur X et d'une distance sur 2^X.

2°) Il y a aussi les distances définies à partir de la structure de treillis de on retrouve comme cas particuliers certaines des distances du 1°), plus la distance induite par "la hauteur".

3°) Dans le cas où l'on se borne à étudier les partitions comprenant un nombre donné de classes, on obtient des distances liées à la notion de "transfert".

PARTITIONS OPTIMALES D'UN HYPERGRAPHE

par : S. REGNIER

Centre de Mathématiques Appliquées et de Calcul
54, bd Raspail - 75270 - PARIS - CEDEX 06
N° de Tél / 544/38/49

RESUME :

Un hypergraphe $H = (E, A_s)$ $(s \in S)$, est caractérisé par sa matrice d'incidence

$$n^{H^m} = (h_i^s),\ (i \in E \text{ et } s \in S)$$

$h_i^s = 1$ ssi le sommet i appartient à l'arête A_s

$h_i^s = 0$ sinon.

Une telle donnée initiale est classique en taxonomie : un ensemble muni d'attributs dichotomiques.

Le problème suivant semble entièrement neuf : trouver des partitions X de E telles que le nombre total des coupures d'arêtes soit minimum ou plus précisément que, si x_s = nombre de classe de la restriction X^s de X à A_s $\varphi(X) = \sum_{s \in S} x_s$ soit minimum.

Il faut naturellement imposer $\left|\frac{E}{X}\right| = x \geqslant k$, sinon $x = 1$ donne $\varphi(X) = S = m$ = minimum absolu. Alors X s'appelle une partition k-optimale de H.

On établit facilement alors que toutes les X optimales ont $x = k$ classes, du moins quand k dépasse le nombre p de composantes connexes de H.

La méthode classique des partitions centrales peut aussi définir des partitions centrales Y de l'hypergraphe. Il suffit que chaque arête A_s soit envisagée comme une partition P_s où la seule classe non atomique est A_s

Cette fois le nombre de Y-classe peut être borné supérieurement. Il n'est pas exclus que les Y puissent être une bonne approximation des X, mais l'étude précise de l'ensemble R(H,k) des partitions k-optimales comporte beaucoup de questions ouvertes.

CLASSIFICATION ET RECONNAISSANCE DES STRUCTURES : L'APPROCHE CATEGORIQUE.

par : Mme M. PAVEL -
Laboratoire de Calcul des Probabilités associé au CNRS de l'Université de Paris VI.
Université de Paris V.

RESUME :

Nous montrons dans cette Communication sous quelles conditions on peut organiser des images $I \in \mathcal{J}$ et des déformations $\delta_{\mu\nu} \in \Delta$ en une catégorie, exprimons les sondages et les fonctions de reconnaissance à l'aide des foncteurs invariants, et étudions la relation existant entre $\mathcal{J}, \mathcal{J}^{\Delta}$, et la catégorie $\mathcal{P}$ des structures ou formes canoniques en utilisant la notion de rétracte. Nous continuons à investiguer la reconnaissance des structures dans des catégories : (I) en exhibant la signification pratique des rétractes et en introduisant les projections et leurs propriétés d'extension ; (2) en étudiant le problème de reconnaissance associé à la catégorie originelle d'images déformées $\mathcal{J}^{\Delta}$ et en introduisant le squelette Σ de $\mathcal{J}^{\Delta}$; (3) en mettant en évidence la relation existant entre squelettes et projections ; et (4) en exhibant la liaison qui existe entre notre formalisme catégorique de classification et de reconnaissance et les applications courantes.

APPROCHE GALOISIENNE DES CLASSIFICATIONS

par : A. DEGENNE* - C. FLAMENT** - P. VERGES***-

* LEST (CNRS) - Chemin du Coton Rouge - 13100 - AIX-EN-PROVENCE -

** Université de Provence - AIX-EN-PROVENCE -

*** IRPEACS (CNRS) - AIX-EN-PROVENCE -

N° de Tél : (91) 26/59/60/ *

Soit un tableau, XxY, de correspondance, en 0,1.

Soit (α , β) la correspondance de Galois induite par ce tableau :

$\alpha : \mathcal{P}(X) \rightarrow \mathcal{P}(Y)$. $\forall A \subset X$; $\alpha(A) = \{y \leftarrow Y;\ \forall x \leftarrow A,\ (x,y)=1\}$

Dualement on définit : $\beta : \mathcal{P}(Y) \rightarrow \mathcal{P}(X)$

On sait que l'application $\beta \circ \alpha$ de $\mathcal{P}(X)$ dans lui même est une fermeture.

Nous considérons une partition G de X en classes $\{C_i\}$, telle que $\forall$ i, $\beta \circ \alpha (C_i) = C_i$

On présentera un algorithme de recherche de telles partitions galoisiennes (il n'y a pas toujours une solution).

Par ailleurs on étudie dans Y les différents types de structures de la famille $\{\alpha (C_i)\}$

Nota : Pour l'étude des fermetures et correspondances de Galois, on se reportera à : Marc BARBUT et Bernard MONJARDET : Ordre et classification Algèbre et Combinatoire, Tome II, Chapitre V, Paris, Hachette, 1970.

METHODE DE CLASSIFICATION CONJOINTE - COCLASS

par : J. BROUSSE - GSI-DITS
J. TOLEDANO - IRIS - Paris IX
G. SLABODSKY - Paris VI - GSI-DITS

GSI-DITS - 69, rue Legendre - 75017 - PARIS
N° de Tél : 627/65/00

La méthode envisagée est une méthode de classification conjointe agissant simultanément sur les lignes et les colonnes d'un tableau pour lequel on cherche une classification des deux dimensions lignes et colonnes, satisfaisant à des contraintes d'homogénéité des classes mises en évidence et de correspondances (identification mutuelle) entre les deux partitions obtenues. L'analyse d'un critère de stabilité permet, par ailleurs, de fixer le niveau auquel on arrête les regroupements des lignes et des colonnes.

Les algorithmes de calcul sont construits sur la distance du χ^2 et l'analyse de l'information mutuelle (Shannon).

On présentera des exemples d'utilisation notamment dans le domaine de l'analyse du corps social pour laquelle cette méthode a été conçue (analyse d'un tableau croisant les 40 CS à 2 chiffres de l'Insée et 200 indicateurs de comportement socio-économique).

AGREGATION SUIVANT LA VARIANCE LOCALEMENT NORMALISEE ET LA CLASSIFICATION HIERARCHIQUE

par P. RIOUX
INSERM U 88 - 91, bd de l'hôpital - 75634 - PARIS CEDEX 13
N° de Tél : 707/67/79 - Poste : 488.

RESUME :

Le critère d'agrégation de deux groupes i et j en un groupe k que nous proposons correspond, pour toutes les réunions possibles de groupes i et j à chaque pas de la construction d'une hiérarchie, au calcul de :

$$\text{Trace } (T_k^{-1} B_k) . \det (T_k)^{1/p}$$

avec T_k - matrice de variance-covariance totale calculée sur la réunion en un groupe k des deux groupes i et j donnés

B_k - matrice de variance-covariance intergroupes correspondante

p - dimension de ces deux matrices

L'utilisation de ce critère nous a posé d'importants problèmes d'analyse numérique et informatique, mais semble maintenant pouvoir permettre d'obtenir des résultats nettement plus satisfaisants que ceux obtenus avec d'autres méthodes de classification utilisables actuellement.

UN ALGORITHME RAPIDE DE CLASSIFICATION ASCENDANTE HIERARCHIQUE D'UN GRAND ENSEMBLE DE DONNEES

par M. BRUYNOOGHE
Département Transport-Logistique - Institut Universitaire de Technologie - Avenue Gaston Berger - 13100 - AIX-EN-PROVENCE

Le nouvel algorithme de classification ascendante hiérarchique, ici présenté, permet le traitement d'ensembles de données composés de quelques milliers d'objets.

La hiérarchie binaire de parties construite par cet algorithme sur l'ensemble des objets à classer est la même que celle qu'aurait engendrée l'algorithme classique.

L'algorithme proposé se caractérise par sa rapidité et le peu de place mémoire demandée, surtout dans le cas de la méthode d'agrégation selon la variance, la métrique de base étant celle du chi-d'eux, les distances entre les objets à classifier étant directement calculées à partir du tableau des facteurs issus d'une analyse factorielle des correspondances.

La version de l'algorithme classique de classification arborescente qui calculerait systématiquement les distances entre objets à partir des facteurs se caractériserait par le peu de place demandée. Cependant, pour édifier une hiérarchie sur un ensemble de 1500 objets, il serait nécessaire de calculer $0,56\ 10^9$ distances à partir du tableau des facteurs.

Le nouvel algorithme ne calcule que $1,59\ 10^6$ distances et construit une classification arborescente sur un ensemble de 1500 objets en 40 secondes, sur un ordinateur IBM 370-168.

Enfin, le temps de calcul de l'algorithme propose, augmente approximativement comme le carré du nombre d'objets à classer et non plus comme le cube ; il est donc possible d'ensivager l'analyse hiérarchique d'un grand ensemble de données en un temps de calcul acceptable.

P. COLLOMB
J.L. MOLLIERE

ELECTRICITE DE FRANCE
Direction des Etudes et Recherches
Service I.M.A.
1, Ave du Général de Gaulle
92141 CLAMART CEDEX
Tél 645 21.61

T I T R E

PRESENTATION D'UN ENSEMBLE DE MODULES DE CLASSIFICATION ET UN EXEMPLE D'APPLICATION

1. DESCRIPTION DES MODULES

Ces modules, basés sur la méthode des nuées dynamiques permettent d'une part de calculer les formes fortes, d'autre part de les décrire de façon très complète. On effectue de plus une classification hiérarchique ascendante de celles-ci : l'arbre hiérarchique obtenu fournit un critère pour le choix d'un nombre de types dans la population. La description desformes fortes ou des types résultant des regroupements peut être demandée sur les variables actives de la classification ou sur des variables passives.

Un grand choix d'options est laissé dans le module des nuées dynamiques (choix de distances, des noyaux de départ, etc...). Un module particulier traite le cas des variables qualitatives avec la distance du χ^2.

Une grande souplesse est prévue dans l'articulation de ces modules pour obtenir aisément les descriptions des types aux niveaux de regroupement des formes fortes désirés.

Des descriptions par analyse factorielle peuvent être enchaînées facilement.

2. APPLICATION

Essai de classification des clients E.D.F. basse tension en fonction de leur courbe de consommation journalière : on dispose, pour 453 clients basse tension issus d'une campagne de mesures, des valeurs de leurs appels de puissance demi-heure par demi-heure pour une journée donnée d'hiver (48 variables quantitatives).

Remarque :

Les modules réunis ici dans une chaîne où ils ont des fonctions complémentaires sont issus de programmes qui existaient séparément par ailleurs.

CLASSIFICATION UTILISANT LA NOTION DE VOISINAGE

par G. GOVAERT - Y/ LECHEVALLIER
IRIA - Domaine de Voluceau - Le Chesnay - 78150 - ROCQUENCOURT
N° de Tél : 954/90/20

Les classifications proposées ici sont uniquement obtenues à partir du tableau des k plus proches voisins (k constante arbitraire). Elles ne tiennent pas compte de la valeur des distances entre ces points. Cela revient à travailler sur un graphe orienté ou non.

La recherche des classes connexes sur un tel graphe permet de bien reconnaître les classes sur différents exemples simples. Ces exemples sont mal reconnus par des méthodes "métriques" habituelles (Méthodes utilisant directement le tableau de distances).

Pour trouver des groupes à l'intérieur de ces classes connexes nous avons utilisé un algorithme d'échange optimisant comme critère l'indice de séparabilité de l'analyse discriminante et une méthode hiérarchique basée sur différents indices de similarité permettant de retrouver systématiquement, si elles existent, les classes connexes dans le graphe induit par les k plus proches voisins.

LA METHODE DES POLES D'ATTRACTION

(Deux algorithmes de classification automatique, un algorithme d'analyse de la sériation)

par H. LEREDDE
Université de Paris-Nord - Département de Mathématiques -
Avenue J.B. Clément - 93430 - VILLETANEUSE -
N° de Tél : 820/61/70

L'étude de la variance des proximités des éléments d'un tableau de données offre la possibilité d'analyser le phénomène de la sériation. Nous avons mis au point un algorithme conduisant à une représentation graphique autour de deux axes non-orthogonaux qui permet la visualisation du phénomène de sériation. Ce système de représentation, qui peut être étendu à plusieurs axes pris deux à deux, présente certaines analogies avec les représentations en analyse factorielle ou en analyse des correspondances. Cet algorithme ne nécessite aucune diagonalisation de matrice, d'où sa très grande rapidité.

Dans la continuité de ce travail est née une nouvelle méthode de classification automatique : la méthode des Pôles d'Attraction. Cette méthode opère une classification soit sur les variables, soit sur les observations d'un tableau de données descriptives ou numériques. Elle génère une suite de partitions des éléments à classer. A chaque étape le nombre de classes d'une partition est augmenté de un ; pour former une nouvelle partition, nous considérons l'ensemble des éléments à classer dans sa globalité et non la partition précédente : il s'agit donc d'une méthode de classification automatique non-hiérarchique. De cette suite de partitions sont extraites les partitions les plus "significatives" au regard de deux critères. Cette méthode se décompose en deux algorithmes distincts : l'un travaille sur une matrice de similarités, l'autre sur une matrice de distances entre éléments de l'ensemble. Le premier de ces deux algorithmes, Attraction-Similarités, inclut l'algorithme décrit pour l'étude de la sériation ; le second, Attraction- Distances, ne possède pas de système de représentation graphique.

Nous avons comparé les résultats produits par ces deux algorithmes avec ceux fournis par l'analyse factorielle, l'analyse des correspondances, la méthode des nuées dynamiques,... Ces tests ont été pratiqués sur une douzaine de tableaux de données réelles (psychologie, sociologie, économie, archéologie, sciences naturelles, médecine,...) et une centaine de tableaux simulés, dont certains atteignent des tailles de 500x5000 et même 1000x10000.

CLASSIFICATION D'UNE FAMILLE D'ECHELLES AU MOYEN D'UN NOUVEL INDICE. APPLICATION A DES DONNEES EN PSYCHO-PEDAGOGIE.

par : I. COHEN
Laboratoire de Météorologie Dynamique du C.N.R.S.
Ecole Polytechnique - R.D. 36 - 91128 - PALAISEAU CEDEX
N° de Tél : 941/82/00

Ce travail porte sur l'évaluation de la méthode de classification "hiérarchique" de I.C. LERMAN pour l'organisation de l'ensemble des items d'un questionnaire en classes et sous classes de proximités. Chacun des items correspond à une échelle d'attitude définissant un préordre total sur l'ensemble que constitue l'échantillon de la population étudiée ; le nombre de modalités par item n'étant pas constant d'une question à l'autre.

Le support concret de la recherche concerne une importante enquête psycho-pédagogique (4000 sujets décrits au moyen d'une centaine d'échelles) portant sur le développement de la petite enfance et les opérations scolaires dans le contexte des relations parentales.

Pour la comparaison de telles variables, il y avait bien un indice qu'avait proposé M. G. Kendall mais où, pour l'établir, on ne retenait qu'une des formes de l'algorithme de calcul de l'indice τ de comparaison d'un couple d'ordres totaux. En fait, l'indice de Lerman qui généralise celui, τ, permet de voir que l'indice de M. G. Kendall pour la comparaison d'un couple de préordres totaux est "biaisé".

Bien que notre optique ne soit pas celle des tests d'hyptothèses il importait de comparer les résultats de la synthèse automatique lorsqu'on remplace l'indice de Kendall par celui, nouveau. On montrera que l'organisation des liens faibles au moyen du nouvel indice rejoint les hypothèses les plus profondes du Psychologue.

D'autre part en "oubliant" la structure sous-jacente à l'ensemble des modalités d'un même item, on a attaché à chaque modalité d'un item un attribut descriptif, ce qui a conduit à un tableau d'incidence 250x4000. Nous avons dans ces conditions cherché à analyser le type de résultat qu'on obtient avec ce nouveau codage par rapport au précédent. On verra que si le premier traitement permet de dégager les principales tendances du comportement de la population étudiée ; ce dernier montre plutôt des profils ou types d'attitudes qui "expliquent" notamment l'apparition des précédentes tendances.

Cette analyse expérimentale n'a pas été sans l'élaboration d'un important programme TAUX qui établit, de façon optimale, le tableau des proximités entre échelles pour de très gros fichiers.

LES DIFFERENTES FORMES DE L'APPREHENSION DES DONNEES SANS L'EXPLORATION FONCTIONNELLE HEPATHIQUE :

par Jean-Yves LAFAYE
Département Statistique
I.U.T. due Montaigne
56 008 VANNES
Tél.: (97) 66-45-46

Le développement des méthodes d'analyse classificatoires met à la disposition du statisticien un très large éventail de techniques, d'algorithmes, et de mesures de la similarité.

Il s'agit ici de proposer une stratégie face à un type particulier de données.

L'étude se développe dans deux cadres particuliers :

- traitement de données numériques, qui donne lieu à la généralisation d'indices basés sur la corrélation. Ceci permet de retrouver la "géographie" classique de la pathologie hépato-biliaire et aboutit à la définition de profils biologiques.

- traitement de variables nominales obtenues après un découpage adéquat des variables numérique en caractères à plusieurs modalités. Les résultats sont alors moins globaux et l'on peut préciser le rôle spécifique des différents enzymes dans le diagnostic biologique.

Une étude comparative permet d'évaluer l'influence de l'appauvrissement du codage sur la pertinence des résultats, ainsi que de tirer des règles générales concernant la forme des hypothèses d'absence de lien à prendre en compte.

-:-:-:-

APPLICATION D'UNE METHODE DE CLASSIFICATION (méthode k-means) SUR LA REPARTITION DE PUCERONS DANS UNE PARCELLE BOCAGERE

par : C. DERVIN* - C. JACOB** - C. LESTY* -

* I.N.R.A. - C.N.R.Z. - JOUY EN JOSAS

** I.N.R.A. - C.N.R.A. - VERSAILLES

N° de Tél : 956/80/80*

950/75/22**

RESUME :

Dans le cadre d'études sur le bocage breton, on désire étudier l'influence des haies sur la répartition spatiale des pucerons. Dans ce but, des pièges (de deux types, jaunes et à succion) sont disposés à 4 hauteurs et 7 distances de l'une des haies d'une parcelle, le long d'un transect perpendiculaire à celle-ci. Les observations sont d'une part les effectifs de pucerons (identifiés par espèce) capturés durant des intervalles de temps variables de piègeage (de l'ordre de 3 heures, la journée et 13 à 14 h la nuit), et d'autre part des mesures climatiques de vitesse de vent et de température. La forme de l'expérimentation impose un niveau d'étude "ponctuel", à la fois spatial (i.e. parcellaire) et temporel (durée d'une semaine). Le problème est complexe, en partie parce qu'il concerne un grand nombre d'effets concomitants (en particulier nombre important d'espèces, pour la plupart peu représentées (ou même pas du tout)). L'énorme masse des données a impliqué un premier travail de "nettoyage" consistant en une classification de tous les tableaux de contingence N_{ij} (effectif capturé à la distance i et hauteur j) relatifs à chaque espèce et à chaque relevé. La méthode est une méthode itérative à partition dont le principe est "l'amélioration" à chaque stade de la partition précédente. La distance utilisée s'adaptant le mieux au type des données est celle du χ^2. La partition finale permet d'une part de fournir des répétitions à chaque tableau type N_{ij} et donc d'augmenter la précision d'analyses ultérieures (cartographie dans le plan distance x hauteur) et d'autre part de comparer les espèces relativement à leur répartition spatiale dans la parcelle.

METHODE DE CLASSIFICATION AUTOMATIQUE SOUS CONTRAINTES SPATIALES

par : P. MONESTIEZ
Centre National de Recherches Forestières
Champenous - 54280 - SEICHAMPS
N° de Tél : 26/61/31/

RESUME

Les individus que l'on cherche à classer sont des points, des parcelles ou des zones d'un champ spatial, décrits par plusieurs variables. En tenant compte de leur positions géographique, la méthode répond à deux problèmes :

1° - déterminer des zones spatiales d'un seul tenant qui soient le plus homogènes possible,

2° - obtenir une description du champ (cartographie) tenant compte de plusieurs variables simultanément.

Le principe de la méthode consiste à n'agréger deux classes que si elles vérifient un critère de voisinage. On comparera les résultats obtenus par cette méthode et par une CAH classique, pour divers critères d'agrégation, en exhibant les avantages et les inconvénients de chacune. On notera aussi que cette méthode permet de traiter des champs spatiaux plus vastes du fait de la rapidité de l'algorithme.

UNE ETUDE DE L'ECONOMIE AGRICOLE DES DEPARTEMENTS FRANCAIS PAR LA METHODE DE LA CLASSIFICATION AUTOMATIQUE.

par : B. TALLUR
Laboratoire de Statistiques - Département de Mathématiques et Informatique - Université de Rennes - I.R.I.S.A. -
B.P. 25 A - 35031 - RENNES CEDEX
N° de Tél : 36/48/15

RESUME

Il s'agit de classifier l'ensemble des départements français, chaque département étant caractérisé par trois groupes de variables qualitatives :

1° surfaces par principales cultures telles que céréales, plantes pérénnes etc.
2° nombre de têtes de bétails-bovins, ovins, porcins et
3° structures d'exploitation (8 modalités).

Le tableau de données, bien que numérique, n'est pas un tableau de mesures mais un tableau de contingence ou plutôt trois tableaux de contingence juxtaposés. L'indice de proximité entre deux départements qui est basé sur le coefficient de corrélation n'est pas approprié dans ce cas.

On a défini l'indice de proximité entre lignes (resp. colonnes) d'un tableau de contingence, respectant la métrique du χ^2, et conformément à la classe des indices de proximités de I. C. Lerman ce qui a rendu adéquate l'application de la suite de la chaine de programmes de classification hiérarchique supposant l'Algorithme de la Vraisemblance du Lien.

Il nous appartient dans un deuxième temps de comparer nos résultats avec ceux qu'on obtiendrait à partir de la chaine de programmes de classification hiérarchique basée sur la distance du χ^2 et le critère de l'inertie expliquée.

Ayant à notre disposition de telles données pour les différentes périodes, on s'est fixé l'objectif d'étudier l'Evolution de la structure agricole dans le temps et d'essayer de retrouver les facteurs de déformations.