REVUE DE STATISTIQUE APPLIQUÉE

F. MARCOTORCHINO
P. MICHAUD

Agrégation de similarités en classification automatique

Revue de statistique appliquée, tome 30, n° 2 (1982), p. 21-44
<http://www.numdam.org/item?id=RSA_1982__30_2_21_0>

© Société française de statistique, 1982, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM
Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

AGREGATION DE SIMILARITES EN CLASSIFICATION AUTOMATIQUE

F. MARCOTORCHINO et P. MICHAUD

Centre Scientifique, IBM, France

1. INTRODUCTION

Cet article a pour but de présenter, par une approche tout à fait différente de celles habituellement utilisées par les spécialistes de l'Analyse des Données, une méthodologie générale, utilisable pratiquement pour résoudre de nombreux problèmes d'analyse des données (en particulier de classification automatique).

L'originalité de cette approche repose essentiellement sur l'utilisation de modèles linéaires généraux unifiant des problèmes a priori aussi différents que la recherche d'un "ordre" sur une population dans le cadre de l'agrégation des préférences ou que la recherche d'une "partition" de cette même population dans le contexte de la classification automatique.

La méthode proposée est une méthode d'**Agrégation de Données** consistant à chercher une "relation collective" sur n objets, soit un ordre total (dans le cas de l'agrégation des préférences), soit une partition (dans les cas de l'agrégation de similarités ou agrégation classificatoire), soit encore d'autres relations plus complexes, "approximant au mieux" un ensemble de relations individuelles de départ représentant les données de m "critères" ou "juges" relatives aux n objets à analyser.

Cette approche a été présentée la première fois dans [12].

A titre d'exemple donnons ici quelques points qui différencient cette approche des méthodes existantes dans le domaine sur lequel nous avons choisi d'insister ici: c'est-à-dire la Classification Automatique.

Rappelons qu'en classification automatique, les méthodes existantes se divisent approximativement en deux grands types:

– les méthodes de classification hiérarchique ascendante;
– les méthodes de classification non hiérarchique.

Pour les premières, le résultat est en général un "arbre de classification" (représentant une suite de partitions emboitées), obtenu pas à pas sur des critères de "distances" minimales entre classes par agglomération successives des classes [5].

Pour les secondes, le résultat est une "partition" de la population obtenue (après fixation a priori du nombre de classes) par diminution d'un critère lié à des opérations de "recentrage – réaffectation" [8].

Mots-Clés : **Relations binaires, Classification automatique, Programmation linéaire**

L'approche proposée ici diffère des deux précédentes tant du point de vue du résultat obtenu que du critère retenu :

- le résultat est certes une "partition", mais sans fixation a priori du nombre de classes, et avec possibilité de mesurer les liaisons entre les classes obtenues.
- le critère repose sur l'utilisation d'une règle majoritaire ancienne puisqu'il s'agit de la fameuse "règle de Condorcet".

II. HISTORIQUE ET POSITION DU PROBLEME

Donnons tout d'abord la définition des termes de base de ce texte : "Agrégation des préférences" et "Agrégation des similarités". On appellera par la suite dans le texte :

Problème d'Agrégation des Préférences par un ordre) : un problème où les données seront la représentation de préférences sur les objets (ordres, préordres, comparaisons préférentielles par paires, etc. . .) et où le résultat cherché, sera un Ordre Total.

De même on appellera :

Problème d'Agrégation de Similarités (par une partition) : un problème où les données seront la représentation de similarités sur des objets (partitions, similarités par paires etc. . .) et où surtout le résultat cherché sera une Partition.

L'approche que nous allons proposer, dont le but est la résolution des deux problèmes précédents, bien que récente de par les techniques employées est fondée sur des concepts anciens introduits par A. de CONDORCET dès 1785 (recherche d'une relation d'ordre collectif à partir de comparaisons par paires). Ce dernier avait déjà mis en évidence à l'époque l'aspect paradoxal de la recherche d'un ordre collectif à partir d'opinions individuelles (Effet Condorcet).

Cet aspect paradoxal se retrouve également dans le cas de l'agrégation de similarités, où le problème consiste alors à trouver une "relation d'équivalence collective" (partition des n objets) la plus "proche possible" de m partitions ou relations de similarités individuelles données au départ par m "juges" ou "critères" sur les n objets à analyser. Le paradoxe ne s'appelle plus dans ce cas "Effet Condorcet", mais "Effet Poincaré".

Comme nous venons de le dire les méthodes que nous allons présenter permettent de résoudre aussi bien les problèmes d'agrégation des préférences que d'agrégation des similarités. Or si, comme nous le montrerons plus loin les modèles associés à ces deux problèmes de nature a priori distincte, ne diffèrent que très légèrement, il semble curieux de constater qu'historiquement ces problèmes ont été étudiés de façon tout à fait séparée et sans aucune connexion apparente.

En effet, alors que l'on peut faire remonter les premières idées de base sur "l'agrégation des préférences" à A. de CONDORCET (1785), le problème de la recherche de "partitions dites centrales" (cas particulier de l'agrégation de similarités) n'a été posé pour la première fois par S. Regnier qu'en 1965 [17].

Quant au développement des connaissances sur ces deux problèmes, tant au niveau théorique et axiomatique qu'au niveau des algorithmes de résolution, l'état d'avancement était loin d'être identique jusqu'à aujourd'hui.

A l'étude du problème de "l'agrégation des préférences" outre Condorcet déjà cité, on peut associer les noms de M.G. KENDALL [11] et G. KEMENY [10] qui ont étudié au niveau théorique le problème sous forme "métrique" dans le cas le plus simple de l'agrégation d'ordres totaux sans pour cela proposer de méthodes de résolution. K. ARROW [1] et D. BLACK en ont poursuivi l'étude axiomatique. M. BARBUT, dès 1966 [2], avait également souligné l'aspect métrique du problème d'agrégation.

En ce qui concerne "l'agrégation des similarités" dont la recherche d'une "partition centrale" n'est qu'un cas particulier (consistant à chercher la partition agrégeant des données elles-mêmes sous forme de partitions), à S. REGNIER déjà cité, on peut associer également le nom de B. MIRKIN [16]. Sous sa forme initiale le problème de la recherche d'une partition centrale n'avait pas été rattaché à l'approche générale.

Ce n'est que récemment que J.P. BARTHELEMY et B. MONJARDET [3] [4] ont attiré l'attention sur le fait que ces deux problèmes avaient une même représentation métrique, sans toutefois proposer de méthodes de résolutions.

En ce qui concerne les méthodes de résolution, la distinction à faire entre "agrégation des préférences" et "agrégation des similarités" était encore plus nette.

Les méthodes de résolution exactes associées à ce type "d'agrégation des préférences" étaient toutes fondées, jusqu'à présent, sur le principe d'énumération implicite "branch and bound", (nous avons donné en [13] une liste de certains de ces algorithmes); mais en pratique il était impossible de les utiliser pour un nombre d'objets à classer supérieur à 20.

Seul J. de CANI [6] avait vu l'intérêt de l'utilisation de la programmation linéaire pour résoudre un problème voisin de celui dont nous traitons ici; après quelques essais infructueux limités à l'agrégation d'ordres totaux de taille très faible $n = 4$, il a de lui même renoncé à cette approche et a proposé à son tour ultérieurement [7] un algorithme d'énumération implicite.

En ce qui concerne les méthodes de résolution associées à "l'agrégation des similarités", exceptée l'heuristique de S. REGNIER [17] pour les "partitions centrales" aucun algorithme a fortiori exact n'avait été proposé jusqu'à ce jour (tout au moins à notre connaissance).

Aucun des auteurs précités n'avait, tout au moins à notre connaissance, fait le lien entre l'agrégation des préférences et l'agrégation des similarités au niveau des méthodes de résolution, et proposé un modèle général.

A ce niveau notre contribution a permis de synthétiser et d'améliorer cette situation en proposant l'approche méthodologique suivante:

i) Sur le plan théorique agrégation de relations arbitraires par une relation particulière.

ii) Construction d'un modèle linéaire général de l'agrégation de relations de tous types.

iii) Création de méthodes de résolution exactes et approchées adaptées aux deux types de problèmes d'agrégation considérés ici. Adjonction de procédures de validation et de visualisation associées à la présentation des résultats.

En effet nous avons modélisé pour la première fois [12] ces divers problèmes de recherche de relations centrales sous une forme linéaire générale permettant ainsi de développer des méthodes exactes de résolution en des temps de calcul raisonnables, même pour des tailles de n relativement importantes ($n \leqslant 80$).

Enfin, nous avons montré que l'ensemble de ces problèmes pouvait être considéré comme l'utilisation de la règle de la majorité de Condorcet (fonction critère linéaire), sous des contraintes linéaires dépendant de la nature de la relation cherchée [12], [15].

En résumé, le problème général que nous allons traiter consiste à rechercher une relation collective (ordre total ou partition entre autre) "approximant au mieux" un ensemble de relations individuelles de départ.

La modélisation et la résolution du problème repose sur les 6 concepts fondamentaux suivants:

1) Présentation des données sous forme relationnelle. (On montrera que bon nombre de données usuelles se présentent sous forme de relations).
2) Utilisation d'un critère linéaire d'agrégation qui n'est autre que la règle majoritaire de Condorcet, dont on montrera qu'elle est équivalente à une approche métrique;
3) Linéarisation des contraintes portant sur la relation cherchée (la relation cherchée que nous appelerons Y par la suite vérifiera un certain nombre de contraintes linéaires);
4) Utilisation de la Programmation Linéaire pour la résolution des problèmes ainsi posés (en particulier utilisation du système MPSX/370);
5) Utilisation d'indices de validation du problème, tant au niveau des données que des résultats; ceci afin de répondre aux questions fondamentales suivantes:
   - est-il raisonnable de vouloir classer ou classifier les données de départ ?
   - le résultat obtenu est-il significatif ?
6) Visualisation des résultats en tenant compte des niveaux de validation.

III. DEFINITION DU PROBLEME GENERAL DE L'AGREGATION RELATIONNELLE

3.1. Présentation générale des données

Afin de définir la méthode proposée, nous supposons ici que les données se présentent déjà sous forme relationnelle. Nous verrons dans un paragraphe ultérieur comment la prise en compte des données usuelles les plus variées peut se faire sous cette forme.

On a donc au départ m relations individuelles correspondant soit aux opinions de m juges soit aux caractéristiques des n objets par rapport à m "critères".

a) Tableaux de données individuels

Chaque relation initiale, correspondant aux données individuelles du "juge" ou "critère" numéro k sera représentée par un tableau de "comparaisons par paires"

noté C^k. Ce tableau de comparaisons est un tableau carré (nXn) à valeurs (0 ou 1) défini comme suit :

$c_{ij}^k = 1$ Si i "est en relation avec" j pour le "juge" k

$c_{ij}^k = 0$ Sinon.

Par convention les valeurs c_{ii}^k du tableau ne seront pas considérées.

Compte tenu du problème à traiter et de la nature des données, la signification des valeurs c sera différente. Par exemple :

- i "est en relation avec" j signifie : i "est préféré à" j si C^k est un tableau de données de préférences.
- i "est en relation avec" j signifie : i "est semblable à" j si C^k est un tableau de similarités.

Chaque tableau C^k peut donc représenter n'importe quelle relation binaire.

A chaque relation binaire représentée par un tableau C^k, on peut associer un autre tableau de comparaisons, noté C'^k, ce nouveau tableau est un tableau carré (nXn) à valeurs $(-1, +1)$, défini par :

$c_{ij}'^k = 1$ Si i "est en relation avec" j pour le "juge" ou "critère" k

$c_{ij}'^k = -1$ Sinon.

Les valeurs $c_{ij}'^k$ peuvent etre exprimées au moyen des c_{ij}^k par :

$$c_{ij}'^k = 2\, c_{ij}^k - 1$$

Remarque : Dans le cas du codage C' on a la possibilité d'introduire la valeur $c_{ij}'^k = 0$ qui signifie une "non-réponse" du juge k. (Voir la signification de ce codage dans la thèse d'Etat de P. MICHAUD [14]).

Comme nous le verrons au § VI les tableaux C'^k et C^k peuvent être obtenus directement ou indirectement par l'intermédiaire de tableaux annexes (tableaux de rangs de notes, de distances etc. . .).

b) Tableaux collectifs

En sommant tous les tableaux de comparaisons individuels C^k et C'^k, on obtient des tableaux de données collectifs notés :

$$C = \sum_k C^k \quad \text{et} \quad C' = \sum_k C'^k$$

$$C' = 2\,C - mU$$

où U est la matrice (n x n) dont tous les éléments sont égaux à 1. Le terme (i , j) du tableau C est donné par :

$$c_{ij} = \sum_k c_{ij}^k$$

Le terme (i , j) du tableau C′ est donné par :

$$c'_{ij} = \sum_k c'^k_{ij}$$

c'_{ij} = Nombre de "juges" ou "critères" pour lesquels i "est en relation avec" j moins Nombre de "juges" ou "critères" pour lesquels i "n'est pas en relation avec" j.

De ce fait $c'_{ij} \geqslant 0$ signifie qu'une majorité de "juges" ou "critères" considèrent que i "est en relation avec" j.

Remarque : On peut généraliser l'approche précédente au cas où les critères ou juges C^k sont pondérés par un coefficient $p_k \geqslant 0$. On a alors :

$$C = \sum_k p_k C^k \quad \text{et} \quad C' = \sum_k p_k C'^k .$$

La relation liant C′ à C est alors donnée par :

$$C' = 2C - \left(\sum_k p_k\right) U$$

3.2. Définition de la relation collective cherchée Y.

Pour les deux problèmes (Agrégation des Préférences et de Similarités) nous recherchons une relation collective transitive (soit un ordre total, soit une relation d'équivalence "partition") approximant de la meilleure façon possible les m relations binaires individuelles de départ. Cette procédure de recherche d'une relation dite "centrale", conduira à l'obtention d'une relation Y.

Cette relation collective cherchée sera l'inconnue de notre modèle et sera représentée elle aussi par un tableau binaire carré (n x n) défini par :

$y_{ij} = 1$ Si i est "en relation avec" j dans le résultat final ;

$y_{ij} = 0$ Sinon.

Comme pour les données le terme "est en relation avec" dépend du contexte de l'agrégation que l'on désire effectuer et de la nature du résultat attendu.

Exemples :

$y_{ij} = 1$ signifie : i est classé avant j si Y représente un ordre total ;

$y_{ij} = 1$ signifie : i est dans la même classe que j si Y est une partition.

En règle générale Y peut représenter n'importe quel autre type de relation binaire : quasi-ordre, semi-ordre, préordre total, etc. . . [14] et [15].

Généralisation à d'autres types de solutions Y

Il est bien évident que l'Agrégation Relationnelle Générale permet d'autres combinaisons entre type de résultat cherché et type de données à agréger que celles rencontrées en "agrégation des préférences" et en "agrégation des similarités".

Exemple : agréger un ensemble de préordres totaux par une partition.

3.3. Définition d'un "critère" d'agrégation

Nous avons maintenant à définir ce que nous entendons par:

"approximer de la meilleure façon possible".

Pour cela considérons la fonction de Y suivante:

$$F(Y) = \sum_i \sum_{j \neq i} c'_{ij} y_{ij}$$

où c'_{ij}, comme nous l'avons déjà vu, représente le nombre de "juges" ou "critères" mettant i "en relation avec" j moins le nombre de "juges" ou "critères" ne mettant pas i "en relation avec" j.

La fonction F(Y) est alors évidemment maximum pour Y^c définie par:

$$y^c_{ij} = 1 \quad \text{Si} \quad c'_{ij} > 0$$
$$y^c_{ij} = 0 \quad \text{Si} \quad c'_{ij} < 0$$
$$y^c_{ij} = 0 \text{ ou } 1 \quad \text{Si} \quad c'_{ij} = 0$$

Comme nous l'avons déjà vu $c'_{ij} > 0$ implique qu'une majorité de "juges" a mis i "en relation avec" j ; en conséquence Y^c n'est rien d'autre que le résultat de l'application directe d'une règle majoritaire au tableau collectif de comparaisons C'; c'est justement cette règle qu'avait proposée Condorcet dès 1785 pour l'agrégation d'ordres totaux.

Cependant ce critère peut également s'interpréter de façon "métrique".

3.4. Interprétation métrique de la fonction F(Y)

La règle de Cordorcet, peut s'interpréter également comme une fonction d'éloignement, à partir du moment où l'on définit la distance $d(C^k, Y)$ entre la relation cherchée Y et un tableau relationnel et individuel de données C^k par:

$$d(C^k, Y) = \sum_{i,j} |c^k_{ij} - y_{ij}|$$

On a montré en [12] et [15] l'équivalence entre la minimisation de la fonction de Y:

$$F'(Y) = \sum_k d(Y, C^k)$$

et la règle de la majorité de Condorcet.

Minimiser en Y cette fonction F'(Y) reviendra à chercher la relation Y la plus "proche possible" de l'ensemble des m relations binaires de départ.

Dès lors, il apparaît la relation évidente entre les fonctions de Y, F(Y) et F'(Y):

$$F'(Y) = \text{Cte} - F(Y)$$

Et l'on obtient les deux résultats fondamentaux suivants :

i) Les fonctions objectifs $F(Y)$ et $F'(Y)$ sont linéaires en y_{ij} ;

ii) Les problèmes :

$$\min_{Y} F'(Y) \quad \text{(présentation métrique)}$$

et

$$\max_{Y} F(Y) \quad \text{(présentation au moyen de la règle de Condorcet)}$$

sont deux problèmes équivalents.

Comme nous l'avons déjà vu ces deux problèmes Min $F'(Y)$ et max $F(Y)$ sont trivialement résolus en appliquant la règle majoritaire de Condorcet.

Remarque : *Il faut noter qu'ici Y est solution d'une fonction économique sans contraintes et peut donc représenter n'importe quelle relation binaire. Bien entendu sauf si le tableau C' a une structure particulière, la solution Y ne sera pas en général transitive. Or nous cherchons une relation qui devra être soit une relation d'ordre total soit une relation d'équivalence, c'est-à-dire transitive ; la seule utilisation de la règle majoritaire de Condorcet ne nous garantit donc pas contre une solution dénuée de fondement (vis-à-vis du problème posé).*

Ce fait est du en particulier à l'existence de "paradoxes" inhérents à ce type de situations.

L'un de ces paradoxes est le fameux "paradoxe de Condorcet". Le principe de ces paradoxes repose sur le fait suivant : l'agrégation de données individuelles "transitives" peut donner un résultat Y "intransitif" si on applique la règle de la majorité.

En supposant que des "juges" ou "critères" aient donné des opinions individuelles transitives (ordres totaux ou partitions) sur 3 objets x, y et z, alors il y aura (vis-à-vis de la règle de Condorcet) :

Effet Condorcet

Si, une majorité de "juges" ayant préféré collectivement x à y : $(x \geqslant y)$; une majorité de "juges" ayant également préféré collectivement y à z $(y \geqslant z)$, on a une majorité de "juges" qui préfère collectivement z à x $(z \geqslant x)$, au lieu de la transitivité attendue $(x \geqslant z)$, x préféré à z par une majorité de juges).

De même il y aura :

Effet Poincaré

Si, sachant que pour majorité de "critères" x est considéré comme similaire à y $(x \simeq y)$ y comme similaire à z $(y \simeq z)$, mais que x n'est pas considéré comme similaire à z par une majorité de "critères" $(x \not\simeq z)$; au lieu de la transitivité attendue $(x \simeq z)$ x considéré comme similaire à z par une majorité de "critères".

Comme l'application directe (fort simple) de la règle de Condorcet ne nous permet pas d'obtenir un résultat transitif à coup sûr, la relation Y cherchée doit donc vérifier un certain nombre de contraintes pour nous garantir l'obtention soit d'un ordre total, soit d'une "partition", relations auxquelles nous nous sommes volontairement limités dans cet article. La vérification de ces contraintes sur Y va

rendre les problèmes précédents "combinatoires", c'est-à-dire aboutir à des calculs inextricables si l'on ne possède pas d'algorithmes performants.

- En effet, dans le cas de l'agrégation des préférences il faudrait choisir parmi les n! ordres totaux Y celui ou ceux qui maximisent F(Y).
- Dans le cas de l'agrégation des similarités c'est parmi les B(n) partitions de n objets (nombre de Bell) qu'il faudrait choisir la ou les partitions Y maximisant F(Y).

A titre d'illustration disons que $(17)! = 3{,}5 \times 10^{14}$ et $(70)! > 10^{100}$ et que $B(71) = 4{,}7 \times 10^{14}$.

Mais là encore, nous allons le voir au § suivant, ces contraintes s'exprimeront linéairement en fonction des valeurs y_{ij}, et un algorithme approprié nous permettra d'éviter l'énumération de toutes les solutions Y possibles (procédé impraticable pour $n > 15$).

Cependant bien qu'exceptionnels, il existe des cas où de par la structure particulière des données, la simple application de la règle de la majorité par paires de Condorcet donnera soit un ordre total, soit une relation d'équivalence ; c'est-à-dire la solution optimale du problème posé. (Voir dans [13] la description de certaines de ces configurations exceptionnelles).

IV. LES MODELES LINEAIRES ASSOCIES

4.1. Les contraintes linéaires sur la relation cherchée Y

- Si la relation cherchée Y est un ordre total, Y doit être (Transitive, Antisymétrique, Totale) ;
- Si la relation cherchée Y est une partition (relation d'équivalence), Y doit être : (Transitive, Symétrique).

Nous avons alors le résultat suivant :

Toutes ces propriétés peuvent être formulées comme des expressions linéaires des valeurs y_{ij} (voir [12] et [15] pour plus de détails).

1) Transitivité : Elle s'exprime par les contraintes suivantes :

$$y_{ij} + y_{jk} - y_{ik} \leqslant 1 \quad \forall \quad (i, j, k) \text{ tous différents.}$$

Ces inégalités expriment que si iRj et jRk alors iRk pour la relation Y.

2) Symétrie : Elle s'exprime par les contraintes suivantes :

$$y_{ij} - y_{ji} = 0 \quad \forall \quad (i \neq j)$$

Cette contrainte signifie simplement que pour un problème de classification automatique si i est dans la même classe que j, alors j est dans la même classe que i.

3) Antisymétrie : Elle s'exprime par les contraintes suivantes :

$$y_{ij} + y_{ji} \leqslant 1 \quad \forall \quad (i \neq j)$$

Ces inégalités signifient qu'au plus iRj ou jRi est vraie pour la relation Y.

4) Totalité : Elle s'exprime par les contraintes suivantes :

$$y_{ij} + y_{ji} \geqslant 1 \quad \forall \quad (i \neq j)$$

Ces inégalités signifiant qu'au moins iRj ou jRi est vraie pour la relation Y.

Remarque : Si Y est un ordre total (antisymétrique et total), Y doit vérifier simultanément les contraintes 3) et 4) ce qui conduit aux nouvelles relations linéaires en y_{ij} :

5) $y_{ij} + y_{ji} = 1 \quad \forall \quad (i \neq j)$

Ces relations signifient que soit i est classé avant j, soit j est classé avant i dans la relation cherchée Y.

De ce fait nous pouvons donner une interprétation générale des problèmes d'agrégation des préférences ou d'agrégation de similarités :

Ces deux problèmes peuvent être considérés comme l'application de la règle majoritaire sous contraintes :

– contraintes 1) et 5) pour la recherche d'un ordre total ;
– contraintes 1) et 2) pour la recherche d'une partition Y.

4.2. Les modèles bivalents

Il découle des résultats précédents que les problèmes d'agrégation des préférences et d'agrégation des similarités peuvent être formulés comme des problèmes de programmation linéaire en nombres entiers en variables bivalents (0-1).

Les variables (0-1) sont les valeurs y_{ij} de la relation collective cherchée Y.

La fonction objectif à maximiser est la fonction F(Y) que nous avons déjà définie comme la formulation de la règle majoritaire de Condorcet ; les contraintes sont les différentes propriétés que la relation Y doit vérifier.

Selon le type de relation Y cherchée nous obtenons donc les modèles linéaires suivants :

a) Modèle pour une relation d'ordre total

$$\text{Max} \sum_i \sum_{j \neq i} c'_{ij} y_{ij}$$

$$y_{ij} + y_{ji} = 1 \quad i \neq j$$

$$y_{ij} + y_{jk} - y_{ik} \leqslant 1 \quad i \neq j,\ j \neq k,\ i \neq k$$

$$y_{ij} = 0 \text{ ou } 1$$

b) Modèle pour une relation d'équivalence

$$\text{Max} \sum_i \sum_{j \neq i} c'_{ij} y_{ij}$$

$$y_{ij} - y_{ji} = 0 \quad i \neq j$$

$$y_{ij} + y_{jk} - y_{ik} \leqslant 1 \quad i \neq j,\ j \neq k,\ i \neq k$$

$$y_{ij} = 0 \text{ ou } 1$$

Il est tout à fait remarquable que des problèmes primitivement très différents quant aux résultats cherchés: Recherche d'une agrégation des préférences (classement) et recherche d'une agrégation de similarités (classification), ne diffèrent du point de vue de la modélisation que par les équations:

$$5)\ y_{ij} + y_{ji} = 1 \quad \text{et} \quad 2)\ y_{ij} - y_{ji} = 0$$

On trouvera dans [13] et plus spécialement dans [15] un certain nombre de modèles linéaires associés à des relations Y plus complexes.

Le nombre de contraintes de ces problèmes est de l'ordre $O(n^3)$, tandis que le nombre de variables est de l'ordre $O(n^2)$.

D'autre part, on trouvera dans [13] et [15] la signification de quelques variantes de l'expression de la fonction économique F(Y), qui lui sont équivalentes (c'est-à-dire donne des résultats collectifs identiques).

V. LA METHODE DE RESOLUTION DU MODELE PROPOSE

5.1. Quelques méthodes existantes de résolution exacte

a) Cas de l'agrégation des préférences

Dans le cas de l'agrégation d'ordres il existe un certain nombre de méthodes exactes; elles sont toutes fondées sur le principe d'énumération implicite des n! permutations (Branch and Bound). La plupart de ces méthodes sont détaillées ou citées dans [16]. Malheureusement on peut considérer que ces méthodes ne sont applicalbes que pour des tailles de n (nombre d'objets) inférieures à 20.

b) Cas de l'agrégation de similarités

Pour ce problème il n'existait pas (à notre connaissance) de méthodes exactes de résolution jusqu'à aujourd'hui.

5.2. La méthode exacte proposée

Les deux modèles a) et b) du § V sont des modèles de programmation bivalente, c'est-à-dire en variable (0-1) et de ce fait sont a priori des problèmes excessivement compliqués vus leur tailles. Cependant nous avons pu obtenir des solutions optimales en des temps de calcul raisonnables, même pour des tailles de n relativement importantes.

Le principe de l'algorithme de résolution exacte que nous avons proposé repose sur l'utilisation de la programmation linéaire continue avec adjonction de coupes (cutting planes). Cet algorithme consiste alors à résoudre une succession

de programmes linéaires continus associés aux modèles a) ou b), c'est-à-dire où l'on a négligé au préalable les contraintes d'intégrité $y_{ij} = 0$ ou 1, transformées en $0 \leqslant y_{ij} \leqslant 1$ et où l'on a ajouté à chaque étape les contraintes définissant les coupes successives. C'est un résultat bien connu de programmation linéaire qu'une telle méthode est une méthode finie – chaque PL nécessite un nombre fini d'itérations et le nombre de programmes linéaires est fini – Pourtant, à ce niveau deux questions se posent :

i) Combien de programmes linéaires faudra-t-il résoudre pour aboutir à une solution optimale bivalente ?

ii) La résolution d'un seul de ces programmes linéaires ainsi générés est-elle possible, compte tenu des tailles mises en œuvre ?

Contrairement à l'habitude pour ce type de problème, ce n'est pas le nombre de programmes linéaires successifs à résoudre qui sera la principale difficulté (puisqu'en fait le plus souvent on n'aura à traiter qu'un seul programme linéaire), mais bien plutôt la résolution d'un seul de ces programmes linéaires continus.

En effet la résolution des modèles a) et b) rendus continus en transformant la contrainte $y_{ij} = 0$ ou 1 par $0 \leqslant y_{ij} \leqslant 1$ est a priori impossible si on essaye de traiter directement ces programmes, car le programme sera automatiquement mis sous forme standard par adjonction de variables d'écart. Le nombre des contraintes étant alors de l'ordre de n^3, oblige a générer des matrices de bases de tailles $O(n^3) \times O(n^3)$, ce qui bloque rapidement la place mémoire disponible, même pour n faible et rend impossible les calculs effectifs.

A ce niveau l'intérêt et l'originalité de la méthode que nous proposons est de pouvoir traiter le programme linéaire continu précédent avec des bases de tailles $O(n^2) \times O(n^2)$ seulement. Dans [12] nous avons détaillé le processus qui nous a conduit à ce résultat. C'est la procédure "duale" du dual qui a été utilisée, avec le code produit-programme IBM MPSX/370 dont les avantages sont également détaillés dans [15].

Ceci nous a permis de résoudre de façon exacte des problèmes de tailles relativement grandes (jusqu'à $n = 80$ pour l'instant).

Cependant, et ceci est un point fondamental, pour tous les exemples pratiques que nous avons traités (plus de 200 à l'heure actuelle, tant d'agrégation des préférences que d'agrégation des similarités), les solutions optimales des premiers programmes continus (c'est-à-dire sans adjonction de coupes) étaient toujours à valeurs (0-1), c'est-à-dire toujours solutions des modèles linéaires bivalents a) ou b). Le fait de n'avoir à résoudre la plupart du temps qu'un seul programme linéaire continu est un résultat empirique tout à fait remarquable, qui justifie amplement à lui tout seul l'approche "optimisation" que nous avons choisie.

En effet le nombre de programmes linéaires continus successifs que nous aurons à résoudre étant extrêmement faible, sauf cas exceptionnels, l'algorithme exact que nous proposons permettra des temps de calculs très raisonnables.

5.3. Temps de calcul

Nous donnons ici quelques temps de résolution (avec MPSX/370 sur 370/168 IBM incluant : Génération-Résolution-Edition).

Type du PB	Taille	Nature	Temps (Méth. Exact.)	Temps (heuri.)
Archéologie	21	P	7 s	0.1 s
Sociologie	20	P	13 s	0.1 s
Medecine	23	P	15 s	0.2 s
Linguistique	25	C	7 s	0.2 s
Linguistique	28	C	11 s	0.2 s
Medecine	29	P	22 s	0.2 s
Linguistique	29	C	12 s	0.2 s
Esthetique	36	C	22 s	0.2 s
Psychologie	35	P	36 s	0.25 s
Sociologie	40	C	31 s	0.2 s
Test Compar.	40	P	295 s	0.35 s
Linguistique	45	C	60 s	0.25 s
Linguistique	45	C	58 s	0.25 s
Esthétique	48	C	65 s	0.30 s
Esthétique	54	C	117 s	0.30 s
Politique	54	P	220 s	0.90 s
Sociologie	72	C	157 s	0.90 s
Linguistique	77	C	331 s	1.20 s

(Dans la colonne NATURE "P" signifie que le résultat obtenu est une "partition", "C" que le résultat obtenu est un "classement").

Temps Heuristiques: Ils proviennent de l'utilisation de deux heuristiques différentes. La première relative à l'Agrégation des Préférences est donnée dans le livre [12], la deuxième relative à l'Agrégation des Similarités (amélioration de celle de S. Régnier) est détaillée dans la thèse d'Etat de F. MARCOTORCHINO [13].

Remarque: *Les méthodes d'Agrégation des données présentées dans cet article vérifient un certain nombre de propriétés de structure que nous ne faisons que citer ici et qui sont détaillées dans la thèse [13].*

1. Propriété de consistance additive

2. Propriété de neutralité totale

3. Propriété de non fixation à priori du nombre de classes (pour les problèmes d'agrégation de similarités seulement)

VI. PRISE EN COMPTE DES DONNEES

La présentation relationnelle, sous forme de tableaux de comparaisons par paires C permet une grande souplesse dans la prise en compte des données sans arbitraire dans le codage et surtout sans perte d'information.

Dans la mesure où nous avons présenté dans d'autres articles (voir [12] et [15] par exemple) comment transformer en données relationnelles des données de préférences, nous allons insister presque exclusivement dans cet article sur les données de similarités.

En fait les données sur lesquelles nous allons travailler se divisent en deux types distincts:

– les données sont déjà directement sous forme de tableaux de comparaisons par paires;

– elles ne sont pas initialement sous forme de tableaux de comparaisons par paires et il s'agit alors de les mettre sous forme de tableaux C^k (binaires) ou du tableau C.

6.1. Données directement sous forme de tableaux de comparaisons

a) Cas de l'agrégation des préférences

Se présentent directement sous forme de tableaux de comparaisons par paires les types de données suivants:

1 - les tableaux de tournois (football, escrime, rugby, . . .);

2 - les graphes de "dominations" ou de hiérarchies dans les populations animales ou humaines (voir exemple sur les chimpanzés dans [12];

3 - les graphes de relations hiérarchiques dans les entreprises (la distribution des 1 et des 0 dans de tels tableaux peut être quelconque);

4 - les graphes représentation de "systèmes";

5 - les données de préférences par paires (enquêtes psychosociologiques, enquêtes d'opinions, etc. . .).

b) Cas de l'agrégation de similarités

Se présentent directement sous forme de tableaux de comparaisons par paires.

1 - Les tableaux individuels de similarités

Chaque "juge" définit directement ses "similarités" sur une population de n objets en posant:

soit: $c_{ij}^k = 1$ Si pour lui i et j sont "similaires" ou "semblables";

$c_{ij}^k = 0$ Sinon.

ou si l'on admet la possibilité de non-réponses (voir [14]):

$c_{ij}^k = 1$ Si pour lui les objets i et j sont "similaires":

$c_{ij}^k = 1/2$ S'il ne répond pas;

$c_{ij}^k = 0$ Dans les autres cas.

(Voir dans [13] un exemple de tels tableaux de similarités relatifs à l'opinion de 51 personnes sur les ressemblances entre 12 machines à écrire).

Remarque: En général ces tableaux sont symétriques mais rarement transitifs, c'est-à-dire ne représentent pas forcément une partition. Ceci correspond à la prise en compte de l'EFFET POINCARE individuel: où un "juge" k trouve que X et Y sont similaires, Y et Z similaires mais pas forcément X et Z.

2 - Les tableaux carrés (n x n) de distances, d'éloignements, de similarités, ou de noircissement du type BERTIN-CZEKANOWSKY.

Ce sont en général des tableaux résultats de calculs préalables pour lesquels nous ne connaissons pas "a priori" le tableau de données brutes, croisant les n objets avec eux-mêmes. Le terme (i , j) de ce tableau c'est-à-dire s_{ij} représente :

- soit une mesure de distance ou d'éloignement entre i et j (non nécessairement symétrique);
- soit une mesure de similarité entre i et j ;
- soit une valeur de "grisé" caractérisation visuelle d'une liaison entre i et j.

i) Cas d'un tableau de distance ou d'éloignement S. On peut définir un tableau de similarité individuel C en posant :

$c_{ij} = 1$ Si $s_{ij} \leqslant s$ (s étant un seuil donné);

$c_{ij} = 0$ Sinon.

C peut très bien ne pas représenter une partition ou n'être pas symétrique si S ne l'est pas lui-même.

ii) Soit S un tableau de similarités (construit à partir d'un indice) tel que $0 \leqslant s_{ij} \leqslant 1$, alors on fabriquera directement le tableau C' en posant :

$$c'_{ij} = 2 s_{ij} - 1$$

et l'on maximisera la fonction économique suivante :

$$F(Y) = \sum_{i} \sum_{j=i} (2 s_{ij} - 1)\ y_{ij}$$

mais cette fonction n'est pas le résultat de l'application d'une règle de majorité collective et ne vérifiera pas sauf exception les propriétés déjà présentées auparavant [13].

iii) Dans le cas des tableaux de BERTIN-CZEKANOWSKY c'est le noircissement des cases qui fait office de tableaux de similarités, on posera par exemple pour un tableau C' :

$c'_{ij} = 1$ Si la case (i , j) est noire;

$c'_{ij} = 0$ Si elle est grise;

$c'_{ij} = -1$ Si elle est blanche.

Plus généralement s'il y a k niveaux de gris on construira directement le tableau collectif C en posant :

$c_{ij} = k$ Si la case (i , j) est totalement noire;

$c_{ij} = k - 1$ Si la case est gris foncé;

. . . . .

$c_{ij} = 0$ Si elle est complètement blanche.

L'idée sous-jacente ici étant que la valeur d'un grisé est additive et correspond à l'addition de plusieurs critères (virtuels) jouant sur du noir et blanc. Ainsi un gris de valeur (k – 2) par exemple pour la case (i , j) correspond au fait que k – 2 "critères" ont jugé (i , j) noirs (c'est-à-dire ressemblants) alors que deux "critères" (virtuels) les ont trouvés blancs c'est-à-dire "non similaires".

6.2. Données non directement sous forme "comparaisons par paires"

Dans ce cas il s'agit de transformer les données de façon à obtenir des tableaux C^k (représentant un "juge" ou un "critère") sous forme de comparaisons par paires afin de pouvoir utiliser les algorithmes précédemment décrits.

a) Cas de l'agrégation des préférences

Nous ne donnerons ici, à titre d'exemple que la façon dont on transforme en tableaux de comparaisons par paires des données initialement sous forme de "classements de n objets" (ordres totaux).

Supposons que m "juges" ou "critères" classent n objets. Alors en notant r_{ik} = rang de l'objet i pour le juge k, on obtient le tableau C^k associé de la façon suivante:

$$c_{ij}^k = 1 \quad \text{Si } r_{ik} \leqslant r_{jk};$$
$$c_{ij}^k = 0 \quad \text{Sinon.}$$

Dans ce cas particulier, si l'on permute les lignes et les colonnes du tableau C^k obtenu en fonction du classement du juge k on obtient des "1" au dessus de la diagonale de C^k et des "0" en dessous.

On peut d'ailleurs généraliser ce codage aux préordres, quasi-ordres etc. [15].

b) Cas de l'agrégation de similarités

1. Les données initiales sont des partitions

Le cas le plus simple correspond à la situation où un "juge" ou un "critère" représente une variable à plusieurs "modalités". Le tableau de données brutes associé est alors un tableau de modalités, où le terme général r_{ik} représente la valeur de la modalité attribuée à l'objet i par le juge ou critère k. Le tableau C_k est alors défini de la façon suivante:

$c_{ij}^k = 1$ Si $r_{ik} = r_{jk}$ c'est-à-dire si i et j ont même modalité pour le juge k;

$c_{ij}^k = 0$ Sinon.

C^k représente alors une relation d'équivalence (partition).

Un autre codage possible, plus souple, utilisant directement le tableau C'^k permet à ce niveau de prendre en compte des variables pour lesquelles il existe une hiérarchie sous-jacente dans les modalités:

$$c_{ij}'^k = 1 \quad \text{Si } |r_{ik} - r_{jk}| < s;$$
$$c_{ij}'^k = 0 \quad \text{Si } |r_{ik} - r_{jk}| = s;$$
$$c'^k = -1 \quad \text{Si } |r_{ik} - r_{jk}| > s.$$

s est un seuil fixé à l'avance en général faible (s = 1 ou 2). Le tableau C'^k dans ce cas ne représente pas une partition, puisqu'il contient par définition des configurations où l'effet Poincaré se produit.

Ce type de données (variables à modalités) est extrêmement fréquent en analyse des données lorsqu'il y a hétérogénéité des variables.

Cette configuration de problème, pathologique ou délicate pour les autres méthodes d'analyse des données, est le cas le plus adéquat et le plus simple dans le cadre de l'approche proposée ici.

2. Cas où les données initiales sont des données binaires de (présence - absence) ou de réponses (oui-non)

Dans le cas où le "juge" ou "critère" k représente une propriété ou une question, le tableau de données brutes associé est alors un tableau de (présence-absence) ou de réponses (oui-non) où :

$r_{ik} = 1$ Si l'objet i à la propriété k, ou a répondu "oui" à la question k

$r_{ik} = 0$ Sinon.

La notion de similarité peut se définir de plusieurs façons dépendant du contexte du problème.

i) 1er Cas

$c'^{k}_{ik} = 1$ Si $r_{ik} - r_{jk} = 0$ c'est-à-dire que i et j ont ou n'ont pas simultanément la propriété k

$c'^{k}_{ij} = -1$ Si $r_{ik} - r_{jk} \neq 0$

ii) 2e Cas

$c'^{k}_{ij} = 1$ Si $r_{ik} \cdot r_{jk} = 1$ c'est-à-dire que i et j ont tous les deux la propriété k

$c'^{k}_{ij} = -1$ Sinon c'est-à-dire si $r_{ik} \cdot r_{jk} = 0$

iii) 3e Cas

$c'^{k}_{ij} = 1$ Si $r_{ik} = 1$, i et j ont simultanément la propriété k,

$c'^{k}_{ij} = 0$ Si $r_{ik} = r_{jk} = 0$, i et j n'ont pas simultanément la propriété k,

$c'^{k}_{ij} = -1$ Si $r_{ik} - r_{jk} = 0$.

Cette possibilité de définir le tableau C'^{k} de plusieurs façons différentes correspond au poids ou à la signification que l'on doit accorder suivant le contexte aux configurations du type $(0 - 0)$. Problème pour lequel la multitude d'indices de similarités définis par les spécialistes de la classification montre le caractère ambigu.

Interprétation de ces différentes configurations

i) Le 1er cas correspond à la configuration où chaque "propriété k" induit une partition en 2 classes seulement (propriété SEXE, par exemple).

ii) Le 2e cas correspond à la configuration où pour deux individus i et j ne pas avoir une propriété k équivaut à une dissimilarité entre eux. La propriété k n'induit pas une partition en 2 classes.

iii) Le 3e cas correspond à la configuration où le fait que i et j n'aient pas la propiété k n'implique pas forcément que i et j se ressemblent mais n'implique pas plus leur non ressemblance, c'est-à-dire implique le doute (propriété AVOIR LES YEUX VERTS).

3. Le tableau de données brutes est un tableau quantitatif

Ce tableau correspond à des données représentant des "mesures" sur des variables continues mais pas forcément du même type. Dans ce cas, comme dans d'autres méthodes, une étude statistique descriptive préalable et nécessaire (fabrication d'histogrammes). Ces histogrammes associés à chaque "variable" ou "critère" k induisant alors des classes sur la population. On peut alors comme précédemment définir la similarité par les tableaux C ou C′ suivants:

$c_{ij}^k = 1$ Si i et j sont dans une même classe de l'histogramme associé à la variable k

$c_{ij}^k = 0$ Sinon.

Dans ce cas C^k représente une partition.

Autre codage possible moins discontinu:

$c_{ij}'^k = 1$ Si i et j sont dans la même classe ou dans des classes adjacentes de l'histogramme associé à k, c'est-à-dire si |Num classe de i – Num classe de j| $< s$

$c_{ij}'^k = 0$ Si |Num classe de i – Num classe de j| $= s$

$c_{ij}'^k = -1$ Si |Num classe de i – Num classe de j| $> s$

(s étant un seuil faible, fixé *a priori*). Dans ce cas C'^k n'est pas un tableau représentant une partition.

Exemple: Mesures sur des squelettes en zoologie (craniométrie par exemple).

Comme nous venons de le voir, l'utilisation de tableaux de données sous forme de comparaisons par paires permet donc de prendre en compte de données aussi différentes que les tournois sportifs ou les résultats de tests psychométriques (données 0 – 1).

Les tableaux C^k ou C'^k eux-mêmes peuvent tout aussi bien représenter des données aussi simples que des ordres totaux ou des partitions, que des données complexes ou paradoxales (Effet Condorcet ou Poincaré, impossible à prendre en compte autrement).

Outre cette souplesse d'utilisation, la prise en compte des données sous forme de tableaux C^k ou C'^k permet d'éviter des codages "*a priori*" de données souvent arbitraires et sans signification.

Cette affirmation peut être illustrée par la mise en évidence des relations existant entre les différents indices de similarité utilisés habituellement dans le cas de données binaires et les tableaux C^k, C'^k, C ou C′ (voir [13] T1 pp. 33-47).

Remarque: *Mesures de la qualité des données et de la validité des résultats obtenus*

Il faut distinguer ici deux notions tout-à-fait différentes:

– La notion de "Concordance" liée à la qualité des données,

– La notion de "Cohérence" liée à la validité du résultat Y optimal obtenu.

Pour ces deux notions nous avons défini des indices permettant de les mesurer dans [13] et dans [14].

VII. EXEMPLES D'APPLICATION

Nous ne présenterons ici qu'un seul exemple, mais on pourra trouver dans [12] des exemples d'agrégation des préférences et dans [13] des exemples d'agrégation des similarités.

L'exemple (réel) traitera de la classification des félidés suivant des caractéristiques morphologiques externes ou internes, et des caractéristiques de comportement telles que J. DORST les présente dans [9];

Présentation du problème

Il s'agit dans cet exemple de classifier 31 félins provenant des trois continents: Afrique, Eurasie, Amérique, à l'aide de 14 paramètres (ou caractéristiques) morphologiques ou de comportement.

Les 30 félins se répartissent comme suit:

Afrique	Eurasie	Amérique
Lion	Tigre	Jaguar
Léopard	Once	Puma
Guépard	Léopard	Ocelot
Serval	Pant. Nébul.	Jaguarundi
Caracal	Lynx	Ch. Marguay
Ch. Viverrin	Ch. de Chine	Ch. Tigrin
Ch. Chaus	Ch. du Bengale	Ch. des Andes
Ch. Doré	Ch. Rouilleux	Ch. Marbré
Ch. Marguerite	Ch. Malais	
Ch. Cafer	Ch. de Bornéo	
Ch. Nigripes	Ch. Manul	
	Ch. Temminck	

Ces différents félins sont de genre et d'allure différents (voir certains d'entre-eux sur la planche n° 1). Certains sont du genre panthera, d'autres neofelis, d'autres félis, un du genre acynonix; leurs poids varient de 5 à 150 kilos, etc. . .

Les différents paramètres choisis ont été extraits de la liste descriptive généralement utilisée par J. DORST pour décrire une espèce animale ou une variété dans son livre [9].

A partir de ces variables on obtient le tableau de données (30 x 14) reproduit ci-après.

Chacune des 14 variables déterminant une partition de l'ensemble des félins en autant de classes que la variable a de modalités. La recherche de la "partition centrale" consistera à chercher le Maximum de F(Y) sous des contraintes de transitivité et de symétrie définies au Chapitre IV.

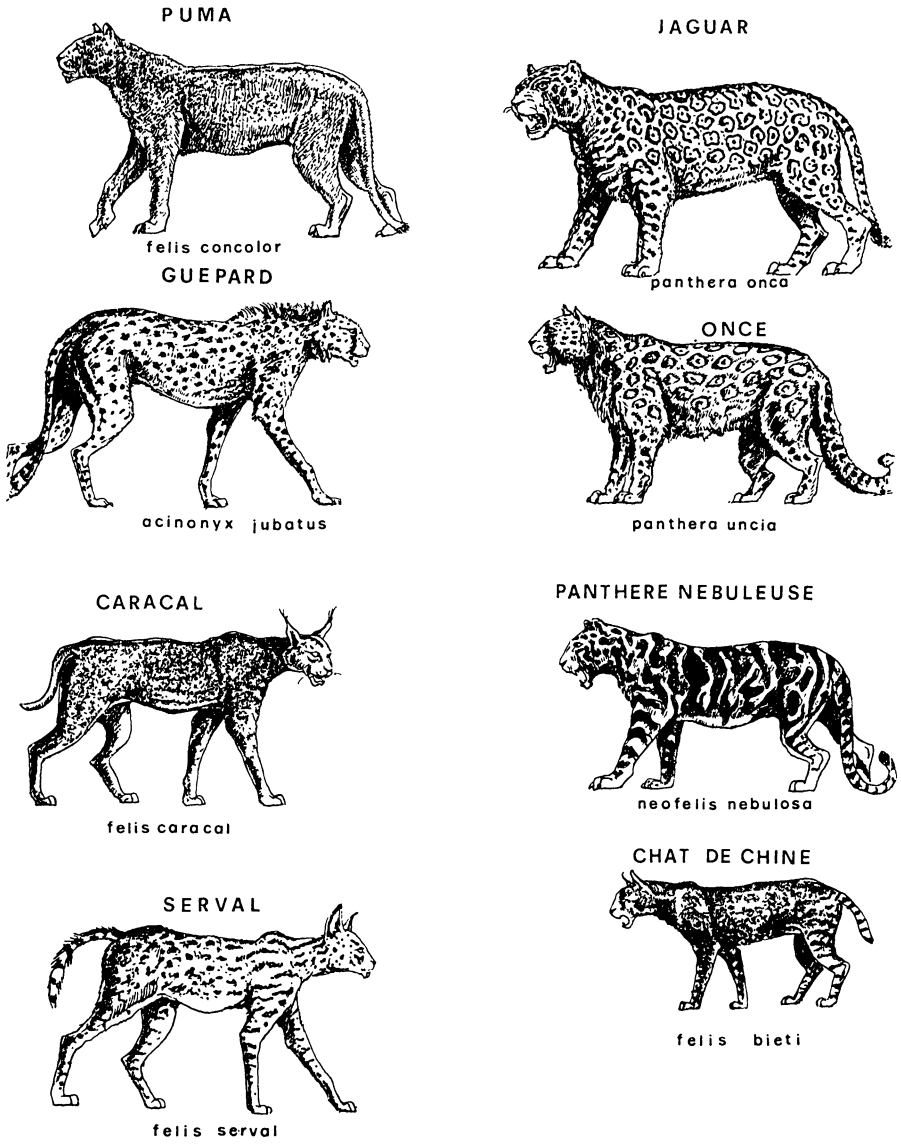

PLANCHE 1
jfm (1980).

	TYPPEL	LONGPOIL	RETRACT	COMPORT	ORELLES	LARYNX	TAILLE	POIDS	LONGUEUR	QUEUE	DENTS	TYPPROIE	ARBRES	CHASSE
LION	1	0	1	1	1	1	3	3	3	2	1	1	0	1
TIGRE	3	0	1	3	1	1	3	3	3	2	1	1	0	0
JAGUAR	2	0	1	2	1	1	3	3	2	1	1	1	1	0
LEOPARD	2	0	1	3	1	1	3	3	2	2	1	2	1	0
ONCE	2	1	1	1	1	1	2	2	2	3	1	2	1	0
GUEPARD	2	0	0	1	1	0	3	2	2	3	0	2	0	1
PUMA	1	0	1	2	1	0	2	3	2	3	1	2	1	0
NEBUL	4	0	1	3	1	1	2	2	2	3	1	3	1	0
SERVAL	2	0	1	1	2	0	2	2	2	1	0	3	1	1
OCELOT	2	0	1	2	1	0	2	2	2	2	0	3	1	0
LYNX	2	1	1	2	2	0	2	2	2	1	1	2	1	0
CARACAL	1	0	1	2	2	0	2	2	1	1	0	3	1	1
VIVERRIN	2	0	1	2	1	0	1	1	2	2	0	3	0	0
YAGUARUN	1	0	1	2	1	0	1	2	2	3	0	3	1	0
CHAUS	1	1	1	3	2	0	1	2	1	2	0	3	1	0
DORE	1	0	1	3	1	0	1	1	1	2	0	3	1	0
MERGUAY	2	0	1	3	1	0	1	1	1	2	0	3	1	0
MARGERIT	1	1	1	2	1	0	1	1	1	2	0	3	0	0
CAFER	3	0	1	3	1	0	1	1	1	2	0	3	1	1
CHINE	1	0	1	2	2	0	1	1	1	1	0	3	1	0
BENGALE	2	0	1	3	1	0	1	1	1	2	0	3	1	0
ROUILLEU	2	0	1	2	1	0	1	1	1	2	0	3	1	0
MALAIS	1	1	1	3	1	0	1	1	1	1	0	3	1	0
BORNEO	1	0	1	3	1	0	1	1	1	2	0	3	1	0
NIGRIPES	2	0	1	2	1	0	1	1	1	1	0	3	1	1
MANUL	1	1	1	3	1	0	1	1	1	1	0	3	1	0
MARBRE	4	0	1	3	1	0	1	1	1	3	0	3	1	0
TIGRIN	2	0	1	3	1	0	1	1	1	2	0	3	1	0
TEMMINCK	1	0	1	3	1	0	1	1	1	2	0	3	1	0
ANDES	2	1	1	3	1	0	1	1	2	2	0	2	1	0

Ces différents paramètres se répartissent comme suit :

Code	Paramètres	Nbr. modalités
	1) Paramètres morphologiques	
1	(TYPEL) Aspect du pelage	4
2	(LONG POIL) Fourrure	2
3	(OREILLES) Oreilles	2
4	(TAILLE) Taille au garrot	3
5	(POIDS) Poids	3
6	(LONGUEUR) Longueur du corps	3
7	(QUEUE) Longueur de la queue	3
8	(DENTS) Canines développées	2
9	(LARYNX) Os hyaoide	2
10	(RETRACT) Griffes rétractiles	2
	2) Paramètres de comportement	
11	(COMPORT) Comportement prédateur	3
12	(TYPPROIE) Type de la proie	3
13	(ARBRES) Monte ou non aux arbres	2
14	(CHASSE) Chasse (courre ou affut)	2

b) Résultats

On a alors le résultat suivant :

$$\text{Max } F(Y) = F(Y^*) = 2\,608$$

Y^* correspondant à la partition solution :

Classe 1 : Lion, tigre.

Classe 2 : Jaguar, léopard, once, puma, pant. nébul, lynx

Classe 3 : Serval, caracal, ocelot, viverrin, jaguarundi, chaus, doré, marguay, margerit, cafer, chine, bengale, rouilleux, malais, borneo, nigripes, manul, marbre, tigrin, temminck, andes.

Classe 4 : Guepard.

Cette partition optimale donne un résultat en accord avec les connaissances généralement admises par les zoologues sur la classification des félins par genres et par espèces.

En effet :

La Classe 1 regroupe les 2 plus gros félins (le tigre et le lion) (genre panthera), que leur corpulence et leur comportement distinguent de façon évidente de l'ensemble des autres.

La Classe 3 (classe d'intermédiarité), regroupe d'une part les 3 plus petits félins du genre "panthera" (Jaguar, léopard, once) d'autre part les 2 plus grand félins du genre "felis" (le puma et le lynx) ainsi qu'un félin d'un genre intermédiaire (la panthère nébuleuse du genre "neofelis"). Cette classe caractérise donc les félins de taille intermédiaire entre les "grands" et les "petits" félins. Ceci d'autant plus

qu'il existe des variations intraspécifiques pour certains des félins de cette classe compliquant la difficulté d'affectation à une classe précise, (ainsi le léopard des savanes africaines peut-il peser plus de 60 kg et être considéré alors comme un félin de grande taille alors que son homologue des forêts asiatiques ne pèse lui qu'une vingtaine de kilos).

La Classe 3 caractérise les petits félins, c'est-à-dire à l'exception du puma et du lynx tous les félins du genre "felis". Ce sont en général des espèces que l'on peut schématiquement affubler du nom de "chat x...".

La Classe 4 enfin contient le guépard tout seul. Ce résultat est très satisfaisant du point de vue de la classification zoologique des félins, car pour la plupart des zoologues le guépard ne devrait pas être considéré comme un félin. Il est en effet morphologiquement très différents des autres félidés; d'ailleurs son nom latin d'acinonyx jubatus traduit cette ambiguité.

Remarque:

Il faut noter qu'ici ce résultat a été obtenu d'une part sans fixation a priori du nombre de classes, d'autre part sans transformation du tableau des données brutes, en particulier sans être obligé de passer par une mise sous forme disjonctive complète.

VIII. REFERENCES

[1] K. ARROW (1963). – Social Choice and individual value, Wiley, New-York.

[2] M. BARBUT (1966). – Note sur les ordres totaux à distance minimum d'une relation binaire donnée, *Math. et Sciences humaines,* n° 17, pp. 47-48.

[3] J.P. BARTHELEMY et B. MONJARDET (1981). – The median procedure in cluster analysis and social choice theory", *Math. social sciences,* Vol 1, n° 3, North Holland, Amsterdam.

[4] J.P. BARTHELEMY et B. MONJARDET (1980). – Ajustement et Résumé de données relationnelles: les relations centrales, in Data Analysis and Informatics, North-Holland.

[5] J.P. BENZECRI et coll. (1973). – L'analyse des Données, Tome I: *la Taxinomie,* Paris, Dunod,

[6] J. de CANI (1969). – Maximum likelihood paired comparison ranking by linear programming, *Biometrika,* Vol. 56, n° 3, 537.

[7] J. de CANI (1972). – A branch and bound algorithm for maximum likelihood paired comparison ranking, *Biometrika,* Vol. 59, n° 1, pp. 131-135.

[8] E. DIDAY (1972). – Nouvelle méthode et nouveaux concepts en Classification Automatique et Reconnaissance des Formes, Thèse de Doctorat d'Etat, Université Paris VI.

[9] J. DORST et P. DANDELOT (1972). – Guide des grands mammifères d'Afrique, Delachaux et Niestlé, Neuchâtel.

[10] J. KEMENY et L. SNELL (1962). – Mathematical Models in the Social Sciences, The M.I.T. Press, Cambridge U.S.A..

[11] M.G. KENDALL (1962). – Rank Correlation Methods, Griffin, Londres,

[12] F. MARCOTORCHINO et P. MICHAUD (1979). – Optimisation en Analyse Ordinale des Données, Masson, Paris.

[13] F. MARCOTORCHINO (1981). – "Agrégation des similarités en classification automatique", Thèse de Doctorat d'Etat, Université Paris VI.

[14] P. MICHAUD (1981). – "Agrégation des préférences", Thèse de Doctorat d'Etat, Université Paris VI.

[15] P. MICHAUD et F. MARCOTORCHINO (1979). – Modèles d'organisation en analyse des données relationnelles, *Math. et Sciences Humaines*, n° 67, pp. 7-38, Dunod, Paris,

[16] B. MIRKIN (1979). – Group Choice, édité par P. Fishburn, V. Winston and Sons, J. Wiley, New-York.

[17] S. REGNIER (1965). – Sur quelques aspects mathématiques des problèmes de classification automatique, *I.C.C.* Bulletin, Rome.