

# CAHIERS DU BURO

JEAN-PIERRE MAILLES

## **Analyse des tableaux de proximités**

*Cahiers du Bureau universitaire de recherche opérationnelle.*

*Série Recherche*, tome 33 (1980), p. 5-96

[http://www.numdam.org/item?id=BURO\\_1980\\_\\_33\\_\\_5\\_0](http://www.numdam.org/item?id=BURO_1980__33__5_0)

© Institut Henri Poincaré — Institut de statistique de l'université de Paris, 1980, tous droits réservés.

L'accès aux archives de la revue « Cahiers du Bureau universitaire de recherche opérationnelle. Série Recherche » 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/>

## INTRODUCTION

Dans ce cahier, on traite de l'analyse des tableaux de dissimilarités. Dans une première partie, nous développons ce qu'on appelle l'analyse factorielle sur tableau de distances, technique intéressante dans la mesure où toutes les analyses factorielles, aussi bien celles qui relèvent de l'analyse canonique que celles qui relèvent de l'analyse en composantes principales, peuvent être considérées comme des analyses factorielles sur tableau de distances particulières. Dans la deuxième partie, exploitant une équivalence entre tableaux qui relève de l'analyse canonique, nous montrons comment on peut utiliser l'information contenue dans un tableau de dissimilarités comme si elle était équivalente à celle fournie par un tableau "individus×caractères". Enfin, dans la troisième partie, nous abordons des techniques d'analyses factorielles qui n'exploitent que les préordonnances associées aux tableaux de dissimilarités ; ces techniques, issues des premiers travaux de R.N. Shepard, connaissent, à l'heure actuelle, une grande vogue aux Etats-Unis.

Les résultats décrits dans le second chapitre sont pour la plupart désormais classiques, certains ayant été publiés dans le livre de F. Cailliez et J.P. Pagès, auquel nous avons contribué. En exploitant la formule qui donne les coordonnées des points supplémentaires, nous proposons une procédure qui apporte une solution aux problèmes numériques posés par l'analyse de tableau de grande dimension.

Dans le troisième chapitre, les résultats sont pour la plupart originaux, même s'ils empruntent souvent à des idées venant d'outre-Atlantique : l'utilisation de l'analyse canonique pour décrire des tableaux de dissimilarités soulève des questions fort intéressantes, les descripteurs (caractères) pouvant être extraits des tableaux de dissimilarités n'ayant pas *a priori* la même importance.

Nous nous sommes efforcés, dans le quatrième chapitre, de dresser un bilan des multiples propositions qui ont été faites, principalement aux Etats Unis, pour construire une image euclidienne à partir de la seule connaissance de l'ordre sur les dissimilarités ; s'étant attaché à retrouver les idées à la base des différentes techniques proposées, ce bilan peut être considéré comme critique. Il l'est d'autant plus que les essais comparatifs qui ont été effectués conduisent à penser que, si les données sont suffisamment homogènes, et c'est en général le cas, les procédures basées sur la notion de préordonnances ne conduisent pas, et ce malgré un coût plus élevé, à des résultats plus intéressants que ceux fournis par l'analyse factorielle sur tableau de distances.



## CHAPITRE I

### RAPPELS ET NOTATIONS

#### 1. MESURES DE PROXIMITES

Pour mesurer les proximités entre éléments de  $I$ , on utilise en général, soit :

- *un indice de dissimilarité  $d$  :*

$d$  est une application de  $I \times I$  dans  $\mathbf{R}^+$  :

$$\begin{array}{ccc} I \times I & \longrightarrow & \mathbf{R}^+ \\ (i, j) & \rightsquigarrow & d(i, j) \end{array}$$

qui vérifie les axiomes :

$$d(i, j) = d(j, i) \text{ pour tout couple } (i, j) \text{ de } I \times I, \quad (1)$$

$$d(i, i) = 0 \quad \text{pour tout élément } i \text{ de } I. \quad (2)$$

- *un indice de distance  $d$  :*

Un indice de distance  $d$  sur  $I$  est un indice de dissimilarité qui vérifie :

$$d(i, j) = 0 \Leftrightarrow i = j. \quad (3)$$

- *une distance  $d$  :*

Une distance  $d$  sur  $I$  est un indice de distance sur  $I$  qui vérifie l'inégalité triangulaire (4) :

$$d(i, j) \leq d(i, k) + d(k, j) \text{ pour tout triplet } (i, j, k) \text{ de } I^3. \quad (4)$$

- *un écart :*

Un écart est un indice de dissimilarité vérifiant (4).

- *une distance ultramétrique :*

Une distance ultramétrique sur  $I$  est une distance sur  $I$  qui vérifie :

$$d(i, j) \leq \max_{k \in I} \{d(i, k), d(j, k)\}. \quad (5)$$

*Remarque :*

- La propriété (5) implique la propriété (4).
- Lorsque la distance  $d$  définie sur  $I$  est ultramétrique, tous les triangles sont "isocèles pointus".

On trouve, dans la littérature, de nombreux exemples d'indices de dissimilarité, par exemple dans [4].

## 2. PREORDONNANCE ET ORDONNANCE SUR UN ENSEMBLE $I$

Rappelons qu'un préordre  $R$  défini sur un ensemble  $I$  est une relation binaire réflexive et transitive ; si cette relation est de plus antisymétrique, le préordre  $R$  est un ordre.

**Définition**

*Une préordonnance sur  $I$  est un préordre total défini sur l'ensemble*

$$\{(i, i') | i \in I, i' \in I, i < i'\} \text{ des } \frac{n(n-1)}{2} \text{ couples de } I \times I.$$

A tout indice de dissimilarité  $d$  défini sur  $I$  est associée la préordonnance définie par :

$$(i, j) \underset{d}{\leq} (k, l) \Leftrightarrow d(i, j) \leq d(k, l).$$

Si  $n_{ij}$  est le nombre de sujets qui estiment que les objets  $i$  et  $j$  se ressemblent, on peut mesurer la similitude entre les objets  $i$  et  $j$  par :

$$s_{ij} = \frac{n_{ij}}{n_{i.} + n_{.j} + n_{.i} + n_{.j}}$$

avec  $n_{i.} = \sum \{n_{ik} | k = 1, \dots, n\}$

$n_{.j} = \sum \{n_{kj} | k = 1, \dots, n\}.$

Stefflire [3] propose alors une méthode simple pour obtenir une préordonnance sur  $I$  :

$$(i, j) < (i', j') \Leftrightarrow s_{ij} > s_{i'j'}$$

### 3. SCHEMA DE DUALITE

On considère :

- Deux ensembles  $I$  et  $J$ ,  
où  $I$  est appelé ensemble des individus avec  $|I| = n$ ,  
 $J$  ensemble des caractères avec  $|J| = p$ .
- Une application  $x$  de  $I \times J$  dans  $Q$ ,

$$(i, j) \xrightarrow{x} x(i, j) = x_j^i,$$

où  $Q$  est appelé ensemble des réponses ou des modalités des caractères.

Dans la suite, sauf précision contraire, l'ensemble  $Q$  est identifié à  $\mathbf{R}$  ou à une partie de  $\mathbf{R}$ .

En considérant les applications coordonnées :

$$- \text{à } i \text{ fixé, } \quad x(i, \cdot) : J \longrightarrow \mathbf{R}$$

$$\underline{x}_i = \{x(i, j) \mid j \in J\} \in \mathbf{R}^p \text{ (produit cartésien de } \mathbf{R})$$

caractérise l'individu  $i$  ;

$$E = \mathbf{R}^p \text{ est l'ensemble des individus.}$$

$$- \text{à } j \text{ fixé, } \quad x(\cdot, j) : I \longrightarrow \mathbf{R}$$

$$\underline{x}^j = \{x(i, j) \mid i \in I\} \in \mathbf{R}^n \text{ (produit cartésien de } \mathbf{R})$$

caractérise la variable  $j$  ;

$$F = \mathbf{R}^n \text{ est l'ensemble des caractères.}$$

*Hypothèse H1*

*$E$  et  $F$  sont munis de leur structure d'espace vectoriel naturel.*

$E$  et  $F$  sont munis de leur base canonique :

$$\{\underline{e}_j \mid j = 1, \dots, p\} \text{ et } \{\underline{f}_i \mid i = 1, \dots, n\}.$$

Dans  $E$ , chaque direction de la base est associée à un caractère ; dans  $F$ , chacune d'elles est associée à un individu.

Si  $E^*$  et  $F^*$  sont les espaces duaux de  $E$  et  $F$ , munis de leurs bases canoniques :

$$\{\underline{e}_j^* \mid j = 1, \dots, p\} \text{ et } \{\underline{f}_i^* \mid i = 1, \dots, n\},$$

on a :

$$\underline{x}_i = \begin{pmatrix} x_i^1 \\ x_i^2 \\ \vdots \\ x_i^p \end{pmatrix} = \Sigma \{x_i^j \underline{e}_j \mid j = 1, \dots, p\}$$

$$x_i^j = e_j^*(\underline{x}_i) = \langle \underline{e}_j^*, \underline{x}_i \rangle$$

$$\underline{x}^j = \begin{pmatrix} x_1^j \\ x_2^j \\ \vdots \\ x_n^j \end{pmatrix} = \Sigma \{x_i^j \underline{f}_i \mid i = 1, \dots, n\}$$

$$x_i^j = f_i^*(\underline{x}^j) = \langle \underline{f}_i^*, \underline{x}^j \rangle.$$

Alors on peut associer :

- à un caractère  $j$  :

$\Delta e_j$  axe de  $E$  engendré par  $\underline{e}_j$

$\underline{e}_j^*$  forme linéaire de  $E^*$

$\underline{x}^j$  vecteur de  $F$ .

- à un individu  $i$  :

$\Delta f_i$  axe de  $F$  engendré par  $f_i$

$f_i^*$  forme linéaire de  $F^*$

$x_i$  vecteur de  $E$ .

On note  $X = \{x_j^i \mid i \in I ; j \in J\}$ , le tableau des données, du type individus  $\times$  caractères.

La matrice  $X$  de format  $(p, n)$  peut alors être considérée comme la matrice associée à l'application linéaire de  $F^*$  dans  $E$ , qui aux vecteurs de base de  $F^*$  associe les colonnes de  $x_i$  repérées dans la base  $\{e_j\}$  :

$$X(f_i^*) = x_i.$$

De même la matrice transposée  $X^t$  est la matrice de l'application linéaire de  $E^*$  dans  $F$  munis de bases  $\{e_j^*\}$  et  $\{f_i\}$  telle que

$$X^t(e_j^*) = x_j.$$

Pour mesurer les "ressemblances" entre individus ou entre caractères,

*Hypothèse H2 :*

*On munit  $E$  (respectivement  $F$ ) d'une structure d'espace euclidien, par la donnée d'une forme bilinéaire définie positive symétrique  $M$  (respectivement  $N$ ).*

*Remarque :*

A la forme bilinéaire définie positive symétrique définie sur  $E$  et notée  $M$ , on peut associer :

- la forme quadratique notée aussi  $M : M(x) = M(x, x)$
- la norme euclidienne :  $\|x\|_M = \sqrt{M(x, x)}$
- la distance euclidienne :  $d_M(x, y) = \|x - y\|_M$
- un isomorphisme de  $E$  dans  $E^* : x^* = M(x)$   
tel que  $\langle x^*, y \rangle = M(x, y)$ .

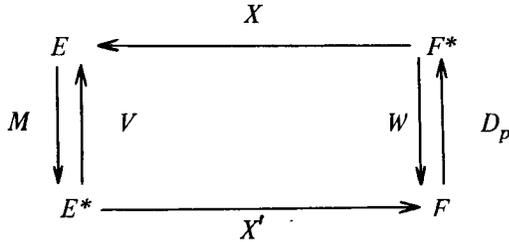
On identifie, dans les notations, forme bilinéaire, forme quadratique et application linéaire associée, ainsi que leur représentation matricielle.

On a des résultats similaires et des notations analogues pour la forme bilinéaire  $N$  dans  $F$ . Usuellement, on choisit pour  $N$ , la métrique des poids  $D_p$  définie par la matrice diagonale

$$D_p = \begin{pmatrix} p_1 & & & \\ & p_2 & & \\ & & \ddots & \\ 0 & & & p_n \end{pmatrix}$$

où  $\sum \{p_i \mid i = 1, \dots, n\} = 1$   
 (distance en moyenne quadratique).

Les applications précédentes sont résumées dans le schéma de dualité suivant :



où  $V$  et  $W$  sont les écarts euclidiens induits respectivement par  $D_p$  sur  $E^*$  et par  $M$  sur  $F^*$  :

$$V = X \circ D_p \circ X' \quad W = X' \circ M \circ X$$

tels que

$$V(\underline{e}_{j1}^*, \underline{e}_{j2}^*) = D_p(\underline{x}^{j1}, \underline{x}^{j2})$$

$$W(\underline{f}_{i1}^*, \underline{f}_{i2}^*) = M(\underline{x}_{i1}, \underline{x}_{i2})$$

$V$  n'est autre que la matrice des moments d'ordre deux, et, si les caractères sont centrés, la matrice de variance-covariance.

On note  $D$  le tableau des distances entre individus :

$$D_{(n, n)} = (d_{i, i'}) = (\|\underline{x}_i - \underline{x}_{i'}\|_M)$$

L'opérateur introduit par Y. Escoufier [11], [12], pour représenter un tableau de données n'est autre que l'opérateur :

$$F \xrightarrow{W \circ D_p} F.$$

Cet opérateur permet d'introduire très simplement des équivalences entre : tableaux de distances, tableaux de données dans une optique de description, paquets de variables quantitatives, un paquet de variables quantitatives et qualitatives ([36], [40]) : cf. § 3.2 du chapitre II.

On note :

- $I = \{1, 2, \dots, n\}$  l'ensemble des individus considérés,
- $\mathfrak{N} = \{x_i \in E \mid i = 1, \dots, n\}$  le nuage des individus dans  $E$ ,
- $\mathfrak{X} = \{x^j \in F \mid j = 1, \dots, p\}$  le nuage des caractères dans  $F$ ,
- $d_I$  l'indice de dissimilarité ou la distance introduite sur  $I$  pour mesurer les proximités.

$(\mathfrak{N}, E, M)$  désigne le triplet constitué du nuage  $\mathfrak{N}$ , du vectoriel  $E$  où est situé le nuage  $\mathfrak{N}$ , et de la métrique  $M$  définie sur  $E$ .

## CHAPITRE II

### ANALYSE FACTORIELLE SUR TABLEAU DE DISTANCES (AFTD)

#### 1. EQUATIONS DE L'ANALYSE FACTORIELLE SUR TABLEAU DE DISTANCES

On dit que  $(\mathfrak{N}, E, M)$  est une image euclidienne de  $(I, d_I)$  si pour tout couple  $(i, i')$  de  $I \times I$

$$d(i, i') \doteq d_{ii'} = \|x_i - x_{i'}\|_M.$$

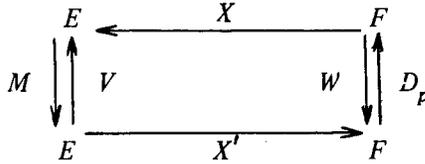
Construire, ayant jugé de son existence, une image euclidienne simple, c'est-à-dire située dans un espace de faible dimension, de  $(I, d_I)$  est l'objectif de l'analyse factorielle sur tableau de distances. Si la dimension de l'espace est égale à 1, 2 ou 3, on pourra juger à l'oeil des proximités entre éléments de  $I$  et construire ainsi par exemple une partition de  $I$  en classes homogènes.

Rappelons deux théorèmes [51] :

#### *Théorème 1 – Unicité*

*Si  $(\mathfrak{N}, E, M)$  est une image euclidienne de  $(I, d_I)$ , l'analyse en composantes principales de  $\mathfrak{N}$  ne dépend que des masses  $p_i$  associées aux éléments de  $I$  et des valeurs prises par  $d_I$ .*

Le schéma de dualité en analyse en composantes principales s'écrivant :



la démonstration du théorème repose sur la remarque, si  $w_{ii'}$  désigne le terme  $(i, i')$  de la matrice  $W$  :

$$w_{ii'} = W(f_i, f_{i'}) = \frac{1}{2} (d_i^2 + d_{i'}^2 - d_{..}^2 - d_{ii'}^2) \quad (1)$$

avec

$$d_{i.}^2 = \sum \{p_k d_{ik}^2 \mid k \in I\}$$

$$d_{.i'}^2 = \sum \{p_i p_{i'} d_{ii'}^2 \mid (i, i') \in I \times I\}.$$

**Théorème 2 : Existence**

Pour qu'il existe une image euclidienne de  $(I, d_i)$ , il faut et il suffit que la forme bilinéaire symétrique  $W$  définie par les équations (1) soit semi-définie positive.

La démonstration n'offre pas de difficultés, on obtient une image euclidienne en tirant les vecteurs propres de  $W \circ D_p$  qui ne sont autres que les composantes principales du nuage  $\mathcal{M}$  recherché.

*Remarque :*

La matrice  $W$  n'est autre, à une constante près, que le tableau  $\bar{D}$  des carrés des distances doublement centré.

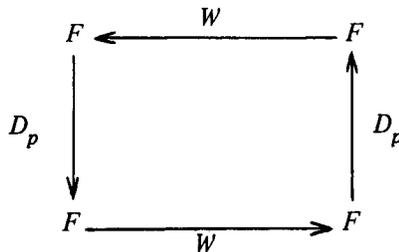
Si  $A = \text{Id} - \underline{j}\underline{j}' D_p$  désigne le  $D_p$ -projecteur sur le sous-espace vectoriel  $\Delta_I^\perp$  de  $F$ ,  $D_p$ -orthogonal à la droite des constantes, avec  $\text{Id}$  application "identité", on a :

$$W = -\frac{1}{2} A \bar{D} A'.$$

**Proposition :**

Les composantes principales du tableau doublement centré  $-1/2 A \bar{D} A'$  sont aussi les vecteurs propres de l'opérateur  $W D_p$ , l'espace des "individus" étant ici de dimension  $n$  et muni de la métrique  $D_p$ .

En effet en considérant le schéma de dualité :



on est amené, pour effectuer l'analyse en composantes principales de  $W$ , à extraire les valeurs propres et les vecteurs propres de  $(W D_p W) D_p = (W D_p)^2$ .

Les valeurs propres de  $W D_p$  sont les racines carrées des valeurs propres de  $(W D_p)^2$ , les composantes principales, vecteurs propres de  $W D_p$ , sont aussi vecteurs propres de  $(W D_p)^2$ .

On note que, dans l'analyse en composantes principales précédente, les valeurs propres qui étaient négatives en analyse factorielle sur tableau de distances, si elles sont de grand module, risquent de conduire à retenir, parmi les premières composantes principales, des vecteurs qui sont associés aux plus petites valeurs propres de l'analyse factorielle sur tableau de distances.

### *Proposition*

*Les vecteurs colonnes de  $W$  engendrent le même sous-espace vectoriel que les composantes principales associées aux valeurs propres non nulles.*

En effet, on sait que, dans une analyse en composantes principales, ce sous-espace vectoriel est engendré à la fois par les composantes principales et les vecteurs lignes de  $X$ . Or d'après ce qui précède, les vecteurs  $\underline{c}$  sont composantes principales de  $(W D_p W D_p) = (W D_p)^2$  en même temps que de  $W D_p$ , donc ils engendrent le même sous-espace vectoriel que les lignes (ou les colonnes) de  $W$ .

## 2. PRATIQUE DE L'ANALYSE FACTORIELLE SUR TABLEAU DE DISTANCES

On vient de voir que l'on sait reconnaître l'existence et reconstruire l'image euclidienne par l'analyse factorielle sur tableau de distance du couple  $(I, d_I)$ . On conçoit que si  $W$  est semi-définie positive, aucun problème ne se pose au niveau de la pratique. Par contre, si  $W$  n'est pas semi-définie positive, comment trouver et apprécier une image euclidienne de  $(I, d_I)$  ?

### 2.1. $W$ est semi-définie positive

On reconstruit plus ou moins parfaitement la proximité  $d_{ii'}$ , entre éléments de  $I$ , suivant que l'on retient 2,3 ou  $k$  vecteurs propres  $\underline{c}$  de  $W \circ D_p$  (vecteurs propres associés aux plus grandes valeurs propres).

La meilleure image euclidienne approximative de  $(I, d_I)$  de dimension 2 est obtenue, par exemple, en considérant dans un plan, le nuage de points  $(c_i^1, c_i^2)$ , où  $c_i^j$  désigne la valeur prise par la  $j^{\text{ème}}$  composante principale pour l'individu  $i$ ; les vecteurs de base  $\underline{u}_1$  et  $\underline{u}_2$  étant placés orthogonalement à une distance unité de l'origine  $\underline{O}$ , on peut ainsi à l'oeil, si la qualité de l'image est bonne, établir un classement de  $I$ , les longueurs des segments séparant les points étant sensiblement égales aux valeurs prises par la distance  $d_I$ .

La part d'inertie expliquée par le plan principal, comme en analyse en composantes principales, permettra de juger la qualité globale de l'image obtenue. Rappelons que cette part qui est égale à

$$\frac{\lambda_1 + \lambda_2}{\text{tr}(W \circ D_p)}$$

(où  $\lambda_i$  est la  $i^{\text{ème}}$  valeur propre de  $W \circ D_p$ ) est comprise entre 0 et 1, et n'est égale à 1 que si les proximités sont rigoureusement reconstruites dans le plan.

*Remarque :*

L'inertie au centre de gravité  $I_g$  vérifie :

$$I_g = \text{tr}(W \circ D_p) = \sum \{\lambda_i \mid i = 1, \dots, n\} = \frac{1}{2} d^2.$$

## 2.2. Cas général : $W$ n'est pas semi-définie positive

$W \circ D_p$  admet alors des valeurs propres négatives, et d'après le théorème 2 il n'existe pas d'image euclidienne parfaite de  $(I, d_I)$ . On pourrait se contenter encore d'une image euclidienne approximative que l'on obtient en retenant les 2,3 ou  $k$  premiers vecteurs propres de  $W \circ D_p$  dont les valeurs propres associées, ordonnées par valeurs décroissantes, sont positives; si les valeurs propres négatives sont petites en valeur absolue, on peut penser que la description des individus obtenue peut ne pas être trop mauvaise. Mais comment juger de façon précise de la qualité globale de l'image ?

La part d'inertie expliquée, qui fait intervenir la trace de  $W \circ D_p$  n'a plus de sens ici puisqu'elle peut dépasser largement 1. On va s'interroger, maintenant, sur les transformations qu'il est possible d'opérer sur  $d_I$  de façon à rendre  $W$  semi-définie positive : cette réflexion permettra en particulier de se faire une idée précise de l'interprétation qui peut être donnée aux valeurs propres négatives de  $W \circ D_p$ .

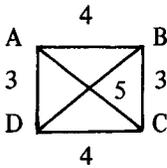
Rappelons que, pour décrire les proximités entre éléments d'un ensemble, on utilise en général des indices de dissimilarité qui ne vérifient pas l'inégalité triangu-

laire et qui sont moins contraignants que les distances. La forme quadratique  $W$  ne peut être semi-définie positive que si  $d_I$  vérifie l'inégalité triangulaire ; il est bien connu que cette condition n'est pas suffisante. On cite ci-dessous un contre-exemple de cette propriété. La matrice associée à la distance  $d_I$  définie sur

$$I = \{A, B, C, D\}$$

par le tableau ci-dessous n'est pas définie positive.

	A	B	C	D
A	0	2	5	3
B	2	0	3	5
C	5	3	0	4
D	3	5	4	0



Le tableau de distance a été obtenu en "raccourcissant" le côté AB du rectangle ci-contre.

### 2.2.1. Transformation sur $d_I$ permettant d'obtenir un indice $\delta_I$ vérifiant l'inégalité triangulaire

Les transformations que l'on applique en général sur les indices de dissimilarité permettent de conserver la préordonnance associée à l'indice considéré ; on sait en effet depuis R.N. Shepard [42], l'importance de la notion de préordonnance. En effet, la connaissance seule de la préordonnance associée à une distance euclidienne permet, si le nombre de points considérés est grand, de retrouver avec une bonne précision la dimension de l'espace dans lequel se trouvait le nuage considéré, et la position dans cet espace, à des similitudes près, des points de ce nuage.

On considère alors l'équivalence suivante entre indices :  $d_I$  et  $\delta_I$ , définis sur  $I$ , sont équivalents si et seulement si :

$$d_{ij} < d_{kl} \Leftrightarrow \delta_{ij} < \delta_{kl} \quad \forall \{i, j, k, l\} \subset I$$

Les transformations les plus communes conservant la pré-ordonnance appartiennent aux deux familles suivantes :

$$1) \quad \begin{cases} \delta_{ii'} = 0 & \text{si } i = i' \\ \delta_{ii'} = (d_{ii'}^r + c)^{1/r} & \text{si } i \neq i' \end{cases}$$

où  $r$  désigne un nombre positif quelconque et  $c$  une constante positive choisie de façon que l'inégalité triangulaire soit respectée par  $\delta_I$ . Quand  $r$  est pris égal à 1, on choisit la constante  $c$  de façon à modifier au minimum  $d_I$ ; on trouve alors :

$$c = \text{Max} \{d_{ij} - d_{ik} - d_{jk} \mid (i, j, k) \in I^3\}$$

$$2) \quad \delta_{ii'} = d_{ii'}^r \quad \forall (i, i') \in I \times I$$

où  $r$  est le plus grand nombre compris entre 0 et 1 qui soit tel que  $\delta_{ii'}$  vérifie l'inégalité triangulaire.

On rencontre, dans la littérature anglo-saxonne, bien des développements à propos des transformations sur les indices ; il nous paraît plus intéressant de travailler sur les carrés des indices eux-mêmes, compte-tenu de l'expression de  $W$  donnée précédemment (équation (1)).

### 2.2.2. Transformations sur $d_I$ qui rendent $W$ semi-définie positive

Les dissimilarités  $d_{ii'}$  interviennent par leur carré dans les équations (1). Aussi, si  $d_I$  et  $c_I$  sont deux indices de dissimilarité définis sur  $I$  dont  $W_d$  et  $W_c$  sont respectivement les formes quadratiques, à l'indice de dissimilarité  $\delta_I$ , défini par :

$$\delta_{ii'}^2 = d_{ii'}^2 + c_{ii'}^2 \quad \forall (i, i') \in I \times I$$

est associée la forme quadratique

$$W_\delta = W_d + W_c .$$

La transformation la plus simple est celle qui consiste à poser :

$$\begin{cases} c_{ii'} = 0 & \text{si } i = i' \\ c_{ii'} = c & \text{si } i \neq i' \end{cases}$$

Si tous les poids  $p_i$  sont pris égaux à  $1/n$ , la forme quadratique  $W_c$  associée à l'indice de dissimilarité  $c_I$  est définie par :

$$W_c(\underline{f}_i, \underline{f}_{i'}) = \begin{cases} \frac{n-1}{2n} c^2 & \text{si } i = i' \\ -\frac{1}{2n} c^2 & \text{si } i \neq i' \end{cases}$$

Si  $\underline{j}$  désigne le vecteur de  $F$  dont toutes les coordonnées rangées en colonne sont égales à 1, et Id est la matrice identité, l'expression matricielle de  $W_c$  est donnée par :

$$W_c = -\frac{1}{2n} c^2 \underline{j} \underline{j}' + \frac{1}{2} c^2 \text{Id} .$$

Le noyau de  $W_c$  coïncide avec la droite  $\Delta_{\underline{j}}$  engendrée par  $\underline{j}$ , et le sous-espace  $\Delta_{\underline{j}}^\perp$ , orthogonal à  $\Delta_{\underline{j}}$  (donc en particulier à  $\underline{j}$ ) dans  $F$  est le sous-espace propre de  $W_c$  associé à la valeur propre  $1/2 c^2$ .

Notons  $\mu_1 \geq \mu_2 \geq \dots \geq \mu_n$  les valeurs propres de  $W_d$  et remarquons que :

–  $\underline{j}$  vecteur propre de  $W_d$  associé à la valeur propre 0 est aussi vecteur propre de  $W_\delta$ , associé à la valeur propre 0.

– Tout vecteur propre de  $W_d$  associé à la valeur propre 0 et orthogonal à  $\Delta_{\underline{j}}$  est vecteur propre de  $W_\delta$  associé à la valeur propre  $1/2 c^2$ .

– Tout vecteur propre de  $W_d$  associé à la valeur propre  $\mu_i$  non nulle est vecteur propre de  $W_\delta$  associé à la valeur propre  $\mu_i + 1/2 c^2$ .

Pour que la matrice symétrique  $W_\delta$  soit semi-définie positive, il faut et il suffit que ses valeurs propres soient toutes positives ou nulles. D'où le théorème :

### ***Théorème 3***

*La valeur minimum qu'il faut donner à la constante additive  $c$  pour que  $W_\delta$  soit semi-définie positive,  $W_d$  ne l'étant pas, est égale à  $\sqrt{2 |\mu_n|}$  où  $\mu_n$  désigne la plus petite valeur propre de  $W_d$ .*

*Remarque :*

On sait que si, pour un système de poids donné, la matrice  $W$  est semi-définie positive, elle l'est pour tout autre système de poids. Aussi, la constante que l'on rajouté aux  $d_{ii}^2$ , rend  $W$  semi-définie positive, quel que soit le système de poids. Il est évident que si les poids ne sont pas tous égaux à  $1/n$ ,  $W_c$  n'a pas l'expression donnée par la formule (2), mais s'écrit :

$$W_c = |\mu_n| (\text{Id} - A) (\text{Id} - A'),$$

où  $A = \underline{j} \underline{j}' D_p$  est la matrice associée au  $D_p$ -projecteur sur l'axe  $\Delta_{\underline{j}}$ .

On rappelle que :  $W_\delta = W_d + W_c$ .

### 2.2.3. Pratique de l'analyse factorielle sur tableau de distance quand $W$ n'est pas semi-définie positive

Il ne paraît pas utile d'opérer des transformations du type de celles décrites au paragraphe 2.2.1. ci-dessus et qui sont préconisées par certains auteurs.

Si toutes les masses  $p_i$  sont égales à  $1/n$ , en effectuant la transformation déterminée au paragraphe 2.2.2., les sous-espaces propres de l'opérateur (d'Y. Escoufier)  $W \circ D_p$  associés aux valeurs propres  $\lambda_i = \mu_i/n$  non nulles ne sont pas modifiées. Les valeurs propres correspondantes sont toutes augmentées de  $|\lambda_n|$ , où  $\lambda_n$  est la plus petite valeur propre de  $W_d \circ D_p$ .

$$\lambda_i \neq 0 : W_d \circ D_p \underline{c}^i = \lambda_i \underline{c}^i \Leftrightarrow W_\delta \circ D_p \underline{c}^i = (\lambda_i + |\lambda_n|) \underline{c}^i.$$

Si la valeur absolue de  $|\lambda_n|$  est grande relativement aux deux ou trois plus grandes valeurs propres positives, l'image euclidienne approximative de  $(I, d_I)$  obtenue en ne retenant que les deux ou trois premières composantes principales  $\underline{c}^i$  sera très imparfaite, et on ne devra lui porter qu'un intérêt très relatif compte tenu de l'importance de la déformation subie par le tableau des  $d_{II}^2$ .

Pour mesurer la qualité globale de la représentation obtenue par l'image euclidienne à deux dimensions dans le plan principal, on utilisera l'indice :

$$\frac{\lambda_1 + \lambda_2 + 2|\lambda_n|}{\text{tr}(W_d D_p) + (n-1)|\lambda_n|}$$

qui n'est autre que la part d'inertie expliquée par le plan principal dans l'analyse factorielle du couple  $(I, \delta_I)$ .

Les composantes principales  $\underline{c}^i$ , vecteurs propres de  $W_d \circ D_p$ , doivent être théoriquement de norme  $\sqrt{\lambda_i}$  ; effectuer la transformation sur  $d_I$  conduit à les normer à  $\sqrt{\lambda_i + |\lambda_n|}$ .

### 2.2.4. Cas particulier : $d_I$ est une distance ultramétrique

On donne dans ce paragraphe une démonstration très courte d'un théorème récent énoncé par Holmann [19] puis par Y. Escoufier.

**Théorème 4 :**

*Si  $d_I$  est une distance ultramétrique, il existe une image euclidienne de  $(I, d_I)$ .*

La démonstration s'appuie sur les deux lemmes suivants :

**Lemme 1 :**

Soit une partition de  $I$  et  $c$  un nombre positif ; définissons  $d$  par :

$$\begin{cases} d_{ii'} = 0 & \text{si } i \text{ et } i' \text{ appartiennent à la même classe de la partition} \\ d_{ii'} = c & \text{sinon} \end{cases}$$

Alors il existe une image euclidienne de  $(I, d_I)$ .

En effet, on a une image euclidienne évidente dans le cas où tous les poids sont égaux à  $1/n$ , donc dans tous les cas (voir remarque 2.2.2.), en plaçant tous les individus d'une classe en un même point et les points représentant chaque classe aux sommets d'un polyèdre régulier de côté  $c$ .

**Lemme 2 :**

Si  $d, d_1$  et  $d_2$  sont trois indices définis sur  $I$  tels que :

$$d^2(i, i') = d_1^2(i, i') + d_2^2(i, i') \quad \forall (i, i') \in I \times I,$$

si il existe des images euclidiennes de  $(I, d_1)$  et de  $(I, d_2)$ , alors il existe une image euclidienne de  $(I, d)$ .

La démonstration résulte de l'étude faite au paragraphie 2.2. pour déterminer  $W$  comme somme de  $W_1$  et  $W_2$  et de la propriété pour  $W_1$  et  $W_2$  d'être semi-définies positives (théorème 2).

Soit  $c_1 = 0 < c_2 < \dots < c_k$  les  $k$  valeurs distinctes prises par l'indice ultramétrique  $d_I$ .

Définissons la famille d'indices  $\{d_\alpha \mid \alpha = 1, 2, \dots, k-1\}$  par :

$$d_\alpha(i, i') = \begin{cases} 0 & \text{si } d_{ii'} \leq c_\alpha \\ \sqrt{c_{\alpha+1}^2 - c_\alpha^2} & \text{sinon} \end{cases}$$

D'après le lemme 1, il existe une famille d'images euclidiennes des  $(I, d_\alpha)$ .

Or on a :  $d_{ii'}^2 = \sum \{d^2(i, i') \mid \alpha = 1, \dots, k-1\} \quad \forall (i, i') \in I \times I.$

Donc d'après le lemme 2, on en déduit qu'il existe une image euclidienne de  $(I, d_I)$ .

### 2.3. Points supplémentaires

La matrice qu'il faut diagonaliser est de format  $(n, n)$ , il paraît donc difficile d'effectuer une analyse factorielle sur tableau de distances dès que le nombre d'individus est trop grand. En réalité, on peut toujours ne faire intervenir dans un premier temps que  $k$  des  $n$  individus qui sont considérés comme représentatifs de l'ensemble, les individus restant intervenant en éléments supplémentaires, avec un poids nul.

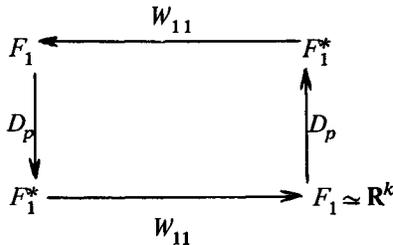
#### 2.3.1. Coordonnées des points supplémentaires

Notons

$$W = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} = -\frac{1}{2} A \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} A'$$

la matrice "doublement centrée" associée à l'ensemble des individus considérés,  $W_{22}$  désignant la matrice associée aux individus supplémentaires. On a évidemment  $W_{12} = W_{21}'$ .

En se basant sur la propriété de la remarque du § 1, on est alors amené à calculer les coordonnées  $\underline{c}_2$  des points supplémentaires décrits par  $W_{21}$  dans le système des axes principaux obtenus en effectuant l'analyse en composantes principales du tableau  $W_{11}$  qui est doublement centré.



En effet, si on note  $\underline{c}_1$  une composante principale  $D_p$  — normée à  $\lambda$ , on sait que  $\underline{c}_1$  est vecteur propre de  $W_{11}$ ,  $D_p$  associé à la valeur propre  $\lambda$  et de  $(W_{11} D_p)^2$  associé à la valeur propre  $\lambda^2$ ; alors  $\frac{1}{\sqrt{\lambda}} D_p \underline{c}_1$  est un facteur principal dans l'analyse en composantes principales de  $W_{11}$  et  $\sqrt{\lambda} \underline{c}_1$  est la composante principale associée. Les coordonnées des points supplémentaires sont donc données par :

$$\underline{c}_2 = \frac{1}{\sqrt{\lambda}} W_{21} D_p \frac{\underline{c}_1}{\sqrt{\lambda}} = \frac{1}{\lambda} W_{21} D_p \underline{c}_1$$

où  $\underline{c}_1$  est le vecteur propre de  $W_{11} D_p$ , associé à la valeur propre  $\lambda$  et de norme  $\lambda$ .

Procéder ainsi revient à effectuer l'analyse factorielle sur tableau de distances sur l'ensemble des individus, les individus supplémentaires étant munis d'un poids nul. En effet, les composantes principales sont vecteurs propres de l'opérateur :

$$U = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} D_p & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} W_{11} D_p & 0 \\ W_{21} D_p & 0 \end{bmatrix}$$

soit 
$$U \underline{c} = \lambda \underline{c} \quad \text{avec } \underline{c} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$$

donc 
$$\begin{cases} W_{11} D_p c_1 = \lambda c_1 \\ W_{21} D_p c_1 = \lambda c_2 \end{cases}$$

### 2.3.2. Sélection des points supplémentaires

Etant donné l'économie de calculs qu'impliquent les résultats précédents, on peut essayer d'obtenir une image euclidienne en ne considérant qu'une partie des individus, soit  $k$  parmi les  $n$  initiaux, les  $n-k$  autres étant traités en points supplémentaires. Il s'agit donc d'extraire parmi les  $n$  individus,  $k$  points "représentatifs", c'est-à-dire ceux pour lesquels l'analyse est équivalente à l'analyse factorielle sur le tableau de distances de l'ensemble des  $n$  points. Les deux analyses sont évidemment équivalentes si les deux tableaux de distances reconstruits sont égaux. Les deux analyses sont considérées comme similaires si les inerties reconstruites sont proches.

D'une manière pratique, la sélection des  $k$  points "privilegiés" doit se faire de telle façon que l'analyse sur ceux-ci soit la plus proche possible de l'analyse factorielle sur les  $n$  points au sens du critère précédent.

On sait que l'analyse en composantes principales, en raison du choix de l'inertie comme critère d'adéquation, donne, pour déterminer les axes principaux, une importance relativement grande aux points éloignés de l'origine. On a donc intérêt à choisir les  $k$  points parmi ceux-ci, tout en sélectionnant des points relativement éloignés entre eux pour donner plus de stabilité aux sous-espaces principaux. C'est cette idée qui a amené Lerman et Leredde [31] à proposer la méthode des pôles d'attraction.

En appliquant une méthode analogue, on sélectionne les points nécessaires pour obtenir une bonne image euclidienne des  $n$  points grâce à l'analyse facto-

rielle sur les seuls points "privilegiés" en choisissant les points qui ont un grand "bras de levier".

On procède de manière progressive de la façon suivante :

– le premier point  $i_1$  sélectionné est le point qui a le plus fort bras de levier

$$P_1(i) = p_i (\sum \{p_l d_{il}^2 \mid l = 1, \dots, n\})^{\frac{1}{2}}$$

– le  $(s + 1)^{\text{ème}}$  point sélectionné, après les  $s$  points  $\{i_1, i_2, \dots, i_s\}$ , est celui qui maximise le critère  $P_{s+1}(i)$ .

$P_{s+1}(i)$  mesure l'intérêt de la sélection du point  $i$  pour les résultats de l'analyse.

$$P_{s+1}(i) = p_i \times \left( \frac{1}{\Pi} \sum \{p_l d_{il}^2 \mid l \neq i, \dots, i_s\} \right)^{\frac{1}{2}} \times \text{Min} \{d_{ij} \mid j = i_1, \dots, i_s\}$$

où

$$\Pi = \sum \{p_l \mid l \neq i_1, \dots, i_s\}$$

Cette quantité  $P_{s+1}(i)$  est un compromis entre le bras de levier de  $i$  mesuré par  $\sum \{p_l d_{il}^2 \mid l \neq i_1, \dots, i_s\}$ , l'éloignement de  $i$  aux  $s$  points déjà sélectionnés et le poids  $p_i$  du point considéré.

L'examen des quantités  $P_s(i)$  permet en plus d'estimer le nombre de points à choisir en tenant compte de deux facteurs pour arrêter la sélection :

– si, au cours d'une itération, les quantités  $P_s(i)$  sont à peu près constantes, tous les points ont le même bras de levier et aucun n'est plus intéressant que les autres ;

– si, entre deux itérations, les quantités  $P_s(i) - P_{s+1}(i)$  sont à peu près nulles, aucun nouveau point n'améliore l'image obtenue.

On peut opérer plus simplement de la façon suivante :

– tirer au hasard, si on ne peut faire mieux,  $k$  points parmi les  $n$  ( $k/n$  de l'ordre de 0.1) ;

– effectuer une analyse factorielle sur le tableau de distances des  $k$  points choisis, les  $n-k$  autres intervenant comme points supplémentaires ;

– choisir  $k'$  nouveaux points sur les graphiques obtenus, de préférence parmi les points extrêmes dans les sous-espaces principaux (cf. ci-dessus) ;

– effectuer une deuxième analyse factorielle sur le tableau de distances des  $k'$  points, les  $n-k'$  autres intervenant toujours comme points supplémentaires.

Si on met en parallèle les opérations à effectuer dans les trois stratégies :

- analyse factorielle sur le tableau de distances complet,
- analyse factorielle sur le tableau de distances des seuls points à plus grands bras de levier,
- analyses factorielles successives sur deux sous-ensembles de points,

les économies de calculs apparaissent clairement, quand on sait que l'extraction des valeurs et vecteurs propres d'une matrice a un coût fortement croissant, de l'ordre du cube de  $n$ .

En effet, dans le premier cas, on doit calculer la matrice  $W$  de format  $(n, n)$ , puis diagonaliser l'opérateur  $U = W \circ D_p$ , également de format  $(n, n)$ . Dans le second cas, on doit calculer les matrices  $W_{11}$  et  $W_{12}$  de format global  $(n, k)$ , puis diagonaliser l'opérateur  $W_{11} \circ D_p$  de format  $(k, k)$ , enfin calculer les coordonnées des  $n - k$  points restants par un produit matriciel. Mais la sélection des  $k$  points privilégiés par la méthode du bras de levier est relativement longue puisqu'elle est basée sur un calcul de moments, puis une série d'optimisations min-max. C'est pourquoi le troisième cas où on effectue deux analyses factorielles partielles sur les  $k$  points privilégiés par tirage au sort puis à vue sur les graphiques, est souvent avantageux.

### 2.3.3. Exemple

On donne ci-dessous un exemple de cette méthode.

On a considéré le tableau des distances routières de 36 villes de France (source : Guide Michelin 1978) et on effectue les trois traitements suivants :

1) Image euclidienne obtenue à partir du tableau des distances complet (figure 1).

2) On sélectionne les villes à plus fort bras de levier, soit, dans l'ordre du choix, Nice, Brest, Bayonne, Metz, Perpignan, Calais et Cherbourg, pour obtenir l'image euclidienne à partir de ces seules villes (figure 2).

La figure 3 montre la décroissance des bras de levier  $P_s(i)$ , à la fois entre les différentes villes au cours d'une étape et d'une étape à l'autre, ce qui nous a amené à sélectionner les sept premières villes.

3) On tire au hasard 4 villes (ici Cherbourg, Limoges, Marseille et Saint-Etienne). On obtient une image euclidienne à partir de ces quatre villes (figure 4).

On remarque que ces quatre villes sont en réalité voisines d'un même axe géographique et que la carte obtenue est très déformée. Néanmoins, elle permet de déterminer les villes situées aux sommets de l'"hexagone".

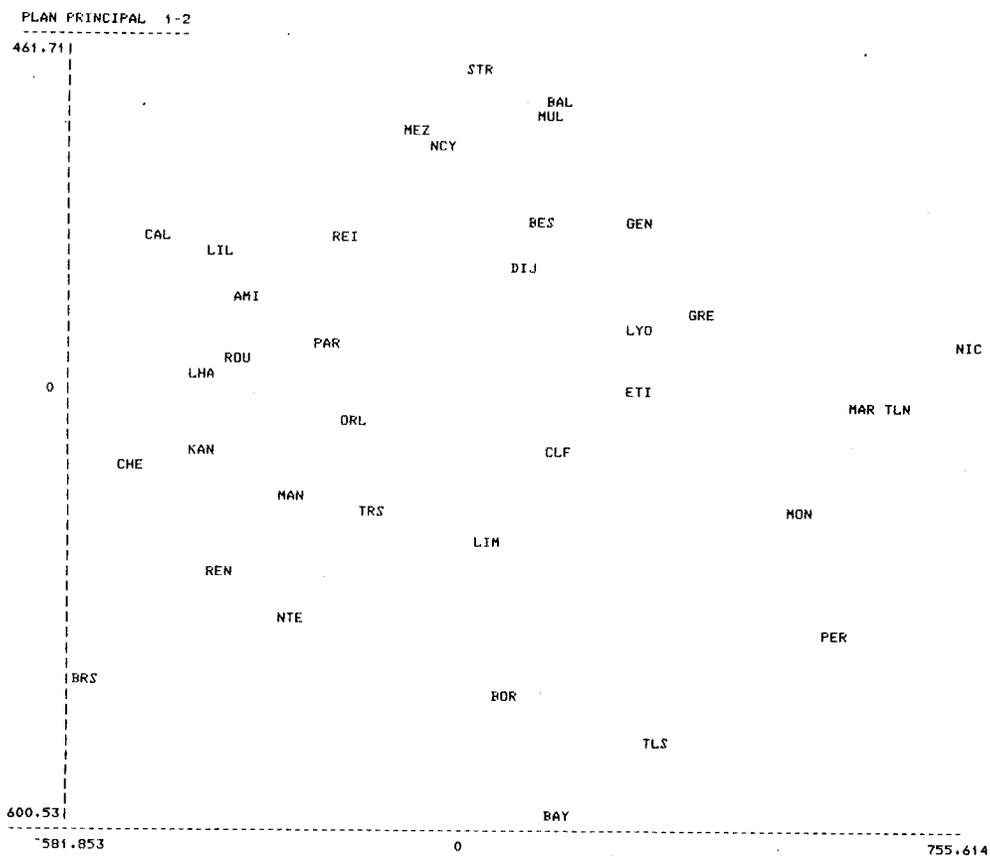


Figure 1 – Image euclidienne obtenue à partir du tableau des distances routières de 36 villes.



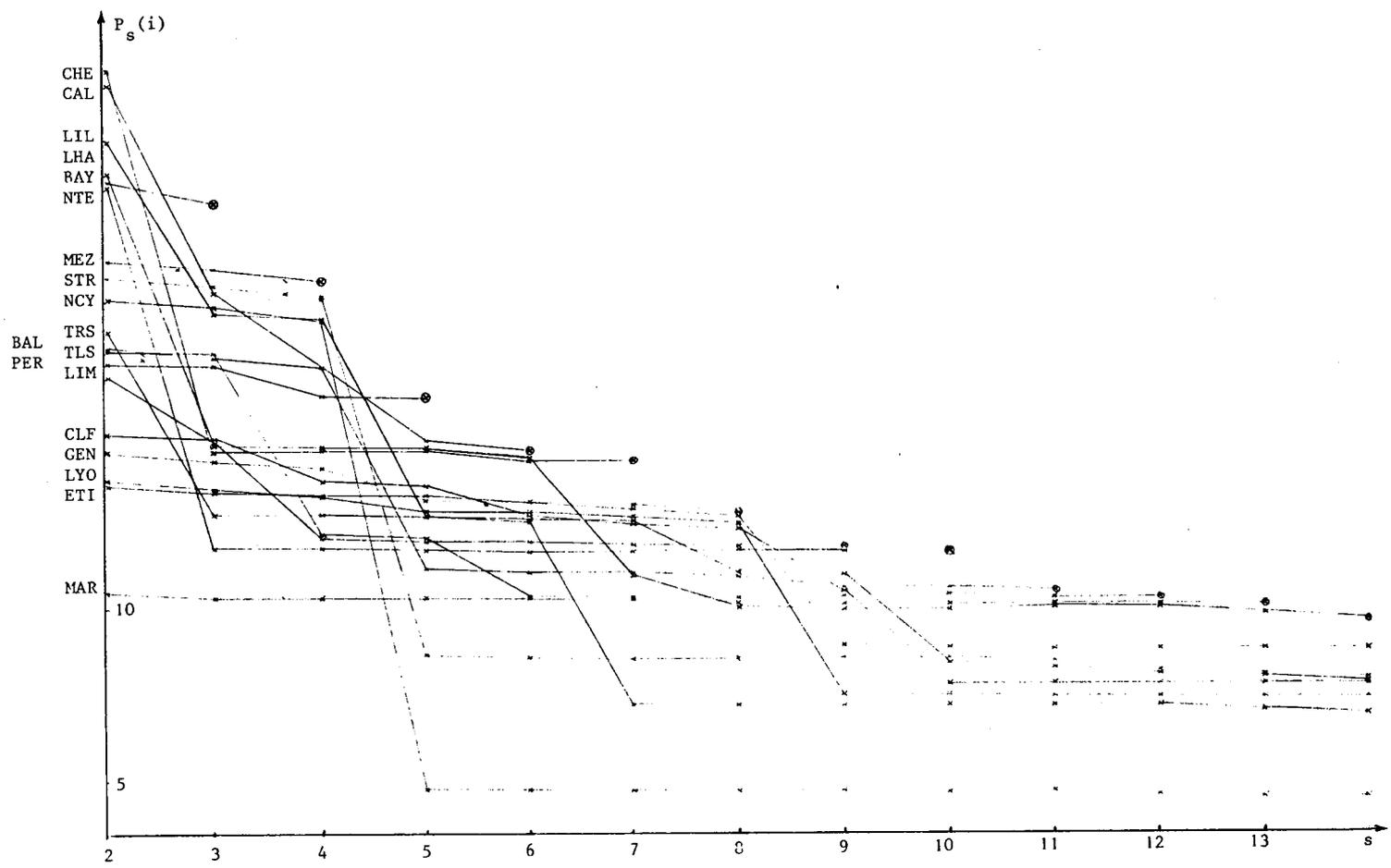


Figure 3 – Variations des bras de levier  $P_s(i)$  de 21 villes parmi les 36, en fonction du nombre  $s$  villes sélectionnées (après sélection de Brest et Nice).

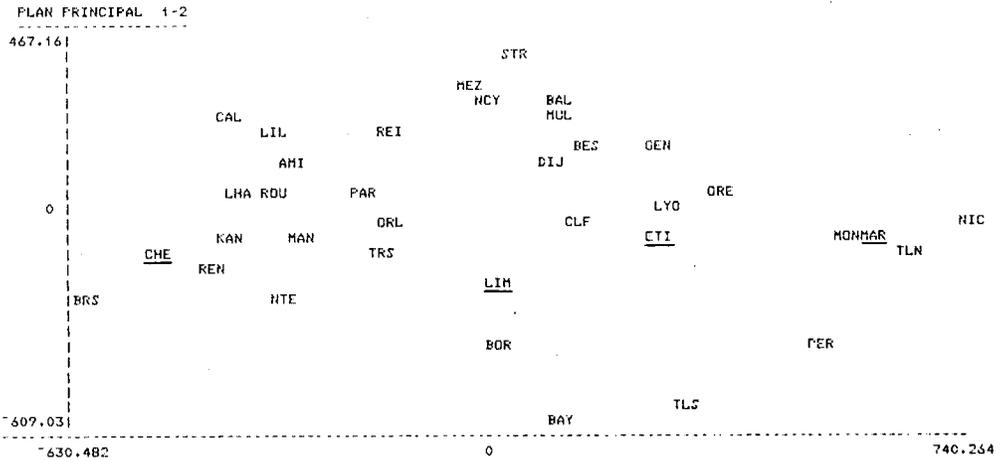


Figure 4 – Image euclidienne obtenue à partir de 4 villes sélectionnées au hasard.

On choisit sur la figure 4, six villes (soit Bayonne, Brest, Calais, Nice, Perpignan et Strasbourg). On obtient la “carte de France des distances routières” sur la figure 5.

En ne choisissant que quatre villes sur la figure 4 pour effectuer la deuxième étape (soit Bayonne, Brest, Nice et Strasbourg), on obtient une carte voisine de la précédente (figure 6).

On constate que, du point de vue des résultats, les quatre “cartes” sont équivalentes. Pour ce faire, on peut observer le critère global de qualité de la représentation, c’est-à-dire l’inertie reconstruite  $\sum \{p_i p_j d_{ij}^2 \mid i = 1, \dots, n; j = 1, \dots, n\}$  par rapport à l’inertie à reconstruire  $\text{tr}(W \circ D_v)$  en tenant compte de l’existence de valeur(s) propre(s) négative(s). On peut aussi calculer les erreurs commises sur chaque segment  $(i, j)$  pour juger de la qualité de la représentation de chaque point.

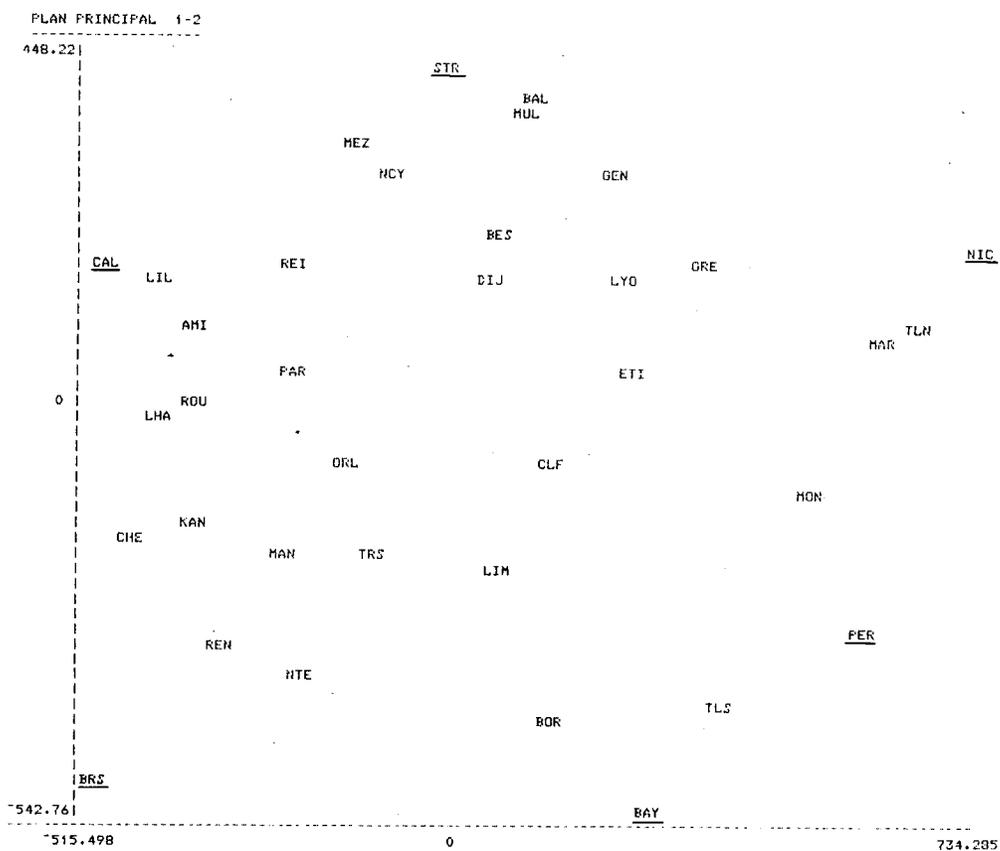


Figure 5 – Image euclidienne obtenue à partir des 6 villes les plus excentrées choisies sur la figure 4.

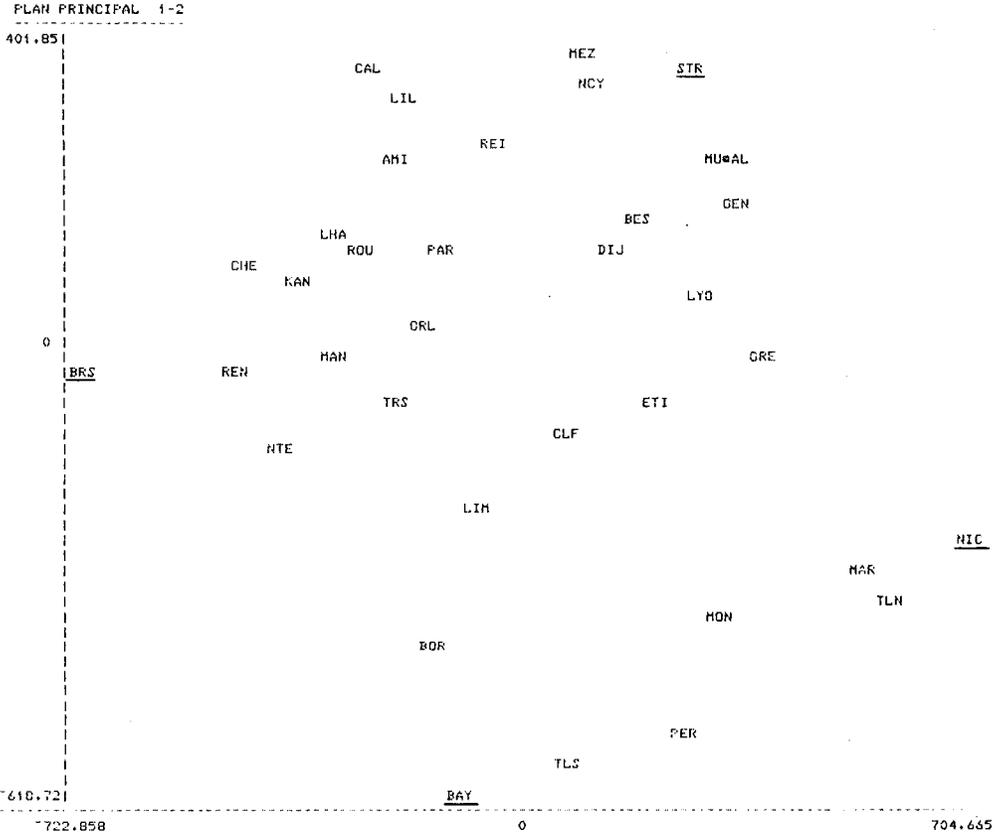


Figure 6 – Image euclidienne obtenue à partir des 4 villes les plus excentrées choisies sur la figure 4.

### 3. CAS PARTICULIERS

#### 3.1. Analyse factorielle sur tableau de distances entre modalités de variables qualitatives. Liaison avec l'analyse factorielle des correspondances.

Soit  $I$  un ensemble de  $n$  éléments munis de poids  $p_i$  ; F. Cailliez [4] a introduit la distance suivante entre parties.

Si  $A$  et  $B$  sont deux parties de  $I$ , on pose

$$d^2(A, B) = \frac{p(A \Delta B)}{p(A) p(B)}$$

où :

- $p(A) = \sum \{p_i \mid i \in A\}$
- $A \Delta B$  est la différence symétrique de  $A$  et  $B$ .

$d$ , qui est bien une distance entre parties, permet de définir une distance entre relations binaires, considérées comme des applications de  $I$  dans  $\mathcal{P}(I)$ . Si  $R_1$  et  $R_2$  sont deux relations binaires définies sur  $I$ , on pose :

$$d^2(R_1, R_2) = \sum \{p_i d^2(C_1(i), C_2(i)) \mid i \in I\}$$

où  $C_1(i)$  (resp.  $C_2(i)$ ) est la classe associée à l'élément  $i$  dans la relation  $R_1$  (resp.  $R_2$ ).

#### Proposition :

Si  $R_1$  et  $R_2$  sont des relations d'équivalence à  $p$  et  $q$  classes respectivement, alors :

$$d^2(R_1, R_2) = p + q - 2(\Phi^2 + 1)$$

où  $\Phi^2 = \frac{\chi^2}{n}$  est le "phi-deux" correspondant au tableau de contingence décrivant l'imbrication entre les deux partitions associées à  $R_1$  et  $R_2$ .

Preuve :

$$d^2(R_1, R_2) = \Sigma \left\{ p_{kj} \left( \frac{p_{k.} + p_{.j} - 2p_{kj}}{p_{k.} p_{.j}} \right) \mid k = 1, \dots, p ; j = 1, \dots, q \right\}$$

où  $k$  et  $j$  décrivent l'ensemble des modalités de  $R_1$  et  $R_2$ .

Donc :

$$\begin{aligned} d^2(R_1, R_2) &= \Sigma \left\{ p_{kj} \left( \frac{1}{p_{k.}} + \frac{1}{p_{.j}} - 2 \frac{p_{kj}}{p_{k.} p_{.j}} \right) \mid \forall (k, j) \right\} \\ &= \Sigma \left\{ \left( p_j^k + p_k^j - 2 \frac{p_{kj}^2}{p_{k.} p_{.j}} \right) \mid \forall (k, j) \right\} \\ &= p + q - 2 \Sigma \left\{ \frac{p_{kj}^2}{p_{k.} p_{.j}} \mid k = 1, \dots, p ; j = 1, \dots, q \right\} \end{aligned}$$

soit

$$d^2(R_1, R_2) = p + q - 2(\Phi^2 + 1) \quad \text{c.q.f.d.}$$

F. Cailliez a alors appelé cette distance, **distance du khi-deux entre parties**. Nous allons l'utiliser pour mesurer les proximités entre toutes les modalités de deux caractères qualitatifs  $x$  et  $y$ .

Si on désigne par  $d_{kk'}$  la distance entre les modalités  $k$  et  $k'$  du caractère  $x$ ,  $d_{jj'}$  la distance entre les modalités  $j$  et  $j'$  du caractère  $y$ , et  $d_{kj}$  la distance entre les modalités  $k$  et  $j$  des caractères  $x$  et  $y$  respectivement, on trouve :

$$d_{kk'}^2 = \begin{cases} \frac{p_{k.} + p_{k'.} - 1}{p_{k.} p_{k'.}} = \frac{1}{p_{k.}} + \frac{1}{p_{k'.}} & \text{si } k \neq k' \\ & \text{modalités de la} \\ & \text{même variable} \\ 0 & \text{sinon} \end{cases}$$

et 
$$d_{kj}^2 = \frac{p_{k.} + p_{.j} - 2p_{kj}}{p_{k.} p_{.j}} = \frac{1}{p_{k.}} + \frac{1}{p_{.j}} - 2 \frac{p_{kj}}{p_{k.} p_{.j}}$$

si  $k$  est modalité de  $x$  et  $j$  modalité de  $y$ .

On remarque d'après ce qui précède, que

$$\Sigma \{ p_{kj} d_{kj}^2 \mid k = 1, \dots, p ; j = 1, \dots, q \} = d^2(R_1, R_2) = p + q - 2(\Phi^2 + 1).$$

**Théorème 5 :**

Faire l'analyse factorielle sur le tableau à  $(p + q)$  lignes et colonnes des distances  $d_{kj}^2$ , les éléments étant munis des poids  $\frac{p_{k.}}{2}$  et  $\frac{p_{.j}}{2}$ , est équivalent à faire une analyse factorielle des correspondances sur le tableau  $Z = \begin{bmatrix} X \\ Y \end{bmatrix}$  à  $(p + q)$  lignes et  $n$  colonnes, constitué par les deux matrices accolées des indicatrices associées à  $x$  et  $y$ , considéré comme un tableau de contingence.

**Démonstration :**

$$X = (x_i^k) \quad Y = (y_i^j)$$

avec

$$x_i^k = \begin{cases} 1 & \text{si l'individu } i \text{ prend la } k^{\text{ème}} \text{ modalité de } x \\ 0 & \text{sinon} \end{cases}$$

$$y_i^j = \begin{cases} 1 & \text{si l'individu } i \text{ prend la } j^{\text{ème}} \text{ modalité de } y \\ 0 & \text{sinon} \end{cases}$$

La matrice  $P$  des probabilités issues de la matrice  $Z = \begin{bmatrix} X \\ Y \end{bmatrix}$  est donc  $P = \frac{1}{2n} Z$  où  $n$  est le nombre des individus.

On en déduit le tableau des lois conditionnelles associées aux colonnes de  $Z$  :

$$B_2' = \frac{1}{2} Z \quad B_1 = \frac{1}{n} M Z$$

avec :

$$M = 2 \left[ \begin{array}{c|c} \frac{1}{p_{k.}} & 0 \\ \hline 0 & \frac{1}{p_{.j}} \end{array} \right]$$

L'analyse factorielle des correspondances du tableau  $Z$  revient à faire une analyse en composantes principales avec le schéma de dualité suivant :

$$\begin{array}{ccc}
 F = \mathbf{R}^{p+q} & \xleftrightarrow{B'_2} & F = \mathbf{R}^n \\
 \updownarrow M & \begin{array}{c} B'_1 \\ D_{1/p} \end{array} & \updownarrow D_p = \frac{1}{n} I_n \\
 E^* & \xleftrightarrow{B_2} & F^* \\
 & B_1 &
 \end{array}$$

Une telle analyse revient donc à chercher les composantes principales du tableau de distances entre "caractères-modalités", qui sont données par :

$$d_{\alpha\beta}^2 = D_{1/p} (\underline{b}_\alpha - \underline{b}_\beta) = \underline{b}'_\alpha D_{1/p} \underline{b}_\alpha + \underline{b}'_\beta D_{1/p} \underline{b}_\beta - 2 \underline{b}'_\alpha D_{1/p} \underline{b}_\beta$$

donc :

$$\bullet \text{ si } \underline{b}_\alpha = \frac{1}{n p_k} x^k \quad \text{et} \quad \underline{b}_\beta = \frac{1}{n p_{k'}} x^{k'}$$

$$d_{kk'}^2 = \begin{cases} \frac{1}{p_k} + \frac{1}{p_{k'}} & \text{si } k = k' \\ 0 & \text{sinon} \end{cases}$$

$$\bullet \text{ si } \underline{b}_\alpha = \frac{1}{p_k} x^k \quad \text{et} \quad \underline{b}_\beta = \frac{1}{p_j} y^j$$

$$d_{kj}^2 = \frac{1}{p_k} + \frac{1}{p_j} + \frac{2 p_{kj}}{p_k p_j} \quad \text{c.q.f.d.}$$

### 3.2. Opérateurs d'Escoufier

La donnée d'un tableau "individus  $\times$  caractères"  $X$  allant de pair avec celle des métriques  $M$  et  $D_p$  permettant de calculer les distances entre individus et entre caractères respectivement, il est logique, en analyse factorielle, de parler non pas du tableau  $X$  mais du triplet  $(X, M, D_p)$ .

Au triplet  $(X, M, D_p)$ , auquel est associé le schéma de dualité du chapitre I, Y. Escoufier [12] a associé l'opérateur

$$U = W \circ D_p .$$

Cet opérateur n'est pas caractéristique d'un triplet donné  $(X, M, D_p)$ , de nombreux triplets pouvant conduire au même opérateur, par contre il est caractéristique de la paire  $(D, D_p)$ , où  $D$  est le tableau des distances entre individus au sens de la métrique  $M$ .

Quelle que soit l'optique dans laquelle on se place (analyse en composantes principales ou analyse canonique), comparer des tableaux revient à comparer des triplets  $(X, M, D_p)$  et donc à comparer des opérateurs :

– *optique analyse en composantes principales* : deux triplets conduisent à la même analyse en composantes principales si les deux opérateurs associés coïncident ;

– *optique analyse canonique* : deux ensembles de caractères (tableaux  $X$  et  $Y$ ) engendrent le même sous-espace si les opérateurs associés aux triplets  $(X, V_x^{-1}, D_p)$  et  $(Y, V_y^{-1}, D_p)$  coïncident,  $V_x$  et  $V_y$  désignant respectivement les matrices de variance-covariance associées à  $X$  et  $Y$  ; les opérateurs associés à ces triplets ne sont autres, en effet, que les projecteurs associés aux deux sous-espaces correspondant à  $X$  et  $Y$  respectivement (d'où l'intérêt de munir l'ensemble des opérateurs d'une métrique).

Les opérateurs de type  $W \circ D_p$  ( $W$  est semi-définie positive) sont des opérateurs  $D_p$ -symétriques, ces derniers engendrent un sous-espace de l'ensemble des applications linéaires du vectoriel  $F = \mathbf{R}^n$  dans  $F$ . Y. Escoufier a muni ce sous-espace vectoriel  $G$  d'un produit scalaire  $P$  défini à partir de la trace (tr) :

$$U \in G, V \in G : P(U, V) = \text{tr}(UV).$$

Remarquons que ce produit scalaire permet de retrouver les indices classiques utilisés par les statisticiens pour mesurer la liaison entre caractères ou paquets de caractères.

Si  $U$  et  $V$  sont, en effet, les opérateurs associés aux triplets  $(X, M, D_p)$  et  $(Y, N, D_p)$ , en utilisant les notations définies au chapitre I :

– si  $M$  et  $N$  sont les *métriques identité*,

$$P(U, V) = \Sigma \{ \text{cov}^2(\underline{x}^k, \underline{y}^l) \mid (k, l) \}$$

où  $\text{cov}(\underline{x}^h, \underline{y}^l)$  désigne la covariance entre  $\underline{x}^k$  et  $\underline{y}^l$ ,

– si  $M$  et  $N$  sont les *métriques*  $D_{1/\sigma^2}$  des inverses des carrés des écarts-type

$$P(U, V) = \Sigma \{ \text{cor}^2(\underline{x}^k, \underline{y}^l) \mid (k, l) \}$$

où  $\text{cor}(\underline{x}^k, \underline{y}^l)$  désigne le coefficient de corrélation linéaire,

– si  $M$  et  $N$  sont les *métriques de Mahalanobis*, c'est-à-dire :  $M = V_x^{-1}$ ,  
et  $N = V_y^{-1}$ ,

$$P(U, V) = \Sigma \{ \rho_k^2 \mid k \}$$

où  $\rho_k$  désigne le  $k^{\text{ème}}$  coefficient de corrélation canonique entre les deux paquets de variables,

– si les tableaux  $X$  et  $Y$  sont les tableaux disjonctifs associés respectivement aux variables qualitatives  $x$  et  $y$ , et si  $M$  et  $N$  sont les *métriques du khi-deux* correspondantes,

$$P(U, V) = \Phi^2 + 1 \quad (\text{cf. } \S 3.1.).$$

Grâce au produit scalaire  $P$ , on peut mesurer les proximités entre opérateurs et donc effectuer des analyses factorielles sur tableaux de distances entre opérateurs pour décrire les proximités entre tableaux de données (cf. par exemple [48]).

### CHAPITRE III

## ANALYSE CANONIQUE SUR TABLEAUX DE DISTANCES

La perception que l'on a d'un ensemble de  $n$  objets peut être dans certains cas décrite directement, c'est-à-dire sans s'appuyer sur des variables définies *a priori*, à l'aide d'un indice (dissimilarité, distance, . . .) précisant les proximités relatives entre ces objets.

Il est alors naturel de se poser les questions suivantes :

*Les proximités entre objets telles qu'elles sont énoncées par celui qui juge, peuvent-elles être considérées comme résultant à la fois d'une sélection implicite de critères (variables) et d'une métrique euclidienne ?*

L'analyse factorielle sur tableau de distances, qui opère comme l'analyse en composantes principales, permet de répondre à cette question en appliquant la métrique "identité" sur les composantes principales fournies par l'analyse factorielle ; le tableau de distances est reconstruit rigoureusement si les valeurs propres obtenues dans l'analyse sont positives ou nulles.

*Les critères précédents sont-ils des combinaisons linéaires de variables données a priori ?*

Un problème du même type a été abordé dans Sandkya par C.R. Rao en 1964 [37] : l'analyse en composantes principales sous contraintes linéaires [41] (les composantes principales doivent appartenir au sous-espace engendré par les variables données *a priori*) permet d'obtenir les combinaisons linéaires des variables données *a priori*, reconstruisant au mieux les proximités.

Un changement de métrique euclidienne revenant à un changement de base, cette question est équivalente à la suivante.

*Les proximités peuvent-elles être considérées comme construites à l'aide d'une métrique euclidienne à partir d'un ensemble de variables données ?*

La métrique euclidienne permettant d'ajuster au mieux les proximités est obtenue par régression.

*Les proximités entre objets, telles qu'elles sont énoncées par différents juges, résultent-elles d'un même ensemble de variables sur lesquelles les juges utiliseraient des métriques euclidiennes différentes ?*

Ce problème a été abordé par J.D. Carroll dans ses modèles INDSCAL et IDIOSCAL.

*Les proximités sont-elles compatibles avec une partition des objets, donnée a priori ?*

Ce problème, en réalité du même type que les précédents, conduit à envisager une nouvelle technique qui sera abordée à la fin de ce paragraphe.

Ces questions seront traitées après avoir introduit des équivalences entre tableaux de distances, entre tableaux de distances et tableau "individus  $\times$  caractères", et entre tableaux de distances et variables qualitatives (partition).

## 1. EQUIVALENCES UTILES DANS L'ANALYSE DES TABLEAUX DE DISTANCES

Deux types d'équivalences peuvent être introduits dans l'ensemble des tableaux de distances définis sur les mêmes objets : le premier emprunte à l'analyse en composantes principales, le second à l'analyse canonique.

**Equivalence 1 :**

$D_1$  est équivalent à  $D_2$  si et seulement si les opérateurs  $U_1 = W_1 \circ D_p$  et  $U_2 = W_2 \circ D_p$  associés respectivement à  $D_1$  et  $D_2$ . coïncident ou sont homothétiques.

L'analyse factorielle appliquée sur les deux tableaux équivalents fournit des axes principaux identiques et des parts d'inertie identiques.

**Equivalence 2 :**

$D_1$  est équivalent à  $D_2$  si et seulement si il existe un tableau  $X$  ( $n \times p$ ) et des métriques  $M_1$  et  $M_2$  de dimension  $p \times p$  tels que, si

$$W_1 = X' M_1 X \quad \text{et} \quad W_2 = X' M_2 X,$$

$W_1 \circ D_p$  et  $W_2 \circ D_p$  coïncident avec les opérateurs associés à  $D_1$  et  $D_2$ .

Cette équivalence, moins contraignante que la précédente, exprime que les dimensions sous-jacentes aux tableaux  $D_1$  et  $D_2$  (les critères utilisés par les juges considérés) engendrent des sous-espaces de  $\mathbf{R}^n$  identiques. Les composantes principales, vecteurs propres de  $W_1 \circ D_p$  et de  $W_2 \circ D_p$ , si elle ne coïncident pas ici, engendrent les mêmes sous-espaces.

Des deux équivalences précédentes, on peut déduire des équivalences <sup>(1)</sup> entre une paire  $(D, D_p)$  et un triplet  $(X, M, D_p)$ .

#### Equivalence 1' :

*La paire  $(D, D_p)$  et le triplet  $(X, M, D_p)$  sont équivalents si et seulement si les opérateurs associés coïncident ou sont homothétiques.*

#### Equivalence 2' :

*La paire  $(D, D_p)$  et le triplet  $(X, M, D_p)$  sont équivalents si et seulement si les composantes principales obtenues dans l'analyse factorielle de la paire  $(D, D_p)$  (vecteurs de l'opérateur correspondant, associés à des valeurs propres non nulles) engendrent le même sous-espace que les caractères définis par les lignes du tableau  $X$ .*

Toute variable qualitative peut être représentée dans  $\mathbf{R}^n$  par le sous-espace engendré par les variables indicatrices de ses modalités ; d'où l'équivalence 2'' <sup>(1)</sup> :

#### Equivalence 2'' :

*La paire  $(D, D_p)$  est équivalente à la variable qualitative  $x$  si et seulement si le sous-espace engendré par les indicatrices de  $x$  coïncide avec le sous-espace engendré par les composantes principales obtenues dans l'analyse factorielle de la paire  $(D, D_p)$ .*

---

(1) En toute rigueur, le terme *équivalence* ne devrait pas être utilisé ici, les êtres mathématiques considérés n'appartenant pas au même ensemble.

## 2. AJUSTEMENT D'UN TABLEAU DE DONNEES A UN TABLEAU DE DISTANCES

Le problème qui se pose est le suivant :

*Etant donné un tableau de distances  $D$  entre  $n$  objets et un tableau individus  $\times$  caractères  $X$ , existe-t-il une métrique euclidienne  $M$  telle que le tableau des distances  $Y$  construit avec la métrique  $M$  à partir du tableau  $X$  soit équivalent au tableau initial  $D$  au sens de l'équivalence 1'?*

Pour résoudre ce problème, on va chercher la métrique  $M$  telle que les opérateurs d'Escoufier associés respectivement à la paire  $(D, D_p)$  et au triplet  $(X, M, D_p)$  soient les plus proches possibles pour la métrique  $P$  entre opérateurs définie à partir de la trace (cf. chap. II, 3.2.).

Soit :

$$\bullet U = W D_p = -\frac{1}{2} A D^2 A' D_p$$

où  $D^2$  est le tableau des carrés des distances, et  $A$  le  $D_p$ -projecteur sur l'hyperplan  $D_p$ -orthogonal à la droite des constantes  $\Delta_j$ ,

$$\bullet \bar{U} = \bar{W} D_p = X' M X D_p.$$

**Théorème 1 :**

*La métrique  $M$  rendant minimum la quantité  $\|U - \bar{U}\|_P$  est définie par :*

$$M = (X D_p X')^{-1} X D_p W D_p X' (X D_p X')^{-1}.$$

En effet :

$$\begin{aligned} \|U - \bar{U}\|_P^2 &= P(U - \bar{U}, U - \bar{U}) \\ &= \text{tr}(U^2 - 2U\bar{U} + \bar{U}^2) \end{aligned}$$

$$\text{Min}_M \|U - \bar{U}\|_P^2 = \text{Min}_M [\text{tr}(\bar{U}^2 - 2U\bar{U})] + \text{tr}(U^2)$$

puisque  $U$  est indépendant de  $M$ .

On doit donc minimiser  $E = \text{tr}(\bar{U}^2 - 2U\bar{U})$ .

$E = \text{tr}(X' M X D_p X' M X D_p) - 2 \text{tr}(W D_p X' M X D_p)$  puisque l'opérateur trace est linéaire.

$$E = \text{tr}(X D_p X' M X D_p X' M) - 2 \text{tr}(W D_p X' M X D_p)$$

car  $\text{tr}(AB) = \text{tr}(BA)$ .

On est donc amené à résoudre l'équation :

$$\frac{\partial E}{\partial M} = 0.$$

*Rappel :*

Soit  $f$  une fonction réelle de  $pq$  variables  $m_{ij}$  réelles. En convenant de noter  $\frac{\partial f(M)}{\partial M}$  le tableau  $(p, q)$  des dérivées partielles  $\frac{\partial f(M)}{\partial m_{ij}}$  ( $i = 1, 2, \dots, p$  et  $j = 1, 2, \dots, q$ ), on a les résultats suivants :

- $\frac{\partial \text{tr} K}{\partial M} = 0_{(p, q)}$  si  $K$  est indépendant de  $M$
- $\frac{\partial \text{tr}(AMB)}{\partial M} = (BA)'$
- $\frac{\partial \text{tr}(AMBM')}{\partial M} = AMB + B'MA'$

Ces résultats, valables à condition que les matrices aient des formats compatibles, s'obtiennent par simple écriture des définitions.

D'où :

$$\frac{\partial E}{\partial M} = 2 X D_p X' M X D_p X' - 2 X D_p W D_p X'.$$

La métrique  $M$  est donc solution de l'équation :

$$X D_p X' M X D_p X' = X D_p W D_p X'.$$

soit, si  $X$  est de plein rang :

$$M = (X D_p X')^{-1} X D_p W D_p X' (X D_p X')^{-1}$$

On peut remarquer que, si  $X'$  est injective,  $M$  définit un écart euclidien si et seulement si la matrice  $W$  est semi-définie positive.

La formule précédente permet d'ajuster non seulement un tableau de données à un tableau de distances, mais aussi un tableau de données à un triplet  $(Y, N, D_p)$ .

En effet, ajuster  $X$  au triplet  $(Y, N, D_p)$  revient à rechercher la métrique  $M$  telle que les opérateurs  $X' M X D_p$  et  $Y' N Y D_p$  soient les plus proches possible ; d'après ce qui précède, la métrique  $M$  vérifie :

$$M = (X D_p X')^{-1} X D_p Y' N Y D_p X' (X D_p X')^{-1}$$

d'où 
$$M = V_{11}^{-1} V_{12} N V_{21} V_{11}^{-1}$$

avec 
$$V_{11} = X D_p X' \quad \text{et} \quad V_{12} = X D_p Y' = V_{21} .$$

On en déduit donc :

### ***Théorème 2 :***

*Si les sous-espaces vectoriels engendrés par les lignes de  $X$  et par celles de  $Y$  coïncident, quelle que soit la métrique  $N$  choisie pour mesurer les proximités entre les individus à l'aide des caractères de  $Y$ , il existe une métrique  $M$  qui permet de calculer les mêmes distances à l'aide des caractères de  $X$ .*

## **3. COMPARAISON ENTRE TABLEAUX DE DISTANCES**

J.D. Carroll a abordé le problème suivant :

*Si à chacun des  $k$  individus (juges) considérés, est associé un tableau de proximités  $D_i$  entre  $n$  objets, existe-t-il un tableau objets  $\times$  caractères  $X$  ( $n \times p$ ) et  $k$  métriques euclidiennes  $M_1, M_2, \dots, M_k$  tels que les opérateurs associés aux triplets  $(X, M_1, D_p), (X, M_2, D_p), \dots, (X, M_k, D_p)$  coïncident avec les opérateurs associés aux paires  $(D_1, D_p), (D_2, D_p), \dots, (D_k, D_p)$  ?*

J.D. Carroll a proposé deux algorithmes pour résoudre ce problème. Dans le premier, INDSCAL (INDividual SCALing) [6], il impose aux métriques  $M_i$  d'être diagonales ; dans le second, IDIOSCAL [7], ces métriques peuvent être quelconques ; les algorithmes mis au point utilisent la méthode itérative NIPALS introduite par H. Wold [52], [7].

A la recherche d'une solution directe, J.D. Carroll fait la remarque que, si l'on note  $U$  la moyenne des opérateurs  $W_i \circ D_p$  associés aux triplets  $(X, M_i, D_p)$ , cet opérateur s'écrit :

$$U = \frac{1}{k} \Sigma \left\{ W_i D_p \mid i = 1, \dots, k \right\} = \frac{1}{k} \Sigma \left\{ X' M_i X D_p \mid i = 1, \dots, k \right\}$$

$$U = X' \left( \frac{1}{k} \Sigma \left\{ M_i \mid i = 1, \dots, k \right\} \right) X D_p$$

Si  $C$  désigne le tableau des composantes principales associées au triplet  $(X, M, D_p)$ , avec  $M = (1/k) \Sigma M_i$ , on a

$$U = X' M X D_p = C' C D_p$$

Tout système de caractères engendrant le même sous-espace que les lignes de  $X$  pouvant jouer le rôle de  $X$ , on choisit pour  $X$  le tableau  $C$  (on ne retient que les vecteurs propres de  $U$  associés à des valeurs propres positives).

Si la matrice de changement de base associée au système des axes principaux est notée  $T$  ( $C = TX$ ), on a alors, pour chaque juge  $i$ , (cf. [4]) :

$$W_i = X' M_i X = C' N_i C$$

avec

$$N_i = (T^{-1})' M_i T^{-1}$$

Ayant retenu les vecteurs propres associés aux valeurs propres positives de l'opérateur moyen  $U$ , pour retrouver les métriques  $N_i$ , il suffit d'appliquer la formule démontrée au paragraphe précédent.

On peut vérifier que la moyenne des métriques  $N_i$  respecte bien la contrainte :

$$\frac{1}{k} \Sigma \{ N_i \mid i = 1, \dots, k \} = I$$

En effet :

$$\begin{aligned} \frac{1}{k} \Sigma N_i &= (C D_p C')^{-1} C D_p \frac{1}{k} \Sigma W_i D_p C' (C D_p C')^{-1} \\ &= (C D_p C')^{-1} C D_p U C' (C D_p C')^{-1} \\ &= (C D_p C')^{-1} C D_p C' C D_p C' (C D_p C')^{-1} = I \end{aligned}$$

Remarquons que, si il existe effectivement un tableau  $X$  respectant la condition de Carroll, les tableaux  $D_1, D_2, \dots, D_k$  sont équivalents au sens de l'équivalence  $2'$  introduite précédemment ; cette équivalence, rappelons-le, est issue de l'analyse canonique.

Si les sous-espaces vectoriels engendrés par les vecteurs propres des opérateurs associés aux tableaux  $D_i$  ne sont pas confondus, c'est-à-dire si ces tableaux ne sont pas "canoniquement" équivalents, pour trouver le sous-espace (le tableau X) représentant aux mieux ces sous-espaces, il ne nous semble pas approprié d'opérer sur l'opérateur moyen  $U$ . On sait en effet, et J.D. Carroll le sait aussi [5], [26], qu'une façon d'étudier les positions relatives de  $k$  sous-espaces vectoriels (analyse canonique généralisée) consiste à diagonaliser la somme  $\Sigma A_i$  des projecteurs associés à chacun de ces sous-espaces.

Nous allons développer des propositions dans ce sens, propositions qui restent dans la logique de J.D. Carroll. Une amélioration de cette nouvelle procédure utilisant des idées récentes d'Y. Escoufier, développées successivement par L'Herminier des Plantes [30], puis par P. Cazes, S. Bonnefous et coll. [9] sera ensuite décrite.

### 3.1. Résolution du problème de Carroll par l'analyse canonique

On sait [8] qu'il y a deux façons d'aborder l'analyse des positions relatives de deux sous-espaces vectoriels.

La première, introduite par H. Hotelling [20] en 1936, consiste à rechercher les couples successifs, orthogonaux entre eux, de variables ( $\underline{\xi}$ ,  $\underline{\eta}$ ), dites variables canoniques, les plus corréllées, où  $\underline{\xi}$  et  $\underline{\eta}$  appartiennent respectivement au premier et au deuxième sous-espace.

La seconde consiste à rechercher les vecteurs  $\underline{\beta}$  successifs, orthogonaux entre eux, équidistants des deux sous-espaces considérés et les plus proches de ces sous-espaces.

La première méthode conduit, si  $A_1$  et  $A_2$  désignent les projecteurs sur les sous-espaces vectoriels considérés, aux vecteurs  $\underline{\xi}$  et  $\underline{\eta}$ , solutions des équations :

$$A_1 \circ A_2 \underline{\xi} = \lambda \underline{\xi}$$

$$\underline{\eta} = \frac{1}{\sqrt{\lambda}} A_2 \underline{\xi}$$

$$\|\underline{\xi}\| = \|\underline{\eta}\| = 1.$$

La deuxième méthode conduit aux équations :

$$(A_1 + A_2)\underline{\beta} = \mu \underline{\beta}$$

$$\|\underline{\beta}\| = 1.$$

Si  $\lambda$  est égal à 1, on a :

$$\mu = 2 \quad \text{et} \quad \underline{\beta} = \underline{\xi} = \underline{\eta}.$$

Si  $\lambda$  est strictement compris entre 0 et 1, on a :

$$\mu = 1 \pm \sqrt{\lambda} \quad \text{et} \quad \underline{\beta} = \underline{\xi} \pm \underline{\eta}.$$

Les vecteurs du premier sous-espace (respectivement du deuxième sous-espace), orthogonaux à tous les vecteurs propres  $\underline{\xi}$  (respectivement  $\underline{\eta}$ ) associés à des valeurs propres  $\lambda$  différentes de 0 sont des vecteurs  $\underline{\beta}$  de valeur propre  $\mu$  égale à 1.

La deuxième méthode a l'avantage de se généraliser immédiatement : si on a à préciser les positions relatives de  $k$  sous-espaces vectoriels, pourquoi ne pas diagonaliser l'opérateur  $\Sigma A_i$  ?

Cette méthode a été proposée en particulier par J.D. Carroll [5].

Il est intéressant de noter que, lorsque l'on procède ainsi en diagonalisant l'opérateur  $\Sigma A_i$ , on est beaucoup plus dans l'optique de l'analyse en composantes principales que dans celle de l'analyse canonique.

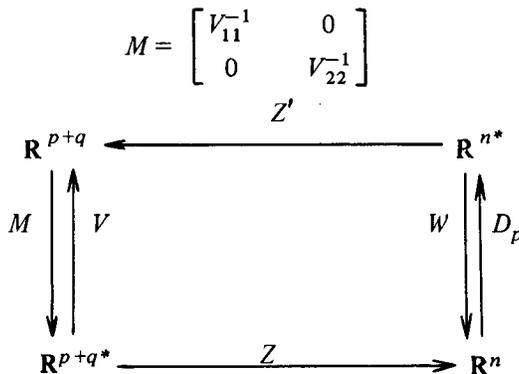
Plaçons-nous, en effet, dans le cas où les sous-espaces considérés sont engendrés par deux paquets de caractères quantitatifs (tableaux  $X$  et  $Y$ ) et notons :

$$Z = \begin{bmatrix} X \\ Y \end{bmatrix} \text{ le tableau des données, de format } (n, p + q),$$

$$V = Z D_p Z' = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \text{ la matrice de variance associée à } Z \text{ avec :}$$

$$V_{11} = X D_p X', \quad V_{22} = Y D_p Y' \text{ et } V_{12} = X D_p Y'.$$

Considérons le schéma de dualité associé au triplet  $(Z, M, D_p)$  où  $M$  est la métrique de Mahalanobis généralisée, de matrice associée :



L'opérateur  $W \circ D_p$ , associé au triplet  $(Z, M, D_p)$  n'est autre que l'opérateur  $A_1 + A_2$ . Aussi, effectuer l'analyse en composantes principales du triplet  $(Z, M, D_p)$  revient à diagonaliser la somme  $A_1 + A_2$ .

Le résultat se généralise immédiatement. Nous appellerons donc analyse en composantes principales réduite généralisée, et non pas analyse canonique généralisée, la procédure qui consiste à diagonaliser  $\Sigma A_i$ , procédure qui généralise l'analyse en composantes principales sur variables réduites (métrique  $D_{1/\sigma^2}$ ).

Nous avons vu que le problème de Carroll conduit à considérer les sous-espaces propres engendrés par les vecteurs propres des opérateurs associés aux paires  $(D_i, D_p)$ . Il reste à indiquer comment obtenir les projecteurs  $A_i$  représentatifs de ces sous-espaces pour appliquer la procédure précédente.

Désignons par  $C_i$  le tableau des composantes principales (le spectre est noté  $\lambda(i)$ ) associées au tableau de distances  $D_i$  (on retiendra le maximum de vecteurs propres). On a :

$$A_i = C_i D_{1/\lambda(i)} C_i' D_p.$$

La diagonalisation de  $\Sigma A_i$  fournit les vecteurs  $\underline{\beta}$  ; on ne sélectionne évidemment que les vecteurs propres  $\underline{\beta}$  associés aux plus grandes valeurs propres, si une valeur propre est égale à  $k$ , le vecteur  $\underline{\beta}$  correspondant est dans l'intersection de tous les sous-espaces. En utilisant les graphiques que l'on considère en général en analyse canonique (ils sont analogues à ceux que l'on considère en analyse factorielle des correspondances et en analyse factorielle discriminante), on peut visualiser bien des faits : par exemple estimer, pour chacun des juges, dans quelle mesure les vecteurs  $\underline{\beta}$  retenus coïncident avec les dimensions qui permettent de bien reconstruire le tableau de distances. Pour cela, on projette sur les "cercles de corrélations" définis à partir des premiers vecteurs  $\underline{\beta}$ , les composantes principales de chacun des juges : plus ces composantes principales sont excentrées sur le cercle, plus il est justifié de considérer les caractères  $\underline{\beta}$  correspondants comme étant les dimensions utilisées par le juge pour donner son avis (le tableau de distances).

Ayant sélectionné les vecteurs  $\underline{\beta}$ , c'est-à-dire ayant trouvé le tableau commun  $X$ , on utilise, comme précédemment, la formule donnée au paragraphe 2 pour calculer les métriques  $M_i$ .

Si la procédure décrite ici est plus compliquée que celle de Carroll au niveau des calculs (elle est, sur le plan théorique, aussi simple), elle a l'avantage de donner une solution unique  $X$ . En effet, on considère ici qu'une variable a d'autant plus de chances d'être commune à tous les juges que le cosinus qu'elle fait avec l'ensemble des sous-espaces est grand ; le tableau  $X$  est donc composé de variables sélectionnées de façon hiérarchique par rapport à ce critère.

Si on utilise la technique de Carroll, rien ne garantit que les premiers vecteurs propres de l'opérateur moyen  $\Sigma W_i D_p$  soient plus intéressants que les suivants pour reconstruire les tableaux  $D_i$  ; ceci n'est vrai que si les métriques  $M_i$  utilisées par les différents juges sont assez proches les unes des autres.

### 3.2. Amélioration de la procédure : résolution optimale du problème de Carroll par l'analyse canonique.

Avant de rechercher le tableau "commun"  $X$ , il est bon d'avoir une idée sur la dispersion de l'avis des juges. Pour cela, on dispose de la technique des opérateurs.

Quel opérateur associer au couple  $(D_i, D_p)$  ?

Si on suit la voie empruntée par Carroll, c'est l'opérateur  $W_i \circ D_p$ , où  $W_i$  est calculé par les formules du § 1 - 2. qu'il faut associer au couple  $(D_i, D_p)$  ; par contre, si c'est notre démarche que l'on retient, il est logique d'associer au couple  $(D_i, D_p)$  l'opérateur de projection  $A_i$  sur le sous-espace vectoriel engendré par les vecteurs propres de  $W_i \circ D_p$ .

Ayant choisi son opérateur, on décrit les proximités entre les avis en effectuant l'analyse factorielle sur le tableau des distances entre opérateurs.

Si ce sont les projecteurs  $A_i$  que l'on associe aux couples  $(D_i, D_p)$ , plus que leurs distances ce sont les angles entre ces opérateurs qui sont intéressants ; en effet :

$$P(A_i, A_j) = \Sigma \{\rho_i^2 \mid i = 1, \dots, n\}$$

où  $\rho_i$  désigne le  $i^{\text{ème}}$  cosinus (corrélation) canonique entre les sous-espaces images des projecteurs.

Pour dresser un bilan de ces angles, on est conduit, par analogie avec ce qu'on fait en analyse en composantes principales pour décrire les caractères, à diagonaliser la matrice  $V$  des produits scalaires entre opérateurs considérée comme une matrice de variance<sup>(1)</sup>. Tous les termes de cette matrice sont positifs ; d'après le théorème

---

(1) On utilise les produits scalaires plutôt que les cosinus car généralement les opérateurs ont des normes comparables ; en effet, la norme d'un projecteur n'est autre que son rang, ici égal au nombre de valeurs propres positives de l'opérateur  $(D_i, D_p)$ .

de Frobenius, le vecteur propre  $\underline{\alpha}$  associé à la plus grande valeur propre, c'est-à-dire le premier facteur principal, a toutes ses coordonnées  $\alpha_i$  positives :

$$V\underline{\alpha} = \lambda_1 \underline{\alpha} ; \alpha_i \geq 0 .$$

La première composante principale correspondante  $\underline{c}$  s'écrit :

$$\underline{c} = \Sigma \alpha_i A_i .$$

Cette composante principale, renormée à l'unité, n'est autre que l'opérateur  $D_p$ -symétrique rendant maximum la quantité :

$$J_c = \Sigma \{ \cos^2 (\underline{c}, A_i) \mid i = 1, \dots, k \} [36] .$$

Au sens de l'indice  $J_c$ , l'opérateur  $\underline{c}$  peut être considéré comme le plus représentatif des  $k$  opérateurs  $A_i$ . Aussi, plutôt que de diagonaliser  $\Sigma A_i$ , il semble préférable de diagonaliser l'opérateur  $B = \Sigma \alpha_i A_i$ .

La technique consistant à diagonaliser  $B$  fournit ce que l'on appelle la résolution optimale du problème de Carroll par l'analyse canonique.

### 3.3. Bilan

De façon idéale, il faudrait procéder ainsi pour traiter le problème de Carroll :

#### 3.3.1. Description des proximités entre tableaux de distances

On effectue successivement l'analyse factorielle sur les tableaux de distances entre opérateurs  $W_i \circ D_p$  et opérateurs  $A_i$ .

Les plans principaux fournis par la première analyse nous renseignent sur les proximités entre tableaux de distances ; ceux fournis par la seconde nous indiquent dans quelle mesure les dimensions sous-jacentes aux tableaux de distances sont identiques.

#### 3.3.2. Description des cosinus entre projecteurs

On extrait les deux premiers vecteurs propres  $\underline{v}_1 = \alpha$  et  $\underline{v}_2$  de  $V$  ; l'opérateur  $B = \Sigma \alpha_i A_i$  est l'opérateur le plus proche des opérateurs  $A_i$ .

Les opérateurs  $A_i$  sont représentés par leurs coordonnées dans la base des composantes principales associées aux facteurs  $\underline{v}_1$  et  $\underline{v}_2$ . Elles sont données par (cf. [4] p. 264) :

$$(\sqrt{\lambda_1} v_1, \sqrt{\lambda_2} v_2).$$

### 3.3.3. Représentation moyenne des objets

Les coordonnées des vecteurs propres de l'opérateur moyen  $B$  permettent de construire directement une image moyenne des objets. Le tableau des vecteurs propres  $\beta$  rangés en lignes les uns sous les autres n'est autre que le tableau  $X$  cherché.

### 3.3.4. Adéquation entre $X$ et les tableaux $D_i$ .

Pour juger de l'hypothèse "le juge  $i$  a émis son jugement à partir du tableau  $X$ ", on représente les composantes principales normées de l'opérateur  $W_i \circ D_p$  dans la base orthonormée des vecteurs  $\underline{\beta}$  (les coordonnées sont des corrélations).

### 3.3.5. Calcul des métriques propres à chaque juge

La formule donnée au paragraphe 2 fournit les métriques  $M_i$  cherchées.

### 3.3.6. Comparaison entre les triplets $(X, M_i, D_p)$ et les couples $(D_i, D_p)$

On représente en points supplémentaires sur les plans principaux obtenus en 3.3.1., les opérateurs  $W \circ D_p$  associés aux triplets  $(X, M_i, D_p)$ ; le graphique obtenu permet de juger de façon plus sensible l'adéquation entre la matrice  $X$  et les différents tableaux  $D_i$  (cf. 3.3.4.).

## 4. COMPARAISON ENTRE UN TABLEAU DE DISTANCES ET UNE PARTITION

On caractérise un ensemble de  $n$  objets par une paire  $(D, D_p)$  et une variable qualitative  $y$  à  $k$  modalités. On associe à  $y$  le tableau  $Y(k, n)$  des  $k$  indicatrices des modalités de  $y$ .

Peut-on "prévoir" les modalités de  $y$  à partir des distances de  $D$  ?

Ce problème, qui relève de la discrimination, peut être résolu en se basant sur l'équivalence 2" (cf. § 1).

Dans le cas particulier où  $y$  est obtenu par une méthode de classification automatique, peut-on apprécier la validité et la cohérence des résultats obtenus ?

#### 4.1. Résultats préliminaires

##### 4.1.1. Adéquation entre un tableau de distances et une partition

Remarquons d'abord que :

$$I_{\underline{g}} = \frac{1}{2} d_{..}^2 = \frac{1}{2} \Sigma \{p_i p_j d_{ij}^2 \mid i \in I ; j \in I\}$$

soit :

$$I_{\underline{g}} = \frac{1}{2} \underline{j}' D_p \bar{D} D_p \underline{j}.$$

Calculons  $U = Y D_p \bar{D} D_p Y'$ .

soit  $u_{rs} = \Sigma \{p_i p_j d_{ij}^2 \mid i \in G_r ; j \in G_s\}$

où  $G_r = \{i \in I \mid y_i^r = 1\}$ .

Si on note  $\{\underline{x}_i \mid i = 1, \dots, n\}$  une image euclidienne de la paire  $(D, D_p)$ , on a :

$$\begin{aligned} d_{ij}^2 &= \|\underline{x}_i - \underline{x}_j\|^2 \\ &= \|\underline{x}_i - \underline{g}_r\|^2 + \|\underline{x}_j - \underline{g}_s\|^2 + \|\underline{g}_r - \underline{g}_s\|^2 \\ &\quad + 2 \langle \underline{x}_i - \underline{g}_r, \underline{g}_r - \underline{g}_s \rangle + 2 \langle \underline{x}_j - \underline{g}_s, \underline{g}_s - \underline{g}_r \rangle \\ &\quad - 2 \langle \underline{x}_i - \underline{g}_r, \underline{x}_j - \underline{g}_s \rangle \end{aligned}$$

si  $i$  appartient à  $G_r$ ,  $j$  à  $G_s$ , et si  $\underline{g}_r$  est le centre de gravité du groupe  $G_r$ , c'est-à-dire :

$$\underline{g}_r = (1/P_r) \Sigma \{p_i \underline{x}_i \mid i \in G_r\}$$

avec  $P_r = \Sigma \{p_i \mid i \in G_r\}$

(donc  $\Sigma \{p_i (\underline{x}_i - \underline{g}_r) \mid i \in G_r\} = \underline{0}$ ).

D'où :

$$\begin{aligned} u_{rs} &= P_r P_s \|\underline{g}_r - \underline{g}_s\|^2 \\ &\quad + P_r \Sigma \{p_i \|\underline{x}_i - \underline{g}_s\|^2 \mid i \in G_s\} + P_s \Sigma \{p_i \|\underline{x}_i - \underline{g}_r\|^2 \mid i \in G_r\}. \end{aligned}$$

Soit, en posant :

$$\delta_{rs}^2 = \|\underline{g}_r - \underline{g}_s\|^2$$

$$v_r = \frac{1}{P_r} \sum \{p_i \|x_i - \underline{g}_r\|^2 \mid i \in G_r\}$$

on a :

$$u_{rs} = P_r P_s \delta_{rs}^2 + P_r P_s (v_r + v_s)$$

soit :

$$U = D_p (\underline{j}' \Gamma + \Delta + \Gamma \underline{j}'') D_p$$

où  $\Delta = (\delta_{rs}^2)$  est la matrice des distances entre les centres de gravité des groupes ;

$$\Gamma = \begin{bmatrix} v_1 & & & \\ & \ddots & & \\ 0 & & \ddots & 0 \\ & & & v_k \end{bmatrix} \text{ est la matrice des inerties intra-groupes.}$$

On peut ainsi mesurer l'adéquation globale de la partition induite par  $y$ , au tableau de distances de la façon suivante :

$$- \text{ on a vu que : } I_{\underline{g}} = \frac{1}{2} \underline{j}' D_p \bar{D} D_p \underline{j} = \frac{1}{2} \text{tr} (\underline{j}' D_p \bar{D} D_p \underline{j})$$

– on constate que la variance intra-classe est liée à la trace de la matrice  $U$  par la relation :

$$V_{\text{intra}} = \frac{1}{2} \text{tr} (D_{1/P} U)$$

où

$$V_{\text{intra}} = \sum \{P_r v_r \mid r = 1, 2, \dots, k\}$$

– en calculant le rapport de corrélation :

$$\rho^2 = 1 - \frac{\text{tr} (D_{1/P} Y D_p \bar{D} D_p Y')}{\text{tr} (\underline{j}' D_p \bar{D} D_p \underline{j})}$$

on a une mesure de la cohérence entre la variable qualitative et la dispersion du nuage de points.

On remarque l'analogie entre le numérateur et le dénominateur en considérant la matrice  $\underline{j}'$  comme la matrice de l'indicatrice unique de la variable qualitative triviale à une modalité c'est-à-dire constante sur tous les individus.

Ce coefficient  $\rho$  permet donc de juger le degré de liaison entre un tableau D et une variable qualitative, ou bien de mesurer la qualité d'une partition, ce qui peut permettre la comparaison de méthodes de classification.

#### 4.1.2. Distances des individus aux centres de groupes

D'après ce qui précède, on connaît le tableau des distances entre centres de gravité :

$$\delta_{st}^2 = \frac{1}{P_s} \frac{1}{P_t} u_{st} - v_s - v_t$$

avec

$$v_t = (1/2P_t^2) u_{tt}$$

$$u_{st} = \sum \{p_i p_j d_{ij}^2 \mid i \in G_s ; j \in G_t\}$$

$$U = Y D_p \bar{D} D_p Y'$$

On peut calculer, de façon analogue, les distances  $e_{ir}$  entre un individu  $i$  et le centre de gravité du groupe  $r$  :

$$e_{ir}^2 = \frac{1}{P_r} \sum \{p_j d_{ij}^2 \mid j \in G_r\} - \frac{1}{P_r^2} \sum \{p_j p_l d_{jl}^2 \mid j \in G_r ; l \in G_r\}$$

Cette formule "trigonométrique" se démontre par récurrence sur le nombre de points du groupe  $G_r$ .

Soit  $\underline{x}$  un point de  $\mathbf{R}^n$ , et  $\underline{a}_1, \underline{a}_2, \dots, \underline{a}_k$  un groupe de points auxquels sont associés les poids  $p_1, p_2, \dots, p_k$ .

On pose :  $Q_j = \sum \{p_l \mid l = 1, 2, \dots, j\}$

$g_j$  centre de gravité de  $\{\underline{a}_1, \underline{a}_2, \dots, \underline{a}_j\}$

$$d_{j1}^2 = \|\underline{a}_j - \underline{a}_1\|^2$$

$$\alpha_j^2 = \|\underline{x} - \underline{a}_j\|^2$$

$$\theta_j^2 = \|\underline{x} - \underline{g}_j\|^2$$

Pour  $j = 2$ , on a :

$$\|\underline{x} - \underline{a}_1\|^2 = \|\underline{x} - \underline{g}_2\|^2 + \|\underline{g}_2 - \underline{a}_1\|^2 + 2 \langle \underline{x} - \underline{g}_2, \underline{g}_2 - \underline{a}_1 \rangle$$

$$\|\underline{x} - \underline{a}_2\|^2 = \|\underline{x} - \underline{g}_2\|^2 + \|\underline{g}_2 - \underline{a}_2\|^2 + 2 \langle \underline{x} - \underline{g}_2, \underline{g}_2 - \underline{a}_2 \rangle$$

En remarquant que :

$$\underline{g}_2 - \underline{a}_1 = (p_2/Q_2)(\underline{a}_2 - \underline{a}_1)$$

$$\underline{g}_2 - \underline{a}_2 = (p_1/Q_2)(\underline{a}_1 - \underline{a}_2),$$

on obtient, en multipliant les deux égalités par  $p_1$  et  $p_2$  respectivement, puis en sommant :

$$\begin{aligned} p_1 \alpha_1^2 + p_2 \alpha_2^2 &= Q_2 \theta_2^2 + \frac{p_1 p_2}{Q_2} d_{12}^2 \\ + 2 \langle \underline{x} - \underline{g}_2, p_1 (\underline{g}_2 - \underline{a}_1) + p_2 (\underline{g}_2 - \underline{a}_2) \rangle \end{aligned}$$

d'où :

$$\theta_2^2 = \frac{1}{Q_2} \Sigma \{p_j \alpha_j^2 \mid j=1, 2\} - (1/Q_2)^2 p_1 p_2 d_{12}^2.$$

Supposons que :

$$\begin{aligned} \theta_j^2 &= (1/Q_j) \Sigma \{p_l \alpha_l^2 \mid l=1, \dots, j\} \\ &- (1/Q_j)^2 \Sigma \{p_l p_m d_{lm}^2 \mid l=1, \dots, j; m=1, \dots, j\} \end{aligned}$$

comme précédemment, on a :

$$\|\underline{x} - \underline{g}_j\|^2 = \|\underline{x} - \underline{g}_{j+1}\|^2 + \|\underline{g}_{j+1} - \underline{g}_j\|^2 + 2 \langle \underline{x} - \underline{g}_{j+1}, \underline{g}_{j+1} - \underline{g}_j \rangle$$

$$\|\underline{x} - \underline{a}_{j+1}\|^2 = \|\underline{x} - \underline{g}_{j+1}\|^2 + \|\underline{g}_{j+1} - \underline{a}_{j+1}\|^2 + 2 \langle \underline{x} - \underline{g}_{j+1}, \underline{g}_{j+1} - \underline{a}_{j+1} \rangle$$

soit :

$$Q_j \theta_j^2 + p_{j+1} \alpha_{j+1}^2 = Q_{j+1} \theta_{j+1}^2 + \frac{p_{j+1} Q_j}{Q_{j+1}} \|\underline{g}_j - \underline{a}_{j+1}\|^2.$$

D'après la récurrence, on a :

$$\begin{aligned} \|\underline{g}_j - \underline{a}_{j+1}\|^2 &= \frac{1}{Q_j} \Sigma \{p_l d_{lj+1}^2 \mid l=1, \dots, j\} \\ &- \frac{1}{Q_j} \Sigma \{p_l p_m d_{lm}^2 \mid l=1, \dots, j; m=1, \dots, j\}, \end{aligned}$$

ce qui donne :

$$Q_{j+1} \theta_{j+1}^2 = \Sigma \{p_l \alpha_l^2 \mid l = 1, \dots, j+1\} \\ - \left[ \frac{1}{Q_j} - \frac{p_{j+1}}{Q_j Q_{j+1}} \right] \Sigma \{p_l p_m d_{lm}^2 \mid l = 1, \dots, j; m = 1, \dots, j\} \\ - \frac{p_{j+1}}{Q_{j+1}} \Sigma \{p_l d_{lj+1}^2 \mid l = 1, \dots, j\}.$$

D'où on déduit :

$$\theta_{j+1}^2 = \frac{1}{Q_{j+1}} \Sigma \{p_l \alpha_l^2 \mid l = 1, \dots, j+1\} \\ - \frac{1}{Q_{j+1}^2} \Sigma \{p_l p_m d_{lm}^2 \mid l = 1, \dots, j+1; m = 1, \dots, j+1\}$$

c . q . f . d.

Ayant ainsi le tableau des distances mutuelles entre les individus initiaux et les centres de groupes, on peut effectuer une analyse factorielle sur tableau de distances qui permet de visualiser les liens entre la variable qualitative, représentée par les centres de groupes, et les variables résumées dans le tableau de distances. On peut, par exemple, utiliser les centres de groupes comme points de base, et projeter les individus initiaux en points supplémentaires. Cette analyse n'est évidemment pas une analyse factorielle discriminante puisque, si le nuage dont on cherche les composantes principales est bien celui des centres de gravité, la métrique utilisée est celle exprimée par le juge qui a construit le tableau D et non pas la métrique de Mahalanobis induite par l'ensemble des individus initiaux.

#### 4.2. Analyse factorielle discriminante sur tableau de distances

Pour juger de l'adéquation entre une partition et un tableau de distances, on peut effectuer une analyse factorielle sur ce tableau de distances et faire apparaître en points supplémentaires sur les plans principaux, les centres de groupes associés aux éléments de la partition. On dit que la partition reflète le tableau de distances si ces centres de groupes apparaissent comme bien séparés, compte tenu des dispersions à l'intérieur des groupes.

Cette technique privilégie la reconstruction des distances. Il peut se trouver que, dans certains cas, les dernières composantes principales permettent de mieux

retrouver la partition que les premières. Si on privilégie la notion de liaison (liaison entre les dimensions sous-jacentes au tableau de distances et le caractère qualitatif associé à la partition), c'est-à-dire si on se place dans l'optique de l'analyse canonique, il faut non pas tenter de reconstruire au mieux les distances entre points à partir de la partition (optique analyse en composantes principales et analyse factorielle sur tableau de distances) mais comparer, dans  $\mathbf{R}^n$ , les positions relatives des sous-espaces vectoriels engendrés respectivement par le tableau de distances (cf. chap. II, § 1) et les variables indicatrices de la partition. Procéder ainsi revient à effectuer ce que l'on peut appeler une analyse factorielle discriminante sur tableau de distances.

Pour effectuer cette analyse factorielle discriminante sur tableau de distances, logiquement, on devrait retenir, dans la précédente analyse factorielle sur tableau de distances, toutes les composantes principales associées aux valeurs propres positives ; puis effectuer une analyse factorielle discriminante classique, le rôle des variables quantitatives étant tenu par les composantes principales retenues. On procède bien ainsi, mais en ne retenant que les composantes principales associées à des valeurs propres nettement positives. En effet, il ne faut pas oublier que quand on effectue classiquement une analyse factorielle discriminante, toutes les variables quantitatives considérées ont *a priori* des importances identiques. Or la situation est ici tout à fait différente : les composantes principales ont *a priori* des importances qui sont fonction de l'inertie qu'elles permettent de reconstruire, exprimant ainsi l'influence qu'elles ont dans la détermination des distances ; il n'est donc pas intéressant de constater qu'une composante principale de faible inertie soit bien discriminante.

Pour justifier la procédure ci-dessus, on remarque que, si le rang de l'opérateur  $W \circ D_p$  associé au tableau de distances est égal à  $n-1$ , le sous-espace de  $\mathbf{R}^n$  associé au tableau de distances est l'hyperplan orthogonal à la droite des constantes. Cet hyperplan contient donc la partie orthogonale à la droite des constantes du sous-espace associé aux indicatrices de toute partition. Il est donc trivial, dans ce cas, de discriminer à partir du tableau de distances si on retient toutes les composantes principales ; par contre, si on ne retient que les premières composantes principales, le problème de la discrimination garde tout son sens.

La procédure précédemment décrite qui tient bien compte de l'information initiale peut être comparée à l'analyse factorielle discriminante pas à pas [38] effectuée sur les colonnes de  $W$  considérées comme caractères ; on exploite ici le fait que le sous-espace engendré par les colonnes de  $W$  est identique à celui engendré par les composantes principales (cf. chap. III, § 1). Après  $k$  itérations, la procédure fournit l'ensemble des  $k$  individus (colonnes de  $W$ ) les plus "discriminants".

## CHAPITRE IV

### ANALYSE ORDINALE DES PROXIMITES

#### 1. EXPOSE DU PROBLEME

Dans ce chapitre, nous allons rappeler des méthodes permettant d'obtenir une image euclidienne tenant compte de l'ordre des distances ou dissimilarités et non pas de leur valeur absolue. On recherche une configuration des individus dans un espace vectoriel euclidien qui respecte au mieux leur ordonnance.

On rappelle qu'une ordonnance (respectivement préordonnance) sur un ensemble  $I$  de cardinal  $n$ , est une relation d'ordre (respectivement de préordre) sur l'ensemble  $I \times I$ .

A tout indice de distance ou de dissimilarité est associée la préordonnance :

$$\delta_{ij} < \delta_{kl} \iff (i, j) < (k, l).$$

Rappelons d'abord deux théorèmes qui nous assurent que le problème possède une solution.

#### *Théorème de Shepard [42]*

*Une préordonnance sur  $I$  admet une image euclidienne respectant la préordonnance dans un espace vectoriel de dimension  $n - 1$ .*

J.P. Benzecri donne une démonstration de ce théorème en raisonnant par continuité : à partir du polyèdre régulier à  $n$  sommets dans  $\mathbf{R}^{n-1}$ , il raccourcit ou allonge les arêtes de façon à respecter la préordonnance.

#### *Théorème de Guttman [16]*

*Une préordonnance qui n'est pas ultramétrique, admet une image euclidienne dans un espace vectoriel de dimension inférieure ou égale à  $n - 2$ . De plus le carré des distances entre deux points est une fonction affine du rang  $r_{ij}$  du couple  $(i, j)$  dans la préordonnance.*

*Remarque :*

Si la préordonnance est ultramétrique, le résultat du § 2.2.4 du chapitre II nous assure de l'existence d'une image euclidienne.

*Preuve :*

Considérons, comme Gutman, le tableau  $R(n, n)$  des rangs des couples  $(l, k)$  dans la préordonnance. A ce tableau considéré comme s'il était un tableau de dissimilarités (qui respecte l'ordonnance) est associée une forme quadratique  $W$  qui n'est pas en général semi-définie positive. On sait, d'après le théorème 3 du chapitre II, qu'en ajoutant à chaque  $r_{kl}$  la constante  $2|\mu_n|$ , où  $\mu_n$  est la plus petite valeur propre de la matrice  $W$ , d'une part on conserve la préordonnance, et que d'autre part, on rend semi-définie positive la matrice  $W$  qui admet alors un noyau de dimension supérieure ou égale à 2.

J.P. Benzecri [1] a proposé une méthode, basée sur une autre idée, pour associer une image euclidienne à une préordonnance.

En effet, si on trouve dans l'espace  $\mathbb{R}^p$  muni de la métrique  $M$  un nuage  $\{\underline{x}_i \in \mathbb{R}^p \mid i = 1, 2, \dots, n\}$  respectant la préordonnance

$$\delta_{ij} < \delta_{kl} \Rightarrow M(\underline{x}_i - \underline{x}_j) < M(\underline{x}_k - \underline{x}_l),$$

on peut considérer que  $\underline{x}_i$  est une réalisation d'un vecteur aléatoire  $\underline{x}$  suivant une loi de Laplace-Gauss multidimensionnelle, de moyenne  $\mu$  et de matrice de variance-covariance  $\Sigma = M^{-1}$ ; on sait alors que  $\underline{x}_i - \underline{x}_{i'}$  suit une loi de Laplace-Gauss de moyenne 0 et de matrice de variance-covariance  $2\Sigma$ .

On en déduit que :

$$-\frac{1}{2} d_{ii'} = \frac{1}{2} M(\underline{x}_i - \underline{x}_{i'}) \text{ est la réalisation d'une variable aléatoire } D^2 \text{ qui suit}$$

une loi du khi-deux à  $p$  degrés de liberté,

$$-E(D^2) = I_\mu = \text{tr}(\Sigma M) = p \text{ où } I_\mu \text{ est le moment d'inertie du vecteur aléatoire } \underline{x} \text{ par rapport au centre de gravité : } I_\mu = E(M(\underline{x} - \underline{\mu})).$$

On peut donc considérer que  $\frac{1}{2} d_{ii'}$  est une réalisation d'une variable aléatoire qui suit une loi du khi-deux à  $p$  degrés de liberté et dont l'espérance mathématique est égale à  $p$ .

D'où une méthode pour calculer à partir de la préordonnance un système de  $d_{ii'}$  :  $r_{ii'}$  étant le rang de la paire  $(i, i')$  dans la préordonnance  $\left(\frac{n(n-1)}{2}\right)$  paires

sont considérées si  $I$  a  $n$  éléments), on pose :

$$P \left[ D^2 < \frac{d_{ii'}^2}{2} \right] = \frac{r_{ii'} - \epsilon}{n(n-1)} = P_{ii'} ;$$

il suffit de consulter une table du khi-deux à  $p$  degrés de liberté pour obtenir les valeurs  $d_{ii'}^2/2$ .

Une analyse factorielle sur tableau de distances effectuée sur le tableau des  $d_{ii'}$  permet alors de construire une image euclidienne.

Si les résultats précédents nous assurent d'une solution au problème posé, ils ne permettent pas de le résoudre d'une manière pratique, car le nombre d'individus étant généralement grand, l'image euclidienne obtenue est inexploitable. Il s'agit donc maintenant de trouver une image euclidienne de "faible" dimension  $r$ , telle que la préordonnance induite par les distances euclidiennes soit la plus "proche" possible de la préordonnance de départ.

D'abord, la contrainte  $r \ll n - 1$  ne permet pas d'obtenir une solution respectant la préordonnance de départ. Ensuite, il faut définir un indice mesurant la proximité des ordonnances. Enfin, le couple de contraintes,  $r$  petit – préordonnances proches, est contradictoire et un compromis doit être trouvé entre les valeurs de  $r$  et celle de l'indice.

Remarquons que deux facteurs pratiques mais non statistiques vont influencer la méthode :

- $r$  doit être choisi de façon à permettre la lecture et l'interprétation des résultats. On retrouve ici un problème qui n'est pas propre à l'analyse des proximités.

- Le "coût" des calculs est un élément du choix de la méthode utilisée et du type de solution recherchée.

Les techniques existantes diffèrent par leur critère d'adéquation des préordonnances. Nous nous proposons d'exposer un certain nombre de ces critères (§ 2), puis d'évoquer la procédure qui en découle (§ 3). Dans un dernier paragraphe, nous discuterons de la validité des résultats obtenus.

## 2. PRINCIPE DES DIFFERENTES TECHNIQUES

Puisque le but recherché est de respecter la préordonnance et non pas les valeurs des dissimilarités, l'idée principale est de construire un indice de proximité

entre préordonnances invariant par transformation monotone des dissimilarités, ou du moins dont l'optimum est invariant par transformation monotone.

Pour juger de la proximité entre la préordonnance induite par  $\delta$ , indice de dissimilarité initial, et celle induite par  $d$ , distance des individus dans l'image euclidienne obtenue, il est commode d'associer à un indice de distance, en plus du tableau  $D$  à  $n$  lignes et  $n$  colonnes des  $d_{ij}$ , la colonne  $\underline{d}$  de  $\mathbf{R}^N$ , formée des  $N = \frac{n(n-1)}{2}$  valeurs  $d_{ij}$ , ordonnées suivant le préordre induit par l'indice. Alors un indice de distance compatible avec la préordonnance initiale est un élément de  $\mathbf{R}^N$  contenu dans le cône convexe des éléments à coordonnées croissantes

$$C = \{ \underline{x} \in \mathbf{R}^N \mid 0 \leq x_1 \leq x_2 \leq \dots \leq x_N \}.$$

### 2.1. Stress de Kruskal [23]

Pour mesurer l'adéquation d'une image euclidienne à la préordonnance de départ, Kruskal utilise le stress défini par :

$$S = \left[ \frac{\sum \{(d_{ij} - \hat{d}_{ij})^2 \mid i < j\}}{\sum \{d_{ij}^2 \mid i < j\}} \right]^{\frac{1}{2}}$$

où :

- $d_{ij}$  est la distance entre  $i$  et  $j$  dans l'image euclidienne obtenue,
- $\hat{d}_{ij}$  est une distance entre  $i$  et  $j$  respectant la préordonnance :

$$\delta_{ij} < \delta_{kl} \Rightarrow \hat{d}_{ij} < \hat{d}_{kl}$$

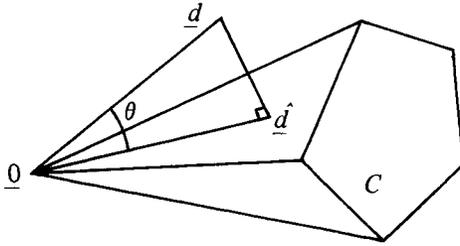
et la plus proche de  $d$  au sens des moindres carrés :

$$\sum \{(d_{ij} - \hat{d}_{ij})^2 \mid i < j\} \text{ est minimum.}$$

En raisonnant dans  $\mathbf{R}^N$  on voit que

$$S = \frac{\| \underline{d} - \hat{\underline{d}} \|}{\| \hat{\underline{d}} \|} = \sin \theta,$$

et le vecteur  $\hat{\underline{d}}$  est le point du cône  $C$  le plus proche de  $\underline{d}$ . On obtient donc le vecteur  $\hat{\underline{d}}$  en projetant le vecteur  $\underline{d}$  sur ce cône ; en procédant ainsi on effectue une "régression monotone".



Kruskal, après Shepard, a montré que la distance  $\hat{d}$  possédait la propriété, évidente d'après ce qui précède, de rendre le stress minimum quand on considère toutes les transformations monotones de l'indice  $\delta$ . Il a raisonné en considérant le graphique obtenu en portant les valeurs de  $\delta_{ij}$  en fonction des valeurs de  $d_{ij}$ . Si la distance  $d$  respecte la préordonnance, la courbe qui joint les points est non décroissante (fig. 1). Dans le cas contraire, on détermine les valeurs de  $\hat{d}$  en traçant la courbe monotone croissante la plus proche du nuage de points au sens des moindres carrés (fig. 2).

On note aussi que le stress est invariant par transformation orthogonale (symétrie, rotation), et par translation des images euclidiennes. D'autre part, cet indice est une fonction continue et différentiable des distances.

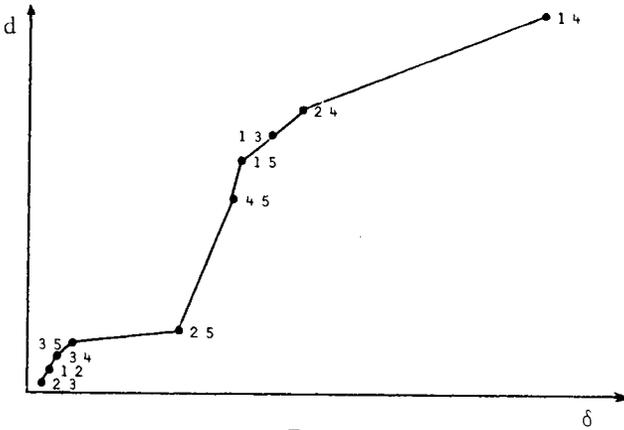


Figure 1

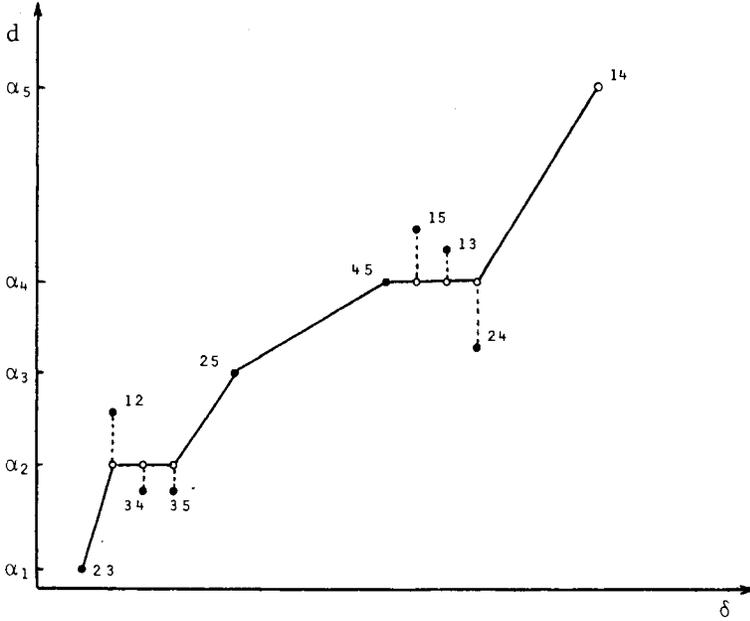


Figure 2 -  $\hat{d}_{23} = d_{23} = \alpha_1$   
 $\hat{d}_{12} = \hat{d}_{34} = \hat{d}_{35} = \frac{1}{3}(d_{12} + d_{34} + d_{35}) = \alpha_2$   
 $\hat{d}_{25} = d_{25} = \alpha_3$   
 $\hat{d}_{45} = \hat{d}_{15} = \hat{d}_{13} = \hat{d}_{24} = \frac{1}{4}(d_{45} + d_{15} + d_{13} + d_{24}) = d_{45} = \alpha_4$   
 $\hat{d}_{14} = d_{14} = \alpha_5$

*Remarque :*

Dans la méthode TORSCA, proposée par Young et Torgerson [55], l'indice d'adéquation est :

$$I = \frac{1}{2} \left( 1 + \frac{\sum \{d_{ij} \hat{d}_{ij} \mid i < j\}}{(\sum \{d_{ij}^2 \mid i < j\} \sum \{\hat{d}_{ij}^2 \mid i < j\})^{1/2}} \right)$$

où  $d_{ij}$  est toujours la distance mesurée sur l'image euclidienne obtenue et  $\hat{d}$  les transformées des valeurs de  $d$  qui rendent  $I$  maximum.

En raisonnant dans  $\mathbf{R}^N$ , on a la même interprétation que précédemment et l'indice  $I$  s'écrit,  $\hat{d}$  appartenant à  $C$  :

$$I = \frac{1}{2} \left( 1 + \frac{\langle \underline{d}, \hat{\underline{d}} \rangle}{\|\underline{d}\| \|\hat{\underline{d}}\|} \right) = \frac{1}{2} (1 + \cos \theta).$$

On constate donc que les optimums de ces deux indices sont obtenus pour le même vecteur  $\hat{\underline{d}}$ .

## 2.2. Indice de McGEE [34]

McGee propose un autre indice d'adéquation :

$$W = \sum \left\{ c^2 \left( \frac{d_{ij} - \hat{d}_{ij}}{\hat{d}_{ij}} \right)^2 \mid i < j \right\}$$

où les  $d_{ij}$  et les  $\hat{d}_{ij}$  sont définis comme dans le paragraphe précédent.

Il raisonne de la façon suivante :

Si chaque couple de points est relié par un ressort de longueur initiale  $\hat{d}_{ij}$ ,  $W$  est le travail nécessaire pour transformer la configuration initiale en une configuration où les longueurs des ressorts sont devenues  $d_{ij}$ ,  $c$  étant interprété comme la constante d'élasticité de chaque ressort, indépendante du couple  $(i, j)$  considéré [35].

Il cherche alors la transformation qui minimise le travail.

A l'aide d'une hypothèse probabiliste, il obtient également une estimation de la dimension  $r$  de l'espace. En effet, il suppose que chaque terme  $\frac{d_{ij} - \hat{d}_{ij}}{\hat{d}_{ij}}$  suit une loi normale centrée réduite, donc que  $W$  suit une loi du  $\chi^2$ . Ainsi, ayant fixé la dimension  $r$  de l'espace où il cherche une solution, il n'y accepte la configuration que si la statistique  $W$  a une probabilité de réalisation suffisamment grande ; sinon il cherche la configuration solution dans un espace de dimension  $r + 1 \dots$

Il justifie son raisonnement en remarquant que :

– Le ressort  $(i, j)$  n'a aucune raison *a priori* d'être allongé plutôt que raccourci, donc l'espérance de  $d_{ij}$  est égale à  $\hat{d}_{ij}$ .

– La variance de l'allongement du ressort  $(i, j)$  est proportionnelle à  $\hat{d}_{ij}$ , c.a.d. que la variabilité de la distance des individus très dissemblables est plus grande que celle des individus proches.

*Remarque :*

D'une manière symétrique, McGee propose une deuxième stratégie très voisine de la précédente et analogue dans ses principes en considérant l'indice :

$$W' = \sum \left\{ c^2 \left( \frac{d_{ij} - \hat{d}_{ij}}{d_{ij}} \right)^2 \mid i < j \right\}.$$

Contrairement à Kruskal qui tolère la même transformation sur une distance  $d_{ij}$  grande ou petite, l'indice de McGee pénalise les couples en désaccord avec la préordonnance proportionnellement à la distance  $d_{ij}$ . Cette propriété n'est d'ailleurs pas un avantage pour tout type de données, comme le font remarquer Gregson et Russel [15].

### 2.3. Indice de Guttman-Lingoes [18], [17]

L'originalité de la méthode de Guttman-Lingoes provient non pas du choix de l'indice d'adéquation entre distances, mais de la distance servant de référence.

L'indice d'adéquation est analogue à celui de Young-Torgerson et par conséquent au stress de Kruskal puisque l'indice appelé *coefficient d'accord de monotonie* est défini par :

$$\mu = \frac{\sum \{ d_{ij} \tilde{d}_{ij} \mid i < j \}}{\sqrt{\sum d_{ij}^2 \sum \tilde{d}_{ij}^2}}$$

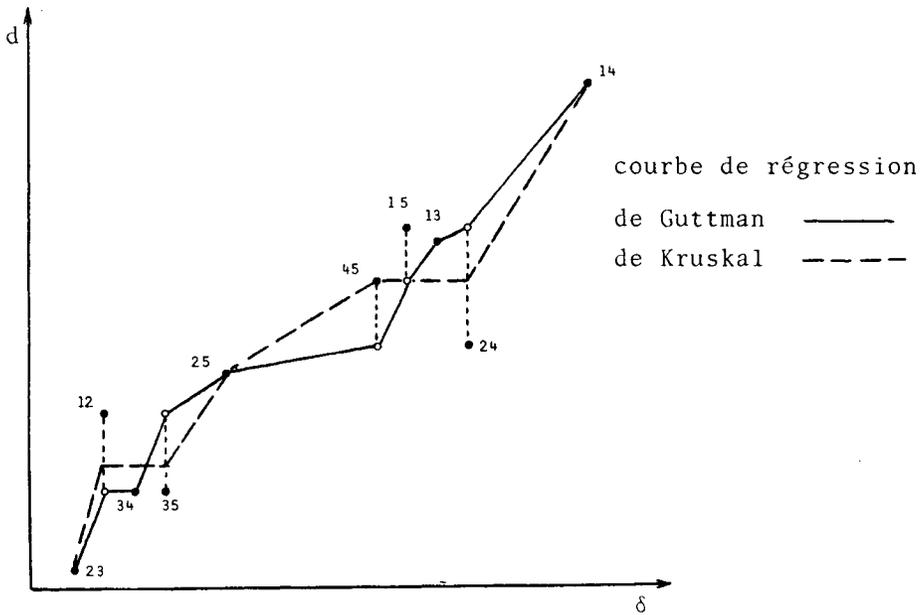
En raisonnant dans  $\mathbf{R}^N$  comme précédemment (§ 2), on constate que :

$$\mu = \frac{\langle \underline{d}, \underline{\tilde{d}} \rangle}{\| \underline{d} \| \| \underline{\tilde{d}} \|}.$$

Par conséquent, ce coefficient possède les mêmes extremums que le stress de Kruskal.

La différence provient du choix de la distance  $\tilde{d}$  considérée comme la plus proche de la distance  $d$  reconstruite.

La distance  $\tilde{d}$  est obtenue par la transformation appelée *image de rang* par Guttman et définie de la façon suivante : si  $r_{ij}$  est le rang de la paire  $(i, j)$  dans l'ordonnance induite par l'indice de dissimilarité initial, on donne à  $d_{ij}$  la valeur de rang  $r_{ij}$  des distances  $d$  ordonnées par ordre croissant, sans référence au couple  $(i, j)$  dont elle provient.



Préordonnance  $\delta$  - (23) (12) (34) (35) (25) (45) (15) (13) (24) (14)  
 associée à:  $d$  - (23) (35) (34) (12) (25) (24) (45) (13) (15) (14)

Figure 3

En reprenant l'exemple du paragraphe 2.1, on voit sur le graphique joint la courbe obtenue par la transformation de rang (fig. 3) et on peut la comparer à la courbe obtenue par régression monotone (algorithme des blocs de Kruskal).

Un avantage du coefficient d'accord de monotonie est d'être un *simili-coefficient de corrélation de rang de Spearman* continu par rapport aux distances. En effet, il est obtenu comme le coefficient de corrélation des rangs entre la préordonnance de départ et la préordonnance reconstruite, mais en remplaçant les valeurs des rangs par les valeurs situées aux mêmes rangs dans le tableau des distances reconstruites (contrairement aux rangs, celles-ci ne sont pas centrées).

La transformation de rang diffère aussi l'analyse factorielle sur tableau de distances et l'analyse ordinaire des proximités au sens de Guttman-Lingoes, car la maximisation de l'inertie reconstruite faite en analyse factorielle sur tableau de distances revient également à maximiser le coefficient de corrélation linéaire mais entre les tableaux des distances initiales et des distances reconstruites.

Cette transformation de rang implique une différence fondamentale entre l'indice de Guttman-Lingoes et les indices précédents. En effet, la distance  $\hat{d}$  prise comme distance de référence par la régression monotone est la plus proche au sens des moindres carrés parmi les distances en accord avec la préordonnance, c'est-à-dire que cette distance corrige le plus faiblement possible les contradictions de la distance  $d$ . Au contraire, par la transformation de rang, Guttman et Lingoes obtiennent une distance  $\hat{d}$  qui n'a rien à voir avec la distance  $d$  de départ, même si elle prend les mêmes valeurs numériques. Plus précisément, pour Kruskal, la valeur de  $\tilde{d}_{ij}$  est la moyenne de  $d_{ij}$  et d'autres distances qui entourent, au sens de la préordonnance,  $d_{ij}$  (algorithme des blocs) tandis que pour Guttman-Lingoes, la valeur de  $\tilde{d}_{ij}$  est la valeur d'un couple  $(k, k')$  qui peut être totalement différent. Ceci implique que globalement la distance au sens des moindres carrés entre les deux tableaux peut être très grande.

#### 2.4. Indice de Johnson [21]

Pour mesurer la qualité de la représentation, Johnson utilise l'indice  $\theta$  suivant :

$$\theta^2 = \frac{\sum \{ \Delta_{ij}^{kl} (d_{ij}^2 - d_{kl}^2)^2 \mid i < j, k < l, (i, j) \neq (k, l) \}}{\sum \{ d_{ij}^2 - d_{kl}^2 \mid i < j, k < l, (i, j) \neq (k, l) \}}$$

$$\Delta_{ij}^{kl} = \begin{cases} 1 & \text{si } \text{sgn}(d_{ij} - d_{kl}) = \text{sgn}(\delta_{ij} - \delta_{kl}) \\ 0 & \text{sinon} \end{cases}$$

Le coefficient  $\theta$  est compris entre 0 et 1. En effet, apparaissent au numérateur seulement les termes du dénominateur pour lesquels la préordonnance est respectée :

- $\theta = 1$  si et seulement si la préordonnance est parfaitement respectée dans l'image euclidienne obtenue,
- $\theta = 0$  si et seulement si la préordonnance est inversée.

On peut remarquer l'analogie de cet indice avec le coefficient de corrélation des rangs de Kendall  $\tau$ . Si on définit  $\tilde{\theta}$  par

$$\tilde{\theta} = \frac{\sum \{ \Delta_{ij}^{kl} \mid i < j, k < l, (i, j) \neq (k, l) \}}{n(n-1)/2},$$

on démontre que  $\tilde{\theta} = \frac{1}{2}(1 - \tau)$ .

Ainsi la caractéristique du coefficient  $\theta$  est que l'influence des termes qui ne respectent pas la préordonnance est pondérée par le carré de l'écart de leur distance reconstruite. Et l'avantage du coefficient  $\theta$  par rapport au coefficient  $\tilde{\theta}$  est d'être une fonction continue des distances des points de l'image. Ceci permet d'améliorer l'image obtenue par un algorithme du gradient.

On retrouve, sous-jacent dans le choix de cet indice, le critère des moindres carrés car la différence ( $a_{ij}^2 - a_{kl}^2$ ) intervient à la puissance 2. En faisant intervenir cette différence d'une autre façon (i.e. par une fonction croissante quelconque), on peut construire d'autres indices de proximité.

### 3. METHODES DE RESOLUTION

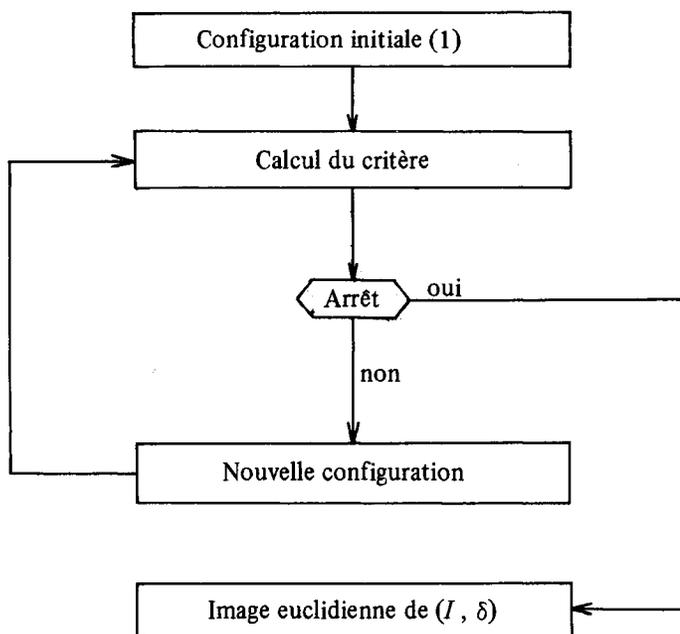
#### 3.1. Généralités

On cherche à obtenir une image euclidienne dans un espace de dimension  $r$ , qui respecte au mieux la préordonnance induite par  $\delta$ . Cette image est fonction de  $n \times r$  variables, les  $r$  coordonnées de chacun des  $n$  points. Le respect de la préordonnance est jugé par la valeur de l'un des indices précédents.

Résoudre ce problème revient donc à chercher le minimum d'une fonction de  $n \times r$  variables. En règle générale, on ne sait pas calculer la solution de ce problème si la fonction à minimiser ne présente pas des propriétés particulières. On se contente de construire itérativement une image à l'aide d'un algorithme du gradient. Celui-ci n'est pas un algorithme optimal, mais seulement heuristique. Il peut conduire d'abord à un optimum local, mais de plus la convergence étant souvent trop lente, on est amené à arrêter la procédure prématurément.

Chaque itération de l'algorithme comporte deux phases :

- 1) Calcul de l'indice d'adéquation de l'image euclidienne, obtenue à cette itération, à la préordonnance.
- 2) Amélioration, si nécessaire, de cet indice en modifiant l'image euclidienne par déplacement des points.



Chacune de ces phases présentent des difficultés propres. Dans la première, elles sont d'ordre numérique et généralement résolues de façon satisfaisante par un algorithme de calcul. Pour la seconde phase, la principale difficulté réside dans le choix du pas d'incrémentation de l'algorithme du gradient utilisé. Celui-ci conditionne la convergence de l'algorithme et sa rapidité.

Pour chacun des indices précédents, les différents auteurs ont précisé la méthode utilisée, basée plutôt sur une connaissance empirique du problème considéré que sur un raisonnement théorique.

### 3.2. MDSCAL

Kruskal mesure par le stress (cf. § 2.1 de ce chapitre), l'accord entre la préordonnance de départ et celle associée à l'image euclidienne obtenue :

$$S = \frac{\sum \{(d_{ij} - \hat{d}_{ij})^2 \mid i < j\}}{\sum \{d_{ij}^2 \mid i < j\}} .$$

---

(1) Le choix de la configuration initiale est traitée au § 4.1. de ce chapitre.

Les disparités  $d_{ij}$  sont obtenues par régression monotone à l'aide de l'algorithme *des blocs*, à partir des distances  $d_{ij}$ , distances calculées sur l'image euclidienne considérée.

Cet algorithme procède de la manière suivante :

1) On ordonne les couples  $(i, i')$  suivant les valeurs croissantes des dissimilarités  $\delta_{ii'}$ . Soit  $(i_1, i'_1), (i_2, i'_2), \dots$ , la préordonnance obtenue.

2) On procède par va-et-vient :

– on commence par comparer  $d_{i_1 i'_1}$  à  $d_{i_2 i'_2}$  :

- si  $d_{i_1 i'_1} < d_{i_2 i'_2}$ , on compare  $d_{i_2 i'_2}$  à  $d_{i_3 i'_3}$
- sinon on remplace  $d_{i_1 i'_1}$  et  $d_{i_2 i'_2}$  par leur moyenne.

– on compare  $d_{i_2 i'_2}$  à  $d_{i_3 i'_3}$  :

- si  $d_{i_2 i'_2} < d_{i_3 i'_3}$ , on compare  $d_{i_3 i'_3}$  à  $d_{i_4 i'_4}$
- sinon on remplace  $d_{i_2 i'_2}$  et  $d_{i_3 i'_3}$  par leur moyenne et on compare  $d_{i_2 i'_2}$  à  $d_{i_1 i'_1}$ .

etc. . .

Cette procédure de va-et-vient conduit en fin d'algorithme à une partition des couples  $(i, i')$  en "blocs" tels que les disparités  $\hat{d}_{ii'}$  calculées dans chacun de ces blocs soient constantes, et égales à la moyenne des distances constituant le bloc.

Kruskal ayant observé un certain nombre d'exemples, a déterminé empiriquement une échelle de valeurs pour le stress. Celle-ci dépend évidemment du nombre de points considérés et de la dimension de l'espace où est reconstruite l'image euclidienne.

Lorsque le stress est jugé trop élevé, on améliore la configuration par la méthode du gradient.

Toute solution qui rend le stress S minimum vérifie les équations :

$$g_i^j = \frac{\partial S}{\partial x_i^j} = 0 \quad i = 1, 2, \dots, n \quad j = 1, 2, \dots, r$$

où  $x_i^j$  est la  $j^{\text{ème}}$  coordonnée du point  $x_i$ .

Cette condition est une condition nécessaire mais pas suffisante (cas d'un minimum local par exemple).

Pour obtenir une telle solution, on procède par approximations successives. Considérons le nuage  $\mathfrak{N}_K$  obtenu à l'étape  $K$  dont le stress est égal à  $S$ , et le moment d'inertie rendu égal à 1 par homothétie. Au nuage  $\mathfrak{N}_K$  est associé le tableau  $X$  de dimension  $(r, n)$  des  $r$  coordonnées des  $n$  points du nuage.

Les corrections qu'il faut apporter à ces coordonnées pour maximiser la diminution du stress sont définies par le gradient  $dX$  de composantes :

$$g_i^j = S \sum \{(\delta_i^k - \delta_i^l) \left[ \frac{d_{kl} - \hat{d}_{kl}}{S^*} - \frac{d_{kl}}{T^*} \right] \frac{x_k^j - x_l^j}{d_{kl}} \mid (k, l)\}$$

où

- $S^* = \sum \{(d_{kl} - \hat{d}_{kl})^2 \mid (k, l)\}$
- $T^* = \sum \{d_{kl}^2 \mid (k, l)\}$
- $\delta_i^k$  est le symbole de Kronecker.

La correction suivant  $dX$  se fait avec un pas de longueur  $\alpha$  et détermine un nouveau nuage :

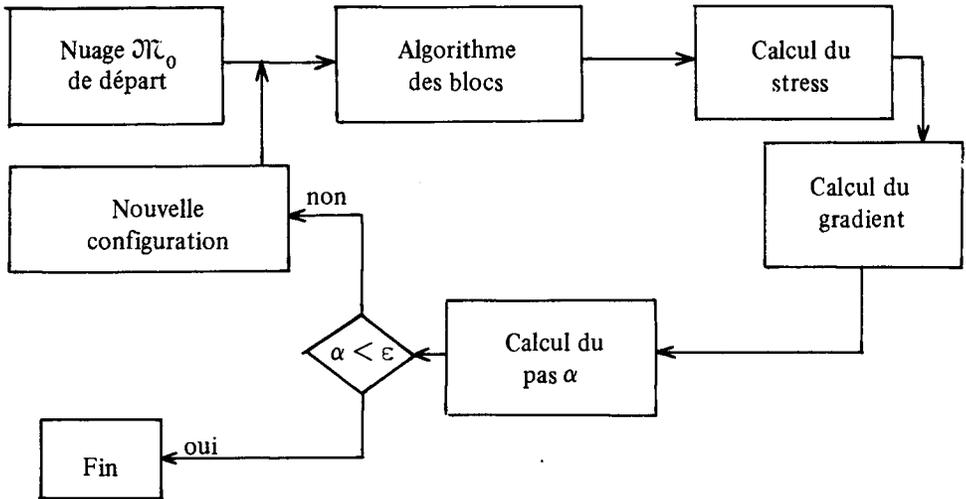
$$x_i^j = x_i^j - \frac{\alpha}{g_i^*} g_i^j \quad \text{où} \quad g_i^* = \frac{\sum \{(g_k^j)^2 \mid (k, l)\}}{\sum \{(x_k^l)^2 \mid (k, l)\}}$$

On obtient le nuage  $\mathfrak{N}_{K+1}$  de l'étape  $K + 1$  en effectuant une homothétie qui rend le moment d'inertie égal à 1.

Kruskal définit le pas  $\alpha$  de façon empirique comme une fonction :

- de la norme du gradient à l'itération précédente,
- de l'angle entre les gradients de l'itération en cours et de l'itération précédente,
- du pas de l'itération précédente,
- du stress de l'itération précédente.

Le processus pour obtenir un nuage de points à stress minimum est résumé dans le schéma suivant :



Les critères qui arrêtent la procédure sont basés sur :

- la valeur du stress,
- la valeur du pas,
- le nombre d'itérations.

### 3.3. Méthodes de Johnson et de McGee

Les principes sont exactement identiques, chaque auteur utilise l'algorithme du gradient sur son indice d'adéquation pour améliorer la configuration obtenue. Ainsi, seules les formules et les valeurs du pas proposé diffèrent d'un auteur à l'autre.

McGee fait sa correction dans la direction du gradient :

$$dX = \left( \frac{\partial W}{\partial x_i^j} \right) = (g_i^j) \text{ avec un pas } \alpha = \frac{W_k}{\tilde{g} \cdot Z_k},$$

où

- $W_k$  est la valeur de l'indice d'adéquation à l'étape  $k$
- $\tilde{g} = \sum \{(g_i^j)^2 \mid (i, j)\}$
- $Z_k$  est une constante dont la valeur initiale est 1 et qui double chaque fois que  $W_k$  est supérieur à  $W_{k-1}$ .

Le but du coefficient  $Z_k$  est d'amortir les oscillations autour d'un optimum ; en effet le pas diminue quand  $W$  augmente, c'est-à-dire quand  $W$  est passé par un minimum.

Johnson détermine le nuage  $\mathfrak{N}_k$  à partir du nuage  $\mathfrak{N}_{k-1}$  par la formule :

$$X_k = X_{k-1} - \theta_{k-1} G_{k-1}$$

où

- $X_k$  est la matrice des coordonnées  $x_i^j$  à l'étape  $k$
- $\theta_k$  est la valeur du coefficient d'accord de monotonie
- $G_k$  est la matrice des gradients calculée de la façon suivante : si on note

$$\theta^2 = \frac{U}{V}$$

et  $E_{ij}$  la matrice  $(n, n)$  définie par :

$$E_{ij} = (e_{kl}) \text{ avec } e_{kl} = \begin{cases} 1 & \text{si } (k, l) = (i, i) \text{ ou } (j, j) \\ -1 & \text{si } (k, l) = (i, j) \text{ ou } (j, i) \\ 0 & \text{sinon} \end{cases}$$

alors, pour toute itération, la formule générale du gradient est donné par :

$$G = (\Pi_U - \theta^2 \Pi_V) X$$

avec

$$\Pi_U = \frac{\partial U}{\partial X} = 4 \sum \{ \Delta_{ij}^{kl} (d_{ij}^2 - d_{kl}^2) (E_{ij} - E_{kl}) X \mid i < j ; k < l ; (i, j) \neq (k, l) \}$$

$$\Pi_V = \frac{\partial V}{\partial X} = 4 \sum \{ (d_{ij}^2 - d_{kl}^2) (E_{ij} - E_{kl}) X \mid i < j ; k < l ; (i, j) \neq (k, l) \}$$

### 3.4. Méthode de Guttman-Lingoes

La procédure de Guttman-Lingoes est de nature différente à cause des propriétés de l'indice d'adéquation.

On cherche le tableau  $X$  des coordonnées des  $n$  points de façon que les distances  $d$  et leurs images  $\tilde{d}$  par la transformation de rang maximisent le coefficient d'accord de monotonie  $\mu(d, \tilde{d})$ . Ce coefficient  $\mu$  est une forme quadratique des distances  $d_{ii}$ , puisque les  $\tilde{d}$  ne sont qu'une permutation des valeurs de  $d$ .

Classiquement, Guttman-Lingoes proposent un algorithme dit *en une phase* qui, itérativement, va conduire à un tableau  $X$  solution des équations :

$$\frac{\partial \mu}{\partial X} = 0.$$

En posant :

$$u = \langle d, d \rangle$$

$$v = \|\underline{d}\|^2 \quad (\|\tilde{d}\|^2 = \|\underline{d}\|^2 \text{ dans le cas de la transformation de rang})$$

alors

$$\mu(d, d) = \frac{u}{v} \quad (1),$$

et le tableau  $X$  formé des  $r$  vecteurs  $x$  des coordonnées des  $n$  points est solution de l'équation matricielle :

$$\frac{\partial \mu}{\partial X} = \frac{2}{v} (\Gamma - D_\mu) \cdot X = 0_{n \times r}$$

avec :

$$\Gamma = 2\mu(\underline{j}\underline{j}' - I) + D_{\bar{r}} - R$$

où

- $\underline{j}$  est le vecteur de  $\mathbf{R}^n$  dont toutes les coordonnées sont égales à 1
- $R$  est la matrice de terme général  $r_{kl} = (d_{\sigma_{kl}} + d_{\sigma_{kl}^{-1}}) / d_{kl}$ ,  $\sigma$  étant la permutation de rang
- $D_{\bar{r}}$  est la matrice diagonale de terme général

$$\bar{r}_i = \sum \{r_{ij} \mid j = 1, 2, \dots, n\}$$

- $D_\mu$  est la matrice scalaire de paramètre  $\mu$ .

---

(1) Remarque : les calculs développés ici supposent que l'information est complète, i.e. qu'il n'y a pas de données manquantes, et que la métrique euclidienne classique est utilisée pour mesurer les distances sur l'image reconstruite. On trouve dans [18] les formules établies dans le cas général (tableau de distances incomplet et non-symétrique, métrique quelconque) ; on peut constater que les formules deviennent très compliquées. Il nous a semblé plus intéressant de déterminer ces formules "simplifiées" qui mettent en évidence le processus utilisé.

Pour obtenir la solution, il faut donc déterminer la matrice  $X$  de dimension  $(n, r)$  telle que :

$$D_\mu \cdot X = \Gamma \cdot X$$

Le problème fait penser à la recherche du régime stationnaire d'une matrice, mais les matrices  $\Gamma$  et  $D_\mu$  sont, dans ce cas, fonction de  $X$ . La solution est obtenue par une procédure itérative :

$$X^{(t+1)} = D_{1/\mu} (t) \cdot \Gamma^{(t)} \cdot X^{(t)}$$

où  $\mu^{(t)}$  et  $\Gamma^{(t)}$  sont calculés à partir de  $X^{(t)}$ .  $\Gamma^{(t)}$  étant une matrice de Gram dans le cas euclidien, la convergence est ainsi assurée. Dans le cas général, la convergence n'est pas automatique car  $\Gamma^{(t)}$  n'a pas forcément ses  $r$  plus grandes valeurs propres en valeur absolue, positives.

La solution ainsi obtenue n'est pas nécessairement optimale (maximum relatif). Pour remédier à cet inconvénient, Guttman et Lingoes proposent la méthode dite *en deux phases*, qu'ils assurent être optimale. C'est une procédure doublement itérative. A la fin de l'itération  $K - 1$  on a obtenu un tableau  $X^{(K-1)}$ , donc un tableau de distances  $D^{(K-1)}$ . Dans la première phase, on calcule les distances  $\tilde{d}^{(K)}$  par la transformation de rang à partir de  $d^{(K-1)}$ . Dans la deuxième phase, on cherche  $\tilde{d}^{(K)}$  qui minimise  $\mu(d^{(K)}, \tilde{d}^{(K)})$  en considérant  $\tilde{d}^{(K)}$  comme constant, ce qui entraîne que  $\mu^{(K)}$  n'est plus maintenant qu'une forme linéaire des distances  $d^{(K)}$ . Pour obtenir la solution de cette seconde phase, une procédure itérative, analogue à celle exposée ci-dessus dans la méthode en une phase, est nécessaire. L'équation matricielle qui permet de calculer la matrice  $X^{(K)}$ , solution de l'étape  $K$ , est de la même forme que ci-dessus :

$$\frac{\partial \mu}{\partial X} = 2 (v_K w)^{-1/2} (\Gamma^{(K)} - D_\alpha^{(K)}) \cdot X^{(K)} = 0_{n \times r},$$

mais maintenant :

$$\Gamma^{(K)} = \frac{u_K}{w} (\underline{j} \underline{j}' - I) - R^{(K)} + D_{r_K}^-$$

avec

- $u_K = \langle d^{(K)}, \tilde{d}^{(K)} \rangle$
- $R^{(K)}$  matrice de terme général  $r_{kl}^{(K)} = \tilde{d}_{kl}^{(K)} / d_{kl}^{(K)}$
- $D_{r_K}^-$  matrice diagonale de terme général

$$\bar{r}_i^{(K)} = \sum \{ r_{il}^{(K)} \mid l = 1, 2, \dots, n \}$$

- $D_\alpha^{(K)}$  matrice scalaire de paramètre  $(n - 1) \frac{u_K}{w} = \alpha$ .

La solution est alors donnée itérativement par une méthode du gradient :

$$X^{(K,t+1)} = X^{(K,t)} + \frac{1}{2} (u_{K,t} w_{K,t})^{1/2} D_{1/\alpha}^{(K,t)} \left[ \frac{\partial \mu}{\partial X} \right]_{K,t}$$

Les mêmes problèmes de convergence que dans le cas d'une seule phase se posent, toujours résolus dans le cas euclidien.

L'intérêt caractéristique de cette procédure en deux phases est de converger vers une solution qui satisfait deux conditions :

- dans la première phase,  $\tilde{d}$  doit être voisin de  $d$ , donc la distance reconstruite être presque en accord avec la préordonnance initiale ;
- dans la deuxième phase,  $\tilde{d}$  doit être voisin de  $d$ , de manière à optimiser l'indice  $\mu$ .

Si on atteint effectivement un maximum pour  $\mu$  au cours de la seconde phase, et qu'il reste invariant par la première phase, Guttman prétend que ce maximum est absolu puisque la préordonnance est respectée.

## 4. PROBLEMES PARTICULIERS

### 4.1. Configuration initiale

Toutes les méthodes précédentes impliquent le choix d'une configuration initiale qui sert de point de départ. Ce choix peut conduire à l'obtention d'un optimum relatif éloigné de l'optimum absolu. Plus la configuration initiale est éloignée de la configuration optimum sous-jacente, plus le processus est en général long et coûteux.

Le choix de la configuration initiale résulte d'un compromis entre son coût d'obtention et sa qualité mesurée par le critère d'adéquation.

Deux solutions sont généralement envisagées :

- a) configuration initiale tirée au hasard par tirage aléatoire des coordonnées dans une loi uniforme sur  $[-1, +1]$ ,
- b) configuration fournie par l'analyse factorielle sur tableau de distances des dissimilarités brutes.

Cette dernière solution permet de conserver au mieux les dissimilarités de départ, puisque son objectif est précisément de déterminer une configuration telle

que les distances entre points soient proches des dissimilarités initiales. Cette procédure est évidemment plus longue que le tirage au hasard mais donne en général une configuration meilleure au sens du critère de référence.

D'autre part, les solutions initiales obtenues par tirage au hasard conduisent souvent à des optimum locaux du fait de leur particularité. On donne sur la figure 4 une illustration de cette situation où, à partir d'un tirage aléatoire, MDSCAL construit une carte de France "vrillée".

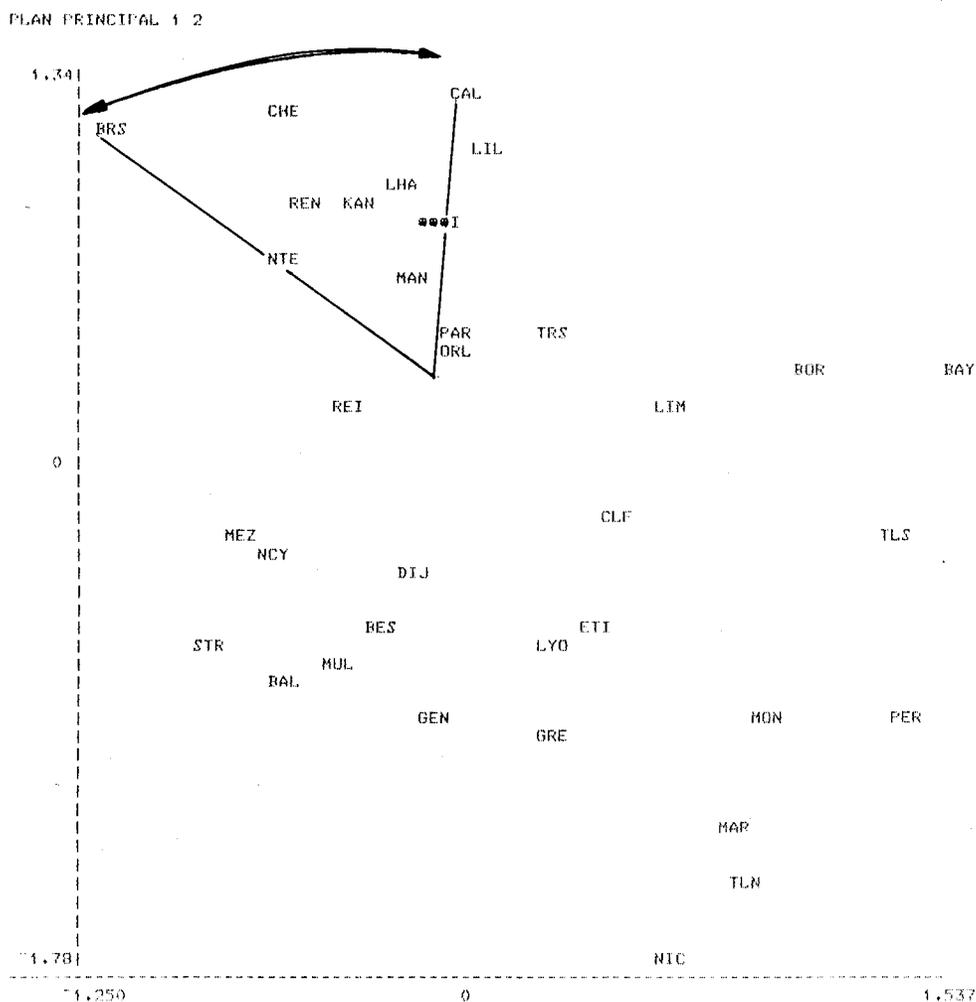


Figure 4 – Optimum local obtenu à partir d'une configuration initiale tirée au hasard

Le tableau 1 permet de comparer les valeurs du stress suivant que la configuration initiale a été obtenue par tirage au hasard ou par l'analyse factorielle sur tableau de distances, pour huit traitements différents. On constate que, sauf dans un cas, on a toujours obtenu un stress meilleur en partant de l'image euclidienne du tableau initial des dissimilarités.

Enfin, si on compare les nombres d'itérations nécessaires pour atteindre cette valeur du stress suivant la configuration initiale adoptée, on constate que leur rapport varie de 1 à 5 avec une valeur moyenne de 3,3 (tableau 1).

Tableau 1

Exemples	Solution initiale par AFTD		Solution initiale aléatoire	
	Stress	Nombre d'itérations	Stress	Nombre d'itérations
1	.04	4	.30	16
			.30	15
2	.05	4	.10	15
			.14	20
3	.06	17	.18	19
4	.08	4	.27	20
5	.08	5	.13	9
6	.04	4	.29	12
			.10	15
7	.24	7	.30	16
			.11	15
8	.13	5	.21	18

Le coût d'obtention d'une solution initiale tirée au hasard est négligeable (40 à 60 U.C.(1)). Celui d'une itération de MDSCAL est beaucoup plus important (13 000 à 16 000 U.C.). Le coût d'obtention de l'image initiale par l'analyse factorielle sur tableau de distances est toujours inférieur, sur les traitements-tests, à celui d'une itération, son coût moyen étant égal à la moitié de celui d'une itération.

(1) U.C. : temps d'unité centrale mesuré en soixantièmes de seconde sur I.B.M. 360-75, système A.P.L.

Les résultats précédents montrent qu'on a donc intérêt à préférer une configuration initiale obtenue par l'analyse factorielle sur tableau de distances, à la fois du point de vue du coût des calculs et de celui de la qualité de la représentation, sauf si la dimension du tableau de dissimilarités est trop importante. Dans ce dernier cas, on peut d'ailleurs remarquer que l'analyse factorielle peut ne porter que sur une partie des individus (cf. chap. II, 2.3.).

Guttman [17] a proposé de construire la configuration initiale en faisant l'analyse factorielle du tableau  $C$  défini par :

$$c_{ij} = 1 - \frac{r_{ij}}{N} + \Delta_{ij} \sum \left\{ \frac{r_{ij}}{N} \mid j = 1, 2, \dots, n \right\}$$

où

- $r_{ij}$  est le rang du couple  $(i, j)$  dans la préordonnance de départ,
- $N = n(n-1)/2$  est le nombre d'éléments de la préordonnance,
- $\Delta_{ij}$  est le symbole de Kronecker.

Ce tableau  $C$  lui a déjà permis de démontrer son théorème prouvant l'existence d'une image euclidienne (cf. § 1 de ce chapitre) ; il lui sert aussi pour estimer la dimension de l'espace où se trouve la configuration (cf. § 4.2. de ce chapitre). Il aboutit ainsi à une configuration évidemment semblable et de qualité équivalente à celle obtenue par l'analyse factorielle sur le tableau des dissimilarités brutes.

#### 4.2. Dimension de l'espace où est recherchée l'image

Dans toutes les méthodes, la dimension  $r$  de l'espace dans lequel est reconstruite l'image euclidienne respectant au mieux la préordonnance, doit être fixée au départ.

Il est évident que plus  $r$  est grand, meilleure est l'adéquation de la configuration obtenue. On sait, d'après les théorèmes de Shepard et de Guttman que, si  $r$  est égal à  $n-1$  ou  $n-2$  suivant la nature des données, le problème possède une solution exacte.

En pratique,  $r$  est choisi de faible valeur (souvent 2 ou 3) car, comme dans toutes les méthodes d'analyse de données, l'interprétation des résultats n'est possible que dans un espace de faible dimension.

La valeur brute de l'indice d'adéquation ne peut être un critère suffisant pour le choix de  $r$ , car celui-ci est également fonction du nombre  $n$  d'individus.

Un certain nombre d'auteurs ont proposé une méthode probabiliste. Ils construisent un test en supposant l'indice d'adéquation aléatoire, obéissant à une loi de probabilité qu'ils déterminent. Généralement, la loi est déterminée en supposant que les carrés des distances suivent une loi du khi-deux ; c'est le cas par exemple de McGee (cf. 2.2. ci-dessus). La valeur  $r$  est alors déterminée comme le plus petit nombre qui permet à l'indice d'adéquation d'avoir une probabilité suffisamment grande d'être réalisé.

Guttman, lui, base son estimation sur les valeurs propres de sa matrice  $C$  construite à partir des rangs dans la préordonnance (cf. § 1 et 4.1. de ce chapitre). Cette méthode repose donc sur le pourcentage d'inertie reconstruite et se rapproche ainsi de l'évaluation de la qualité de la représentation faite en analyse factorielle sur tableau de distances.

D'autres auteurs ont procédé en effectuant des simulations. Ainsi, à propos de la technique MDSCAL pour laquelle Kruskal n'a fait aucune hypothèse probabiliste, Klahr [22] a obtenu une série d'abaques qui permettent de juger de la qualité de l'image d'après la valeur du stress en fonction du nombre de points. Sherman [46] a repris cette étude plus récemment. Spence [47] a fait le même genre d'étude simultanément à propos de MDSCAL, de SSA de Guttman et de TORSCA de Torgerson.

Guttman donne un majorant pour la valeur de  $r$ , en prétendant qu'il ne doit pas être supérieur au nombre de valeurs propres positives de la matrice  $C$ , ce qui est en accord avec les résultats de l'analyse factorielle sur tableau de distances. Pratiquement (voir [51], [47], [46], [22]), on constate que l'amélioration de l'indice d'adéquation devient négligeable lorsque  $r$  est supérieur à 4.

### 4.3. Nombre d'itérations

Le nombre d'itérations est généralement un critère d'arrêt des procédures. Il doit être un compromis entre le coût des calculs et la qualité de l'image obtenue. On observe que cette qualité augmente très lentement au voisinage d'un optimum. Dans ces conditions, il semble préférable d'effectuer plusieurs essais à partir d'images initiales différentes avec un petit nombre d'itérations afin d'obtenir les configurations voisines de différents optimums locaux plutôt que d'effectuer un seul essai avec un grand nombre d'itérations qui donnerait exactement la solution correspondant à un optimum local.

L'expérience prouve que la convergence est rapide : on peut donc se contenter d'un nombre maximum d'itérations faible. Dans le programme correspondant à la méthode MDSCAL que nous avons mis au point, ce nombre maximum a été fixé à 20.

#### 4.4. Données manquantes

Les différentes méthodes n'imposent pas la connaissance de toutes les distances. Dans le cas où un certain nombre de ces données manquent, tous les auteurs ne calculent la qualité de la représentation que sur les seules données existantes.

En principe, l'aménagement de ces méthodes ne présente pas de difficultés théoriques : par exemple, on modifie simplement les formules en multipliant chaque distance par le "symbole de Kronecker" qui prend la valeur 1 si la distance a été mesurée et 0 dans le cas contraire. Toutefois les formules obtenues sont plus compliquées, elles cessent d'être linéaires par exemple.

On peut aussi estimer les données manquantes. Cette méthode a l'avantage d'éviter les complications des calculs. D'autre part, comme nous le verrons plus loin, les méthodes de positionnement multidimensionnel sont stables et robustes. Donc il nous semble préférable d'estimer les données manquantes, par exemple à l'aide de la formule :

$$d_{ij} = \frac{1}{2} (\text{Min} \{ |d_{ik} + d_{kj}| \mid k \neq i; k \neq j \} + \text{Max} \{ |d_{ik} - d_{kj}| \mid k \neq i; k \neq j \} )$$

donnant à  $d_{ij}$  une valeur qui satisfait l'inégalité triangulaire et qui, de plus, n'empêche pas *a priori* la distance obtenue d'être euclidienne.

#### 4.5. Données entachées d'erreurs

Pour les mêmes raisons que précédemment (robustesse et stabilité des méthodes, cf. 3.4.) il n'est pas en général très gênant d'avoir des mesures entachées d'erreurs.

En particulier, si les erreurs n'occasionnent aucune perturbation dans la préordonnance sur les individus, d'après le principe des méthodes, l'image obtenue est semblable.

Dans le cas où les erreurs n'occasionnent pas de permutations importantes dans cette préordonnance, l'image obtenue est en général équivalente.

Notons enfin une conséquence importante de ces méthodes.

Si une ou quelques distances sont anormalement grandes, ou petites (cas d'un individu aberrant, d'une erreur de mesure ou de retranscription, . . .), l'image obtenue n'est pas fortement modifiée, ou du moins les données anormales ne jouent pas un rôle prépondérant dans le résultat obtenu.

En analyse factorielle sur tableau de distances, au contraire, un point isolé peut fixer un axe principal à lui seul (cf. chap. II, 2.3.).

Les figures 5 et 6 illustrent ce phénomène ; on a cherché les images euclidiennes obtenues par l'AFTD et MDSCAL, pour le tableau des distances des 36 villes, où la distance Rennes-Caen a été arbitrairement multipliée par 10.

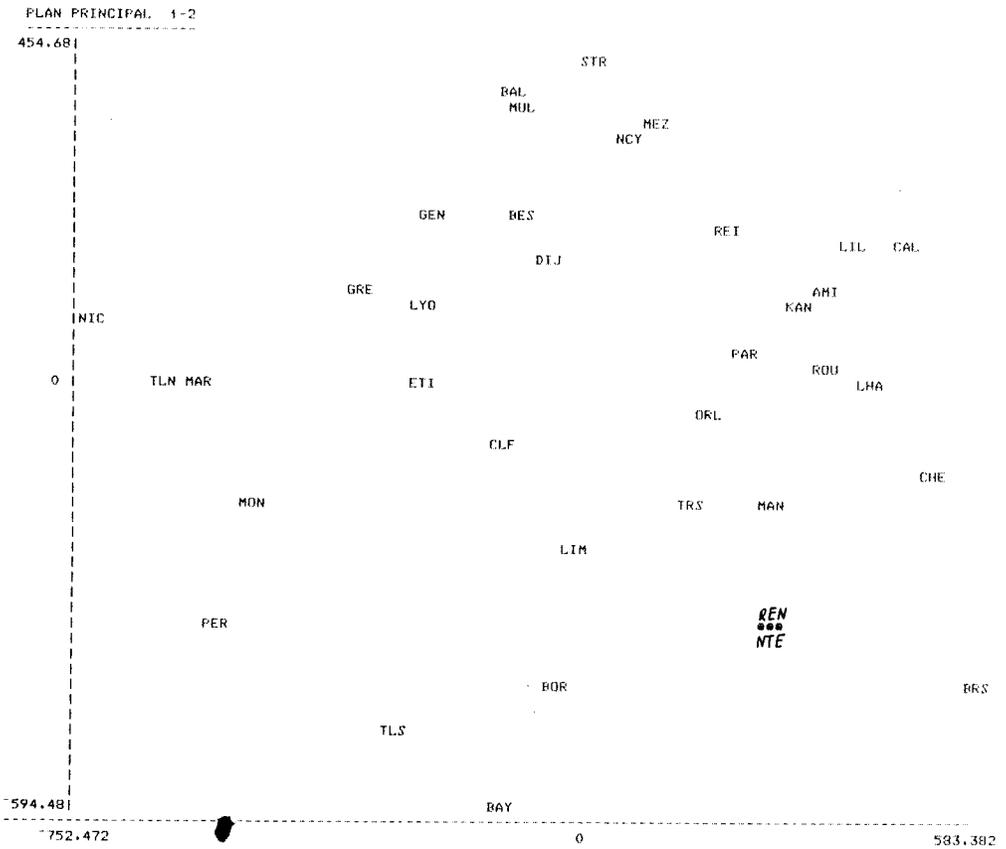


Figure 5 – AFTD sur le tableau des distances routières de 36 villes, la distance Rennes-Caen étant multipliée par 10.

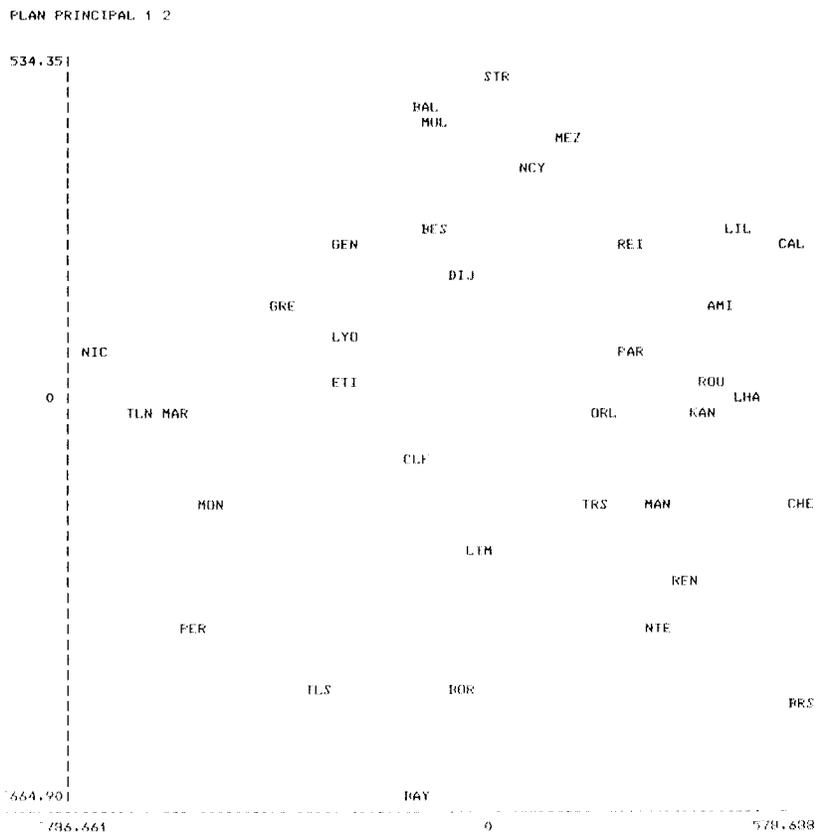


Figure 6 – MDSAL sur le tableau des distances routières de 36 villes, la distance Rennes-Caen étant multipliée par 10.

#### 4.6. Dissymétrie du tableau de dissimilarités

Dans le cas où les dissimilarités initiales  $\delta_{ij}$  et  $\delta_{ji}$  ne sont pas égales, on peut soit construire un tableau de dissimilarités symétrique en moyennant les dissimilarités initiales si on accorde peu d'importance à la dissymétrie, soit appliquer une méthode de positionnement en faisant intervenir dans les formules non pas les  $n(n-1)/2$  termes de la demi-matrice mais les  $n(n-1)$  termes non-diagonaux du tableau non symétrique.

Pour ce type de données, on conçoit que la qualité de l'image obtenue soit mauvaise puisque les axiomes de distance, sauf peut-être  $\delta_{ii} = 0$ , ne sont pas vérifiés. Une autre méthode, par exemple basée sur la théorie des graphes, serait sans doute préférable pour traduire cette dissymétrie.

#### 4.7. Traitement des ex-æquo

Dans le cas où certains couples de points ont la même dissimilarité (ex-æquo dans la préordonnance), ils nécessitent un traitement particulier. La colonne  $\underline{d}$  des  $n(n-1)/2$  dissimilarités  $\delta_{ij}$  peut, alors, être partitionnée en  $b$  blocs, les dissimilarités étant égales dans chaque bloc et croissantes d'un bloc à l'autre.

Kruskal [23] et [24] n'impose par forcément l'égalité des distances reconstruites pour les couples ex-æquo. Il propose donc deux approches :

– approche 1 : les distances reconstruites sont quelconques dans un bloc. La seule condition imposée est que :

$$\delta_{ij} < \delta_{kl} \Rightarrow \hat{d}_{ij} \leq \hat{d}_{kl}.$$

– approche 2 : les distances reconstruites doivent être égales dans un bloc. Les conditions imposées sont alors :

$$\begin{aligned} \delta_{ij} < \delta_{kl} &\Rightarrow \hat{d}_{ij} < \hat{d}_{kl} \\ \text{et} \quad \delta_{ij} = \delta_{kl} &\Rightarrow \hat{d}_{ij} = \hat{d}_{kl}. \end{aligned}$$

Dans les deux cas, la formule du stress n'a pas à être modifiée puisqu'elle traduit l'inconvénient à ne pas respecter l'ordonnance initiale.

Dans le premier cas, on applique directement l'algorithme des blocs de Kruskal pour effectuer la régression monotone.

Dans la deuxième approche, on effectue la régression monotone non pas sur le vecteur  $\underline{d}$ , mais sur celui des  $b$  valeurs, communes à chacun des blocs, et on donne à chaque couple du bloc la valeur obtenue par cette régression. J. de Leeuw [29] a montré que la solution obtenue dans chacun des cas est bien la solution optimale du problème considéré.

J. de Leeuw propose une troisième approche en imposant seulement que l'ordre des moyennes des distances des blocs soit respecté, mais il n'impose pas de conditions sur les valeurs de ces distances. On trouve dans [29] la solution de la régression monotone pour ce problème.

Johnson [21] adopte les deux approches de Kruskal. Il modifie comme suit le calcul du coefficient  $\theta^2$ .

Pour l'approche 1, il ne fait intervenir la différence  $(d_{ij} - d_{kl})^2$  ni au dénominateur ni au numérateur si on a l'égalité des dissimilarités initiales  $\delta_{ij}$  et  $\delta_{kl}$ . Au contraire, pour l'approche 2, cette différence intervient au dénominateur comme au numérateur dans le calcul de  $\theta^2$  si elle est non nulle.

Guttman [17] considère trois stratégies pour le traitement des ex-aequo, basées sur le nombre  $b_1$  de valeurs distinctes des dissimilarités  $\delta$ , et le nombre  $b_2$  de valeurs distinctes de distances reconstruites  $d$  :

– monotonie forte : on a les deux conditions satisfaites simultanément

$$\delta_{ij} < \delta_{kl} \Rightarrow d_{ij} < d_{kl}$$

$$\delta_{ij} = \delta_{kl} \Rightarrow d_{ij} = d_{kl}$$

i.e.  $b_1$  est égal à  $b_2$ .

– monotonie semi-forte : Guttman permet la non-égalité des distances reconstruites pour les couples ex-aequo. On a alors

$$b_1 \geq b_2.$$

– monotonie faible : Guttman permet l'égalité des distances reconstruites pour des couples situés à des rangs différents dans la préordonnance. On a alors :

$$b_1 \leq b_2.$$

On peut remarquer que, dans ce cas de la faible monotonie, on a une solution triviale qui consiste à placer les  $n$  points aux sommets de l'hyperpyramide régulière. Plus généralement, le problème possède alors des optimum locaux. D'autre part, les solutions obtenues ont tendance à être situées dans un espace de dimension plus importante que dans les autres cas.

En ce qui concerne la transformation de rang, elle est définie dans les trois cas de la façon suivante :

– monotonie forte  $\tilde{d}_{ij} = \Sigma \{\bar{d}_r e_{ijr} \mid r = 1, \dots, b_1\}$  où  $\bar{d}_r$  est la moyenne des distances du  $r^{\text{ème}}$  bloc

$$e_{ijr} = \begin{cases} 1 & \text{si le couple } (i, j) \text{ est dans le bloc } r \\ 0 & \text{sinon} \end{cases}$$

– monotonie semi-forte  $\tilde{d}_{ij} = \Sigma \{\hat{d}_p e_{ijp} \mid p = 1, \dots, b_1 (= b_2)\}$  où  $\hat{d}_p$  est la  $p^{\text{ème}}$  valeur des  $d_{ij}$  reconstruits

$$\text{et } e_{ijp} = \begin{cases} 1 & \text{si le couple } (i, j) \text{ est au } p^{\text{ème}} \text{ rang} \\ 0 & \text{sinon} \end{cases}$$

– monotonie faible  $\tilde{d}_{ij} = \Sigma \{\bar{d}_r e_{ijr} \mid r = 1, \dots, b_2\}$  où  $\bar{d}_r$  est la moyenne des valeurs dans le  $r^{\text{ème}}$  bloc

$$\text{et } e_{ijr} = \begin{cases} 1 & \text{si le couple } (i, j) \text{ est dans le } r^{\text{ème}} \text{ bloc} \\ 0 & \text{sinon} \end{cases}$$

#### 4.8. Robustesse et stabilité

On dira qu'une technique est *robuste* si des perturbations introduites au niveau de certaines données (valeurs aberrantes par exemple) influent peu sur les résultats ; qu'elle est *stable* si des perturbations "aléatoires" des données ont peu d'effets sur les résultats.

MDSCAL est robuste. On en a déjà montré un exemple au § 4.5. de ce chapitre. Cette propriété n'est pas vraie pour l'analyse factorielle sur tableau de distances, où l'effet de "bras de levier", dû à une erreur de mesure par exemple, peut fixer un des axes principaux. On trouve dans [2], p. 96, un exemple particulièrement flagrant de ce cas.

La stabilité de MDSCAL dépend du nombre de points. Dans le cas simplet traité sur les figures 7 et 8, on constate la grande liberté laissée pour le choix du

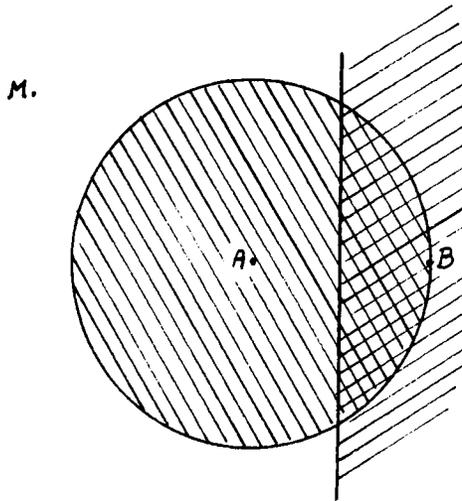


Figure 7 – Possibilité de choix pour le point M vérifiant  $MB > MA > AB$

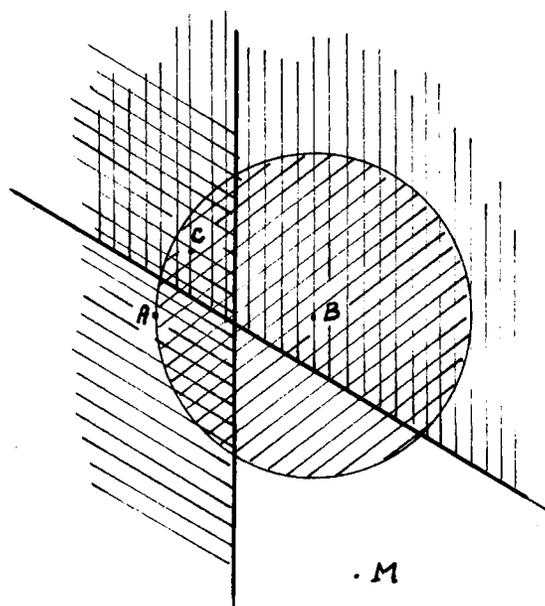


Figure 8 – Possibilité de choix pour le point  $M$  vérifiant  $MC > MA > MB > AB > BC > CA$

dernier point  $M$ . On imagine aisément que, si le nombre de points augmente, les possibilités de choix pour  $M$  diminuent. Sherman [44] a étudié et illustré ce problème à l'aide de simulations.

Signalons pour terminer qu'on peut effectuer une analyse des proximités dans le cas où le tableau des dissimilarités est incomplet. Presque tous les auteurs signalent ce fait, et donnent des formules adaptés à ce cas.

A titre d'exemple, et pour montrer l'avantage de MDSCAL par rapport à l'analyse factorielle sur tableau de distances sur ce point précis, nous exposons le traitement ci-dessous, reproduit de [2] :

25 points ont été tirés au hasard dans un plan et leurs distances (euclidiennes classiques) mutuelles calculées. Le graphique 9 reproduit la position de ces points.

Les valeurs de certaines cases, jouant le rôle de données manquantes, ont été remplacées par la valeur moyenne des distances restantes.

Le tableau 2 résume les résultats obtenus par MDSCAL pour différents tirages au hasard des "trous" (données supprimées) quand on recherche une configuration de dimension 2.

Tableau 2

N° Essai	15 Trous		30 Trous		45 Trous		60 Trous		75 Trous		90 Trous	
	Ité- ration	Stress										
1	16	0.093	16	0.113	15	0.102	15	0.116	17	0.132	15	0.124
2	16	0.075	16	0.064	16	0.105	16	0.110	17	0.133	15	0.139
3	15	0.044	16	0.082	16	0.120	17	0.128	16	0.122	20	0.136
4	17	0.053	15	0.091	15	0.090	15	0.120	16	0.129	15	0.115
5	15	0.057	15	0.098			16	0.113	15	0.129	17	0.125
6	16	0.067	16	0.107			17	0.119	16	0.133	16	0.141
7	15	0.057	16	0.094			16	0.117	16	0.136	16	0.136
8	16	0.076	15	0.080			15	0.132	19	0.123	15	0.147
9	16	0.075	16	0.096			15	0.113	15	0.115		
10			15	0.105			16	0.112				

Pour la configuration initiale on obtient à l'itération numéro 1 une valeur de stress de 0.008

1	2	3	4	5	6
		P			
		J			
7	8	9	10	11	12
		G		R	
			A		W
13	14	15	16	17	18
			M	H	D
	U				F
	K	S			I
19	20	21	22	23	24
	N	B	C		
		V		O	
				Y	
25	26	27	28	29	30
		L	X		
31	32	33	34	35	36
		E			

Figure 9 – Graphique de référence

On constate que, quel que soit l'essai, on obtient, même pour 90 trous (30 % de données manquantes), des configurations acceptables (robustesse) et équivalentes (stabilité).

Les graphiques 10 à 15 permettent de comparer des configurations obtenues par MDSCAL et par analyse factorielle sur le tableau des distances.

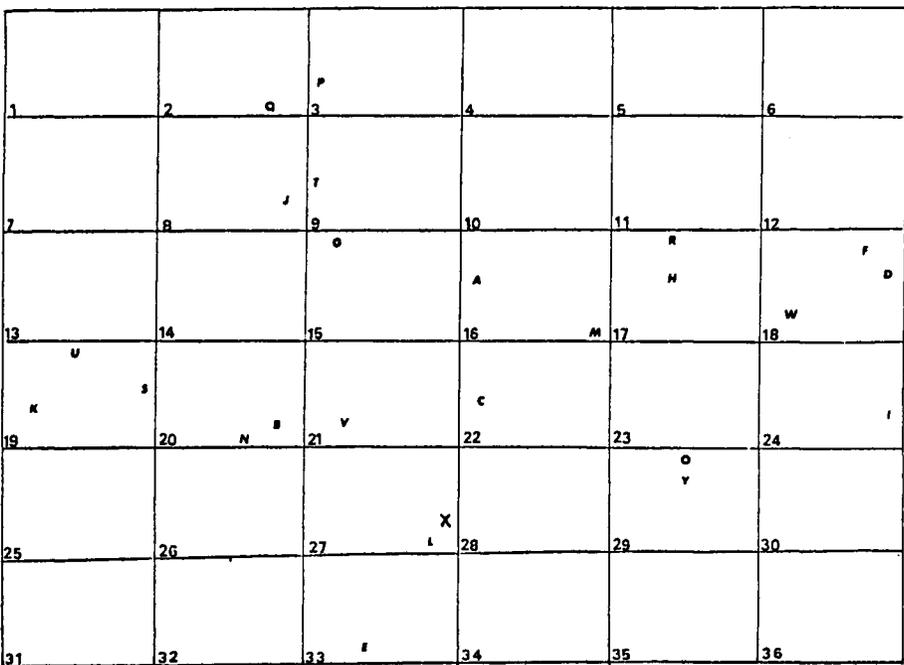


Figure 10 – MDSCAL : 30 Trous

1	Q				
2	P J T			R	W
3	G		A		D
4				M H	
5	S				
6	K U N				F
7					I
8		B V			
9			C		
10			O Y		
11					
12					
13					
14					
15			L X		
16					
17					
18					
19					
20					
21					
22					
23					
24					
25					
26					
27					
28					
29					
30					
31			E		
32					
33					
34					
35					
36					

Figure 11 – AFTD : 30 Trous

1					
2	Q	P			
3					
4					
5					
6					D
7		J T			
8				R	
9					
10				A M H	W
11					F
12					I
13				C	
14					
15					
16					
17					
18					
19					
20		S N B			
21			V		
22					
23				O	
24					
25					
26					
27					
28			X I		
29					
30					
31					
32					
33					
34					
35					
36					

Figure 12 – MDSCAL : 60 Trous

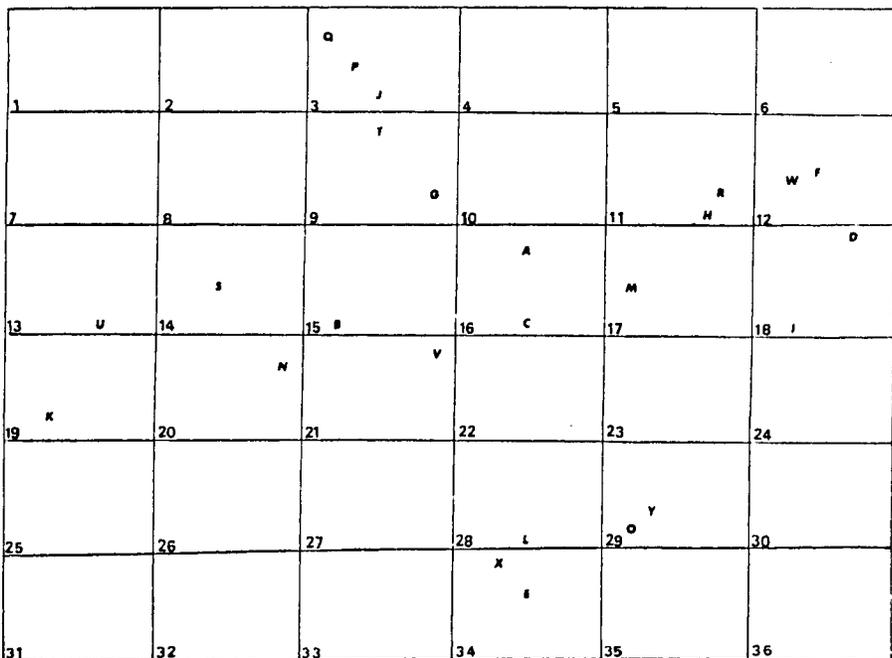


Figure 13 – AFTD : 60 Trous

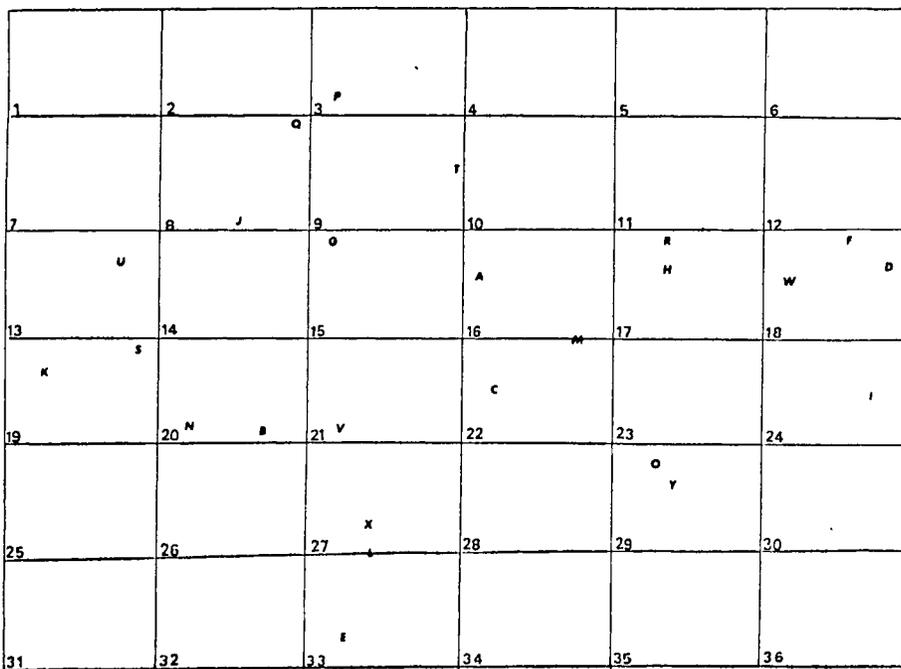


Figure 14 – MDSCAL : 90 Trous

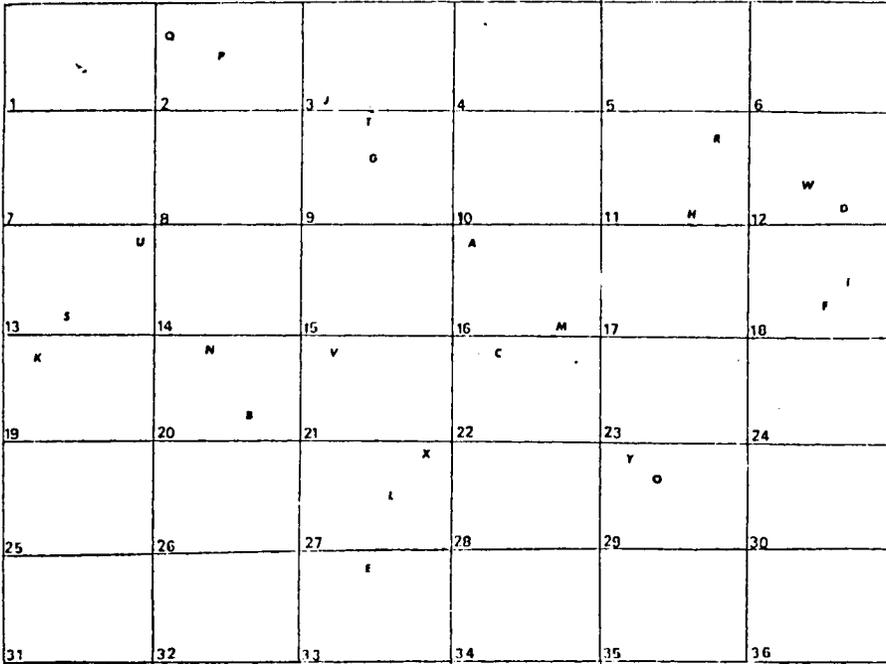


Figure 15 : AFTD : 90 Trous

**BIBLIOGRAPHIE**

- [1] BENZECRI J.P. et Coll. – “L’Analyse des Données”. Dunod, tome 1 et 2, Paris 1973.
- [2] BONNEFOUS S., CLOUTIER E., ROY E. – “Analyse Multidimensionnelle des Proximités : techniques, programmes APL, exemples”. Rapport CEA-R-4910, CEN Saclay, 1978.
- [3] BOUROCHE J.M. et Coll. – “Analyse des Données en Marketing”. *Mono-graphie de l’AFCEt*, Masson, Paris, 1977.
- [4] CAILLIEZ F., PAGES J.P. – “Introduction à l’Analyse des Données”. SMASH, Paris, 1976.

- [5] CARROLL J.D. — “A generalisation of canonical correlation analysis to three or more sets of variables”. *Proc. 76<sup>th</sup> Convention, American Psychology Association*, 1968, p. 227-228.
- [6] CARROLL J.D., CHANG J.J. — “Analysis of Individual Differences in Multidimensional Scaling”. *Psychometrika*, Vol. 35, n° 3, Sept. 1970, p. 283-319.
- [7] CARROLL J.D., WISH M. — “Models and methods for three-way multidimensional scaling”. *Contemporary Developpements in Mathematical Psychology*, W.H. Freeman, San Francisco, 1973.
- [8] CAZES P., BAUMERDER A., BONNEFOUS S., PAGES J.P. — “Codage et analyse des tableaux logiques. Introduction à la pratique des variables qualitatives”. *Cahier du BURQ n° 27*, Université Paris VI, Paris, 1977.
- [9] CAZES P., BONNEFOUS S., BAUMERDER A., PAGES J.P. — “Description cohérente des variables qualitatives prises globalement et de leurs modalités”. *Statistique et Analyse des Données*, Vol. 2, n° 3, 1978, p. 48.
- [10] DROUET D'AUBIGNY G. — “Description statistique des données ordinales : analyse multidimensionnelle”. Thèse de 3<sup>ème</sup> cycle, Université Scientifique et Médicale de Grenoble, Grenoble, 1975.
- [11] ESCOUFIER Y. — “Liaison entre groupes d'aléa”. *Revue de Statistique Appliquée*, Vol. 19, Paris, 1971.
- [12] ESCOUFIER Y. — “Vecteurs aléatoires équivalents du point de vue de l'ACP”. Rapport technique n° 7301, Université de Montpellier, 1973.
- [13] ESCOUFIER Y., PAGES J.P., CAILLIEZ F. — “Géométrie et techniques particulières en analyse factorielle”. Communication à European Meeting of Psychometrics and Mathematical Psychology, Université d'Uppsala (S), Juin 1978.
- [14] GREEN P.E. — “On the robustness of multidimensional scaling techniques”. *Journal of Marketing Research*, Vol. 12, Fev. 1975, pp. 73-81.
- [15] GREGSON R.A., RUSSEL P.N. — “A note on a generating assumption in McGee's multidimensional analysis of "elastic" distances”. *British Journal of Mathematical and Statistical Psychology*, Vol. 20, n° 2, 1967, pp. 239-242.
- [16] GUTTMAN L. — “The development of nonmetric space analysis”. *Multivariate Behavioral Research*, n° 2, 1967, pp. 71-82.
- [17] GUTTMAN L. — “A general non metric technique for finding the smallest coordinate space for a configuration of points”. *Psychometrika*, Vol. 33, n° 4, 1968, pp. 469-506.

- [18] GUTTMAN L., LINGOES J.C. – “Non metric factor analysis”. *Multivariate Behavioral Research*, 1967-2.
- [19] HOLLMAN E.W. – “The relation between hierarchical and euclidean models for psychological distances”. *Psychometrika*, Vol. 37, n° 4, 1972, pp. 472-486.
- [20] HOTELLING H. – “Relations between two sets of variables”. *Biometrika*, Vol. 28, 1936.
- [21] JOHNSON R.M. – “Pairwise non metric multidimensional scaling”. *Psychometrika*, Vol. 38, n° 1, 1973, pp. 11-18.
- [22] KLAHR D. – “A Monte-Carlo investigation of statistical significance of Kruskal’s non metric scaling procedure”. *Psychometrika*, Vol. 34, n° 3, 1969, pp. 319-329.
- [23] KRUSKAL J.B. – “Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis”. *Psychometrika*, Vol. 29, n° 1, 1964, pp. 1-27.
- [24] KRUSKAL J.B. – “Nonmetric multidimensional scaling : a numerical method”. *Psychometrika*, Vol. 29, n° 2, 1964, pp. 115-129.
- [25] KRUSKAL J.B., SHEPARD R.N. – “A nonmetric variety of linear factor analysis”. *Psychometrika*, Vol. 39, n° 2, 1974, pp. 123-157.
- [26] KETTENRING J.R. – “Canonical analysis of several sets of variables”. *Biometrika*, Vol. 58, n° 3, 1966.
- [27] LAFAYE DE MICHEAUX D. – “Analyse factorielle privilégiante”. Communication au Colloque IRIA-Analyse des données et Informatique, Versailles, Sept. 1977.
- [28] LAFAYE DE MICHEAUX D. – “Analyse factorielle privilégiante”. *Thèse de docteur-ingénieur*, Université de Nice, fév. 1978.
- [29] DE LEEUW J. – “Correctness of Kruskal’s algorithms for monotone regression”. *Psychometrika*, Vol. 42, n° 1, 1977, pp. 141-144.
- [30] L’HERMIER DES PLANTES. – “Structuration des tableaux à trois indices de la statistique : théorie et application d’une méthode d’analyse conjointe”. Thèse de 3<sup>ème</sup> cycle, Université des Sciences et Techniques du Languedoc, Montpellier, 1976.
- [31] LERMAN I.C., LEREDDE H. – “La méthode des pôles d’attraction”. *Colloque IRIA-Analyse des Données et Informatique*, Versailles, Sept. 1977, pp. 37-49.
- [32] LINGOES J.C. – “Some boundary conditions for a monotone analysis of symmetric matrices”. *Psychometrika*, Vol. 36, n° 2, 1971, p. 195-203.

- [33] MAILLES J.P., MAILLES D., BONNEFOUS S. — “Analyse des données et APL : techniques, bibliothèque, exemples”. *Rapport CEA-R-4753*, CEN Saclay, 1976.
- [34] McGEE V.E. — “The multidimensional analysis of “elastic” distances”. *British Journal of Mathematical and Statistical Psychology*, Vol. 19, Part 2, 1966, pp. 181-194.
- [35] McGEE V.E. — “A reply to some criticisms of elastic multidimensional scaling”. *British Journal of Mathematical and Statistical Psychology*, Vol. 20, n° 2, 1967, pp. 243-247.
- [36] PAGES J.P., ESCOUFIER Y., CAZES P. — “Opérateurs et analyse des tableaux à plus de deux dimensions”. *Cahier du BUR0 n° 25*, Université Paris VI, Paris, 1976.
- [37] C.R. RAO — “The use and interpretation of principal components analysis in applied research”. *Sankhya. Series A* — Vol. 26, Part 1, July 1964.
- [38] ROMEDER J.M. — *Méthodes et programmes d'analyse discriminante*. Dunod, Paris, 1973.
- [39] ROSKAM E.E. — “The methods of triads for nonmetric multidimensional scaling”. *Nederlands Tijdschrift voor de Psychologie*, Vol. 25, 1970, pp. 404-417.
- [40] SAPORTA G. — “Liaisons entre plusieurs ensembles de variables et codage de données qualitatives”. Thèse de 3<sup>ème</sup> cycle, Université de Paris VI, Paris, 1975.
- [41] SCHEKTMAN Y. — “Propriétés M-extrémales en analyse en composantes principales M — critères”. *Publication du laboratoire de statistique*, n° 01-78. Université Paul Sabatier, Toulouse, 1978.
- [42] SHEPARD R.N. — “The analysis of proximities : multidimensional scaling with an unknown distance function. I.” *Psychometrika*, Vol. 27, n° 2, 1962, pp. 125-140.
- [43] R.N. SHEPARD. — “The analysis of proximities : multidimensional scaling with an unknown distance function. II.” *Psychometrika*, Vol. 27, n° 3, 1962, pp. 219-246.
- [44] SHEPARD R.N. — “Representation of structure in similarity data : problems and prospects”. *Psychometrika*, Vol. 39, n° 4, 1974, pp. 373-421.
- [45] SHEPARD R.N., CARROLL J.D. — “Parametric representation of nonlinear data structures”. *Multivariate analysis*, Academic Press, New-York, 1966, p. 561.

- [46] SHERMAN C.R. — “Nonmetric multidimensional scaling : a Monte-Carlo study of a basic parameters”. *Psychometrika*, Vol. 37, n° 3, 1972, pp. 323-355.
- [47] SPENCE I. — “A Monte-Carlo evaluation of three nonmetric multidimensional scaling algorithms”. *Psychometrika*, Vol. 37, n° 4, 1972, pp. 461-486.
- [48] STEMMELEN E. — “Tableaux d'échanges : description et prévision”. *Cahier du BURO n° 28*, Université Paris VI, Paris, 1977.
- [49] STENSON H.H., KNOLL R.L. — “Goodness of fit for random rankings in Kruskal's nonmetric scaling procedure”. *Psychological Bulletin*, Vol. 71, n° 2, 1969, pp. 122-126.
- [50] TAKANE Y., YOUNG F.W. DE LEEUW J. — “Nonmetric individual differences multidimensional scaling : an alternating least squares method with optimal scaling features”. *Psychometrika*, Vol. 42, n° 1, 1977, pp. 7-67.
- [51] TORGERSON W.S. — “Theory and methods of scaling”. Wiley and Son, 1958.
- [52] WOLD H. — “Estimation of principals components and related models by iterative least squares”. *Multivariate Analysis*, New York Academic Press, New York, 1966, pp. 391-420.
- [53] WOLD H. — “Non linear estimation by iterative least squares procedure”. *Research Papers in Statistics*, Festschrift for J. Neyman, John Wiley, New York, 1966, pp. 411-444.
- [54] WOLD H. — “Non linear iterative partial least squares (NIPALS) estimation procedure”. *Congrès ISI RSS*, 1969.
- [55] YOUNG F.W., TORGERSON W.S. — “TORSCA, a Fortran IV program for Shepard-Kruskal multidimensional scaling analysis”. *Behavioral Science*, n° 12, 1967.

IMPRIMERIE LOUIS-JEAN

Publications scientifiques et littéraires

TYPO OFFSET

05002 GAP - Telephone 51 35 23 -

Dépôt légal 593-1980