REVUE DE STATISTIQUE APPLIQUÉE

L. BONIFAS
Y. ESCOUFIER
P. L. GONZALEZ
R. SABATIER

Choix de variables en analyse en composantes principales

Revue de statistique appliquée, tome 32, nº 2 (1984), p. 5-15
<http://www.numdam.org/item?id=RSA_1984__32_2_5_0>

CHOIX DE VARIABLES EN ANALYSE EN COMPOSANTES PRINCIPALES

L. BONIFAS, Y. ESCOUFIER, P.L. GONZALEZ, R. SABATIER

Unité de Biométrie 9, Place Pierre Viala 34000 Montpellier Cedex

RESUME

Un bref rappel des propriétés d'approximation de l'Analyse en Composantes Principales permet de poser clairement le problème de l'Analyse en Composantes Principales par rapport à des variables instrumentales et d'en expliciter les solutions. On en déduit une démarche de choix de variables en Analyse en Composantes Principales. Les résultats d'une application sont présentés.

INTRODUCTION

Cet article voudrait atteindre deux objectifs: en premier lieu, il souhaite contribuer à une promotion de l'emploi des méthodes "d'analyse en Composantes Principales par rapport á des variables instrumentales", et de "Choix de variables en Analyse en Composantes Principales". Pour cela il apporte les justifications mathématiques de ces méthodes et fait référence aux programmes disponibles auprès des auteurs.

Le second objectif est de rappeler clairement que la mise en œuvre d'une Analyse en Composantes Principales (ACP) repose non seulement sur des données, mais aussi sur le choix d'une métrique. Il en découle que toutes les procédures qui élaborent des cheminements compliqués pour choisir des variables en ACP puis mettent en œuvre l'ACP sur les variables choisies sans se poser la question de la métrique à utiliser avec ces variables, sont améliorables. Les résultats des paragraphes II et III montrent clairement la marche à suivre dans la résolution de ce problème.

I. RAPPEL SUR L'ANALYSE EN COMPOSANTES PRINCIPALES (ACP)

Disposant d'un tableau T, nxp, des mesures de p variables quantitatives faites sur n individus, le chercheur qui veut explorer les liaisons linéaires entre les variables et les ressemblances entre les individus peut faire appel à l'ACP. Pour mettre en œuvre cette méthode, il devra faire des choix que la présentation des techniques d'Analyse des Données popularisée en France par le livre de F. CAILLIEZ et J.P. PAGES [2], rend explicites.

Mots Clés: **Analyse en Composantes Principales, Analyse en Composantes Principales par rapport à des variables instrumentales, choix de Variables, Coefficient RV.**

L'étude des liaisons linéaires entre les variables suppose le calcul de coefficients de covariance et de corrélation, ce qui nécessite l'affectation de poids aux individus. Une matrice diagonale positive D telle que ${}^t\delta_n\, D\delta_n = 1$ où δ_n est un vecteur de $\mathbf{R}^n$ à n composantes égales à l'unité, définit ces poids.

De même, l'étude des ressemblances entre les individus suppose une quantification de ces ressemblances par le calcul de distances euclidiennes entre les individus rendu possible par le choix d'une métrique Q sur R^p. Ces choix faits et X étant le tableau des données centrées pour les poids $D(X = (I_n - \delta_n\, {}^t\delta_n D)T)$, il est possible de calculer les axes principaux de $\mathbf{R}^p$ et $\mathbf{R}^n$ définis par :

$$U_\alpha \in \mathbf{R}^p\ ;\ \alpha = 1, \ldots, p$$

tels que ${}^tX\, DXQU_\alpha = \lambda_\alpha U_\alpha$ avec ${}^tUQU = I_p$

où $$U = (U_1 \vdots U_2 \vdots \ldots U_p)$$

$$V_\alpha \in \mathbf{R}^n\ ;\ \alpha = 1, \ldots, p$$

tels que $XQ^tXDV_\alpha = \lambda_\alpha V_\alpha$ avec ${}^tVDV = I_p$

où $$V = (V_1 \vdots V_2 \vdots \ldots V_p)$$

De ces axes, on déduit les coordonnées principales

$$\varphi_\alpha = \sqrt{\lambda_\alpha}\, U_\alpha \in \mathbf{R}^p \quad \text{et} \quad \psi_\alpha = \sqrt{\lambda_\alpha}\, V_\alpha \in \mathbf{R}^n$$

(aussi appelées composantes principales pour ψ_α) qui vérifient :

$$\| {}^tXDXQ - \sum_{\alpha=1}^{q} \varphi_\alpha\, {}^t\varphi_\alpha Q \|^2 = \sum_{\alpha=q+1}^{p} \lambda_\alpha^2 \quad \text{où} \quad \|AQ\|^2 = \mathrm{Tr}\,[(AQ)^2]$$

$$\| XQ^tXD - \sum_{\alpha=1}^{q} \psi_\alpha\, {}^t\psi_\alpha D \|^2 = \sum_{\alpha=q+1}^{p} \lambda_\alpha^2 \quad \text{où} \quad \|AD\|^2 = \mathrm{Tr}\,[(AD)^2]$$

Ainsi, dans le cas $q = p$, les coordonnées principales dans $\mathbf{R}^p$ permettent une reconstitution exacte de ${}^tXDX$, matrice de variance des variables observées : la covariance des variables i et i' est égale au produit scalaire $\sum_{\alpha=1}^{p} \varphi_{\alpha i}\, \varphi_{\alpha i'}$ des vecteurs $(\varphi_{\alpha i};\ \alpha = 1, \ldots, p)$ et $(\varphi_{\alpha i'};\ \alpha = 1, \ldots, p)$.

De même, dans ce cas, les composantes principales permettent une reconstitution de XQ^tX, matrice des produits scalaires entre individus : la ressemblance entre les individus j et j' est quantifiée par la distance $\sum_{\alpha=1}^{p} (\psi_{\alpha j} - \psi_{\alpha j'})^2$ entre les points de coordonnées $(\psi_{\alpha j};\ \alpha = 1, \ldots, p)$, $(\psi_{\alpha j'};\ \alpha = 1, \ldots, p)$.

Si Q est l'identité de R^p et $D = (1/n) I_n$, les représentations obtenues en se limitant à (φ_1, φ_2) et (ψ_1, ψ_2) fournissent des approximations visuelles des covariances et des ressemblances entre individus.

On sait par ailleurs [10], [13] que pour q donné les approximations de ${}^tXDXQ$ et XQ^tXD fournies respectivement par $\sum_{\alpha=1}^{q} \varphi_\alpha {}^t\varphi_\alpha Q$ et $\sum_{\alpha=1}^{q} \psi_\alpha {}^t\psi_\alpha D$ sont les meilleures qui puissent fournir q vecteurs Q-orthogonaux de R^p d'une part et q vecteurs D-orthogonaux de R^n d'autre part quelles que soient les normes unitairement invariantes utilisées dans R^p ou R^n pour apprécier la qualité de ces approximations [14].

Un résultat analogue, connu sous le nom de "formule de reconstitution des données" existe pour X. Pour $Q = L\,{}^tL$, A de dimensions nxp et $\|A\|^2 = Tr({}^tAA)$, on a :

$$\left\| D^{1/2} XL - \sum_{\alpha=1}^{q} D^{1/2} \frac{\psi_\alpha {}^t\varphi_\alpha}{\sqrt{\lambda_\alpha}} L \right\|^2 = \sum_{\alpha=q+1}^{p} \lambda_\alpha$$

Il sera utile pour la suite de remarquer que :

$$\|XQ^tXD - \sum_{\alpha=1}^{q} \psi_\alpha {}^t\psi_\alpha D\|^2 = \|XQ^tXD\|^2 (1 - RV^2 (XQ^tXD, \sum_{\alpha=1}^{q} \psi_\alpha {}^t\psi_\alpha D))$$

où RV est le coefficient défini entre deux opérateurs D-symétriques de R^n, AD et BD par

$$RV(AD, BD) = \frac{Tr(ADBD)}{\|AD\|\ \|BD\|}$$

On remarque aisément que ce coefficient peut s'interpréter comme le cosinus des vecteurs de R^{nxn} dont les éléments sont respectivement fournis par ceux de AD et BD. De plus, la symétrie de A et B permet d'assurer la positivité de Tr(ABDB). Il en résulte que $0 \leqslant RV(AD, BD) \leqslant 1$. D'autres propriétés de ce coefficient sont exposées en (3).

II. L'ANALYSE EN COMPOSANTES PRINCIPALES PAR RAPPORT A DES VARIABLES INSTRUMENTALES

II.1. – D'après ce qui précède, à une étude (X, Q, D) nous savons associer un opérateur de R^n, $WD = XQ^tXD$ qui en est caractéristique au sens où ses vecteurs propres (ψ_α ; $\alpha = 1, \ldots, p$) convenablement normés permettent une représentation des individus dans laquelle les produits scalaires entre individus sont égaux aux éléments de W, tandis que pour $q < p$, les vecteurs ($\psi_\alpha, \alpha = 1, \ldots q$) permettent la meilleure représentation possible de W.

Ce point étant acquis, on peut alors s'interroger sur le problème suivant : une étude (X, Q, D) étant donnée et des données Y étant observées sur les mêmes individus munis des mêmes poids, peut-on trouver une matrice symétrique définie ou semi-définie positive, c'est-à-dire une métrique "au sens large" M, à utiliser avec Y pour que les études (X, Q, D) et (Y, M, D) conduisent aux mêmes opérateurs, donc aux mêmes composantes principales et par conséquent aux mêmes représentations des individus. La suite montrera que la réponse à cette question est en

général négative. La formulation adoptée permet cependant de bien mettre en évidence ce qu'il est possible de faire.

Supposons en effet que X soit de dimension n x p et Y de dimension n x q. Posons :

$$S_{YY} = {}^tYDY \quad \text{supposée inversible}$$

$$S_{XY} = {}^tXDY \text{ et } S_{YX} = {}^tS_{XY}$$

$$B = S_{YY}^{-1} S_{YX} Q S_{XY} S_{YY}^{-1}$$

Le calcul permet d'établir le résultat suivant

Lemme : Pour toute matrice symétrique R, qxq, on a :

$$\mathrm{Tr}\,(YB^tYDYR^tYD) = \mathrm{Tr}\,(XQ^tXDYR^tYD)$$

Il en découle :

Propriété 1 : Pour toute métrique M, qxq

$$\|XQ^tXD - YM^tYD\|^2 = \|XQ^tXD - YB^tYD\|^2 + \|YB^tYD - YM^tYD\|^2$$

Démonstration : On peut écrire

$$\|XQ^tXD - YM^tYD\|^2 = \|XQ^tXD - YB^tYD\|^2 + \|YB^tYD - YM^tTD\|^2 + 2\,\mathrm{Tr}\,[(XQ^tXD - YB^tYD)(YB^tYD - YM^tYD)]$$

et le résultat s'obtient en appliquant le lemme à R = B − M.

Cette propriété montre que la norme de la différence entre XQ^tXD et YM^tYD dépend de deux termes :

- *d'une part* $\|XQ^tXD - YB^tYD\|^2$ qui ne contient pas la métrique M. C'est dire que ce terme mesure ce qui est présent dans (X, Q, D) et ne pourra jamais être reconstruit à partir d'une étude bâtie sur Y.
- *d'autre part* $\|YB^tYD - YM^tYD\|^2$ qui dépend explicitement du choix de M, le meilleur choix étant B.

II.2.

Propriété 2 $\|XQ^tXD - YB^tYD\|^2 = \|XQ^tXD\|^2 (1 - RV^2\,(XQ^tXD,\ YB^tYD))$

Démonstration : Le lemme appliqué à la matrice B elle-même donne

$$\|YB^tYD\|^2 = \mathrm{Tr}\,[(YB^tYD)^2] = \mathrm{Tr}\,(XQ^tXDYB^tYD)$$

Il en découle :

$$\|XQ^tXD - YB^tYD\|^2 = \|XQ^tXD\|^2 - \|YB^tYD\|^2$$

d'où le résultat.

Posons $P_Y = YS_{YY}^{-1}\,{}^tYD$ le projecteur D-orthogonal sur le sous-espace E_Y de R^n engendré par les variables qui définissent Y.

$$YB^tYD = P_YXQ^tX^tP_YD$$

On trouve donc que la meilleure reconstitution de la représentation des individus fournie par (X, Q, D) sur la base des variables Y est fournie par l'étude $(P_Y X, Q, D)$ c'est-à-dire l'étude faite sur les projections orthogonales des variables X sur le sous-espace E_Y pour la métrique Q. Cette remarque donne un point de vue sur la régression peu souvent mis en avant. De plus si X se réduit à une seule variable, le calcul montre que $RV(X^tXD, YB^tYD)$ est égal au carré du coefficient de corrélation linéaire entre X et $P_Y X$, c'est-à-dire au carré du coefficient de corrélation multiple entre X et Y.

II.3. Intéressons-nous maintenant au second terme du second membre de l'égalité de la propriété 1.

Soit U la matrice q x r des r axes principaux de R^q associés aux plus grandes valeurs propres de l'étude (Y, B, D).

Propriété 3 : Parmi toutes les métriques M, q x q, de rang r le meilleur choix est fourni pour $M = BU^tUB$.

Démonstration : On sait que ${}^tYDYBU = U\Lambda$ avec ${}^tUBU = I_r$. Un calcul classique utilisé dans l'établissement des "formules de transition" permet d'établir que $YBU = \psi$ où ψ est la matrice n x r des r premières composantes principales de (Y, B, D). Choisissons alors $M = BU^tUB$.

On a $YM^tYD = \psi^t \psi D = \sum_{\alpha=1}^{r} \psi_\alpha {}^t\psi_\alpha D$. Un retour aux résultats rappelés dans le paragraphe I établit la propriété.

De plus, la remarque finale du paragraphe I permet d'écrire pour ce choix particulier de M :

$$\|YB^tYD - YM^tYD\|^2 = \|YB^tYD\|^2 (1 - RV^2(YB^tYD, YM^tYD))$$

Corollaire : Pour $M = BU\,{}^tUB$

$$\|XQ^tXD - YM^tYD\|^2 = \|XQ^tXD\|^2 (1 - RV^2(XQ^tXD, YB^tYD)\, RV^2(YB^tYD, YM^tYD))$$

Démonstration : on a

$$\begin{aligned}
&\|YB^tYD - YM^tYD\|^2 \\
&\quad = \|XQ^tXD\|^2 (1 - RV^2(YB^tYB, YM^tYD)) \\
&\quad = \|XQ^tXD\|^2 \frac{\|YB^tYD\|^2}{\|XQ^tXD\|^2} (1 - RV^2(YB^tYB, YM^tYD)) \\
&\quad = \|XQ^tXD\|^2 RV^2(XQ^tXD, YB^tYB)(1 - RV^2(YB^tYD, YM^tYD))
\end{aligned}$$

et le résultat en découle en utilisant les propriétés 1 et 2.

En conclusion, l'Analyse en Composantes Principales des données Y par rapport à l'étude de référence (X, Q, D) revient à réaliser l'ACP du triplet (Y, B, D) qui est équivalente du point de vue de la représentation des individus à l'ACP de $(P_Y X, Q, D)$. Un programme permettant la mise en oeuvre de cette méthode a été

intégré dans la bibliothèque ANADO disponible au CNUSC. Des copies peuvent être fournies par les auteurs.

L'annexe 2 de la thèse de R. SABATIER [13] fournit une description précise de ce programme.

Remarque 1

Posons $Z = S_{YY}^{-1} U$

Alors ${}^tYDYBU = U\Lambda$ avec ${}^tUBU = I_r$

implique $S_{YX}\, QS_{XY} Z = S_{YY} Z\Lambda$ avec ${}^tZS_{YX}\, QS_{XY} Z = I_r$

et ${}^tZS_{YY} Z = \Lambda^{-1}$

Les solutions de l'ACPVI telles que proposées pour $Q = I_p$ par RAO [9] correspondent donc à $Z\,\Lambda^{1/2}$. Elles conduisent à des composantes principales

$$\xi = YZ\,\Lambda^{1/2}.$$

Or

$$\xi = YZ\,\Lambda^{1/2} = Y\,S_{YY}^{-1} S_{YX}\, Q\, S_{XY}\, S_{YY}^{-1} U\,\Lambda^{-1/2}$$
$$= Y\,B\,U\,\Lambda^{-1/2} = \psi\Lambda^{-1/2}$$

Remarque 2

On peut écrire :

$$\mathrm{Tr}\,(YM{}^tYD) = \mathrm{Tr}\,(YBU{}^tUB{}^tYD)$$
$$= \mathrm{Tr}\,({}^tUBU\Lambda)$$
$$= \mathrm{Tr}\,(\Lambda^{1/2}\,{}^tZ{}^tYDXQ{}^tXDYZ\Lambda^{1/2})$$
$$= \mathrm{Tr}\,({}^t\xi DXQ{}^tXD\xi)$$

Si on note $\xi = (\xi_1, \ldots, \xi_q)$, $X = (X_1, \ldots, X_p)$ et Q_{ij} l'élément (i, j) de la métrique Q, on peut écrire :

$$\mathrm{Tr}\,(YM{}^tYD) = \sum_{\alpha=1}^{q} \sum_{i=1}^{p} \sum_{j=1}^{p} Q_{ij}\, \mathrm{cov}\,(\xi_\alpha, X_i)\, \mathrm{cov}\,(\xi_\alpha, X_j)$$

Si $Q = I_p$, on a donc la somme des carrés des covariances entre les variables ξ_α et X_i. On retrouve la présentation faite par OBADIA en termes de "composantes Explicatives" [8].

III. CHOIX DE VARIABLES

L'étude (X, Q, D) étant donnée, nous extrayons r colonnes de X qui permettent la construction d'une matrice notée $Y_{[r]}$. Soit $B_{[r]}$ la matrice à associer à $Y_{[r]}$ pour que les études (X, Q, D) et ($Y_{[r]}$, $B_{[r]}$, D) soient les plus proches possibles. $B_{[r]}$ a été définie au paragraphe précédent.

Soit $\mathcal{R}$ l'ensemble des matrices à r colonnes extraites de X, nous souhaitons résoudre le problème suivant :

$$\left\| \begin{array}{l} \text{Trouver } Y^*_{[r]} \in \mathcal{R} \text{ telle que} \\ RV(XQ^tXD, Y^*_{[r]} B^*_{[r]} {}^tY^*_{[r]} D) = \sup_{Y_{[r]} \in \mathcal{R}} RV(XQ^tXD, Y_{[r]} B_{[r]} {}^tY_{[r]} D) \end{array} \right.$$

D'après la propriété 2, ce problème est équivalent au suivant :

$$\left\| \begin{array}{l} \text{Trouver } Y^*_{[r]} \in \mathcal{R} \text{ telle que} \\ \| XQ^tXD - Y^*_{[r]} B^*_{[r]} {}^tY^*_{[r]} D \|^2 = \inf_{Y_{[r]} \in \mathcal{R}} \| XQ^tXD - Y_{[r]} B_{[r]} {}^tY_{[r]} D \|^2 \end{array} \right.$$

Pour r et p suffisamment petits, on peut toujours envisager le calcul exhaustif sur l'ensemble des possibilités offertes. Lorsque r et p deviennent raisonnablement grands, cette solution n'est plus possible et l'optimum ne pouvant être défini de façon analytique, on propose un algorithme progressif pour sa construction.

Le lemme permet d'écrire :

$$RV^2 (XQ^tXD, Y_{[r]} B_{[r]} {}^tY_{[r]} D) = \frac{Tr (XQ^tXDY_{[r]} B_{[r]} {}^tY_{[r]} D)}{\|XQ^tXD\|^2}$$

Posons ${}^tY_{[r]} D Y_{[r]} = L_{[r]} {}^tL_{[r]}$; on a :

$$Tr (XQ^tXDY_{[r]} B_{[r]} {}^tY_{[r]} D) = Tr [(L^{-1}_{[r]} {}^tY_{[r]} DXQ^tXDY_{[r]} ({}^tL_{[r]})^{-1})^2]$$

Choisissons pour $L_{[r]} {}^tL_{[r]}$ la décomposition de CHOLESKI de ${}^tY_{[r]} D Y_{[r]}$ ($L_{[r]}$ est triangulaire inférieure) et posons

$$A_{[r-1]} = L^{-1}_{[r-1]} {}^tY_{[r-1]} DXQ^tXDY_{[r-1]} ({}^tL_{[r-1]})^{-1}$$

Connaissant les $[r-1]$ premières variables retenues qui constituent $Y_{[r-1]}$, on connait $L^{-1}_{[r-1]}$ et il est facile d'en déduire $L^{-1}_{[r]}$ due à l'introduction d'une $r^{ième}$ ligne de la matrice symétrique $A_{[r]}$.

On a alors :

$$Tr [(A_{[r]})^2] = Tr [(A_{[r-1]})^2] + 2 \sum_{i=1}^{r} (A_{[r]})^2_{ri} - (A_{[r]})^2_{rr}$$

ce qui sert de base à la construction de l'algorithme progressif qui maximisera $Tr (XQ^tXDY_{[r]}B_{[r]} {}^tY_{[r]} D)$ donc $RV^2 (XQ^tXD, Y_{[r]}B_{[r]} {}^tY_{[r]}D)$.

Un programme est également disponible dans la bibliothèque ANADO (utilisable au CNUSC) ou auprès des auteurs. Il permet de s'occuper du choix des variables sans s'occuper du choix de $B_{[r]}$ option métrique identité) de limiter la recherche de la métrique à utiliser avec $Y_{[r]}$ à celle d'une métrique diagonale (option métrique diagonale) ; ou de résoudre le problème général tel qu'exposé ici. Le programme permet d'imposer la prise en compte de certaines variables dans le choix.

IV. EXEMPLE

Dans une étude concernant la population atmosphérique, on dispose d'un réseau de 18 stations de mesures numérotées de 1 à 18. On dispose de 562 mesures journalières de fumées noires concomittantes sur ces 18 stations qui permettent de calculer la matrice des corréaltions fournies à la fin de l'article.

Ayant décidé d'extraire de ce réseau, un réseau restreint à 10 stations, le programme de choix de variables décrit au paragraphe III fournit les résultats consignés dans la table I.

TABLE I

Station introduite	Valeur de RV après l'introduction de la station
14	0,866
17	0,923
18	0,943
1	0,968
5	0,976
12	0,982
15	0,986
4	0,989
16	0,991
11	0,992

La valeur du dernier coefficient RV obtenu est en elle-même une information sur la qualité de l'approximation faite qui s'interprète par référence à la propriété 2.

Dans la pratique, il apparait que donnée sous cette forme, la solution ne satisfait pas pleinement le spécialiste de la pollution qui la trouve peu explicite. Pour l'éclairer, plusieurs possibilités sont envisageables :

1) Utilisant le support visuel de l'implantation géographique du réseau initial, on fait apparaître le réseau restreint. Bien que certaines informations puissent être tirées de cette représentation, elle n'apporte aucune précision sur les raisons du choix.

2) Une seconde solution consiste à calculer les coefficients de corrélation multiple entre les variables non retenues et les variables retenues. On passe ainsi d'un critère RV global à des critères plus individuels. Les valeurs obtenues pour l'exemple étudié sont présentées dans la table II.

3) Une dernière solution, qui a l'avantage de mettre en évidence les contraintes d'évolutions du programme, consiste à réaliser une classification automatique des stations sur la base de la matrice des corrélations puis à positionner sur cette classification les stations retenues. On voit alors que le programme cherche à énumérer les variables dans un ordre tel que, à tout moment, les différents partitions apparues soient représentées de façon équitable.

TABLE II

Stations non retenues	Coefficients de corrélation multiple avec les stations retenues
2	0,91
3	0,89
6	0,86
7	0,94
8	0,93
9	0,89
10	0,96
13	0,83

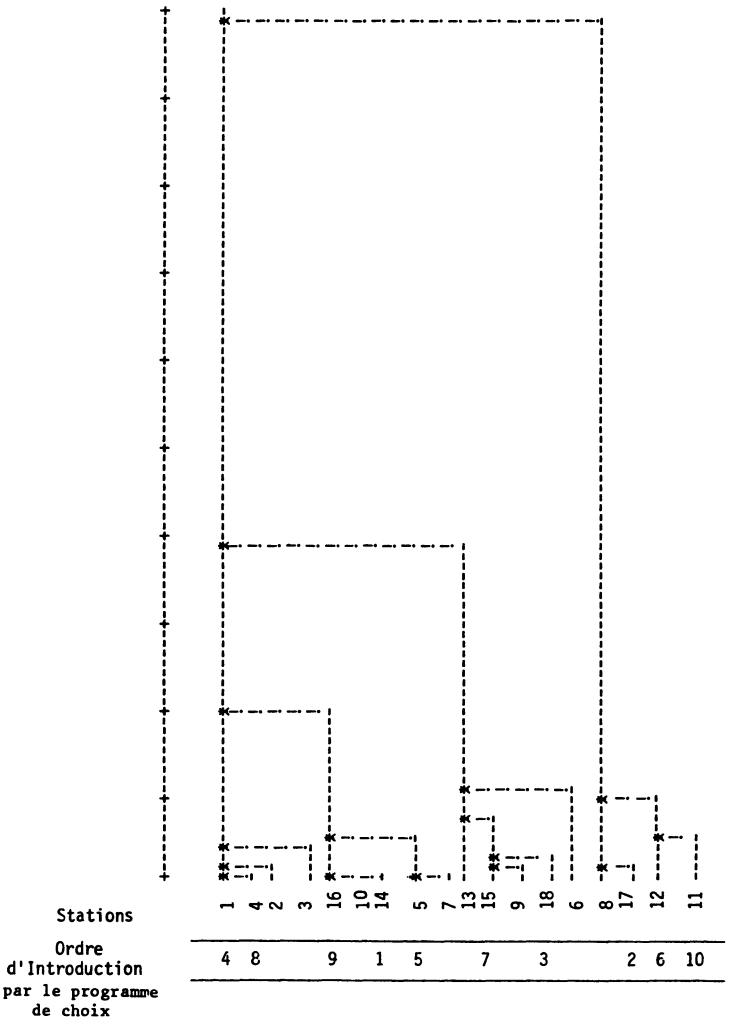

Figure 1. – Classification obtenue en appliquant le programme CAH2COORD du logiciel ADDAD à la matrice des corrélations

MATRICE DES CORRELATIONS

	1	2	3	4	5	6	7	8	9	10	11
1	1.0000	0.8774	0.8821	0.8981	0.8386	0.7466	0.8545	0.7609	0.7481	0.8410	0.7370
2	0.8774	1.0000	0.8328	0.8815	0.8257	0.7213	0.8223	0.7438	0.7476	0.8270	0.6924
3	0.8821	0.8328	1.0000	0.8392	0.7878	0.7049	0.8155	0.7276	0.6529	0.7752	0.7098
4	0.8981	0.8815	0.8392	1.0000	0.8789	0.7590	0.8641	0.7666	0.7830	0.8526	0.7378
5	0.8386	0.8257	0.7878	0.8789	1.0000	0.8551	0.8987	0.7479	0.7999	0.8645	0.6948
6	0.7466	0.7213	0.7049	0.7590	0.8551	1.0000	0.7988	0.6769	0.7179	0.7884	0.5881
7	0.8545	0.8223	0.8155	0.8641	0.8987	0.7988	1.0000	0.7717	0.7966	0.9020	0.7250
8	0.7609	0.7438	0.7276	0.7666	0.7479	0.6769	0.7717	1.0000	0.6544	0.7452	0.7833
9	0.7481	0.7476	0.6529	0.7830	0.7999	0.7179	0.7966	0.6544	1.0000	0.8540	0.5690
10	0.8410	0.8270	0.7752	0.8526	0.8645	0.7884	0.9020	0.7452	0.8540	1.0000	0.6980
11	0.7370	0.6924	0.7098	0.7378	0.6948	0.5881	0.7250	0.7833	0.5690	0.6980	1.0000
12	0.6777	0.6509	0.6617	0.6605	0.6249	0.5570	0.6883	0.8261	0.5358	0.6609	0.7512
13	0.7142	0.6982	0.6605	0.7401	0.7322	0.6217	0.7611	0.6164	0.7142	0.7976	0.6188
14	0.8254	0.8158	0.7623	0.8437	0.8522	0.7749	0.8879	0.7108	0.8603	0.9330	0.6902
15	0.7638	0.7351	0.7348	0.7427	0.7793	0.6885	0.8013	0.6522	0.8109	0.8708	0.5989
16	0.8201	0.7818	0.7715	0.8112	0.8280	0.7459	0.8767	0.7103	0.8331	0.9243	0.6843
17	0.7142	0.7026	0.6939	0.7325	0.7174	0.6499	0.7538	0.8753	0.6467	0.7202	0.7056
18	0.7860	0.7509	0.7023	0.7809	0.8152	0.7146	0.8098	0.6058	0.7840	0.8358	0.6118

	12	13	14	15	16	17	18
1	0.6777	0.7142	0.8254	0.7638	0.8201	0.7142	0.7860
2	0.6509	0.6982	0.8158	0.7351	0.7818	0.7026	0.7509
3	0.6617	0.6605	0.7623	0.7348	0.7715	0.6939	0.7023
4	0.6605	0.7401	0.8437	0.7427	0.8112	0.7325	0.7809
5	0.6249	0.7322	0.8522	0.7793	0.8280	0.7174	0.8152
6	0.5570	0.6217	0.7749	0.6885	0.7459	0.6499	0.7146
7	0.6883	0.7611	0.8879	0.8013	0.8767	0.7538	0.8098
8	0.8261	0.6164	0.7108	0.6522	0.7103	0.8753	0.6058
9	0.5358	0.7142	0.8603	0.8109	0.8331	0.6467	0.7840
10	0.6609	0.7976	0.9330	0.8708	0.9243	0.7202	0.8358
11	0.7512	0.6188	0.6902	0.5989	0.6843	0.7056	0.6118
12	1.0000	0.5358	0.6192	0.5530	0.6319	0.7588	0.5291
13	0.5358	1.0000	0.7898	0.7552	0.7802	0.6009	0.7429
14	0.6192	0.7898	1.0000	0.8384	0.9003	0.6836	0.8289
15	0.5530	0.7552	0.8384	1.0000	0.8603	0.6268	0.7657
16	0.6319	0.7802	0.9003	0.8603	1.0000	0.6908	0.8205
17	0.7588	0.6009	0.6836	0.6268	0.6908	1.0000	0.5869
18	0.5291	0.7429	0.8289	0.7657	0.8205	0.5869	1.0000

TRACE DE RR= 192.60521

Pour interpréter les résultats obtenus sur l'exemple traité et présentés dans la figure 1, on notera que les stations 14, 17 et 18 avaient été imposées pour des raisons extérieures à la logique mathématique.

*

L'application qui illustre cet article est tirée d'un travail réalisé dans le cadre d'une collaboration avec la direction environnement de la Société Nationale Elf Aquitaine. Les auteurs tiennent à remercier cette société pour le soutient qu'elle leur apporte.

BIBLIOGRAPHIE

[1] BERNARD C. DOCHI. – *Choix de variables en Analyse des Données*, 29 juin 1979, Thèse USTL Montpellier.

[2] F. CAILLIEZ, J.P. PAGES (1976). – *Introduction à l'Analyse des Données*, SMASH, 9, rue Duban, 75016 Paris.

[3] Y. ESCOUFIER (1973). – Le traitement des variables vectorielles, *Biometrics*, 29, p. 751-760.

[4] Y. ESCOUFIER, P. ROBERT (1979). – *Choosing variables and metrics* by optimizing the RV coefficient, *In: Optimizing methods in Statistic*. Ed. J.S. Rustagi Academic Press INC, p. 205-219.

[5] P.L. GONZALEZ (1981). – *Choix de variables, applications au choix de Stations dans un réseau*, Rapport Technique 8104, CRIG Montpellier.

[6] P.L. GONZALEZ. – *Analyse statistique de données psycho-sensorielles*, 19 avril 1982. Thèse USTL Montpellier.

[7] L. LEBART, A. MORINEAU, J.P. FENELON (1979). – *Traitement des données statistiques : Méthodes et Programmes*, Dunod.

[8] J. OBADIA (1978). – L'analyse en composantes explicatives, *R.S.A.*, XXVI, n° 4, p. 5-28.

[9] C.R. RAO (1964). – The use and the interpretation of principal component analysis in applied research, *Sankya, Sci. A.*, 26, p. 329-359.

[10] C.R. RAO (1980). – Matrix approximations and reduction of dimensionality in multivariate statistical analysis, *In: Multivariate Analysis V*. Ed. P.R. Krishnaiah, NHPC.

[11] P. ROBERT, Y. ESCOUFIER (1976). – A unifying tool for linear Multivariate statistical Methods: The RV coefficient, *Appl. Stat. C*, 25 (3), p. 257-265.

[12] R. SABATIER (1981). – *Two examples of choosing variables and metrics with the RV coefficient*, Rapport Technique 8102, CRIG Montpellier.

[13] R. SABATIER. – *Approximations d'un tableau de données. Application à la reconstitution des paléoclimats*, 21 mars 1983, Thèse USTL Montpellier.

[14] R. SABATIER, Y. JAN, Y. ESCOUFIER (1983). – *Approximations d'applications linéaires et Analyse en Composantes Principales*, 3e Journées Internationales Analyse des Données et Informatique, Ed. I.N.R.I.A. Domaine de Voluceau-Rocquencourt, B.P. 105, 78153 Le Chesnay Cedex.