

REVUE DE STATISTIQUE APPLIQUÉE

YVES ESCOUFIER

La dépendance de deux aléas vectoriels critères et visualisation

Revue de statistique appliquée, tome 21, n° 2 (1973), p. 5-16

http://www.numdam.org/item?id=RSA_1973__21_2_5_0

© Société française de statistique, 1973, tous droits réservés.

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

NUMDAM

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

LA DÉPENDANCE DE DEUX ALÉAS VECTORIELS CRITÈRES ET VISUALISATION

Yves ESCOUFIER

Département d'informatique, Université de Montréal

RESUME

Trois approches, dont deux classiques, de l'étude de la dépendance de deux vecteurs sont présentées. Chacune de ces méthodes conduit à un critère de dépendance. Les trois critères sont comparés. Une représentation graphique susceptible d'aider à la compréhension de la dépendance est suggérée. Les techniques décrites sont appliquées à un exemple.

I – INTRODUCTION

Supposons que n observations d'un vecteur aléatoire à $p + q$ composantes aient été faites ; les données sont regroupées en une matrice X à $p + q$ lignes et n colonnes ; la colonne j contient les $p + q$ observations faites sur l'individu j ; la ligne i , les n observations du $i^{\text{ème}}$ caractère. La matrice ($p \times n$) des p premières lignes de X sera notée X_1 . De même X_2 sera la matrice ($q \times n$) des q dernières lignes de X . X_{ij} sera l'élément de la ligne i , colonne j de X

$$X = \begin{bmatrix} X_1 \\ \dots \\ X_2 \end{bmatrix} \begin{matrix} n \\ p \\ q \end{matrix}.$$

Les exemples suivants montrent que de nombreuses applications conduisent à des matrices de données possédant cette structure.

Exemple n° 1 : Des relevés ont été effectués sur n placettes forestières : X_1 est la matrice des observations floristiques ; X_2 celle des observations écologiques.

Exemple n° 2 : Pour un mois donné, et pour n années, on connaît les pluies tombées en différents points d'un bassin versant ainsi que les débits des rivières issues de ce bassin. X_1 et X_2 seront respectivement les matrices des pluies et des débits.

Exemple n° 3 : Un cas particulier intéressant est celui où X_2 se réduit à une matrice ($1 \times n$) définissant l'instant auquel les p premières mesures ont été faites ; la $j^{\text{ème}}$ colonne de la matrice X est composée des p observations faites à l'instant j et de j en tant que $(p + 1)^{\text{ème}}$ élément [8].

Le problème à résoudre est celui de la description de l'hétérogénéité de X_1 en une partie expliquée par X_2 et une partie résiduelle. Il est bien sûr très voisin de ceux que résolvent les techniques de régression multilinéaire et des corrélations canoniques mais s'en différencie par la globalité des solutions qu'il recherche. Une première approche consiste à représenter les variables dans l'espace des individus [paragraphe 2] ; elle conduit à prendre le coefficient de corrélation de Hotelling comme critère de quantification de "l'information" apportée par X_2 sur X_1 . La représentation des individus dans l'espace des variables [paragraphe 3] conduit à un autre critère, très voisin du précédent. Ces deux méthodes suggèrent une même représentation graphique [paragraphe 4] susceptible d'aider dans l'interprétation de l'hétérogénéité expliquée ou résiduelle. Une troisième approche [4] fournit un nouveau critère dont on verra les liaisons avec les précédents [paragraphe 5].

II – REPRESENTATION DES VARIABLES DANS L'ESPACE DES INDIVIDUS

IIa –

Après avoir identifié chacun des n individus à l'un des vecteurs d'une base orthogonale (e_1, \dots, e_n) de R^n choisie de telle sorte que $\|e_j\| = \sqrt{m_j}$ où m_j est une masse attachée à l'individu j , on peut représenter la variable définissant la ligne i de X par un vecteur U_i ayant pour origine l'origine des axes et pour extrémité les point de coordonnées

$$X_{ij} = \frac{\sum_{k=1}^n m_k X_{ik}}{m} \quad j = 1, \dots, n$$

où

$$m = \sum_{k=1}^n m_k$$

Le produit scalaire entre deux vecteurs U_i et U_l est alors égal à :

$$\sum_{j=1}^n m_j \left(X_{ij} - \frac{\sum_{k=1}^n m_k X_{ik}}{m} \right) \left(X_{lj} - \frac{\sum_{k=1}^n m_k X_{lk}}{m} \right).$$

On peut simplifier ces notations en substituant à la matrice X la matrice \tilde{X} telle que :

$$\tilde{X}_{ij} = \sqrt{m_j} \left(X_{ij} - \frac{\sum_{k=1}^n m_k X_{ik}}{m} \right)$$

Alors

$$\langle U_i, U_l \rangle = \sum_{j=1}^n X_{ij} X_{lj}$$

Soient $(U_i^1)_{i=1, \dots, p}$ les vecteurs associés aux lignes de \tilde{X}_1 . Ils engendrent un sous-espace de R^n que nous noterons E_1 . De même, les vecteurs $(U_i^2)_{i=1, \dots, q}$ associés aux lignes de \tilde{X}_2 engendrent un espace E_2 .

Il est courant ([1], p. 166) en statistique de mesurer globalement l'hétérogénéité de \tilde{X}_1 par le volume $V(U_1^1, \dots, U_p^1)$ du paralléloétope construit sur les U_i^1 , volume que l'on sait être égal à $|\tilde{X}_1' \tilde{X}_1| = |S_{11}|$ où l'élément de la ligne i et de la colonne j de la matrice S_{11} est $\langle U_i^1, U_j^1 \rangle$. Cette quantité est la variance généralisée estimée des variables définissant les lignes de X_1 . Soit $P(U_i^1)$ la projection orthogonale de U_i^1 dans E_2 et considérons le volume $V(P(U_1^1), \dots, P(U_p^1))$ du paralléloétope bâti sur les $P(U_i^1)$. Si $E_2 > E_1$, le volume initial et le volume des projections sont les mêmes puisque pour tout i $U_i^1 = P(U_i^1)$. Si E_2 et E_1 sont orthogonaux, le volume des projections est nul puisque pour tout i $P(U_i^1) = 0$. Dans tous les cas,

$$\frac{V(P(U_1^1), \dots, P(U_p^1))}{V(U_1^1), \dots, (U_p^1)}$$

donne une indication sur la part de la variation de X_1 qui peut être expliquée par X_2 . Nous allons calculer cette quantité de manière précise.

Iib -

Pour obtenir la valeur de ce rapport de volumes, écrivons que pour tout élément U_i^1 , le vecteur $U_i^1 - P(U_i^1)$ est orthogonal à tout élément U_j^2 , soit :

$$\begin{aligned} \forall i = 1, \dots, p \\ \forall j = 1, \dots, q \end{aligned} \quad \langle U_i^1 - P(U_i^1), U_j^2 \rangle = 0$$

$P(U_i^1)$ est un élément de E_2 , que l'on peut écrire $\sum_{k=1}^q M_k^i U_k^2$, d'où les $p \times q$ égalités

$$\begin{aligned} \forall i = 1, \dots, p \\ \forall j = 1, \dots, q \end{aligned} \quad \langle U_i^1, U_j^2 \rangle - \sum_{k=1}^q M_k^i \langle U_k^2, U_j^2 \rangle = 0$$

Si on appelle S_{12} la matrice d'éléments $(S_{12})_{ij} = \langle U_i^1, U_j^2 \rangle$; S_{22} la matrice d'éléments $(S_{22})_{ij} = \langle U_i^2, U_j^2 \rangle$ et M la matrice dont la $i^{\text{ème}}$ ligne est formée des M_k^i , les $p \times q$ égalités précédentes se résument en l'égalité matricielle $S_{12} - MS_{22} = 0$

Il en découle
$$M = S_{12} S_{22}^{-1}$$

et
$$P(U_i^1) = \sum_{k=1}^q (S_{12} S_{22}^{-1})_{i,k} U_k^2$$

Par conséquent :

$$\begin{aligned} \langle P(U_i^1), P(U_j^1) \rangle &= \sum_{k=1}^q \sum_{l=1}^q (S_{12} S_{22}^{-1})_{i,k} (S_{12} S_{22}^{-1})_{j,l} \langle U_k^2, U_l^2 \rangle \\ &= (S_{12} S_{22}^{-1} S_{21})_{i,j} \end{aligned}$$

et $V(P(U_1^1), \dots, P(U_p^1)) = |S_{12} S_{22}^{-1} S_{21}|$

Le rapport des volumes est donc égal à $\frac{|S_{12} S_{22}^{-1} S_{21}|}{|S_{11}|}$ qui est le carré du coefficient de corrélation vectorielle introduit par Hotelling [5].

IIc –

Il est aisé de voir que le coefficient de corrélation vectorielle est invariant dans des transformations non singulières affectant l'une ou l'autre des matrices X_1 et X_2 . Il est donc possible de le calculer à partir des variables canoniques ce qui établit aussitôt qu'il est égal au produit des carrés des coefficients de corrélation canonique. Pour $p = 2$ Hotelling a donné la distribution d'échantillonnage du coefficient de corrélation vectorielle pour des échantillons issus d'une population multi-normale dans laquelle le coefficient de corrélation vectorielle est nul. Il a montré comment étendre ce résultat pour $2 < p \leq q$ à la condition que tous les coefficients de corrélation canonique sauf un s'annulent. Ces résultats sont difficilement utilisables en pratique et les apports plus récents sur la distribution conjointe des coefficients de corrélation canonique ([1], p. 323-324) ne le sont pas plus.

Remarque : Notons au contraire que la distribution d'échantillonnage de $\frac{|S_{11} - S_{12}^{-1} S_{21}|}{|S_{11}|}$, "coefficient d'aliénation" estimé est connue ([1], p. 237) pour un plus grand nombre de valeurs de p et de q .

III – REPRESENTATION DES INDIVIDUS DANS L'ESPACE DES VARIABLES

IIIa –

En identifiant chacune des variables à un des vecteurs d'une base ortho-normale de R^{p+q} , il est possible de représenter chacun des n individus par un point de R^{p+q} ; l'individu j sera associé au point M_j de coordonnées

$$(X_{ij})_{i=1, \dots, p+q}.$$

Dans ce contexte ([2], ou [3] p. 126), les constructions habituelles consistent à considérer le centre de gravité du nuage des M_j affectés des masses m_j , soit G de coordonnées $G_i = \frac{1}{m} \sum_{j=1}^n m_j X_{ij}$ ($i = 1, \dots, p + q$) puis à mesurer l'hétérogénéité de la matrice X_1 par le moment d'inertie du nuage par rapport au sous-espace défini par $(X_i = G_i)_{i=1, \dots, p}$. Pour le faire, on remarque que la distance d'un point M_j à ce sous-espace est $\left(\sum_{i=1}^p (X_{ij} - G_i)^2 \right)^{\frac{1}{2}}$ si bien que l'inertie I_T cherchée est, avec les notations du paragraphe précédent :

$$I_T = \sum_{j=1}^n \sum_{i=1}^p m_j (X_{ij} - G_i)^2 = \sum_{i=1}^p \left(\sum_{j=1}^n m_j (X_{ij} - G_i)^2 \right) = \text{Tr } S_{11}$$

Appelons X_j^1 le vecteur $(p \times 1)$ de composantes $X_{i,j}$ ($i = 1, \dots, p$) et X_j^2 le vecteur $(q \times 1)$ de composantes $X_{i,j}$ ($i = p + 1, \dots, p + q$). Le point M_j a donc pour vecteur de coordonnées ${}^t(X_j^1 | X_j^2)$. Pour toute matrice A $(p \times q)$ et tout vecteur B $(p \times 1)$, nous pouvons considérer le sous-espace de R^{p+q} défini par $Y_{ij} = A_i X_j^2 + B_i$ ($i = 1, \dots, p$) où A_i et B_i sont respectivement les $i^{\text{èmes}}$

lignes de A et B. La distance de M_j à ce sous-espace est égale à $\left(\sum_{i=1}^p (X_{ij} - Y_{ij})^2\right)^{\frac{1}{2}}$ et on peut considérer l'inertie

$$I_R(A, B) = \sum_{j=1}^n m_j \left(\sum_{i=1}^p (X_{ij} - Y_{ij})^2 \right) = \sum_{i=1}^p \left(\sum_{j=1}^n m_j (X_{ij} - A_i X_j^2 - B_i)^2 \right).$$

Nous allons établir que les valeurs A^* et B^* qui rendent $I_R(A, B)$ minimum permettent de décomposer $\text{Tr } S_{11}$ en la somme des traces de deux matrices.

IIIb -

Pour trouver le minimum, nous devons évaluer à zéro les dérivées de $I_R(A, B)$. Appelons G^1 le vecteur $(p \times 1)$ dont les composantes sont les G_i ($i = 1, \dots, p$) et G^2 le vecteur $(q \times 1)$ dont les composantes sont les G_i ($i = p+1, \dots, p+q$). Alors on tire de

$$\frac{\partial I_R(A, B)}{\partial B_i} = 0 \quad (i = 1, \dots, p)$$

$$B_i = G_i - A_i G^2 \quad (i = 1, \dots, p)$$

soit

$$B = G^1 - A G^2$$

En introduisant ce résultat dans $I_R(A, B)$ et en écrivant

$$A_i X_j^2 = \sum_{k=1}^q a_{ik} X_{p+k,j}, \quad \text{on a :}$$

$$I_R(A, B) = \sum_{i=1}^p \sum_{j=1}^n m_j \left[(X_{ij} - G_i) - \sum_{k=1}^q a_{ik} (\tilde{X}_{p+k,j} - G_{p+k}) \right]^2$$

d'où, en écrivant que $\frac{\partial I_R(A, B)}{\partial a_{ik}}$ doit être nul pour tout $i = 1, \dots, p$ et tout $k = 1, \dots, q$:

$$\sum_{j=1}^n m_j (X_{i,j} - G_i) (X_{p+1,j} - G_{p+1}) = \sum_{k=1}^q a_{ik} \sum_{j=1}^n m_j (X_{p+k,j} - G_{p+k}) (X_{p+1,j} - G_{p+1})$$

Ces $p \times q$ inégalités se résument avec les notations du paragraphe précédent en l'égalité matricielle :

$$S_{12} = A S_{22}$$

d'où

$$A = S_{12} S_{22}^{-1}$$

et la valeur minimale de $I_R(A, B)$ (que nous écrivons I_E) est :

$$I_R = \sum_{i=1}^p \sum_{j=1}^n m_j [(X_{ij} - G_i) - (S_{12} S_{22}^{-1})_i (X_j^2 - G^2)]^2$$

Posons

$$I_E = \sum_{i=1}^p \sum_{j=1}^n m_j [(S_{12} S_{22}^{-1})_i (X_j^2 - G^2)]^2$$

Alors, parce que $\sum_{j=1}^n m_j (X_j^2 - G^2) (X_j^2 - G^2) = S_{22}$, il est aisé de voir que :

$$I_R = \text{Tr} (S_{11} - S_{12} S_{22}^{-1} S_{21})$$

$$I_E = \text{Tr} (S_{12} S_{22}^{-1} S_{21})$$

et donc $I_T = I_R + I_E$

Si pour $i = 1, \dots, p$, il existe j tel que $X_{ij} = A_i X_j^2 + B_i$, le point M_j apporte une contribution nulle dans l'inertie I_R . Ceci permet de voir que I_E représente la part d'hétérogénéité de X_1 expliquée par X_2 tandis que I_R est la part résiduelle. Cette approche conduit donc à prendre $Q_1 : \frac{\text{Tr} (S_{12} S_{22}^{-1} S_{21})}{\text{Tr} (S_{11})}$

ou le conjugué $\frac{\text{Tr} (S_{11} - S_{12} S_{22}^{-1} S_{21})}{\text{Tr} (S_{11})}$ comme critère pour quantifier la part d'hétérogénéité de X_1 expliquée par X_2 .

IIIc -

On peut noter tout de suite que ce critère est invariant dans des transformations non singulières affectant X_2 mais ne possède la propriété d'invariance que pour des transformations orthogonales de X_1 .

Dans le cas particulier où $p = 1$, le carré du coefficient de corrélation vectorielle et le critère Q_1 se ramènent au carré du coefficient de corrélation multiple estimé. Ces deux critères peuvent donc être considérés comme deux généralisations différentes de cette même quantité.

En dehors de ce cas particulier, la distribution d'échantillon de Q_1 n'est pas connue. Les travaux de James [6] sur la distribution des valeurs propres d'une matrice de covariance pourraient servir de point de départ à une recherche ardue sur ce point.

IV - REPRESENTATION GRAPHIQUE

Nous conserverons dans ce paragraphe la terminologie du paragraphe précédent. L'inertie I_E peut être interprétée comme l'inertie totale d'un nuage de n points N_j où N_j aurait pour coordonnées $Y_{ij} = (S_{12} S_{22}^{-1})_i (X_j^2 - G^2)$ pour $i = 1, \dots, p$ et serait associé à la masse m_j .

De la même manière, I_R peut être considérée comme l'inertie d'un nuage de n points P_j de coordonnées $Z_{ij} = (X_j^2 - G^2) - (S_{12} S_{22}^{-1})_i (X_j^2 - G^2)$ pour $i = 1, \dots, p$ affectés des masses m_j .

On peut alors effectuer une analyse en composantes principales de ces deux nuages, c'est-à-dire chercher d'une part une base orthonormale η_1, \dots, η_p de R^p telle que pour $p_1 < p$ donné soit maximum la quantité

$$\sum_{i=1}^{p_1} I_{[\eta_i]} = I_E - \sum_{i=p_1+1}^p I_{[\eta_i]}$$

où $I_{[\eta_i]^\perp}$ est l'inertie du nuage des N_j par rapport au sous-espace orthogonal à η_i et d'autre part une base ν_1, \dots, ν_p telle que $p_2 < p$ donné soit maximum la quantité

$$\sum_{i=1}^{p_1} I_{[\nu_i]^\perp} = I_R - \sum_{i=p_2+1}^p I_{[\nu_i]^\perp}.$$

On sait que la solution de ces problèmes est donnée par les vecteurs propres d'une part de $S_{12} S_{22}^{-1} S_{21}$, d'autre part de $S_{11} - S_{12} S_{22}^{-1} S_{21}$ et que si l'on note λ_i ($i = 1, \dots, p$) les valeurs propres de la première matrice supposées rangées dans l'ordre décroissant et μ_i ($i = 1, \dots, p$) les valeurs propres de la seconde, on a :

$$\sum_{i=1}^{p_1} I_{[\eta_i]^\perp} = \sum_{i=1}^{p_1} \lambda_i \quad \text{et} \quad \sum_{i=1}^{p_2} I_{[\nu_i]^\perp} = \sum_{i=1}^{p_2} \mu_i$$

Si les données sont telles qu'existent p_1 et p_2 petits (1, 2 ou 3) tels que $\frac{\sum_{i=1}^{p_1} \lambda_i}{I_E}$ et $\frac{\sum_{i=1}^{p_2} \mu_i}{I_R}$ soient voisins de l'unité, on peut alors obtenir une représentation du nuage des N_j et une représentation du nuage des P_j dans des espaces à p_1 et p_2 dimensions qui conservent au mieux l'inertie et ainsi visualiser indépendamment la partie expliquée et la partie résiduelle de l'hétérogénéité de X_1 .

On peut remarquer que le nuage des N_j n'aura pas plus de points différents qu'il n'y a de colonnes différentes dans X_2 . L'étude des points importants dans la détermination de l'inertie pourra permettre de détecter les variables dont les variations sont importantes dans la configuration du nuage et donc importantes pour expliquer l'hétérogénéité de X_1 .

V – TROISIEME APPROCHE

Il a été montré dans un travail précédent [4] comment on peut représenter un vecteur aléatoire par un élément d'un espace euclidien qui le caractérise en ce sens que l'élément représentatif permet de retrouver les vecteurs propres et les valeurs propres de la matrice des variances et covariances du vecteur aléatoire. Dans ce contexte, la quantité $\text{Tr}(\Sigma_{12} \Sigma_{21})$ joue entre les vecteurs aléatoires un rôle analogue à celui que joue la covariance entre deux variables aléatoires réelles, ce qui conduit à estimer la dépendance de X_1 et X_2 par le "coefficient de corrélation"

$$Q_2 = \frac{\text{Tr}(\Sigma_{12} \Sigma_{21})}{\sqrt{\text{Tr}(\Sigma_{11}) \text{Tr}(\Sigma_{22})}}.$$

L'analogie entre la covariance et la quantité $\text{Tr}(\Sigma_{12} \Sigma_{21})$ est encore renforcée par la remarque suivante : De même que

$$\text{Var}(X + Y) = \text{Var} X + \text{Var} Y + 2 \text{cov}(X, Y), \text{ si } \Sigma_{11} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

$$\text{on a : } \quad \text{Tr}(\Sigma_{11}^2) = \text{Tr}(A_{11}^2) + \text{Tr}(A_{22}^2) + 2 \text{Tr}(A_{21} A_{12})$$

Il est aisé de voir que le critère Q_2 n'est pas invariant dans des transformations non singulières affectant X_1 ou X_2 ; il n'est invariant que pour des transformations orthogonales. Notons également que dans le cas où l'on utilise l'invariance de Q_1 dans les transformations non singulières affectant X_2 pour faire en sorte que $S_{22} = I$, les critères Q_1 et Q_2 ont le même numérateur. Dans ce cas particulier, le problème de la visualisation de I_E évoquée au paragraphe IV trouve une solution intéressante.

En effet :

Soit C une matrice $q \times q$ telle que $C S_{22} {}^t C = I$. Posons $Z_2 = C X_2$. Si \tilde{S}_{12} est la matrice analogue à S_{12} pour Z_{21} on a :

$$I_E = \text{Tr}(\tilde{S}_{12} \tilde{S}_{21}) = \text{Tr}(\tilde{S}_{21} \tilde{S}_{12}) = \text{Tr}(C S_{21} S_{12} {}^t C)$$

Appelons alors C_i la $i^{\text{ème}}$ ligne de C , on a :

$$I_E = \sum_{i=1}^q (C_i S_{21} {}^t C_i)$$

La matrice C étant donnée, il est donc possible pour $q_1 < q$ donné, de choisir q_1 variables telles que $\sum_{i=1}^{q_1} (C_i S_{21} S_{12} {}^t C_i)$ soit maximum : elles seront définies par les q_1 lignes de C qui maximisent $(C_i S_{21} S_{12} {}^t C_i)$ et il est à noter que le choix de l'une d'entre elles ne dépend pas des choix précédents. Il est alors tentant de se demander si pour $q_1 < q$ donné, il existe une matrice C qui rende $\sum_{i=1}^{q_1} (C_i S_{21} S_{12} {}^t C_i)$ maximum sous la condition $C S_{22} {}^t C = I$. La réponse à cette question est connue ([7], p. 505) et est donnée par les solutions de $(S_{21} S_{12} - \lambda S_{22})V = 0$.

VI – UN EXEMPLE

Les données discutées ici sont extraites d'une étude faite par Messieurs Berger et Rouziès du Centre Régional de Productivité et des Sciences économiques de la faculté de droit de Montpellier. Elles concernent 21 communes du Sud de la France. Pour chaque commune sont données la population totale pour 1962 et 1968 ainsi que le nombre de personnes actives travaillant dans les secteurs primaire, secondaire ou tertiaire pour 1968. La table 1 fournit les données originales. Dans le traitement, les données ont été centrées, ce qui n'a aucune importance, mais aussi réduites de manière à ce que les variables "population totale" n'écrasent pas les autres variables dont la variance est plus faible.

Table 1

	1968 Population totale	1968 Secteur primaire (agricole)	1968 Secteur secondaire (Industriel)	1968 Secteur tertiaire (services)	1962 Population totale
Données					
Aigues Vives	1 506.00	123.00	193.00	254.00	1 367.00
Asperes	246.00	58.00	2.00	8.00	259.00
Aubais	900.00	72.00	78.00	104.00	869.00
Aujargues	232.00	50.00	10.00	31.00	223.00
Boissières	154.00	29.00	14.00	8.00	175.00
Calvisson	1 802.00	220.00	150.00	218.00	1 792.00
Caveirac	704.00	94.00	35.00	112.00	644.00
Clarensac	532.00	81.00	40.00	46.00	616.00
Congenies	473.00	65.00	31.00	60.00	505.00
Fontanes	334.00	96.00	5.00	35.00	383.00
Junas	321.00	87.00	11.00	23.00	326.00
Langlade	381.00	23.00	36.00	71.00	366.00
Lecques	203.00	68.00	10.00	10.00	215.00
Nages	327.00	52.00	3.00	33.00	278.00
St Clément	150.00	27.00	2.00	11.00	160.00
St Come	294.00	69.00	10.00	25.00	327.00
St Dionisy	160.00	31.00	8.00	10.00	191.00
Salinelles	261.00	51.00	11.00	27.00	251.00
Sommières	3 541.00	124.00	367.00	482.00	3 323.00
Souviagnargues	406.00	82.00	12.00	13.00	407.00
Villevielle	507.00	55.00	29.00	83.00	343.00
Moyennes	639.71	74.14	50.33	79.24	620.00

Le but de l'étude est de déterminer dans quelle mesure la variabilité des observations de 1968 est explicable par l'information "population totale de 1962".

Table 2

Matrice des corrélations

1.0000	.6673	.9843	.9855	.9975
.6673	1.0000	.6014	.6233	.6862
.9843	.6014	1.0000	.9872	.9792
.9855	.6233	.9872	1.0000	.9758
.9975	.6862	.9792	.9758	1.0000

La table 2 fournit la matrice $S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$, ici matrice des corrélations, pour laquelle S_{12} est un vecteur colonne ; S_{21} le vecteur ligne transposé ; $S_{22} = 1$.

La variabilité des observations de 1968, mesurée par la trace de S_{11} est égale à 4. Elle se décompose en une partie expliquée $I_E = \text{tr}(S_{12} S_{21}) = \text{Tr}(S_{21} S_{12}) = 3,3769$ et une partie résiduelle, $I_r = \text{Tr}(S_{11} - S_{12} S_{21}) = 0,6231$.

La lecture de ces valeurs numériques – ou le calcul des quantités $Q_1 = 0,8442$ et $Q_2 = 0,9660$ – indique que la population totale de 1962 donne une bonne explication de la variabilité des observations de 1968.

L'observation de la figure 1 permet d'apprécier la signification des résidus.

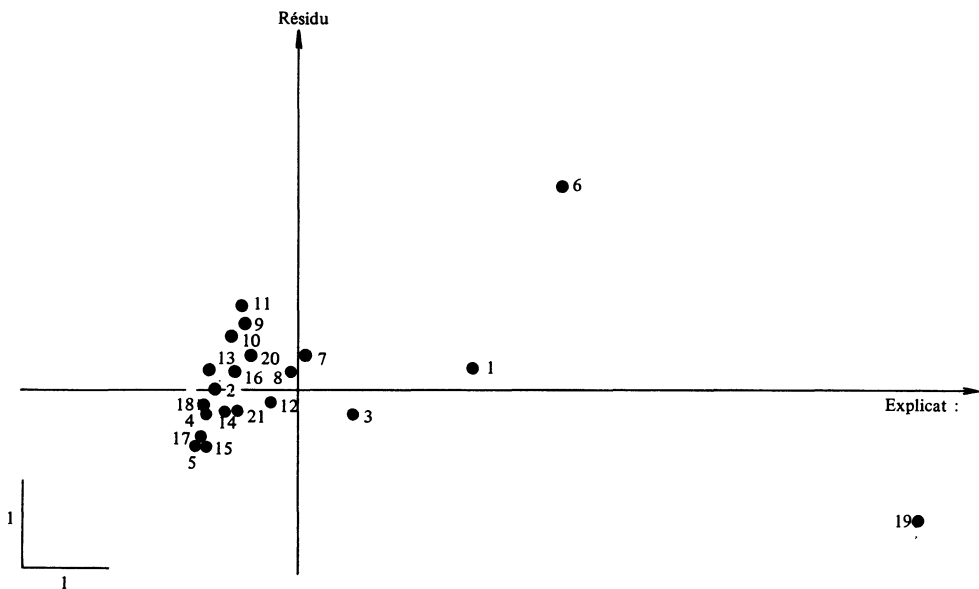


Figure 1

La figure 1 est obtenue en représentant un village j par un point dont les coordonnées sont pour l'abscisse la coordonnée sur le premier vecteur propre de $S_{12} S_{21}$ (unique ici puisque $\text{rang}(S_{12} S_{21}) = \text{rang}(S_{21} S_{12}) = 1$) de N_j et pour ordonnée la coordonnée sur le premier vecteur propre de $S_{11} - S_{12} S_{21}$ de P_j .

Le premier vecteur propre de $S_{12} S_{21}$ a pour composantes

$$(0,5428 ; 0,3741 ; 0,5329 ; 0,5310)$$

On voit donc que les points représentatifs vont être d'autant plus sur la droite de la figure que la population de la commune est grande.

Le premier vecteur de $S_{11} - S_{12} S_{21}$, associé à la valeur propre $0,5449 = 0,871_R$ a pour composantes

$(- 0,0356 ; 0,9836 ; - 0,1446 ; - 0,1016)$.

Il en découle que le résidu est essentiellement dû à la population du secteur primaire pour 1968. Si cette population a été sous-estimée Z_{2j} est positif et le résidu est positif. Si au contraire, cette population a été surestimée, le résidu est négatif. La figure montre qu'il y a eu sous-estimation pour Calvisson (commune n° 6) ce qui n'est pas étonnant puisque Calvisson a une très forte population agricole, la plus forte de toutes les communes et qui représente un pourcentage élevé de sa population totale.

La figure montre également que le résidu est négatif pour Sommières (commune n° 19). Ceci s'explique encore en remarquant que la population agricole de Sommières représente un pourcentage de la population totale bien inférieur au pourcentage moyen pour les autres communes.

CONCLUSION

Il est regrettable que les distributions des critères que nous avons rappelés ne soient pas plus simples et mieux connues car il serait alors possible, au moins sous certaines hypothèses de faire des tests sur la qualité des représentations obtenues. Il n'en reste pas moins que l'utilisation à des fins purement descriptives des critères et des représentations graphiques peut aider à "lire" de grands ensembles de données.

REFERENCES

- [1] ANDERSON T.W. : "An introduction to multivariate statistical analysis", John Wiley - 1958.
- [2] BENZECRI J.P. : "Représentation euclidienne d'un ensemble muni de masses et de distances", Lab. de Stat. Math. Faculté des Sciences de Paris, mai 1970.
- [3] DEMPSTER A.P. : "Elements of continuous multivariate analysis", Addison-Wesley, Publishing Company 1969.
- [4] ESCOUFIER Y. : "Les liaisons entre groupes d'aléas", R.S.A., 1971 Vol. XIX n° 2 pp. 5-17.
- [5] HOTELLING H. : "Relations between two sets of variates", Biometrika Vol. XXVIII, 1936 pp. 321-377.
- [6] JAMES Alan T. : "Distributions of Matrix Variates and Latent Roots derived from Normal samples", the annals of mathematical statistics, 1964, vol. 35, n° 4, pp. 475-501.

- [7] RAO C.R. : “Linear Statistical Inference and its Applications”, John Wiley, 1965.
- [8] TENENHAUS M. et B. PRIEURET : “Analyse des séries chronologiques multi-dimensionnelles, Université d’Ottawa, Département de recherche opérationnelle et planification, mai 1971.