DÉFINITIONS MATRICIELLES DES PROPRIÉTÉS DE L'ANALYSE DES COMPOSANTES PRINCIPALES APPLICATION A LA COMPARAISON DE DIFFÉRENTES COTATIONS D'ABONDANCES DU ZOOPLANCTON DE LA BAIE D'AMBARO (NOSY-BÉ, MADAGASCAR)

F. IBANEZ

Station Zoologique de Villefranche-sur-Mer, 06230, France

RÉSUMÉ

Nous proposons une présentation matricielle de l'analyse des composantes principales. L'originalité provient de la démonstration de formules simplifiées des corrélations entre les composantes d'une analyse sur un ensemble de données avec les variables d'un autre ensemble, ainsi que des corrélations entre les composantes de deux ensembles distincts.

Ces corrélations entre deux ensembles nous ont permis d'apprécier la validité de l'utilisation de trois types de cotation d'abondance : une progression géométrique de raison 4,3 ; un découpage en trois classes d'égale amplitude ; et une cotation réduite à deux classes égales.

ABSTRACT

We give a matricial representation of Principal Component Analysis. The originality is the calculation of the correlations between the components of one set of observations and the variables of another set, and the correlations between the components of two distinct sets of data. The formulae permit us to test the interest in the use of 3 scales of numeration : with a geometric progression of ratio 4.3, and a partition in three or two classes of equal amplitudes.

Les langages de programmation sont de plus en plus orientés vers l'application pure et simple des règles du calcul matriciel. Les langages PL1 ou BASIC permettent souvent d'éviter de fastidieuses boucles de programme en traitant les tableaux comme de simples variables. C'est pourquoi nous avons voulu montrer l'intérêt d'une présentation matricielle de l'analyse en composantes principales.

1. RECHERCHE DES ÉLÉMENTS PROPRES.

Très généralement en Écologie planctonique, l'information recueillie est condensée sous la forme de deux tableaux correspondant l'un à des abondances d'espèces, l'autre à des valeurs de paramètres climatiques ou hydrologiques. On a donc un tableau général A subdivisé en un tableau X (m espèces en n stations) et un tableau Y (p paramètres en n stations) représenté figure 1 a.

Nous avons présenté (IBANEZ, 1973) un traitement complet de ce type de données en effectuant l'analyse en composantes principales d'abord à partir du tableau X, ensuite du tableau Y, enfin en mesurant les relations entre ces deux analyses séparées. Après passage en valeur réduites des valeurs de A la matrice de corrélation générale R s'écrira :

Cah. O.R.S.T.O.M., sér. Océanogr., vol. XIV, n° 4, 1976 : 327-330.

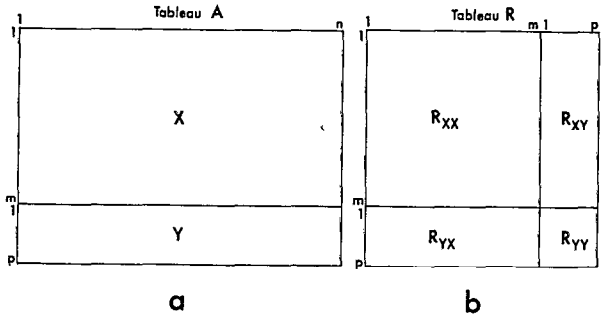

Fig. 1. — *a* : tableau des données ; *b* : matrice des corrélations.

$$R = \frac{1}{n} AA'$$

Cette matrice carrée et symétrique d'ordre m+p, est constituée de trois matrices distinctes : R_{xx}, matrice de corrélation entre les m espèces; R_{yy}, matrice de corrélation entre les p paramètres; R_{xy} et R_{yx}, matrice de corrélation entre les m espèces et les p paramètres et sa transposée (fig. 1 b).

Considérons la matrice de corrélation R_{xx}. Si elle a des racines distinctes, c'est à dire si l'on peut résoudre l'équation :

$$| R_{xx} - \Delta I | = 0$$

on peut écrire

$$R_{xx} = V\Delta V'$$

où V désigne la matrice carrée non symétrique d'ordre m des vecteurs propres, et Δ la matrice carrée diagonale d'ordre m des valeurs propres. Comme V est une matrice orthogonale, les vecteurs propres sont linéairement indépendants et on a la relation : $VV' = V'V = I$. L'analyse des composantes principales consiste à définir les n observations dans l'espace des vecteurs propres c'est à dire un espace ou les m directions sont indépendantes.

Les nouvelles coordonnées des n points stations seront obtenues par :

$$C_x = V'\ X \qquad (1)$$

La covariance entre ces nouvelles coordonnées (les composantes principales) sera :

$$\mathrm{Cov}_{c_x} = \frac{1}{n} V'XX'V = V'R_{xx}V = V'V\Delta V'V = \Delta$$

Ainsi la transformation orthogonale (la rotation des axes) définit de nouvelles variables sans corrélation puisque tous les termes non diagonaux de Δ sont nuls. La variance des composantes est égale aux termes diagonaux de Δ.

Pour obtenir une même échelle graphique sur les axes, on norme à 1 les composantes :

$$C_x = \Delta^{-\frac{1}{2}} C_x = \Delta^{-\frac{1}{2}} V'X$$

Les valeurs de départ X seront données en fonction des composantes par :

$$X = V\Delta^{\frac{1}{2}} C_x$$

Nous pouvons vérifier que :

$$R_{xx} = \frac{1}{n} XX' = V\Delta^{\frac{1}{2}} C_x C'_x \Delta'^{\frac{1}{2}} V'$$

$$R_{xx} = V\Delta^{\frac{1}{2}}\Delta'^{\frac{1}{2}}V' = V\Delta V' = R_{xx}$$

2. CORRÉLATIONS ENTRE LES COMPOSANTES D'UN ENSEMBLE ET LES VARIABLES D'UN AUTRE ENSEMBLE.

Les corrélations entre les composantes C_x et les variables Y s'écriront :

$$R_{C_x}.Y = \frac{1}{n} \Delta^{-\frac{1}{2}} C_x Y' = \frac{1}{n} \Delta^{-\frac{1}{2}} V'XY'$$

d'où :

$$R_{C_x} Y = \Delta^{-\frac{1}{2}} V' R_{xy} \qquad (2)$$

On aurait de la même façon :

$$R_{C_y}.X = \Delta_y^{-\frac{1}{2}} U' R_{xy}$$

Δ_y et U correspondant aux matrices des valeurs et vecteurs propres de R_{yy}.

3. CORRÉLATIONS ENTRE LES COMPOSANTES DE DEUX ENSEMBLES DIFFÉRENTS (Ibanez, 1969).

Soit $C_x = V'X$ la composante du premier ensemble et $C_y = U'Y$ la composante du deuxième. Δ_x et Δ_y les deux matrices des valeurs propres. Les corrélations entre les composantes de R_{xx} et de R_{yy} s'écriront :

$$R_{C_X.C_Y} = \frac{1}{n}\Delta_X^{-\frac{1}{2}} C_X \Delta_Y^{-\frac{1}{2}} C'_Y$$

$$R_{C_X.C_Y} = \frac{1}{n}\Delta_X^{-\frac{1}{2}} V'X \quad (\Delta_Y^{-\frac{1}{2}} U'Y)'$$

$$R_{C_X.C_Y} = \frac{1}{n}\Delta_X^{-\frac{1}{2}} V'XY'U\Delta_Y^{-\frac{1}{2}}$$

d'où d'après (2) :

$$R_{C_X.C_Y} = R_{C_X.Y} U\Delta_Y^{-\frac{1}{2}} \qquad (3)$$

En permutant X et Y on aurait la formule symétrique :

$$R_{C_Y.C_X} = R_{C_Y.X} V\Delta_X^{-\frac{1}{2}}$$

Vérification :

Appelons v' le produit $\Delta_X^{-\frac{1}{2}} V'$

v le produit $V\Delta_X^{-\frac{1}{2}}$

u' le produit $\Delta_Y^{-\frac{1}{2}} U'$

u le produit $U\Delta_Y^{-\frac{1}{2}}$

et R_{XY} le produit $\frac{1}{n} XY'$ (ou $\frac{1}{n} YX'$)

On aura : $R_{C_X.C_Y} = v'R_{XY}u$ et $R_{C_Y.C_X} = u'R_{XY}v$

Donc $R_{C_Y.C_X}$ est bien la matrice transposée de $R_{C_X.C_Y}$

4. APPLICATION : COMPARAISON DE DIFFÉRENTES COTATIONS D'ABONDANCES SUR LES DONNÉES DU ZOOPLANCTON DE LA BAIE D'AMBARO (NOSY-BÉ; MADAGASCAR).

Le comptage du plancton est une opération fastidieuse. De plus même si la numération est parfaite, les données obtenues sont loin de présenter les caractéristiques statistiques théoriquement indispensables pour l'emploi de l'analyse en composantes principales. C'est uniquement à partir d'études empiriques que nous avons pu montrer la remarquable stabilité de cette méthode même avec des données aussi hétérogènes que les numérations planctoniques (IBANEZ 1971). C'est pourquoi nous nous sommes posé le problème suivant : jusqu'ou peut-on économiser le coût des comptages et utiliser valablement l'analyse des composantes principales ?

Nous avons envisagé successivement trois types de cotation d'abondances, les résultats de deux d'entre elles étant déjà publiés :

— La cotisation « 4,3 » (FRONTIER et IBANEZ 1974). Elles est basée sur un découpage en classes dont les limites sont en progression géométrique de raison 4, 3.

— La cotation à 3 niveaux (IBANEZ 1974). Elle comprend seulement trois classes dont les limites sont données par la division en trois parties égales de la distribution des variables après transformation logarithmique. L'estimation *a priori* des valeurs maximum et minimum des distributions est déjà publiée (IBANEZ 1974).

— La cotation à deux niveaux. C'est le même principe que la précédente cotation, mais seulement deux classes divisent l'intervalle entre le maximum et le minimum.

Dans le cas de la cotation « 4,3 » le nombre de classes dépendra de l'amplitude des variations de la variable considérée. Dans le cas des deux autres cotations, ce nombre de classes est invariable.

Les données utilisées dans notre exemple proviennent des comptages de 19 catégories zooplanctoniques, prélevées sur un réseau de 44 stations quadrillant la baie d'Ambaro (côte N. O. de Madagascar). Nous avons ajouté à cette liste d'espèces une variable purement aléatoire « ε » qui nous permet de savoir quels sont les composantes interprétables (IBANEZ 1973).

Pour comparer les trois cotations nous montrerons les corrélations entre les axes significatifs de l'analyse effectuée sur les nombres réels, et ceux obtenus à partir de l'analyse sur les données codées.

Pour chaque comparaison le tableau des données est composé de 44 observations et 40 variables (20 correspondant aux données réelles et 20 au codage considéré. La matrice de corrélation est constitué de trois matrices de corrélations distinctes : R_{XX} la matrice de corrélation entre les espèces comptées; R_{YY} la matrice de corrélation entre les données codées; R_{XY} la matrice de corrélation entre les variables dénombrées et les variables codées.

Les pourcentages de variances contenus dans les trois premiers axes principaux sont donnés par le tableau I (seuls les trois premiers axes sont retenus car la variable « ε » est la plus exprimée par le quatrième axe).

Grâce à la formule (3) nous avons pu calculer les corrélations entre les axes de l'analyse effectuée sur les données réelles et ceux issus des données codées (tabl. II).

Les tableaux I et II mettent clairement en évidence

la validité des codages dans une application future de l'analyse des composantes principales.

Les différences observées ne peuvent absolument pas changer l'interprétation écologique des données réelles.

On peut se demander si une appréciation très approximative des effectifs (abondant d'une part, rare ou absent de l'autre) ne permettrait pas une ordination des variables en temps réel, à bord d'un bateau.

TABLEAU I

Pourcentages de variance contenus dans les trois premiers axes.

Valeurs propres	Données réelles	Codage « 4,3 »	3 niveaux	2 niveaux
λ1.....	38,1	37,6	38,8	38,7
λ2.....	52,5	50,8	48,4	41,8
λ3.....	61,9	60,1	57,9	51,3

TABLEAU II

Corrélations entre les trois premières composantes issues des données réelles et les trois premières composantes issues des données codées.

Données réelles	Cotation « 4,3 »			3 niveaux			2 niveaux		
	C1	C2	C3	C1	C2	C3	C1	C2	C3
C1.....	*1*	— 0,00	— 0,02	*0,99*	0,06	— 0,05	*0,98*	— 0,02	— 0,04
C2.....	— 0,00	*0,98*	0,13	0,02	*0,84*	0,46	0,00	*0,87*	— 0,30
C3.....	0,02	0,12	*0,98*	0,05	— 0,39	*0,80*	0,06	0,28	*0,83*

BIBLIOGRAPHIE

FRONTIER (F.) et IBANEZ (F.), 1974. — Utilisation d'une cotation d'abondance fondée sur une progression géométrique pour l'analyse des composantes principales. *J. Exp. mar. Biol. Ecol.*, 14 (3) : 217-224.

IBANEZ (F.), 1971. — En effet des transformations des données dans l'analyse factorielle en écologie planctonique. *Cah. Océanogr.* 23 (6) : 545-561.

IBANEZ (F.), 1973. — Méthode d'analyse spatio-temporelle du processus d'échantillonnage en Planctologie, son influence dans l'interprétation des données par l'analyse des composantes principales. *Ann. Inst. Océanogr.*, 49 (2) : 83-111.

IBANEZ (F.), 1973. — Un programme FORTRAN IV d'études des structures écologiques marines par un modèle dérivé de l'analyse factorielle. *Doc. sci. Centre O.R.S.T.O.M. Nosy-Bé*, n° 38, *multigr.* 91 p.

IBANEZ (F.), 1974. — Une cotation d'abondance réduite à trois classes, justification de son emploi en analyse des composantes principales. Mise en œuvre pratique en Planctologie. *Ann. Inst. Océanogr.*, 50 (2) : 185-198.