REVUE DE STATISTIQUE APPLIQUÉE

W. IMAM
S. ABDELKBIR
Y. ESCOUFIER

Quantification des effets spatiaux linéaires et non linéaires dans l'explication d'un tableau de données concernant la qualité des eaux souterraines

Revue de statistique appliquée, tome 46, n° 3 (1998), p. 37-52
<http://www.numdam.org/item?id=RSA_1998__46_3_37_0>

Rev. Statistique Appliquée, 1998, **XLVI** (3), 37-52

QUANTIFICATION DES EFFETS SPATIAUX LINÉAIRES ET NON LINÉAIRES DANS L'EXPLICATION D'UN TABLEAU DE DONNÉES CONCERNANT LA QUALITÉ DES EAUX SOUTERRAINES

W. Imam, S. Abdelkbir, Y. Escoufier

Unité de Biométrie (INRA-ENSAM-UMII) 2, Pl. Pierre Viala, 34060 Montpellier.

RÉSUMÉ

On considère une situation expérimentale conduisant pour chaque unité statistique à un ensemble de variables quantitatives actives et à un ensemble de variables quantitatives concomittantes.

La démarche exploratoire la plus immédiate consiste à faire l'analyse en composantes principales des variables actives et à utiliser les variables concomittantes comme variables supplémentaires. L'article montre comment les variables concomittantes peuvent être utilisées de manière plus effective dans l'analyse; on obtient des quantifications de leurs effets linéaires et non linéaires dans la décomposition de la variabilité totale des variables actives et dans celle de la variance de chacune d'entre elles. Dans l'exemple traité, les données concomittantes sont des coordonnées géographiques ce qui permet de quantifier des effets spatiaux.

Mots-clés : *ACPVI, ACPVI Spline Additive, opérateur de projection, décomposition de la variabilité, test de permutation.*

ABSTRACT

We consider an experimental situation that for every statistical unit, there exists two sets of variables; one containing the explicative variables and the other containing the responses.

Our study explores the data in applying, at the first time, the principal components analysis method (PCA) on the responses where the explicative variables are supplementary. We point out that the explicative variables play part in the statistical analysis. The application of the principal components analysis with instrumental variables method (PCAIV) leads for the quantification of their linear and nonlinear effects in the decomposition of the total variability of explicative variables and in the variability of their variance. In our example, the explicative variables are the geographical coordinates that permit the quantification of spatial effects.

Keywords : *PCAIV, PCAIV Additive Spline, matrix of projection, variability's decomposition, test of permutation.*

1. Introduction

L'étude qui est à l'origine du travail présenté porte sur la qualité des eaux de puits d'une région marocaine. Dans le cadre de cet article, pour chacun des 176 puits sont retenues 10 variables physico-chimiques indicatrices de la qualité des eaux et, les coordonnées géographiques des puits. Le but de l'étude est d'identifier la part de la qualité des eaux qui peut être expliquée par la localisation des puits.

L'analyse en composantes principales des données physico-chimiques accompagnée d'une projection en variables supplémentaires des coordonnées géographiques est une méthodologie possible pour aborder le problème. Cette pratique est insuffisante car elle ne fournit pas un critère chiffré mais de simples indications visuelles. De plus deux critiques épistémologiques peuvent lui être faites. Pour Chessel et Mercier (1993) «la critique majeure de cette approche est l'hypothèse implicite que les facteurs de milieu étudiés (ici les coordonnées géographiques) rendent compte d'une importante fraction de la variabilité, or il n'y a aucune garantie que ce soit le cas». On peut légitimement s'interroger par ailleurs sur le paradoxe inhérent à la démarche : on affiche qu'on veut comprendre le rôle des coordonnées géographiques dans la variabilité des variables physico-chimiques et dans la recherche d'une visualisation de cette variabilité, on ne tient aucun compte de l'information géographique.

Pour ces raisons nous nous tournons donc vers des approches qui prennent en compte les localisations géographiques dés le début de l'analyse. On trouve dans (Thioulouse, 1996) une bonne revue bibliographique des différentes approches. Les unes sont fondées sur les relations de voisinages entre les points d'observations. Les autres utilisent directement les coordonnées géographiques de ces points dans des polynômes de degré fixé à priori pour ajuster les données étudiées (ici les variables physico-chimiques). Notre approche relève de cette seconde catégorie. Dans un but pédagogique, nous l'exposons d'abord dans sa forme linéaire, c'est-à-dire la plus simple, puis dans sa forme non-linéaire. Ce plan facilite l'exposition de la méthode; il est légitimement adapté à une compréhension progressive des données.

Dans un premier temps donc, l'étude est engagée par une analyse en composantes principales par rapport à des variables instrumentales (ACPVI) dans laquelle les variables physico-chimiques sont les variables à expliquer et les coordonnées géographiques les variables potentiellement explicatives. Cette méthode fournit une décomposition de l'inertie totale du nuage des variables physico-chimiques en une part expliquée par les coordonnées et une part non expliquée. Cette décomposition globale inclut une décomposition des variances de chacune des variables physico-chimiques et donc une étude fine de la localisation géographique dans le phénomène étudié. Un test de permutation permet de quantifier l'importance de la relation mise en évidence. Reposant sur la technique des projections orthogonales dans le sous-espace engendré par les coordonnées géographiques, l'ACPVI ne met donc en évidence que les relations linéaires entre les variables physico-chimiques et les coordonnées géographiques.

Pour échapper à cette limitation, on recourt à l'ACPVI Spline Additive (Durand, 1993) et (Imam et Durand, 1997). On obtiendra comme précédement des décompositions de l'inertie totale du nuage et de la variance de chacune des variables physico-chimiques. Mais ici, ces décompositions ne s'interprèteront pas en terme des coordonnées géographiques, mais en terme de transformations non-linéaires de ces

coordonnées. Ces transformations sont recherchées par la méthode. Elles peuvent être utilisées pour visualiser les distorsions de la réalité géographique adaptées aux données physico-chimiques. Comme dans l'ACPVI linéaire, un test de permutation peut être mis en œuvre pour quantifier l'importance de la relation trouvée.

2. Les données et leur analyse en composantes principales (ACP)

Les données qui sont utilisées pour illustrer l'article ont été mises à disposition par le Laboratoire de l'Eau et de l'Environnement de la faculté des Sciences d'El Jadida. Elles concernaient la qualité de l'eau pour un ensemble de 176 puits.

Les 10 variables sont les suivantes :

(T) : la température, (TDS) : total des sels dissous, (Ph.) : la mesure du Ph, (Na^+) : sodium à l'extrait saturé, (Cl) : les chlorures, (Ca^{++}) : le calcium, (Mg^{++}) : le magnésium, (Hco_3) : les carbonates, (So_4) : les sulfates, (No_3) : les nitrates.

On dispose en plus des coordonnées (X_1, X_2) géographiques des puits qui constituent les variables supplémentaires.

2.1. Résultats de l'ACP normée

Les données ayant été centrées et réduites, on réalise l'ACP des 10 variables actives et on projette les coordonnées géographiques des puits en variables supplémentaires.

On se contente de discuter les résultats fournis par le premier plan factoriel qui explique d'ailleurs presque les 3/4 de l'inertie globale. La figure 1 donne le premier cercle de correlation; Le premier axe oppose les variables (Ph et T) aux autres variables, avec une contribution moyenne des variables (Mg^{++}, Ca^{++}, Cl, TDS, Na^+, Hco_3, So_4, No_3). Le deuxième axe oppose les variables (Hco_3, So_4, No_3) aux variables (Mg^{++}, Ca^{++}, Cl, TDS, Na^+) avec une contribution importante de la variable No_3. Pour ce qui est des variables supplémentaires, les coordonnées géographiques semblent fortement liées à la variable Hco_3 et s'opposent aux variables T et Ph.

Dans la figure 2, chaque puit est repéré par ses coordonnées géographiques (X_1, X_2). Le puits est représenté par une indication de la valeur de la première composante principale. On a coupé l'étendue de la première composante principale suivant les quatre quartiles : 1 représente le premier, 2 le second et ainsi de suite. La figure 3 est obtenue à partir de la seconde composante principale. Les valeurs des composantes principales montrent une certaine organisation en fonction de X_1 et de X_2. Le phénomène est plus net pour la première composante principale que pour la seconde.

3. L'ACPVI linéaire

3.1. Rappel sur la méthode

Soit ($\boldsymbol{Y}$, $\boldsymbol{Q}$, $\boldsymbol{D}$) une étude statistique portant sur q variables réponses mesurées sur n individus, $\boldsymbol{Y}$ est la matrice $n \times q$ des observations, $\boldsymbol{Q}$ est une matrice $q \times q$

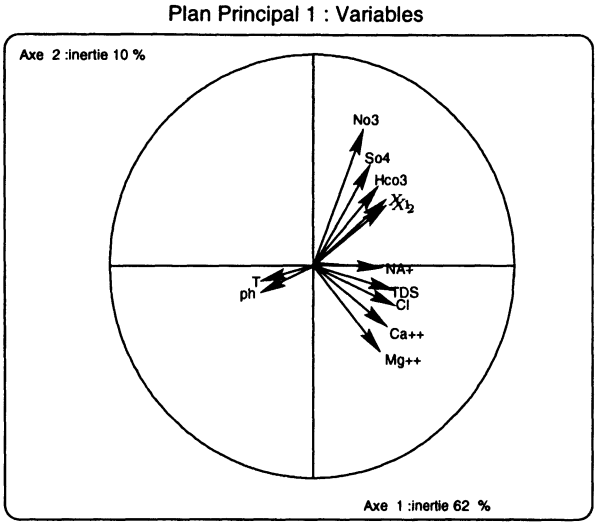

FIGURE 1
Premier cercle de correlation pour une ACP centrée réduite, les variables supplémentaires X_1 et X_2 sont pratiquement confondues en projection sur ce plan

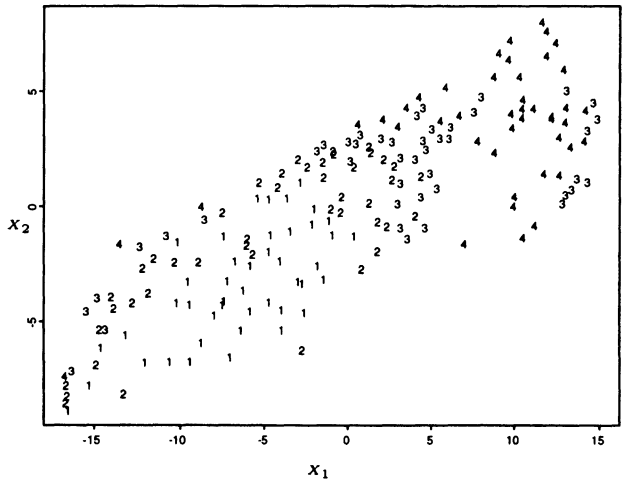

FIGURE 2
Représentation cartographique de la première composante principale

symétrique définie positive qui définit une métrique sur l'espace des individus. La matrice $n \times n$, diagonale $\boldsymbol{D}$ dont les éléments sont positifs et de somme 1, est la matrice des poids associés aux individus. Elle définit une métrique sur l'espace des variables.

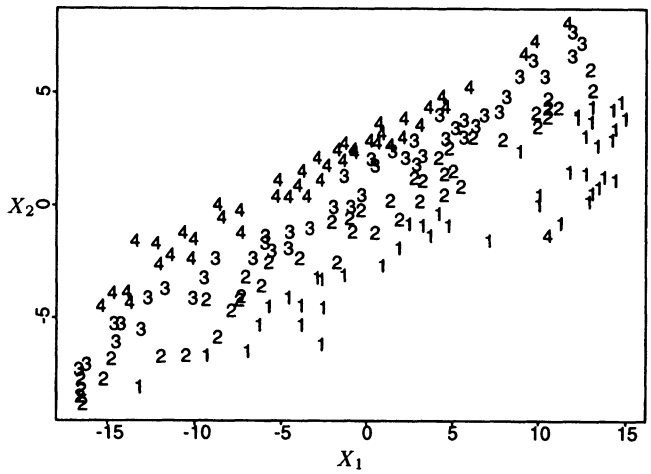

FIGURE 3
Représentation cartographique de la seconde composante principale

Soit $\boldsymbol{X}$ la matrice $n \times p$ des observations sur les p variables instrumentales ou explicatives. On supposera que les n individus sont munis des mêmes poids que précédemment, à savoir les poids définis par la diagonale de $\boldsymbol{D}$.

On utilisera par la suite le D-produit scalaire sur $\mathbb{R}^n$, $(\boldsymbol{x}|\boldsymbol{y})_D = \boldsymbol{x'Dy}$. La matrice $\boldsymbol{YQY'D}$ dont les vecteurs propres sont les composantes principales du triplet $(\boldsymbol{Y}, \boldsymbol{Q}, \boldsymbol{D})$ s'appelle l'opérateur caractéristique de la représentation des individus. Cette matrice est D-symétrique ($\boldsymbol{A}$ est D-symétrique si $\boldsymbol{DA} = \boldsymbol{A'D}$). Robert et Escoufier (1976) ont montré que $\mathrm{tr}(\boldsymbol{AB})$ définit un produit scalaire sur l'espace vectoriel des matrices D-symétriques. On notera la norme associée par $||\boldsymbol{A}|| = (\mathrm{tr}\{\boldsymbol{A}^2\})^{\frac{1}{2}}$.

Définition :

L'ACPVI du triplet $(\boldsymbol{Y}, \boldsymbol{Q}, \boldsymbol{D})$ par rapport à $\boldsymbol{X}$ consiste tout d'abord à déterminer une métrique $\overline{\boldsymbol{R}}$ qui minimise l'écart entre les opérateurs $\boldsymbol{YQY'D}$ et $\boldsymbol{XRX'D}$

$$\overline{\boldsymbol{R}} = \arg\min_{\boldsymbol{R}} ||\boldsymbol{YQY'D} - \boldsymbol{XRX'D}||^2$$

Rappels :

- On trouve dans Bonifas *et al.* (1984) l'expression algébrique de la matrice $\overline{\boldsymbol{R}}$.

$$\overline{\boldsymbol{R}} = (\boldsymbol{X'DX})^{-1}(\boldsymbol{X'DYQY'DX})(\boldsymbol{X'DX})^{-1}$$

- $||\boldsymbol{YQY'D} - \boldsymbol{X\overline{R}X'D}||^2 = ||\boldsymbol{YQY'D}||^2(1 - Rv^2(\boldsymbol{YQY'D}, \boldsymbol{X\overline{R}X'D}))$,

où

$$Rv(\boldsymbol{YQY'D}, \boldsymbol{X\overline{R}X'D}) = \frac{tr(\boldsymbol{YQY'DX\overline{R}X'D})}{\sqrt{\mathrm{tr}(\boldsymbol{YQY'D})^2 \mathrm{tr}(\boldsymbol{X\overline{R}X'D})^2}}$$

En remplaçant $\overline{\boldsymbol{R}}$ dans la dernière équation par son expression, on trouve

$$Rv(\boldsymbol{YQY'D}, \boldsymbol{X\overline{R}X'D}) = \sqrt{\frac{\mathrm{tr}(\boldsymbol{Y'DX}(\boldsymbol{X'DX})^{-1}\boldsymbol{X'DYQ})^2}{\mathrm{tr}(\boldsymbol{YQY'D})^2}}$$

• Soit $\boldsymbol{P_X} = \boldsymbol{X}(\boldsymbol{X'DX})^{-1}\boldsymbol{X'D}$ la matrice de projection D-orthogonale sur le sous espace engendré par les colonnes de $\boldsymbol{X}$.

$\boldsymbol{X\overline{R}X'D}$, opérateur associé au triplet $(\boldsymbol{X}, \overline{\boldsymbol{R}}, \boldsymbol{D})$ est aussi l'opérateur associé au triplet $(\boldsymbol{P_X Y}, \boldsymbol{Q}, \boldsymbol{D})$. Ceci signifie que l'ACP du triplet $(\boldsymbol{X}, \overline{\boldsymbol{R}}, \boldsymbol{D})$ et l'ACP du triplet $(\boldsymbol{P_X Y}, \boldsymbol{Q}, \boldsymbol{D})$ conduisent à la même réprésentation des individus, c'est-à-dire aux mêmes composantes principales associées aux mêmes inerties. De l'écriture suivante $\boldsymbol{Y} = \boldsymbol{P_X Y} + (\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{Y}$, où $\boldsymbol{I}$ désigne la matrice identité, la décomposition

$$\boldsymbol{Y'DYQ} = (\boldsymbol{P_X Y})'\boldsymbol{D}(\boldsymbol{P_X Y})\boldsymbol{Q} + [(\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{Y}]'\boldsymbol{D}[(\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{Y}]\boldsymbol{Q}$$

permet de conclure

$$\mathrm{tr}(\boldsymbol{Y'DYQ}) = \mathrm{tr}((\boldsymbol{P_X Y})'\boldsymbol{D}(\boldsymbol{P_X Y})\boldsymbol{Q} + \mathrm{tr}([(\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{Y}]'\boldsymbol{D}[(\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{Y}]\boldsymbol{Q})$$

Cette équation montre que l'inertie totale du triplet $(\boldsymbol{Y}, \boldsymbol{Q}, \boldsymbol{D})$ peut être décomposée en une part qui correspond aux projections des variables $\boldsymbol{Y}$ dans le sous-espace engendré par $\boldsymbol{X}$ et une part résiduelle qui correspond aux projections des variables $\boldsymbol{Y}$ dans le sous-espace orthogonal.

Le $i^{\text{ème}}$ élément diagonal de la matrice $\boldsymbol{Y'DY}$ est la variance de la variable $\boldsymbol{Y}^i$. Lorsque $\boldsymbol{Q}$ est diagonale, la formule précédente conduit à une décomposition de la variance de chacune des variables en une part expliquée par $\boldsymbol{X}$ et une part résiduelle.

3.2. Test de permutation

Pour juger de la signification de l'influence des variables d'environnement considérées, on a appliqué un test de permutation. C'est une procédure statistique dans laquelle les données sont aléatoirement réaffectées aux différentes unités statistiques. La statistique à laquelle on s'intéresse est calculée pour chacune des permutations; La proportion des valeurs supérieures ou égales à la valeur de la statistique d'avant permutation détermine la valeur P de signification des résultats. Si P est petit, on en déduit que la valeur obtenue avant permutation a peu de chances d'être obtenue par des affectations faites au hasard; dans ce cas l'hypothèse nulle de non relation entre les tableaux $\boldsymbol{Y}$ et $\boldsymbol{X}$ est rejetée.

Si, au contraire P est grand, la valeur obtenue avant permutation ne se différencie des valeurs obtenues par des affectations au hasard; dans ce cas l'hypothèse nulle de non relation entre les tableaux $\boldsymbol{Y}$ et $\boldsymbol{X}$ ne peut pas être rejetée. Les calculs faits ont utilisé pour seuil la valeur 5 %, (voir Edgington, 1987).

L'importance de la relation entre les deux tableaux $\boldsymbol{X}$ et $\boldsymbol{Y}$ est mesurée par l'indice de Stewart et Love

$$T = \frac{\mathrm{tr}(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{Y}\boldsymbol{Q}\boldsymbol{Y}'\boldsymbol{D})}{\mathrm{tr}(\boldsymbol{Y}'\boldsymbol{D}\boldsymbol{Y}\boldsymbol{Q})}$$

Sa distribution est donnée dans Kazi-Aoual F. (1993). On pourrait également utiliser le coefficient Rv présenté dans le paragraphe précédent.

3.3. L'application de l'ACPVI aux données considérées

L'utilisation de l'ACPVI classique sur nos données tient compte de la variation des données analysées et de la corrélation entre les coordonnées géographiques et les variables physico-chimiques.

Le tableau 1 résume la décomposition de la variance des variables initiales. La colonne 3 correspond au sous-espace engendré par les variables instrumentales considérées, alors que la colonne 4 correspond à son orthogonal. La dernière ligne représente la part d'inertie expliquée par cette projection. Cette part est de l'ordre de 25,7 %. Les variables TDS, Ph, cl, Na^+ et Hco_3 ont une part de variance expliquée supérieure ou égale à 25,7 %; on pourra dire que ces variables sont liées à la répartition géographique des puits.

TABLEAU 1
Décomposition de l'inertie par projection
sur le sous-espace engendré par les coordonnées cartésiennes des puits

Variable	Var. initiale	**Proj.**	**Proj. orthogonal**
T	1	0.07647	0.92353
TDS	1	0.39762	0.60238
Ph	1	0.28337	0.71663
Na^+	1	0.50334	0.49666
Cl	1	0.34055	0.65945
Ca^{++}	1	0.11646	0.88354
Mg^{++}	1	0.04631	0.95369
Hco_3	1	0.55530	0.44470
So_4	1	0.11479	0.88521
No_3	1	0.13800	0.86200
Total	10	2.572214	7.427786

La figure 4 montre que le premier axe principal oppose les projections des variables Ph et T aux autres projections de variables.

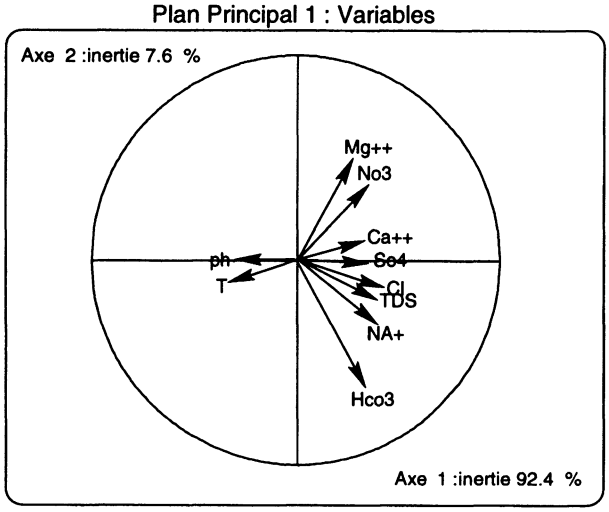

FIGURE 4
Premier cercle de corrélation pour une ACPVI centrée réduite

Le deuxième axe oppose les projections des variables Mg^{++} et No_3 à la projection de la variable Hco_3. On remarque que les projections des variables Ca^{++}, So_4, TDS, Cl et Na^+ sont proches. La corrélation est très forte entre la projection de TDS et la projection de Cl et entre les projections de Ca^{++} et So_4.

Les figures 5 et 6 montrent une organisation des composantes principales sur les coordonnées géographiques. La première composante montre une forte association avec $X_1 + X_2$ tandis que la seconde ne semble liée qu'à X_2. Pour chaque projection, 100 permutations au niveau des unités statistiques ont été effectuées afin de tester la signification de la décomposition de l'inertie. La part d'inertie expliquée n'a jamais été dépassée pour les différentes permutations, ce qui indique une trés grande signification de la décomposition obtenue.

La figure 7 représente les parts d'inertie expliquée pour les différentes permutations concernant les coordonnées cartésiennes des puits.

4. L'approche ACPVI-Spline Additive

Dans la section précédente, l'introduction des composantes spatiales a été faite sous forme linéaire par un processus de projection sur le sous-espace vectoriel qu'elles engendrent. Il est tout à fait naturel de se demander quelle serait la part d'inertie expliquée si l'approche utilisée était non linéaire. On a utilisé pour cela l'ACPVI-Spline Additive (ACPVISA) (voir par exemple Durand et Imam, 1996). Cette méthode

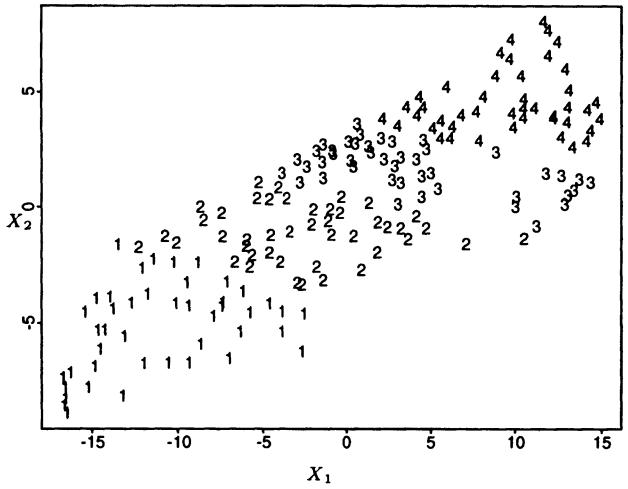

FIGURE 5
Représentation cartographique de la première composante principale. (ACPVI linéaire)

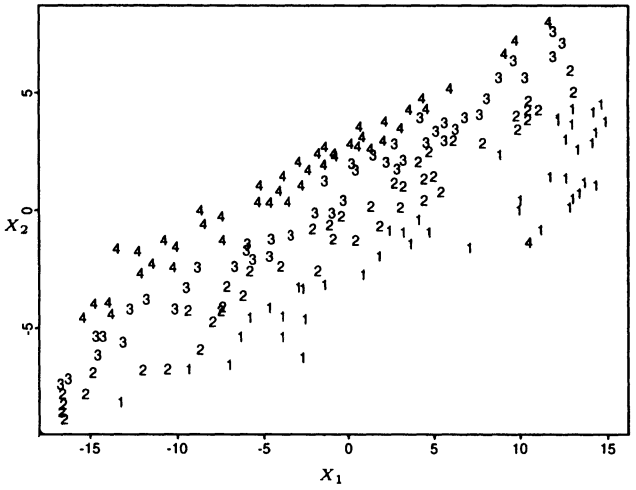

FIGURE 6
Représentation cartographique de la seconde composante principale. (ACPVI linéaire)

combine les caractéristiques de la régression Spline et celles de l'ACP. Elle consiste à rechercher une métrique et une transformation des variables instrumentales par des B-Splines (Schumaker, 1981) de telle sorte que son ACP soit la plus proche possible de celle du tableau initial (Durand, 1993). L'ACPVI linéaire que l'on a utilisée dans le paragraphe précédent en est un cas particulier.

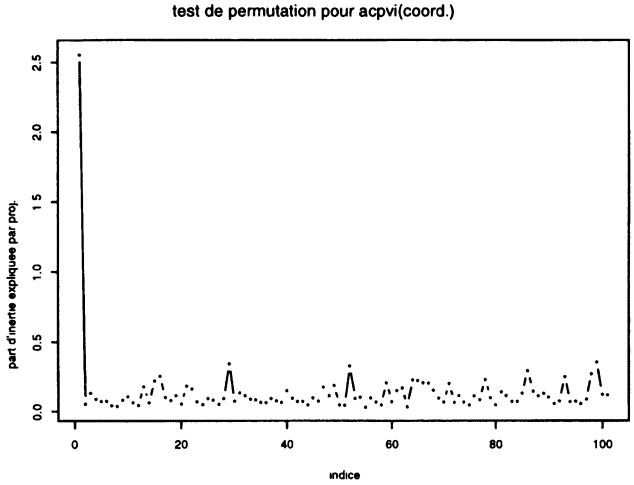

FIGURE 7
Part d'inertie expliquée par projection sur le sous-espace engendré par les variables coordonnées du puits. La première valeur est obtenue avant permutation

Rappel de la méthode :

Les splines sont des fonctions polynomiales par morceaux d'ordre m (de degré $m-1$) qui se raccordent ainsi que certaines de leurs dérivées en K points appelés noeuds intérieurs (De Boor, 1978), (Schumaker, 1981). L'espace des fonctions splines est de dimension $r = m + K$. Une fonction spline $s(z)$ est une combinaison linéaire de fonctions de base appelées B-splines $\{B_j^m(\bullet)\}_{j=1}^r$

$$s(z) = \sum_{j=1}^{r} s_j B_j^m(z)$$

La $j^{\text{ème}}$ colonne de $\boldsymbol{X}$ est remplacée par $\boldsymbol{X}^j(\boldsymbol{s}^j) = \boldsymbol{B}^j\boldsymbol{s}^j$ où $\boldsymbol{B}^j$ est la matrice $n \times r$ de codage de $\boldsymbol{X}^j$, et $\boldsymbol{s}^j$ le vecteur des coefficients splines.

Dans cette méthode, on transforme seulement les prédicteurs (à savoir les coordonnées géographiques X_1 et X_2 ici) par des fonctions splines, les réponses restant inchangées.

Définition :

L'objectif de la méthode est de trouver une matrice $\boldsymbol{X}(\overline{\boldsymbol{s}})$ et une métrique $\overline{\boldsymbol{R}}$ qui minimisent l'écart entre les opérateurs $\boldsymbol{YQY'D}$ et $\boldsymbol{X}(\boldsymbol{s})\boldsymbol{RX'}(\boldsymbol{s})\boldsymbol{D}$

$$(\overline{\boldsymbol{R}}, \overline{\boldsymbol{s}}) = \arg\min_{(\boldsymbol{R},\boldsymbol{s})} ||\boldsymbol{YQY'D} - \boldsymbol{X}(\boldsymbol{s})\boldsymbol{RX'}(\boldsymbol{s})\boldsymbol{D}||^2 .$$

Pour $\boldsymbol{s}$ fixé, une métrique optimale est explicitement calculée en posant $\boldsymbol{X(s)} = \boldsymbol{X}$ dans l'équation

$$\boldsymbol{R} = (\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{Y}\boldsymbol{Q}\boldsymbol{Y}'\boldsymbol{D}\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X})^{-1} \qquad (1)$$

Les équations normales ne fournissent pas de solution explicite pour $\boldsymbol{s}$. Une méthode itérative de relaxation exacte alterne une étape de calcul de $\boldsymbol{R}$ pour $\boldsymbol{s}$ fixé en utilisant la formule (1), et une étape d'un algorithme de la plus forte descente en la variable $\boldsymbol{s}$, $\boldsymbol{R}$ étant fixé jusqu'à la convergence, voir (Durand, 1993) et (Imam et Durand, 1997).

L'ACP du triplet $(\boldsymbol{X}(\overline{\boldsymbol{s}}), \overline{\boldsymbol{R}}, \boldsymbol{D})$ est équivalente à celle du triplet $(\boldsymbol{P}_{\boldsymbol{X}(\overline{\boldsymbol{s}})}\boldsymbol{Y}, \boldsymbol{Q}, \boldsymbol{D})$. Ceci met en évidence le lien avec les lisseurs additifs multiréponses

$$\widehat{\boldsymbol{Y}} = \boldsymbol{P}_{\boldsymbol{X}(\overline{\boldsymbol{s}})}\boldsymbol{Y}$$

Chaque variable explicative contribue additivement à l'approximation de chaque variable réponse.

$$\widehat{Y}^i = \sum_{j=1}^{p} \boldsymbol{B}^j \overline{\boldsymbol{s}}^j \overline{\boldsymbol{M}}_{ji}\,, \qquad i = 1, \ldots, q$$

Avec $\overline{\boldsymbol{M}} = (\boldsymbol{X}'(\overline{\boldsymbol{s}})\boldsymbol{D}\boldsymbol{X}(\overline{\boldsymbol{s}}))^{-1}\boldsymbol{X}(\overline{\boldsymbol{s}})\boldsymbol{D}\boldsymbol{Y}$. On peut alors tracer $\boldsymbol{B}^j\overline{\boldsymbol{s}}^j$ en fonction de $\boldsymbol{X}^j$ ce qui permet de visualiser la transformation de la variable $\boldsymbol{X}^j$ trouvée par la méthode.

On pourrait également, pour chaque variable $\boldsymbol{Y}^i$, tracer $\boldsymbol{B}^j\overline{\boldsymbol{s}}^j\overline{\boldsymbol{M}}_{ji}$ en fonction de $\boldsymbol{X}^j$. On aurait alors une visualisation de l'influence de chacune des variables transformées dans l'explication de la variabilité de $\boldsymbol{Y}^i$.

4.1. L'application de l'ACPVISA sur les données

L'utilisation des transformations splines des coordonnées géographiques est conforme à l'idée d'analyse multivariée introduite en Ecologie par Gittins (1968) et en Géologie par Lee (1969). L'analyse se fait alors en couplant le tableau de données initial avec le tableau des coefficients du polynôme en (x, y) dont le degré a été fixé à priori (Thioulouse, 1985). On fait alors, soit l'analyse canonique (Gittins, 1968), soit, l'analyse canonique des correspondances ou l'analyse des redondances (Brocard *et al.*, 1992).

Pour ce qui est des données que l'on a traitées, les résultats obtenus par l'approche ACPVI-Spline Additive sont dans le tableau 2. On remarque une nette amélioration au niveau de l'inertie expliquée par projection sur le sous-espace des coordonnées transformées : elle est de 44.2 % contre 25.7 % pour l'ACPVI-linéaire. Les variables TDS, Na^+, Cl et Hco_3 ont une part de variance expliquée supérieure à 44.2 %; à la différence du cas linéaire, la variable Ph n'est plus parmi les variables les plus liées aux variables explicatives. Sa part de variance expliquée est, comme Ca^{++}, proche de la moyenne des variances expliquées.

TABLEAU 2
Décomposition de l'inertie par projection sur le sous-espace engendré par les coordonnées transformées (utilisation de l'ACPVI-Spline Additive)

Variable	Var. initiale	**Proj.**	**Proj. orthogonal**
T	1	0.22428	0.77572
TDS	1	0.74721	0.25279
Ph	1	0.41450	0.58550
Na^{+}	1	0.81447	0.18553
Cl	1	0.71263	0.28737
Ca^{++}	1	0.41896	0.58104
Mg^{++}	1	0.30386	0.69614
Hco_3	1	0.47551	0.52449
So_4	1	0.21677	0.78323
No_3	1	0.09691	0.90309
Total	10	4.425097	5.574903

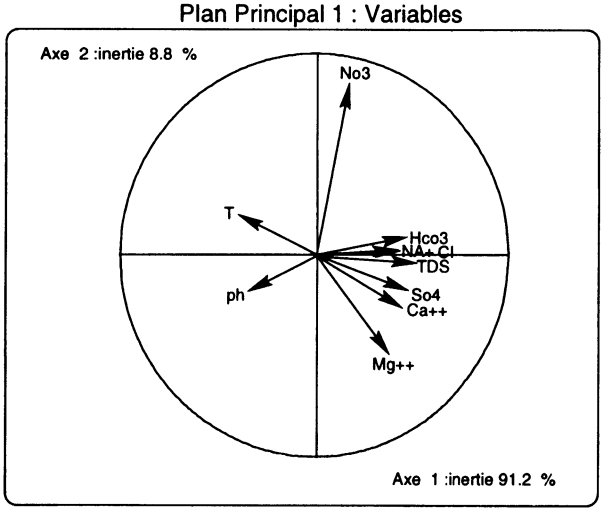

FIGURE 8
Premier cercle de corrélation pour une ACPVI Spline Additive.

L'utilisation du test de permutation dans le cas non-linéaire a donné des résultats significatifs semblables à ceux du cas linéaire; On précise qu'après chaque affectation aléatoire des données au niveau du tableau des variables instrumentales, il y a eu d'abord recherche des meilleures transformations et métrique sur $\boldsymbol{X}(s)$ pour que l'ACP de $(\boldsymbol{X}(s), \boldsymbol{R}, \boldsymbol{D})$ soit la plus proche de celle de $(\boldsymbol{Y}, \boldsymbol{Q}, \boldsymbol{D})$.

La transformation des coordonnées géographiques par des splines favorise la corrélation entre les variables Hco_3, Cl, Na^+ et TDS. En plus on a une contribution très forte pour la variable No_3 et une opposition par rapport au deuxième axe principal entre No_3 et Mg^{++} ce qui n'est pas le cas dans l'analyse linéaire.

Remarque :

Les variances expliquées pour Hco_3 et No_3 sont plus faibles que dans le tableau 1. En effet, l'optimisation globale du critère ne repose pas sur une optimisation individuelle pour chaque variable instrumentale.

4.2. Déformation spatiale

L'ACPVI Spline Additive a fourni le tableau $\boldsymbol{X}(\bar{s})$ des transformées des coordonnées géographiques. La figure 9 montre les deux transformations. Sauf pour les plus grandes valeurs de X_2, la transformation de cette variable est monotone et les plus grandes valeurs ont des transformées qui se confondent avec celles des valeurs intermédiaires de X_2. On peut dire que la transformation tasse les valeurs supérieures de X_2.

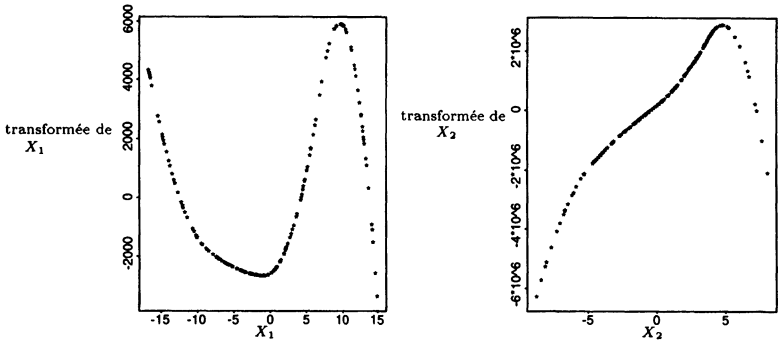

FIGURE 9
Les deux transformations de X_1 et de X_2.

La transformation de X_1 est plus complexe. Elle s'apparente à un polynôme de degré 3. Les quartiles de la transformée regroupent des valeurs issues de quartiles différents de la variable initiale.

Puisque la représentation des individus (les puits) sont les mêmes dans les études $(\boldsymbol{X}(\bar{s}), \overline{\boldsymbol{R}}, \boldsymbol{D})$ et $(\boldsymbol{P}_{\boldsymbol{X}}(\bar{s}), \boldsymbol{Q}, \boldsymbol{D})$, on représente les composantes principales du second triplet par rapport aux coordonnées géographiques transformées. On voit dans la figure 10 que la première composante principales de l'ACPVI spline additive indique une association avec $f(X_1) + g(X_2)$, où $f(X_1)$ et $g(X_2)$ sont les transformations de X_1 et de X_2 montrées dans la figure 9.

En ce qui concerne la figure 11, on remarque que la seconde composante principale est liée à $g(X_2)$.

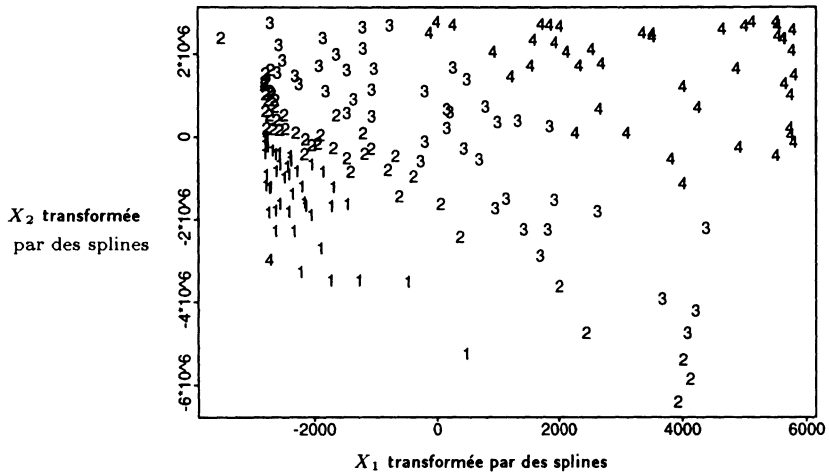

FIGURE 10
Représentation cartographique de la première composante principale. (ACPVI Spline Additive)

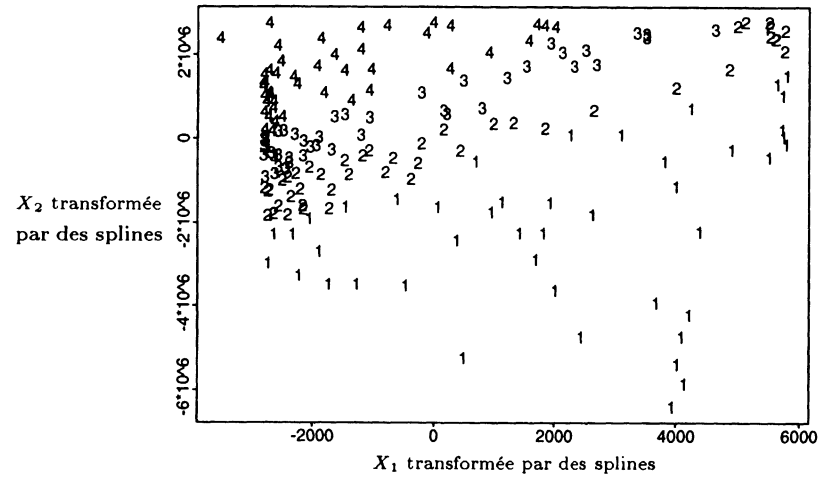

FIGURE 11
Représentation cartographique de la seconde composante principale. (ACPVI Spline Additive)

Si on calcule les variances des $B^j \bar{s}^j \overline{M}_{ji}$ (fonctions coordonnées), on voit alors que la transformée de X_2 a une plus grande variance que celle de X_1 et ceci quelque soit la variable $\mathbf{Y}^i$ à expliquer. On en déduit une prépondérance de la transformée de X_2 dans l'explication des variables physico-chimiques.

5. Conclusion

L'ACPVI linéaire puis l'ACPVI Spline Additive ont permis de mettre en évidence l'influence des coordonnées géographiques sur les variables physico-chimiques. On obtient dans les deux approches une décomposition des variances de chacune des variables physico-chimiques en une part expliquée par les variables géographiques et une part résiduelle.

L'ACPVI-linéaire ne recherche que des explications linéaires des variables physico-chimiques. Son avantage réside dans sa simplicité de mise en oeuvre et de compréhension. L'ACPVI Spline Additive met en évidence, si elles existent, des liaisons plus complexes. La possibilité de visualiser les transformations des coordonnées géographiques donne un élément important d'interprétation.

Dans les deux approches, les tests de permutation permettent de juger de la signification de la variabilité totale. La disponibilité de cette démarche confirmatoire justifie, si besoin est, la nécessité d'une prise en compte de données concomittantes dans une démarche exploratoire pour tout plan expérimental.

Remarque :

La démarche peut se poursuivre par la recherche conjointe de transformations des coordonnées géographiques et des variables physico-chimiques. Ceci sera étudié dans la thèse de Imam Waël au Laboratoire de Biométrie à l'ENSAM de Montpellier.

Remerciements

Nous remercions, très sincèrement, l'ensemble du personnel du laboratoire de Biométrie de l'INRA de Montpellier pour son accueil et la mise à notre disposition de moyens informatiques qui ont permis le traitement statistique et la rédaction de cet article. Un remerciement tout particulier est adressé à J.F. Durand qui accompagne depuis deux ans le travail de Imam Waël.

Bibliographie

[1] BONIFAS L., ESCOUFIER Y., GONZALEZ P.L., SABATIER R., (1984), «Choix de variables en Analyse en Composantes Principales», Revue de Statistique Appliquée vol. XXXII, n° 2, pp. 5-15.

[2] BROCARD D., LEGENDRE P., DRAPEAU P., (1992), «Partialling out the spatial component of ecological variation», Ecology 73(3) pp. 1045-1055.

[3] CHESSEL D., MERCIER P., (1993), «Couplages de triplets statistiques et liaisons espèces-environnement», Biométrie et Environnement, J.D. Lebreton, B. Asselain, Ed. Masson.

[4] DE BOOR, C., (1978), A practical guide to splines, New York : Springer.

[5] DURAND J.F., (1993), «Generalized principal component analysis with respect to instrumental variables via univariate spline transformations», Computational Statistics and Data Analysis, n° 16, pp. 423-440.

[6] DURAND J.F., IMAM W., (1996), «Une extension non linéaire de l'Analyse en Composantes Principales sur Variables Instrumentales, Rapport Technique N° 96-07, Unité de Biométrie.

[7] EDGINGTON E.S., (1987), «Randomization tests», Second édition, M. Dekker Editor.

[8] ESCOUFIER Y., HOLMES S., (1988), «Décomposition de la variabilité dans les Analyses Exploratoires : Un exemple d'Analyse en Composantes Principales en présence de Variables Qualitatives Concomittantes.», Rapport Technique N° 88-04, Unité de biométrie.

[9] FRAILE L., ESCOUFIER Y., RAIBAUT A., (1993), «Analyse des Correspondances de données planifiées : étude de la chémotaxie de la larve infestante d'un parasite.»Biometrics 49 pp. 1142-53.

[10] GITTINS R., (1968), «Trend surface analysis of ecological data», Journal of Ecology 56 pp. 845-869.

[11] IMAM W., DURAND J.F., (1997), «Une extension spline additive de l'Analyse en Composantes Principales sur Variables Instrumentales, XXIX^{e} Journées de Statistique ASU.468-471.

[12] KAZI-AOUAL F., (1993), «Approximations to permutation tests data analysis», Rapport Technique N° 93-06, Unité de Biométrie».

[13] LEE P.J., (1969), «The theory and application of canonical trend surfaces», Journal of Geology 77 pp. 303-318.

[14] ROBERT P., ESCOUFIER Y., (1976), «A unifying tool for linear-multivariate statistical methods : the Rv-coefficient», Applied Statistics, n° 25, pp. 257-265.

[15] SABATIER R., (1994), «Deux nuages de Rn dont un à expliquer», Notes de cours de DEA-Biostatistique (INRA-ENSAM-UMII).

[16] SABATIER R., CHESSEL D., (1993), «Couplages de triplets statistiques et graphes de voisinage». Biométrie et Environnement, J.D. Lebreton, B. Asselain, Ed. Masson.

[17] SABATIER R., LEBRETON J.D., Chessel D., (1989), «Principal Component Analysis with Instrumental Variables as a tool for modelling composition Data», «in Multiway Data Analysis, R. Coppi and S. Bolasco (Editors).

[18] SCHUMAKER L.L., (1981), «Spline Functions : Basic Theory», Wiley, New York.

[19] THIOULOUSE J., (1985), «Structures spatio-temporelles en biologie des populations d'insectes. Application à l'étude de l'altise de Colza : résultats méthodologiques et biologiques», Thèse de 3^{e} cycle, Lyon 1.

[20] THIOULOUSE J., (1996), «Outils logiciels, méthodes statistiques et implications biologiques : Une approche de la biométrie», mémoire d'habilitation à diriger des recherches, Université de Claude Bernard-Lyon I.