REVUE DE STATISTIQUE APPLIQUÉE

L. BELLANGER
D. BAIZE
R. TOMASSONE

L'analyse des corrélations canoniques appliquée à des données environnementales

Revue de statistique appliquée, tome 54, n° 4 (2006), p. 7-40
<http://www.numdam.org/item?id=RSA_2006__54_4_7_0>

Rev. Statistique Appliquée, 2006, **LIV** (4), 7-40

L'ANALYSE DES CORRÉLATIONS CANONIQUES APPLIQUÉE À DES DONNÉES ENVIRONNEMENTALES

L. BELLANGER[1], D. BAIZE[2], R. TOMASSONE[3]

(1) *Département de Mathématiques Jean Leray – UMR 6629, Université de Nantes*
BP 92208, 44322 Nantes Cedex 3.
lise.bellanger@math.univ-nantes.fr
(2) *INRA, Science des Sols, Centre d'Orléans – BP20619, 45166 Olivet Cedex.*
denis.baize@orleans.inra.fr
(3) *Institut National Agronomique, Département de Mathématique, 75231 Paris Cedex 05.*
rr.tomassone@wanadoo.fr

RÉSUMÉ

L'analyse des corrélations canoniques est une vieille méthode statistique surtout connue pour ses qualités théoriques, puisqu'elle englobe de nombreuses autres méthodes. Nous essayons, dans cet article, de montrer que l'interprétation des résultats qu'elle fournit n'est guère plus difficile que celle de méthodes plus largement employées comme la régression multiple ou l'analyse en composantes principales. Dans le domaine des données environnementales, elle pourrait constituer un outil de référence dès qu'il s'agit de mettre en relation des ensembles de variables. L'analyse est illustrée par l'étude des relations entre la teneur en éléments traces métalliques de grains de blé en fonction de certaines caractéristiques des sols où les blés ont poussé, en particulier après épandage de boues d'épuration.

Mots-clés : *analyse des corrélations canoniques, régression, ré-échantillonnage, éléments traces métalliques, sol, blé.*

ABSTRACT

Canonical correlations analysis is an old method well known as a key one of a lot of others. We try to show that the interpretation of results it furnishes is no more difficult than other ones, widely used, as regression or principal components analysis. For environmental data it could be a reference tool when relations between two groups of variates are concerned. Analysis is illustrated by a study of relations between trace metals in wheat and soil characteristics, particularly after sewage sludge spreading.

Keywords : *canonical correlation analysis, regression, resampling methods, trace metals, soil, wheat.*

1. Introduction

Parmi les nombreux problèmes classiques de la Statistique, celui de l'étude de la relation entre variables est sans nul doute l'un des plus fréquents : on calcule le coefficient de corrélation entre deux variables, on estime les paramètres d'un modèle de régression d'une variable à expliquer en fonction d'une ou plusieurs autres (les régresseurs ou variables explicatives) pour tenter d'« expliquer » cette variable et éventuellement de la prédire pour d'autres valeurs des régresseurs. Quand on dispose de deux groupes de variables une méthode, l'*Analyse des Corrélations Canoniques* souvent appelée *Analyse Canonique* (ultérieurement notée *ACC*), existe depuis bien longtemps [11]. Bien que de nombreux logiciels offrent un programme pour réaliser les calculs, elle ne semble pas bénéficier d'une « bonne réputation » : peu de publications avec des applications l'utilisent. Les articles les plus récents de la *Revue de Statistique Appliquée* datent des années 1987-1992, encore que les corpus de données auxquels ils s'appliquent soient de nature différente [7], [20], [21], [15].

Est-ce une méthode sans intérêt, trop difficile d'emploi, pour tout dire « maudite » ? À première vue, il semblerait que ce soit le cas si nous reprenons quelques ouvrages la présentant :

– « *La méthode d'analyse canonique ... présente un intérêt assez limité pour les applications, car elle conduit à de grandes difficultés d'application. Cependant elle joue un rôle théorique important : en effet, elle constitue un cadre général dont la régression multiple, la plupart des méthodes d'analyse des données ... et l'analyse discriminante sont des cas particuliers* » [16], pp.275.

– « *Si les applications directes de l'analyse canonique sont peu nombreuses, elle n'en constitue pas moins une méthode fondamentale car sa démarche se retrouve dans d'autres méthodes comme l'analyse des correspondances ou l'analyse discriminante* » [22] , pp.188.

– « *Canonical analysis is often cooly received despite a lack of suitable alternatives* ». « *Surprisingly, substantive applications of these methods in ecology are few. In practice, all too often other less suitable forms of analysis are pressed into service for the purpose. Among these multiple regression analysis and principal components analysis are frequently encountered* » [10], pp.1.

– En 2004, les auteurs du plus récent des ouvrages tiennent le même discours : « *Canonical correlation analysis is one of the less commonly used multivariate techniques. Its limited use may be due, in part, to the difficulty often encountered in trying to interpret the results* » [1], pp.234.

Pour résumer, l'*ACC* est caractérisée par :

– une interprétation des résultats souvent délicate;

– mais un intérêt théorique essentiel fournissant un cadre unificateur à un certain nombre d'autres méthodes.

Nous ne reviendrons pas sur le second aspect bien connu, mais nous voulons à partir d'un exemple montrer qu'on peut tout de même exploiter les résultats fournis par une analyse canonique, même si l'exploitation peut s'avérer complexe.

Les données que nous allons présenter proviennent d'une étude qui peut s'apparenter à un « cas d'école » pour l'analyse canonique : en 1998, le Ministère de l'Aménagement du Territoire et de l'Environnement a lancé le programme GESSOL (Fonctions environnementales des sols et GEStion du patrimoine SOL[3]). Une des questions fondamentales de ce programme était :

« *Est-il possible de bâtir des modèles permettant de détecter par avance les cas de concentrations excessives en éléments traces métalliques (ETM) dans les grains de blé à partir de données pertinentes acquises sur des échantillons de sol ?* ».

Le problème est d'une extrême importance pour de multiples raisons liées à l'évolution des pratiques agricoles; en particulier celle liée à l'épandage de boues d'épuration riches en *ETM* [8] et aux polémiques qui en découlent [23]. Actuellement, les publications sur le sujet [19], pour intéressantes qu'elles soient, sont des compilations de résultats d'essais agronomiques sur de nombreuses plantes. Les seules méthodes d'analyse utilisées sont la régression linéaire et l'analyse des composantes principales. Les résultats statistiques des régressions se limitent à une équation, une valeur du coefficient de détermination (R^2), mais aucune analyse critique de la validité de ces régressions n'est faite.

Nous allons d'abord présenter le corpus de données qui doit nous aider à répondre à la question posée (§ 2); nous rappellerons ensuite la démarche classique de l'analyse canonique (§ 3) et nous donnerons une première interprétation des résultats (§ 4); enfin nous montrerons que des indices rarement utilisés peuvent faciliter l'interprétation (§ 5).

2. Les données : adéquation à l'objectif du programme de recherche

2.1. Le corpus de données

Il est constitué par un échantillon de $n = 198$ sites étudiés selon le même protocole dans diverses régions de France. Il s'agit de sols agricoles « ordinaires », c'est-à-dire non pollués et n'ayant pas reçu de boues d'épuration (sauf une douzaine de cas particuliers [8]). Ils appartiennent à 18 familles pédo-géologiques contrastées. Sur chaque site, des grains de blé ont été récoltés à maturité sur $1m^2$ (variété « Soissons » ou « Trémie »). Au pied du blé ainsi récolté, l'horizon de surface labouré du sol a été également prélevé. Sur des échantillons séchés et tamisés à $2mm$ de ces horizons de surface, nous avons déterminé :

– 9 variables caractéristiques des propriétés agro-pédologiques classiques : granulométrie 5 fractions (argile : A; limon fin et grossier : LF, LG; sable fin et grossier : SF, SG) [1] , le carbone organique (CS), le pH mesuré après agitation dans l'eau (pH), le calcaire ($CaCO_3$) et la capacité d'échange cationique (CEC); ces variables sont des teneurs, sauf le pH et la CEC.

– 8 variables représentant les concentrations totales des métaux du sol obtenues après mise en solution par les acides fluorhydrique et perchlorique selon la norme NF ISO 14869-1 : FeS, MnS, CdS, CrS, CuS, NiS, PbS et ZnS.

[1] A + LF + LG + SF + SG = 100%, mais naturellement pas la somme de leur logarithme.

– et 8 variables qui sont les concentrations en métaux extraits par deux réactifs, DTPA (DiéthylèneTriamine-PentaAcétique) et NH_4NO_3 (nitrate d'ammonium), choisis pour leur capacité à atteindre seulement les formes chimiques les plus réactives et les plus susceptibles d'être absorbées par les racines des plantes. Les quantités extraites au DTPA correspondraient plutôt aux métaux associés aux matières organiques et aux oxydes de fer, tandis que celles extraites par le NH_4NO_3 correspondraient plutôt aux formes métalliques échangeables, les plus phyto-disponibles. Soit : CdD, CuD, PbD et ZnD (pour DTPA), CdN, CuN, PbN et ZnN (pour NH_4NO_3).

– 7 variables représentant les concentrations dans les grains de blé en CdB, CrB, CuB, FeB, NiB, PbB et ZnB. Notons que deux autres variables potentiellement intéressantes MgB et MnB n'ont pu être mesurées que sur les 162 premiers sites.

Nous avons donc deux groupes de variables :

– 25 variables SOL :

{A, LF, LG, SF, SG, CEC, $CaCO_3$, CS, pH, CdS, CrS, CuS, FeS, MnS, NiS, PbS, ZnS, CdD, CuD, PbD, ZnD, CdN, CuN, PbN, ZnN}

– 7 variables BLE :

{CdB, CrB, CuB, FeB, NiB, PbB, ZnB}+2 {MgB, MnB} sur un échantillon de moindre taille.

2.2. *Difficultés* a priori

Avec ce corpus de données est-il possible de répondre à la question fondamentale du programme GESSOL : prévoir la teneur en *ETM* de grains de blé en utilisant des données analytiques de l'horizon de surface du sol dans lequel ce blé a été cultivé ? Une telle prétention se heurte à d'évidentes difficultés de principe :

– ce qui se passe au champ à l'interface entre les racines et la solution du sol n'est pas bien décrit par des analyses réalisées au laboratoire sur des échantillons de sol séchés et tamisés à 2 mm !

– le rôle des autres couches du sol (horizons profonds) est complètement négligé;

– les processus de redistribution des éléments absorbés au niveau des racines vers les divers organes de la plante ne sont pas pris en compte;

– de même que ne sont pas pris en compte les synergies et les antagonismes intervenant à l'échelle des cellules des végétaux.

Cependant de nombreux scientifiques dans le monde entier utilisent cette approche car elle est très simple à mettre en œuvre. Nous ferons de même tout en sachant que nous ne pourrons pas décrire de façon parfaite la relation entre le sol et le blé.

2.3. Premiers regards sur le corpus de données

Une question préalable à l'analyse est le choix du corpus : doit-on travailler sur les 198 sites ou sur les 162 qui ont l'avantage de contenir toutes les variables importantes disponibles ? Il n'existe pas de régle absolue pour répondre à ce type de question. Si les deux corpus sont assez voisins, il est sans doute préférable de travailler sur celui qui contient davantage de variables, même s'il est de taille plus réduite. Les 36 sites qui sont exclus peuvent ultérieurement servir de données complémentaires pour valider les résultats.

Les distributions de chacune des 34 variables s'étant avérées très dissymétriques (les grandes valeurs sont relativement rares), une transformation logarithmique s'est imposée. La conséquence immédiate pour l'interprétation est que nous devrons penser à des produits ou des rapports de variables et non à des sommes ou des différences. Un examen plus détaillé montre que PbN n'a que 43 valeurs différentes de zéro; en fait, les valeurs nulles sont décrites par le laboratoire d'analyse comme « inférieures à un certain seuil de quantification », ce qui est une difficulté classique dans les dosages chimiques; ce sont des valeurs nulles ou très proches de zéro; mais ce ne sont pas des données manquantes. Dans une étude dans laquelle nous allons analyser des variations simultanées de variables, il est préférable d'éliminer cette variable. Le problème est sensiblement voisin pour $CaCO_3$ (47 valeurs différentes de zéro); à la différence que $CaCO_3$ vaut effectivement zéro dans tous les cas où les sols ne sont pas calcaires. Etant donné l'importance possible du carbonate de calcium, le pédologue souhaite qu'on la conserve.

En conclusion, nous allons travailler sur le corpus de $n = 162$ sites avec :

– 24 variables SOL : {A, LF, LG, SF, SG, CEC, $CaCO_3$, CS, pH, CdS, CrS, CuS, FeS, MnS, NiS, PbS, ZnS, CdD, CuD, PbD, ZnD, CdN, CuN, ZnN}, ensemble défini par une matrice $\mathbf{X}_{162\times 24}$,

– 9 variables BLE : {CdB, CrB, CuB, FeB, NiB, PbB, ZnB, MgB, MnB}, ensemble défini par une matrice $\mathbf{Y}_{162\times 9}$.

La forme des distributions de ces deux ensembles de variables est fournie sur les figures 1 et 2. Nous aurions pu les présenter sous la forme classique d'histogrammes, toutefois elles sont plus « parlantes » avec une version lissée obtenue par estimation de la densité (le lecteur intéressé pourra consulter sur ce sujet [25] (pp.132-138)). Ces distributions sont d'autant plus intéressantes pour la suite de l'analyse que leur forme est voisine de celle de la distribution Normale ou pour le moins symétrique; c'est le cas pour la majorité d'entre elles sauf pour $CaCO_3$ (et nous avons vu pourquoi ci-dessus) et pour PbB qui a aussi de nombreuses valeurs au-dessous du seuil de détection. Certaines présentent une bimodalité généralement peu marquée. Notons que les figures obtenues pour les 198 sites (sauf évidemment pour MgB et MnB) ont toutes la même allure; nous pouvons donc raisonnablement penser que les 36 sites exclus ne sont pas très différents du corpus que nous allons maintenant étudier.

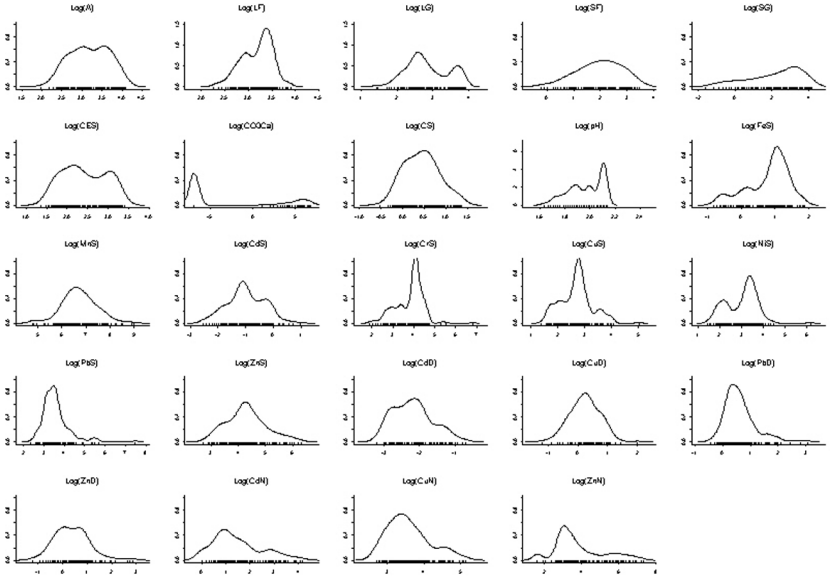

FIGURE 1
Distributions des 24 variables SOL

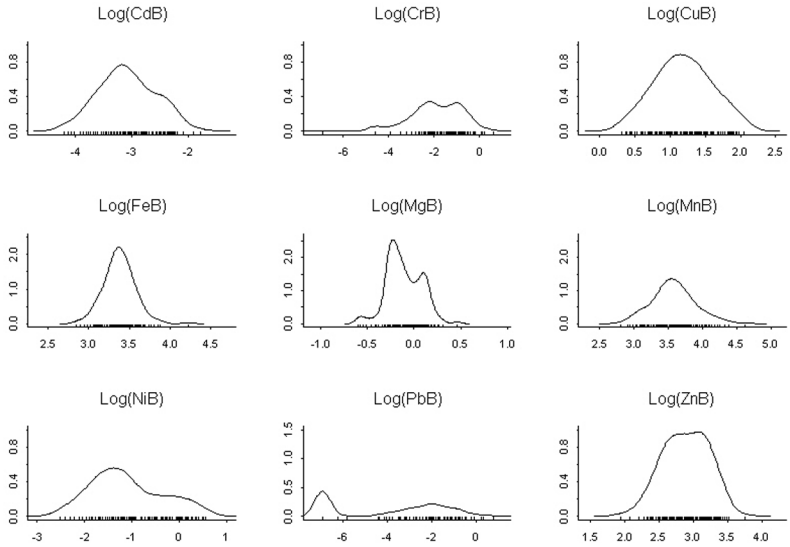

FIGURE 2
Distributions des 9 variables BLE

3. Principe de l'analyse des corrélations canoniques

Pour étudier la relation entre deux ensembles de variables definis chacun par une matrice $\mathbf{X}_{n\times p}$ pour le premier et $\mathbf{Y}_{n\times q}$ pour le second l'*ACC* va être un outil privilégié. Le nombre de lignes n de chaque matrice est identique ; $\mathbf{X}$ a p colonnes et $\mathbf{Y}$ en q; nous supposerons que $rang(\mathbf{X}) = p$ et $rang(\mathbf{Y}) = q$. Les lignes représentent les individus ou les observations : une observation i est représentée par un vecteur séparé en deux :

$$\mathbf{x}_i = [x_{i1}, ..., x_{ip}]^T \text{ et } \mathbf{y}_i = [y_{i1}, ..., y_{iq}]^T, \quad (i = 1, ..., n)$$

Les deux matrices de données sont donc :

$$\mathbf{X}_{n\times p} = [\,\mathbf{x}^1 \quad \dots \quad \mathbf{x}^k \quad \dots \quad \mathbf{x}^p\,] \quad \text{et} \quad \mathbf{Y}_{n\times q} = [\,\mathbf{y}^1 \quad \dots \quad \mathbf{y}^l \quad \dots \quad \mathbf{y}^q\,]$$

où $\mathbf{x}^k$ (resp. $\mathbf{y}^l$) est le vecteur variable de composantes x_{ik} (resp. y_{il}), $(1 \leqslant i \leqslant n)$. Les variables des deux groupes $\mathbf{x}^k$ et $\mathbf{y}^l$, représentées par des vecteurs de $\mathbb{R}^n$, sont supposées centrées. Ainsi, la matrice de covariances expérimentales des $p + q$ caractères s'écrit :

$$\mathbf{S} = \frac{1}{n}\begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Y} \\ \mathbf{Y}^T\mathbf{X} & \mathbf{Y}^T\mathbf{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix}, \quad \mathbf{S}_{12} = \mathbf{S}_{21}^T.$$

3.1. Formulation classique

L'idée initiale de Hotelling [11] a consisté à rechercher deux combinaisons linéaires l'une de $\mathbf{x}^1, \dots, \mathbf{x}^p$ définie par un premier vecteur à p composantes $\mathbf{a}_{p\times 1}$, l'autre de $\mathbf{y}^1, \dots, \mathbf{y}^q$, définie par un second vecteur à q composantes $\mathbf{b}_{q\times 1}$, telle que les vecteurs $\mathbf{a}^T = [\,a_1 \quad \dots \quad a_k \quad \dots \quad a_p\,]$ et $\mathbf{b}^T = [\,b_1 \quad \dots \quad b_l \quad \dots \quad b_q\,]$ maximisent le coefficient de corrélation entre $\mathbf{u} = \underbrace{\mathbf{X}}_{n\times p}\underbrace{\mathbf{a}}_{p\times 1}$ et $\mathbf{v} = \underbrace{\mathbf{Y}}_{n\times q}\underbrace{\mathbf{b}}_{q\times 1}$ vecteurs de $\mathbb{R}^n$. $\mathbf{u}$ et $\mathbf{v}$ sont appelés *variables canoniques*, tandis que les vecteurs de coefficients $\mathbf{a} \in \mathbb{R}^p$ et $\mathbf{b} \in \mathbb{R}^q$ sont appelés *facteurs canoniques*.

On montre que ce problème se résume en fait à :

– obtenir les vecteurs de coefficients $\mathbf{a} \in \mathbb{R}^p$ et $\mathbf{b} \in \mathbb{R}^q$ qui rendent maximal $cor(\mathbf{u}, \mathbf{v}) = r = \frac{1}{n}\mathbf{a}^T\mathbf{X}^T\mathbf{Y}\mathbf{b}$

– de telle sorte que les deux combinaison linéaires soient de variance 1, soit :

$$\begin{cases} \mathbf{a}^T\mathbf{S}_{11}\mathbf{a} = 1 \\ \mathbf{b}^T\mathbf{S}_{22}\mathbf{b} = 1 \end{cases}$$

Lorqu'un premier couple de caractères $\left(\mathbf{u}^1, \mathbf{v}^1\right)$ a été obtenu, on recherche, un deuxième couple $\left(\mathbf{u}^2, \mathbf{v}^2\right)$ tel que $r_2 = cor(\mathbf{u}^2, \mathbf{v}^2)$ soit maximal et $cor(\mathbf{u}^2, \mathbf{u}^1) =$

$cor(\mathbf{u}^2, \mathbf{v}^1) = cor(\mathbf{v}^1, \mathbf{v}^2) = 0$ et ainsi de suite. Il existe au moins $s = \min(p, q)$ couples de tels vecteurs $\left(\mathbf{u}^k, \mathbf{v}^k\right)$.

Matriciellement, le problème se ramène donc à trouver deux matrices de poids :

$$\mathbf{A}_{p\times s} = [\,\mathbf{a}^1 \quad \ldots \quad \mathbf{a}^k \quad \ldots \quad \mathbf{a}^s\,] \text{ et } \mathbf{B}_{q\times s} = [\,\mathbf{b}^1 \quad \ldots \quad \mathbf{b}^l \quad \ldots \quad \mathbf{b}^s\,]$$

permettant de calculer deux matrices $n \times s$, $\mathbf{U} = \underbrace{\mathbf{X}}_{n\times p}\underbrace{\mathbf{A}}_{p\times s}$ et $\mathbf{V} = \underbrace{\mathbf{Y}}_{n\times q}\underbrace{\mathbf{B}}_{q\times s}$, telles que la matrice de covariances de la matrice transformée $[\,\mathbf{U} \quad \mathbf{V}\,]$ ait la forme plus simple :

$$var\,[\,\mathbf{U} \quad \mathbf{V}\,] = \begin{bmatrix} \mathbf{A}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{B}^T \end{bmatrix} \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_s & \mathbf{\Gamma} \\ \mathbf{\Gamma} & \mathbf{I}_s \end{bmatrix}$$

où :

$$\mathbf{\Gamma} = \mathbf{diag}(r_k), \quad 1 \geqslant r_1 \geqslant .... \geqslant r_s \geqslant 0.$$

La recherche des deux matrices de poids fournit simultanément les s coefficients de corrélation; elle s'obtient par la recherche des solutions de :

$$\left(\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{21} - \gamma^2\mathbf{S}_{22}\right)\mathbf{b} = \mathbf{0}$$

ou de :

$$\left(\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21} - \gamma^2\mathbf{S}_{11}\right)\mathbf{a} = \mathbf{0}$$

Les quantités γ_k^2 $(k = 1, ..., s)$, identiques dans les deux équations précédentes, donnent les carrés des coefficients de corrélation $r_k = \sqrt{\gamma_k^2}$ entre les deux variables $\mathbf{u}^k$ et $\mathbf{v}^k$; toutes les valeurs propres γ_k^2 $(k = s+1, ..., p)$ sont nulles (si nous supposons que $p > q$, donc que $s = q$). On peut obtenir les $\mathbf{a}^k$ à partir des $\mathbf{b}^k$ ou inversement les $\mathbf{b}^k$ à partir des $\mathbf{a}^k$:

$$\begin{aligned} \mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{b}^k &= \gamma_k\mathbf{a}^k \\ \mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{a}^k &= \gamma_k\mathbf{b}^k \end{aligned}$$

Ces deux équations ont une interprétation intéressante : $\gamma_k\mathbf{a}^k$ est la projection de $\mathbf{b}^k$ sur l'espace défini par les colonnes de $\mathbf{X}$**,** espace des variables du premier groupe, et $\gamma_k\mathbf{b}^k$ est celle de $\mathbf{a}^k$ sur l'espace défini par les colonnes de $\mathbf{Y}$**,** espace des variables du second groupe. L'obtension des deux matrices de poids des variables de départ $\mathbf{A}_{p\times s}$ et $\mathbf{B}_{q\times s}$ permet alors de calculer les variables canoniques $\mathbf{u}^k$ et les $\mathbf{v}^k$. Les s coefficients de corrélation r_k entre $\mathbf{u}^k$ et $\mathbf{v}^k$ s'appellent *coefficients de corrélation canonique.*

3.2. Autres présentations

Il en existe plusieurs; elles sont d'inégal intérêt mais elles peuvent fournir un éclairage complémentaire pour le calcul ou l'interprétation de l'*ACC*.

Une première consiste dans l'utilisation de la *décomposition singulière* d'une matrice. Partant de la décomposition de Choleski des deux matrices définies positives $\mathbf{S}_{11} = \mathbf{P}_{11}^T \mathbf{P}_{11}$ et $\mathbf{S}_{22} = \mathbf{P}_{22}^T \mathbf{P}_{22}$ et en définissant la nouvelle matrice de dimension $p \times q$: $\mathbf{C} = \left(\mathbf{P}_{11}^T\right)^{-1} \mathbf{S}_{12} \mathbf{P}_{22}^{-1}$, à partir de la décomposition singulière de cette dernière $\mathbf{C}_{p\times q} = \mathbf{L}_{p\times s} \mathbf{\Theta}_{s\times s} \mathbf{M}_{s\times q}$, on obtient directement les variables canoniques $\mathbf{U}_{n\times s} = \mathbf{X}\mathbf{P}_{11}^{-1}\mathbf{L}$ et $\mathbf{V}_{n\times s} = \mathbf{Y}\mathbf{P}_{22}^{-1}\mathbf{M}^T$ dont la matrice de covariances est :

$$\begin{bmatrix} \mathbf{I}_s & \mathbf{\Theta} \\ \mathbf{\Theta} & \mathbf{I}_s \end{bmatrix}$$

Les termes de la matrice diagonale Θ, égale à Γ, donnent directement les coefficients de corrélation canonique r_k.

On peut aussi dériver les variables canoniques d'autres façons par :

– *moindres carrés* [5] : si on cherche les matrices $\mathbf{A}$ et $\mathbf{B}$ qui minimisent la trace de $\left[(\mathbf{XA} - \mathbf{YB})^T (\mathbf{XA} - \mathbf{YB})\right]$, les solutions sont les matrices que nous avons trouvées dans l'approche classique de l'*ACC*.

– *un modèle linéaire général multidimensionnel* : le modèle de régression linéaire multidimensionnel s'écrit :

$$\mathbf{Y}_{n\times q} = \mathbf{X}_{n\times p} \mathbf{\Phi}_{p\times q} + \mathbf{E}_{n\times q}$$

dans lequel $\mathbf{Y}$ est une matrice d'observations de q variables, $\mathbf{X}$ une matrice connue, $\mathbf{\Phi}$ une matrice de paramètres inconnus à estimer et $\mathbf{E}$ une matrice de termes aléatoires. Si on a $\mathbf{V}_{n\times s} = \mathbf{YB}$ et $\mathbf{U}_{n\times s} = \mathbf{XA}$ on a $\mathbf{V} = \mathbf{UD} + \mathbf{E}$, avec $\mathbf{D} = \mathbf{\Gamma} = \mathbf{diag}(r_k)$, ce qui correspond à la régression multidimensionnelle de $\mathbf{V}$ sur $\mathbf{U}$. On en déduit que :

$$\mathbf{V}_{n\times s} = \mathbf{U}_{n\times s} \mathbf{D}_{s\times s} + \mathbf{E}_{n\times s} = \mathbf{X}_{n\times p} \mathbf{A}_{p\times s} \mathbf{D}_{s\times s} + \mathbf{E}_{n\times s}$$

Les s $(= q)$ colonnes de la matrice de paramètres $\mathbf{\Phi}_{p\times s} = \mathbf{AD}$ représentent l'estimation des coefficients de régression de chacune des q variables canoniques du second groupe sur les p variables du premier; c'est la meilleure estimation, au sens des moindres carrés. L'avantage de cette approche tient à ce qu'elle précise, de manière explicite, la structure stochastique du modèle. Toutefois, elle éclipse l'aspect symétrique de la présentation classique, ce peut être un avantage, comme dans l'étude que nous présentons. Plus précisément, avec les p valeurs $\mathbf{x}_0 = [x_{01}, ..., x_{0p}]^T$ et les paramètres $\mathbf{AD}$ nous connaissons les s valeurs des variables canoniques $\mathbf{u}_0$ du premier groupe, donc aussi $\mathbf{v}_0$ celles du second groupe et par conséquent les q valeurs $\mathbf{y}_0 = [y_{01}, ..., y_{0q}]^T$; on obtient donc $\mathbf{y}_0$ en utilisant les paramètres de la régression de $\mathbf{Y}$ sur les variables canoniques du premier groupe. Donc, si la prédiction des variables

du second ensemble par celles du premier est un élément important de l'étude, cette formulation est beaucoup plus intéressante; elle permet, connaissant des valeurs $\mathbf{x}_0$, de prédire l'ensemble des valeurs $\mathbf{y}_0$ du grain de BLE qui aurait pu pousser sur un SOL caractérisé par les valeurs $\mathbf{x}_0$.

Nous avons un outil de prédiction et de simulation de situations nouvelles. Il faut noter que c'est une estimation conjointe des q valeurs du BLE avec une mesure non seulement de la précision de chacune d'elles mais aussi de la corrélation entre ces valeurs; cette formulation donne donc une information plus importante que celles que donneraient de simples régressions multiples [13]. De surcroît, toutes les procédures d'analyse des résidus et de l'influence des observations peuvent être utilisées.

– *minimisation d'une distance euclidienne [12]* : sur des bases purement géométriques on peut aussi rechercher des vecteurs $\mathbf{a}_{p\times 1}$ et $\mathbf{b}_{q\times 1}$ tels que la norme euclidienne $\|\mathbf{Xa} - \mathbf{Yb}\|^2$ soit minimale; l'avantage est ici que les conditions sur les rangs de $\mathbf{X}$ et de $\mathbf{Y}$ ne sont pas nécessaires. En outre, elle met en évidence l'équivalence entre la minimisation de la distance entre deux vecteurs et la minimisation d'une fonction de l'angle entre ces vecteurs.

3.3. Premier bilan

Tout d'abord, on peut voir que les coefficients de corrélation canonique contenus dans la matrice diagonale $\mathbf{diag}(r_k)$ sont invariants par changement de l'échelle des variables observées. C'est la raison pour laquelle on travaille très souvent (et ce que nous ferons ici) non pas sur la matrice des covariances $\mathbf{S}$ partitionnée en deux groupes, mais sur la matrice des coefficients de corrélation :

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{bmatrix}, \text{telle que } \mathbf{R}_{12} = \mathbf{R}_{21}^T.$$

Les éléments des vecteurs de poids $\mathbf{a}^k$ et $\mathbf{b}^k$ sont, eux, exprimés dans l'unité des variables qui leur correspondent. La forme de la matrice de covariance de $\begin{bmatrix}\mathbf{u}^T & \mathbf{v}^T\end{bmatrix}^T$ fournit l'essentiel des résultats utiles :

– Si nous supposons que $rang(\mathbf{X}) = p$ et $rang(\mathbf{Y}) = q$ (c'est-à-dire si nous admettons qu'il n'existe aucune combinaison linéaire stricte à l'intérieur de chaque groupe de variables), il existe $s = \min(p, q)$ couples de variables $\left(\mathbf{u}^k, \mathbf{v}^k\right)$ dont le coefficient de corrélation vaut r_k.

– À l'intérieur de chaque groupe, les couples $\left(\mathbf{u}^l, \mathbf{u}^k\right)$, $\left(\mathbf{v}^k, \mathbf{v}^l\right)$, pour $l, k = 1, ..., s$ et $l \neq k$. ne sont pas corrélés.

– Entre groupes de variables les couples $\left(\mathbf{u}^l, \mathbf{v}^k\right)$, pour $l, k = 1, ..., s$ et $l \neq k$ ne sont pas corrélés.

Le problème essentiel, pour l'interprétation des résultats, est donc de donner un sens aux éléments de $\mathbf{a}^k$ et de $\mathbf{b}^k$. Mais, il faut auparavant se demander si les s couples de variables canoniques sont suffisamment corrélés pour que leur examen ait un intérêt; c'est ce que l'on appelle le *test de la dimension de l'espace de représentation canonique*.

3.4. Dimension de l'espace de représentation canonique

Dans cette étude, il faut auparavant supposer qu'il existe une forme de distribution conjointe des vecteurs $\mathbf{x}_i$ et $\mathbf{y}_i$; pourvu que les distributions soient «raisonnablement» Normales et les observations indépendantes, des tests paramétriques sont disponibles. La matrice de covariances inconnues $\mathbf{\Sigma}$ de la population «théorique» $\mathcal{P}$ d'où sont issus les échantillons est partitionnée comme l'est $\mathbf{S}$. Un test de l'indépendance des deux groupes, est celui de l'hypothèse $H_0 : \mathbf{\Sigma}_{12} = \mathbf{0}$ contre $H_1 : \mathbf{\Sigma}_{12} \neq \mathbf{0}$; le lecteur intéressé pourra consulter de nombreux ouvrages sur le sujet, par exemple [2] (pp.243) ou [18] (pp.253).

Si les coefficients de corrélation canonique de $\mathcal{P}$ sont ρ_k $(k = 1, ..., s)$, le test de l'hypothèse classique pour savoir quelle est la dimension de représentation :

$$H_0^k : \rho_1 \neq 0, ..., \rho_k \neq 0, \rho_{k+1} = ... = \rho_s = 0$$

utilise la statistique :

$$-\left\{(n-1) - k - (p+q+1)/2 + \prod_{i=1}^{i=k} r_i^{-2}\right\} Ln(\prod_{i=k+1}^{i=s}(1-r_i^2) \underset{(H_0^k)}{\sim} \chi^2_{(p-k)(q-k)}.$$

La procédure consiste donc à tester séquentiellement les hypothèses H_0^1,..., H_0^k,...; on décide de la dimension dès qu'une hypothèse H_0^k ne peut pas être rejetée (*cf.* [14]). À notre avis, ces outils sont toutefois à *employer avec précaution*, tout particulièrement les valeurs du niveau des tests. Néanmoins ils constituent, pour le moins, des indices intéressants, c'est le cas du test de la dimension de représentation.

Si la supposition de Normalité est difficilement acceptable, la façon la plus naturelle de s'assurer de la qualité des résultats consiste à employer des méthodes de ré-échantillonnage. Le *jackknife* permet, en particulier, de vérifier leur stabilité et de détecter de possibles observations influentes. Le *bootstrap* permet de déterminer des distributions empiriques des différentes statistiques; en particulier il peut fournir les variances des éléments des vecteurs de poids $\mathbf{a}^k$ et $\mathbf{b}^k$.

3.5. Premiers outils d'interprétation : liens entre variables canoniques et variables observées

Si les coefficients r_k permettent d'établir le nombre (la dimension de représentation) et l'intensité de la relation entre les deux groupes de variables, les coefficients des vecteurs de poids $\mathbf{a}^k$ et $\mathbf{b}^k$ sont essentiels pour interpréter les résultats. La première difficulté que nous rencontrons est analogue à celle de l'interprétation des coefficients d'une équation de régression : si certaines des variables qui constituent la combinaison linéaire, que ce soit $\mathbf{a}^k$ ou $\mathbf{b}^k$, sont fortement corrélées dans leur groupe le sens de leur effet sera délicat à dégager. Ces vecteurs permettent, dans un premier temps de

calculer les coordonnées des n observations dans l'espace des variables canoniques :

$$\mathbf{U}_{n\times k} = \mathbf{X}_{n\times p}\left[\,\mathbf{a}^1 \quad \ldots \quad \mathbf{a}^k\,\right]_{p\times k}, \mathbf{V}_{n\times k} = \mathbf{Y}_{n\times q}\left[\,\mathbf{b}^1 \quad \ldots \quad \mathbf{b}^k\,\right]_{q\times k}$$

Les k graphes des k dimensions de la représentation des relations entre ensembles $\left\{\mathbf{u}^1 = \mathbf{X}\mathbf{a}^1, \mathbf{v}^1 = \mathbf{Y}\mathbf{b}^1\right\}$, ..., $\left\{\mathbf{u}^k = \mathbf{X}\mathbf{a}^k, \mathbf{v}^k = \mathbf{Y}\mathbf{b}^k\right\}$ donnent une bonne image globale de l'intérêt de l'*ACC*. Ils peuvent donner, surtout si le r_k correspondant est élevé, une vision bien meilleure que celle que pourrait donner l'une quelconque des variables observées. L'inconvénient est que si l'image existe, elle ne fournit aucune clé simple d'interprétation. Il faut donc aller plus loin en calculant les coefficients de corrélation entre les variables canoniques et les variables observées pour obtenir des coefficients de corrélation «intra-groupe» et «inter-groupe» dont l'intérêt, comme nous allons le voir, est différent :

– *coefficients de corrélation «intra-groupe»* : ce sont les coefficients de corrélation entre les variables observées ou canoniques du *même groupe*; pour chacun des groupes on a :

  $x_{jk}^{(u)} = cor(\mathbf{x}^j, \mathbf{u}^k) = (\mathbf{R}_{11}\mathbf{a}^k)_j$, ou pour l'ensemble de p variables observées et des k variables canoniques du même groupe

$$\mathbf{X}^{(u)} = cor(\mathbf{X}, \mathbf{U}) = \mathbf{R}_{11}\mathbf{A},$$

  $y_{jk}^{(v)} = cor(\mathbf{y}^j, \mathbf{v}^k) = (\mathbf{R}_{22}\mathbf{b}^k)_j$, ou pour l'ensemble de q variables observées et des k variables canoniques du même groupe

$$\mathbf{Y}^{(v)} = cor(\mathbf{Y}, \mathbf{V}) = \mathbf{R}_{22}\mathbf{B}.$$

  Remarque : le carré d'un des coefficients intra-groupe d'une quelconque variable est la proportion de cette variable qui est expliquée par une variable canonique de son groupe. D'un point de vue pratique, l'interprétation des variables canoniques du premier groupe peut se faire sur $\mathbf{a}^k$ ou sur $\mathbf{X}^{(u)}$, du second groupe sur $\mathbf{b}^k$ ou sur $\mathbf{Y}^{(v)}$. Si les variables sont peu corrélées à l'intérieur de leur groupe, les résultats seront voisins. Par contre, si certains coefficients de corrélation intra-groupe sont élevés, les résultats peuvent être très différents, mieux vaudra utiliser $\mathbf{X}^{(u)}$ et $\mathbf{Y}^{(v)}$. Dans ce cas, il sera bon de voir si la suppression de certaines variables n'est pas une meilleure solution.

– *coefficients de corrélation «inter-groupe»* : ce sont les coefficients de corrélation entre les variables observées ou canoniques de *l'autre groupe*;

  $x_{jk}^{(v)} = cor(\mathbf{x}^j, \mathbf{v}^k) = (\mathbf{R}_{12}\mathbf{b}^k)_j = r_k(\mathbf{R}_{11}\mathbf{a}^k)_j = r_k x_{jk}^{(u)}$, ou pour l'ensemble de p variables observées et des k variables canoniques de l'autre groupe

$$\mathbf{X}^{(v)} = cor(\mathbf{X}, \mathbf{V}) = \mathbf{R}_{11}\mathbf{A}\,\mathbf{diag}(\mathbf{r_k}),$$

$y_{jk}^{(u)} = cor(\mathbf{y}^j, \mathbf{u}^k) = (\mathbf{R}_{21}\mathbf{a}^k)_j = r_k(\mathbf{R}_{22}\mathbf{b}^k)_j = r_k y_{jk}^{(v)}$, ou pour l'ensemble de q variables observées et des k variables canoniques de l'autre groupe

$$\mathbf{Y}^{(u)} = cor(\mathbf{Y}, \mathbf{U}) = \mathbf{R}_{22}\mathbf{B}\,\mathbf{diag}(\mathbf{r_k}).$$

Remarque : le carré d'un des coefficients inter-groupe d'une quelconque variable est la proportion de cette variable qui est expliquée par une variable canonique de l'autre groupe. Il ressemble donc beaucoup à un coefficient de détermination d'une régression linéaire multiple, mais ce n'en pas un ! En effet, dans une régression multiple on recherche le maximum de la corrélation entre *une seule* variable de la matrice $\mathbf{Y}$, alors que dans l'*ACC* c'est avec une combinaison linéaire des colonnes de $\mathbf{Y}$.

4. Premiers résultats

4.1. Liaisons intra-groupes

L'*ACC* va donc porter sur le couple de matrices $\mathbf{X}_{162\times 24}$ pour le SOL et $\mathbf{Y}_{162\times 9}$ pour le BLE. Nous ne donnons que les matrices $\mathbf{R}_{12}$ et $\mathbf{R}_{22}$. Nous ne donnons pas la matrice $\mathbf{R}_{11}$ des coefficients de corrélation des variables SOL, sauf les coefficients de corrélation avec le pH (Tableau 2). Cette variable, on le sait, joue un rôle important : les sols acides (pH faible) ont généralement des teneurs plus élevées en *ETM* dans le blé. Un examen détaillé de $\mathbf{R}_{11}$ montre que peu de variables sont très corrélées; si c'était le cas, l'interprétation serait plus délicate et il serait opportun de supprimer des variables SOL et de n'en garder qu'une pouvant remplacer celles qui sont le plus corrélées avec elle. Parmi les coefficients de corrélation dont la valeur absolue est de 0.9, on ne trouve que les couples $\{NiS, CrS\}$, $\{NiS, FeS\}$, $\{CdD, CdS\}$; ceux dont la valeur absolue est supérieure à 0.8 sont : $\{SG, LG\}$, $\{CEC, A\}$, $\{CdS, A\}$, $\{FeS, CrS\}$, $\{NiS, CEC\}$, $\{ZnS, CuS\}$, $\{PbD, PbS\}$, $\{CdN, pH\}$, $\{ZnN, pH\}$ et $\{ZnN, CdN\}$. Bien que ces valeurs ne nous aient pas paru suffisamment élevées pour éliminer l'une d'entre elles, nous devrons nous souvenir qu'elles risquent de jouer un rôle très voisin. Ainsi des modèles de régression permettant d'expliquer CdB pourraient introduire indifféremment CdN ou le pH; la présence des deux serait vraisemblablement inutile et ne pourrait qu'introduire une instabilité néfaste pour la précision des effets prédits. On peut penser que la suppression de NiS, CrS, CdS et CEC ne pourrait avoir que des effets bénéfiques et ne modifierait pas nos résultats.

Les coefficients de corrélation entre les variables BLE ne révèlent pas de fortes liaisons; ils sont pratiquement tous positifs sauf celui, faible, du couple $\{PbB, CrB\}$ (*cf.* Tableau 1)

4.2. Liaisons inter-groupes

La plupart des variables BLE sont négativement liées au pH (*cf.* Tableau 2), ce qui confirme que la teneur en *ETM* dans les grains de blé croît lorsque les sols sont acides (transferts accrus du sol vers la plante). Ces teneurs sont aussi liées aux

TABLEAU 1
Matrice des coefficients de corrélation BLE R_{22}

$Blé$	CdB	CrB	CuB	FeB	MgB	MnB	NiB	PbB	ZnB
CdB	1.0								
CrB		1.0							
CuB		0.3	1.0						
FeB				1.0					
MgB		0.4	0.5	0.4	1.0				
MnB			0.3	0.4	0.5	1.0			
NiB		0.4	0.4	0.3	0.5	0.6	1.0		
PbB	0.4	-0.3						1.0	
ZnB			0.4	0.3	0.7	0.5	0.4		1.0

(Les valeurs inférieures à 0.3, en valeur absolue, ne sont pas reportées)

propriétés agro-pédologiques classiques : négativement avec l'argile (A), le limon (LF, LG) et le calcaire ($CaCO_3$); positivement avec le sable (SF, SG). On peut aussi noter que les relations des *ETM* BLE sont relativement peu marquées avec les éléments correspondant du SOL (CdS, ZnS), mais beaucoup plus avec les concentrations obtenues par les deux réactifs, DTPA et NH_4NO_3, (CdN, ZnD, ZnN). Il faut aussi relever que la valeur la plus élevée, en valeur absolue, des coefficients de corrélation est $r(pH, MnB) \simeq -0.7$ (dans le tableau 2), les suivantes égales à 0.6 sont $r(SG, ZnB), r(SG, MgB), r(CdN, MnB)$ et $r(ZnN, MnB)$; les autres valeurs se situent entre -0.5 et 0.5.

Remarque : le premier coefficient de corrélation r_1 est, naturellement, supérieur à la valeur maximale. Rien ne s'oppose à ce que plusieurs le soient, nous le vérifierons dans l'exemple.

4.3. Coefficients de corrélation canonique : dimension de la représentation

Les résultats (Tableau 3) montrent qu'en fixant le niveau de confiance du test à $5\,\%$, une valeur possible de la dimension de l'espace de représentation est 6, la probabilité critique (ou p-valeur) $p_{H_0^k}$ du test de H_0^k que la dimension soit égale à k (voir § 3.4) valant $0.09\,\% < 5\,\%$ pour $k = 6$ et $7.5\,\% > 5\,\%$ pour $k = 7$. De surcroît, 4 couples ont un coefficient de corrélation canonique supérieur à 0.74, valeurs toutes supérieures à la valeur maximale (en valeur absolue) 0.7 des coefficients de corrélation inter-groupes. Il existe donc plusieurs combinaisons linéaires fortement liées entre SOL et BLE traduisant des relations entre les deux domaines; ces relations sont non corrélées entre elles : le problème fondamental va maintenant être de les interpréter.

TABLEAU 2
Matrice des coefficients de corrélation SOLxBLE (R_{12}) et $pH \times$ SOL

Blé	CdB	CrB	CuB	FeB	MgB	MnB	NiB	PbB	ZnB		pH
A				-0.3	-0.3	-0.3			-0.3		0.6
LF			-0.3	-0.4	-0.5	-0.4	-0.4		-0.5		0.4
LG			-0.4		-0.5		-0.3		-0.6		
SF		0.3	0.4		0.5	0.3	0.3	-0.3	0.5		-0.5
SG			0.4		0.6	0.3	0.4		0.6		-0.4
CEC									-0.4		0.4
CO_3Ca				-0.3		-0.5	-0.3				0.7
CS			0.3								0.4
pH				-0.3	-0.4	-0.7	-0.5		-0.5		1.0
CdS				-0.3							0.5
CrS											
CuS			0.4		0.3	0.3	0.3				
FeS			0.3								
MnS		0.3	0.3		0.4	0.4	0.4		0.3		
NiS											
PbS			0.3		0.3	0.4	0.3		0.4		-0.3
ZnS			0.3								
CdD	0.3				-0.3						0.4
CuD			0.3				0.3				
PbD			0.3						0.3		
ZnD									0.5		
CdN	0.5					0.6	0.3		0.3		-0.8
CuN				-0.3	-0.3	-0.5	-0.3				0.7
ZnN	0.3					0.6	0.3		0.5		-0.8

(Les valeurs inférieures à 0.3, en valeur absolue, ne sont pas reportées)

TABLEAU 3
Ensemble des $s = \min(24, 9)$ *coefficients de corrélation canonique* r_k

k	r_k	χ^2	dl	$p_{H_0^k}$
1	0.8903	848.83	216	0.0000
2	0.8334	623.39	184	0.0000
3	0.7900	453.65	154	0.0000
4	0.7409	313.55	126	0.0000
5	0.6481	198.79	100	0.0000
6	0.5455	120.32	76	0.0009
7	0.4616	69.30	54	0.0785
8	0.3527	34.21	34	0.4577
9	0.2932	14.44	16	0.5660

Ce sont les graphiques de la figure 3 qui, sans expliquer le pourquoi des relations, les traduisent pour les 162 sites. Ces graphiques permettent, dans un premier temps, de s'assurer qu'il n'existe pas d' *observations suspectes* (des «outliers»). Si, de plus, la forme des nuages est relativement elliptique, c'est une bonne raison pour penser que la Normalité de la distribution est acceptable, même si ces graphiques ne fournissent pas des preuves au sens mathématique du terme.

Nous verrons au paragraphe suivant que les deux premières variables canoniques SOL sont liées toutes les deux négativement au pH; on doit donc retrouver la majorité des sols acides (pH faible) dans la partie haute des deux graphiques. C'est bien le cas pour les sols dénommés sur les graphiques X (pH moyen 6) et T (pH moyen 6.5) opposés aux sols H, J, K M et F qui ont des pH supérieurs à 8. On sait aussi que les sols les plus sableux sont ceux qui étaient initialement les plus acides et qui ont la plus grande probabilité de l'être restés.

4.4. L'espace canonique des variables du premier groupe (SOL)

Un regard sur les corrélations internes du SOL révèle, si nous nous limitons aux valeurs supérieures à (en valeur absolue) ou très voisines de 0.5 (Tableau 4 et figures 4, 5 et 6, à gauche) pour :

– la première variable canonique $\mathbf{u}^1$: une opposition $\{SG, SF, ZnN, ZnD, MnS, PbS\}$ / $\{A, LG, LF, CEC, pH\}$, où les variables à l'intérieur de la première accolade ont des coefficients de corrélation positifs élevés et celles à l'intérieur de la seconde négatifs élevés. On retrouve ici le fait bien connu que les matériaux sableux sont acides et peu limoneux et inversement les matériaux limoneux sont peu sableux; on sait aussi que plus il y a d'argile, plus il y a de carbone représentant les matières organiques, et la valeur de CEC est en relation directe positive avec la teneur en argile et avec la teneur en carbone;

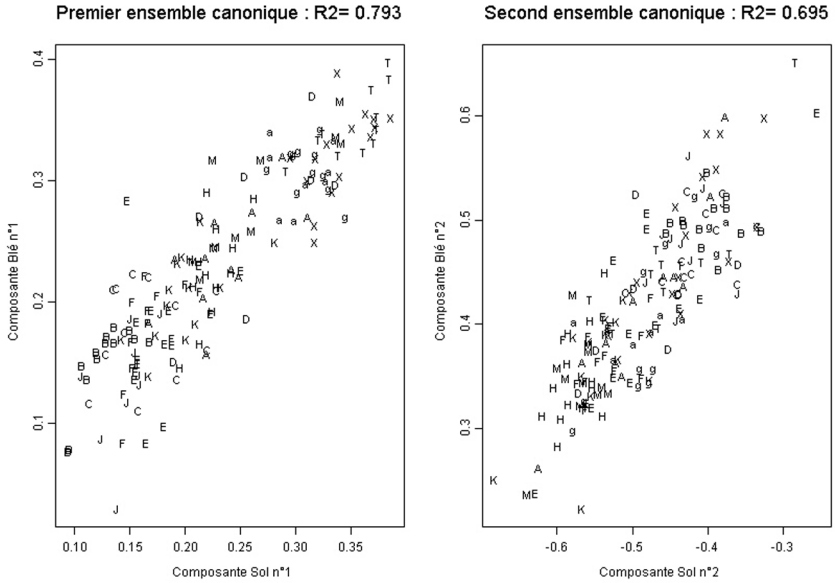

FIGURE 3
Représentations des observations sur $\{\mathbf{u}^1, \mathbf{v}^1\}$ *et* $\{\mathbf{u}^2, \mathbf{v}^2\}$.

- la seconde $\mathbf{u}^2$: une autre opposition $\{CdN, ZnN\}$ / $\{pH, CaCO_3, CuN\}$, autre fait connu puisque pH et calcaire sont étroitement liés : quand un sol est calcaire, son pH est toujours élevé, compris entre 7.9 pour les sols à peine calcaires jusqu'à 8.5 pour les sols «hypercalcaires». À replacer dans la gamme générale de pH qui sont compris entre 4.0 et 8.5. Il en est de même pour CuN dont on sait par d'autres études qu'il est proportionnel au pH ;
- la troisième $\mathbf{u}^3$: un ensemble de variables regroupant la majorité des teneurs totales en métaux ressort un peu $\{CrS, CuS, FeS, MnS, NiS\}$ opposées à $\{ZnD, ZnN\}$, ce regroupement a sans doute une signification, mais laquelle ?
- la quatrième $\mathbf{u}^4$: fait ressortir le couple $\{CS, CdD\}$,
- au-delà : les valeurs sont généralement peu élevées, d'où une interprétation difficile, sûrement liée à des observations suspectes ; en particulier CuD pour la dernière, $\mathbf{u}^9$, non significative.

Les corrélations inter-groupes entre les variables du sol $\mathbf{x}^j$ et les variables canoniques $\mathbf{v}^k$ sont reportées dans le tableau 5.

TABLEAU 4

Variables du premier groupe SOL. Corrélations $x_{jk}^{(u)}$ entre les variables du sol $\mathbf{x}^j$ et les variables canoniques $\mathbf{u}^k$: matrice $\mathbf{X}^{(u)}$ (cf.§ 3.5) et variances internes U_k^2 définies plus loin au § 5.1.

Blé	u^1	u^2	u^3	u^4	u^5	u^6	u^7	u^8	u^9	$\sum_{k=1}^{k=s} x_{jk}^{(u)2}$
A	−0.4			−0.3	0.3	−0.3				0.502
LF	−0.5				0.4	−0.4				0.688
LG	−0.6			0.3						0.561
SF	0.7					0.4				0.642
SG	0.7									0.693
CEC	−0.4						−0.3			0.377
CO_3Ca		−0.5		−0.4	0.3					0.556
CS				−0.6		−0.4				0.610
pH	−0.6	−0.6								0.784
CdS				−0.4	0.3	−0.4	−0.3			0.513
CrS			0.3							0.323
CuS	0.3		0.4							0.624
FeS			0.4		0.3		−0.3			0.439
MnS	0.5		0.4		0.3	−0.3				0.677
NiS			0.4							0.282
PbS	0.5	0.3				−0.3	−0.4			0.659
ZnS	0.3				0.3	−0.3	−0.3			0.451
CdD	−0.3			−0.5			−0.3			0.576
CuD	0.3								−0.7	0.597
PbD	0.3						−0.3			0.340
ZnD	0.5	−0.4	−0.4							0.677
CdN	0.3	0.7	−0.3							0.783
CuN	−0.3	−0.4								0.523
ZnN	0.5	0.5	−0.4							0.754
U_k^2	0.164	0.086	0.060	0.063	0.043	0.053	0.043	0.019	0.038	0.568

(Les valeurs inférieures à 0.3, en valeur absolue, ne sont pas reportées)

TABLEAU 5

Variables du premier groupe SOL. Corrélations $x_{jk}^{(v)}$ entre les variables du sol $\mathbf{x}^j$ et les variables canoniques $\mathbf{v}^k$: matrice $\mathbf{X}^{(v)}$ et communautés $R^2_{\mathbf{x}^j,9} = \sum (x_{jk}^{(v)})^2$

Blé	v^1	v^2	v^3	v^4	v^5	v^6	v^7	v^8	v^9	$R^2_{X_j,9}$
A	-0.3									0.248
LF	-0.4									0.375
LG	-0.6									0.400
SF	0.6									0.427
SG	0.6									0.458
CEC	-0.4									0.208
CO_3Ca		-0.4								0.308
CS				-0.4						0.274
pH	-0.5	-0.5								0.538
CdS										0.193
CrS			0.3							0.164
CuS	0.3		0.3							0.275
FeS			0.3							0.214
MnS	0.4		0.3							0.397
NiS			0.3							0.156
PbS	0.4									0.319
ZnS	0.3									0.201
CdD				-0.4						0.276
CuD	0.3									0.140
PbD	0.3									0.161
ZnD	0.4	-0.3	-0.3							0.405
CdN		0.6								0.512
CuN	-0.3		-0.3							0.286
ZnN	0.5	0.4	-0.3							0.507
$V^2_{X\|v_k}$	0.130	0.060	0.038	0.035	0.018	0.016	0.009	0.002	0.003	0.310

(Les valeurs inférieures à 0.3, en valeur absolue, ne sont pas reportées)

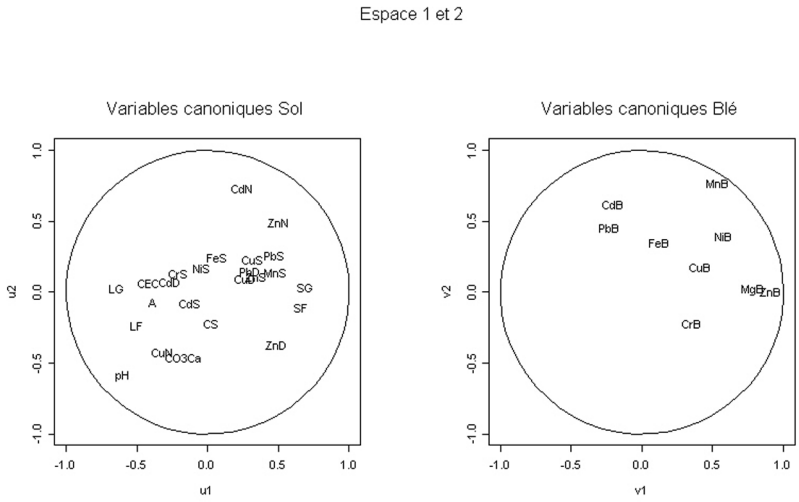

FIGURE 4
Cercles des corrélations intra-groupe (axes 1 et 2)

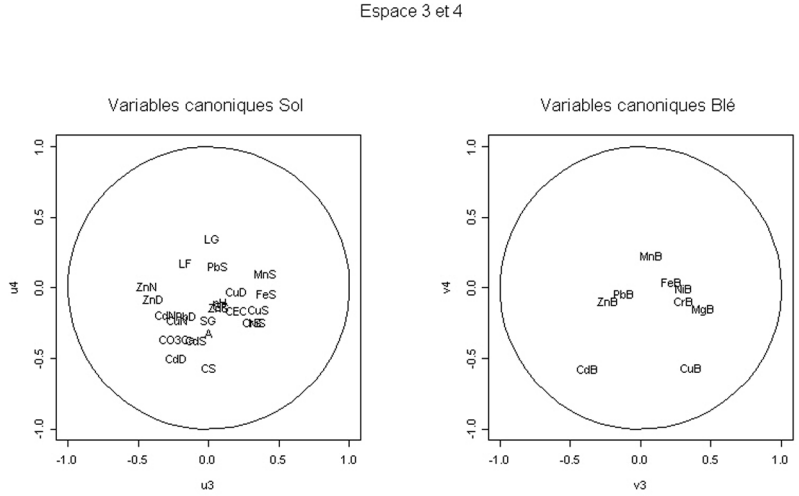

FIGURE 5
Cercles des corrélations intra-groupe (axes 3 et 4)

4.5. L'espace canonique des variables du second groupe (BLE)

Un regard sur les corrélations internes du BLE révèle, (Tableau 6 et figures 4, 5, 6, à droite) pour :

– la première variable canonique $\mathbf{v}^1$: des coefficients de corrélation intra-groupe tous positifs, deux très élevés $\{ZnB, MgB\}$ mais aussi $\{MnB, NiB$, voire $CuB\}$,

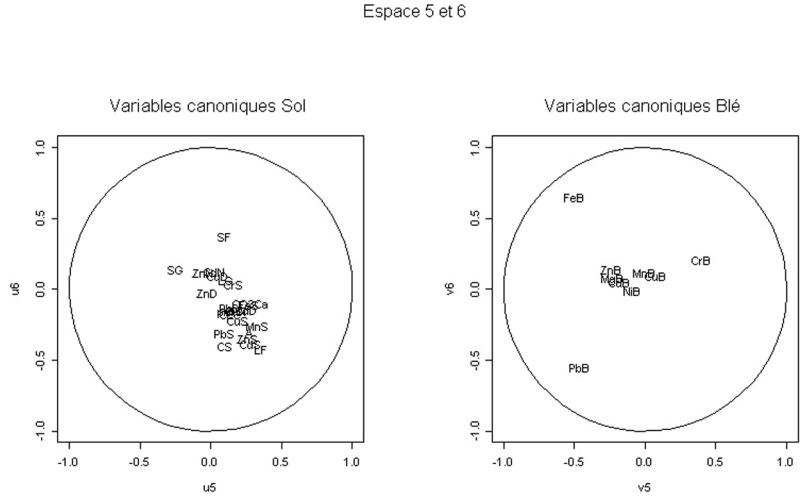

FIGURE 6
Cercles des corrélations intra-groupe (axes 5 et 6)

– la seconde $\mathbf{v}^2$: deux sont dominantes $\{MnB, CdB\}$,
– la troisième $\mathbf{v}^3$: MgB émerge associée à $\{CuB, CrB, NiB\}$ et opposées à CdB,
– la quatrième $\mathbf{v}^4$: essentiellement $\{CuB, CdB\}$.

Au-delà, $\mathbf{v}^5$ et $\mathbf{v}^6$, FeB et PbB apparaissent toutes les deux associées et opposées à CrB dans la première, simplement opposées dans la seconde. Ce fait traduit que les résultats peuvent mettre en évidence une synergie dans l'une et une opposition dans l'autre.

C'est le tableau 7, et les deux graphiques de la figure 7, qui doivent nous donner une partie de l'explication des relations SOL*BLE :

– le couple d'*ETM* du blé $\{ZnB, MgB\}$, et à un moindre degré NiB, est fortement lié à $\mathbf{u}^1$, c'est-à-dire à des variables comme $\{SG, SF, ZnN, ZnD, MnS, PbS\}$ / $\{LG, LF, pH\}$;
– le couple $\{CdB, MnB\}$ est lié à $\mathbf{u}^2$, c'est-à-dire à des variables comme $\{CdN, ZnN\}$ / $\{pH, CaCO_3, CuN\}$;
– la variable MgB est liée à $\mathbf{u}^3$, c'est-à-dire à des variables comme $\{CrS, CuS, FeS, MnS\}$ / $\{ZnD, ZnN\}$;
– quant au couple $\{CdB, CuB\}$ il est lié à $\mathbf{u}^4$, c'est-à-dire à des variables comme $\{CS, CdD\}$.

Ce que nous venons de dire soulève deux remarques :

– «lié à un groupe de variables» ne signifie pas que toutes ces variables vont apparaître dans un modèle de régression linéaire multiple d'une des variables du BLE, mais que les variables de ce groupe sont des prédicteurs plausibles. Leur

TABLEAU 6

Variables du second groupe BLE. Corrélations $y_{jk}^{(v)}$ entre les variables du blé $\mathbf{y}^j$ et les variables canoniques $\mathbf{v}^k$: matrice $Y^{(v)}$ (cf.§ 3.5) et variances internes V_k^2 définies plus loin au § 5.1

Blé	v^1	v^2	v^3	v^4	v^5	v^6	v^7	v^8	v^9	$\sum_{k=1}^{k=s} y_{jk}^{(v)2}$
CdB		0.6	−0.4	−0.6						1
CrB	0.4		0.3		0.4			−0.5	0.5	1
CuB	0.4		0.4	−0.6			−0.5			1
FeB		0.4			−0.5	0.7			0.3	1
MgB	0.8		0.5							1
MnB	0.5	0.8								1
NiB	0.6	0.4	0.3				0.4	−0.5		1
PbB		0.5			−0.5	−0.6		−0.3		1
ZnB	0.9									1
V_k^2	0.279	0.173	0.088	0.082	0.085	0.092	0.060	0.079	0.061	1

(Les valeurs inférieures à 0.3, en valeur absolue, ne sont pas reportées)

TABLEAU 7

Variables du second groupe BLE. Corrélations $y_{jk}^{(u)}$ entre les variables du blé $\mathbf{y}^j$ et les variables canoniques $\mathbf{u}^k$: matrice $Y^{(u)}$ et communautés $R^2_{\mathbf{y}^j,9} = \sum (y_{jk}^{(u)})^2$

Blé	u^1	u^2	u^3	u^4	u^5	u^6	u^7	u^8	u^9	$R^2_{Y_j,9}$
CdB		0.5	−0.3	−0.4						0.588
CrB	0.3		0.3		0.3					0.338
CuB	0.4		0.3	−0.4						0.484
FeB		0.3			−0.3	0.4				0.377
MgB	0.7		0.4							0.672
MnB	0.5	0.6								0.695
NiB	0.5	0.3								0.504
PbB		0.4			−0.3	−0.3				0.404
ZnB	0.8									0.732
$U^2_{Y\|u_k}$	0.221	0.120	0.055	0.045	0.036	0.027	0.013	0.010	0.002	0.533

(Les valeurs inférieures à 0.3, en valeur absolue, ne sont pas reportées)

liaison avec d'autres variables du même groupe peut les éliminer d'un tel modèle; ainsi le pH fortement corrélé avec CdN et ZnN, peut jouer un rôle en tant que prédicteur, mais il peut très bien être remplacé par une autre variable qui contient sensiblement la même information.

– certains *ETM* sont liés presqu'uniquement à une variable canonique SOL (c'est le cas de ZnB et à un moindre degré MnB); d'autres (comme CdB, MgB ou NiB) sont liés à deux ou trois; la question intéressante est de voir que ces derniers sont influencés par des variables non corrélées dont les effets sont additifs.

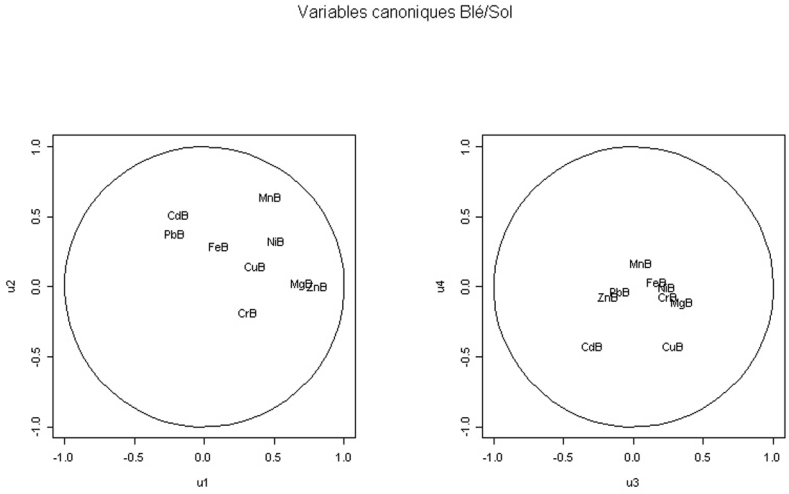

FIGURE 7
Corrélations $y_{jk}^{(u)}$ *des variables BLE* y^j
sur les variables canoniques SOL u^k $(1 \leqslant k \leqslant 4)$

5. Approfondissement de l'analyse

Dans ce qui suit, nous allons considérer quelques informations supplémentaires obtenues à partir des quatre matrices de coefficients de corrélation entre les deux groupes de variables observées et les deux groupes de variables canoniques : $\mathbf{X}^{(u)}$, $\mathbf{Y}^{(v)}$, $\mathbf{X}^{(v)}$ et $\mathbf{Y}^{(u)}$. Nous raisonnerons sur les variables observées standardisées, c'est-à-dire sur les matrices de coefficients de corrélation.

5.1. Variances extraites par une variable canonique

Tout d'abord pour les corrélations intra-groupes.

DÉFINITION 1. – *On appelle variance extraite par la $k^{ème}$ variable canonique la quantité :*

$$U_k^2 = \sum_{j=1}^{j=p} x_{jk}^{(u)2}/p \text{ pour le premier groupe,}$$

$$V_k^2 = \sum_{j=1}^{j=q} y_{jk}^{(v)2}/q \text{ pour le second.}$$

Les variables étant centrées et réduites, ces quantités peuvent être considérées comme des % de variance extraits par la $k^{ème}$ variable canonique.

Remarques :

– Dans le cas général $U_k^2 \neq V_k^2$.

– Si $s = q$, toute l'information sur les variables observées est conservée dans l'ensemble des variables canoniques de ce groupe et la quantité $\sum_{k=1}^{k=s} V_k^2 = 1$ (Tableau 6). C'est un résultat analogue que l'on a lorsque l'on conserve toutes les composantes principales de ce groupe; mais les variables canoniques sont, généralement, différentes des composantes principales. Les deux ensembles (variables canoniques et composantes principales) sont deux ensembles différents de variables orthogonales, non corrélées. Pour l'autre groupe, si $p > q$, généralement $\sum_{k=1}^{k=s} U_k^2 < 1$. Ceci signifie qu'une partie de l'information de ce groupe n'est d'aucune utilité pour mesurer la relation entre les groupes.

Si nous regardons les résultats des tableaux 4 et 6, nous observons :

– pour SOL, la variance moyenne extraite par ses 9 composantes canoniques ($\mathbf{u}^k, \quad k = 1, ..., s$) est $0.568 : 56.8\,\%$ de la variabilité du SOL sert donc à l'étude de la relation entre les deux groupes, $43.2\,\%$ est donc « inutile ». Mais si la première $\mathbf{u}^1$ en représente $16.4\,\%$, la seconde $\mathbf{u}^2$ n'en représente plus que $8.6\,\%$ et les suivantes encore moins.

– pour BLE, on extrait des composantes canoniques de ce groupe ($\mathbf{v}^k, \quad k = 1, ..., s$) toute la variabilité utile, puisqu'elle vaut 1. Le première v_1 en représente $27.9\,\%$, la seconde $\mathbf{v}^2$ $17.3\,\%$, les suivantes moins de $10\,\%$ chacune.

5.2. Redondance ou variance expliquée

DÉFINITION 1. – *La redondance est la proportion de la variance d'un groupe que l'on peut prédire par les variables de l'autre groupe, c'est la variance expliquée d'un groupe par l'autre. On a (compte tenu des résultats du § 3.5) les relations suivantes :*

$$V^2_{\mathbf{X}|\mathbf{v}^k} = r_k^2 U_k^2 = \sum_{j=1}^{j=p} x_{jk}^{(v)2}/p \text{ pour le premier groupe,}$$

$$U^2_{\mathbf{Y}|\mathbf{u}^k} = r_k^2 V_k^2 = \sum_{j=1}^{j=q} y_{jk}^{(v)2}/q \text{ pour le second.}$$

Dans la seconde formulation, la redondance apparaît comme la moyenne des variances de toutes les variables observées d'un groupe, expliquée par une variable canonique de l'autre : c'est un indice de la capacité d'une variable canonique d'un groupe à expliquer ou prédire les variables de l'autre groupe. Dans chacun des groupes, les redondances totales sont :

$$V^2_{\mathbf{X}|\mathbf{v}^1,\ldots,\mathbf{v}^s} = \sum_{k=1}^{k=s} V^2_{\mathbf{X}|\mathbf{v}^k} \text{ pour le premier et } U^2_{\mathbf{Y}|\mathbf{u}^1,\ldots,\mathbf{u}^s} = \sum_{k=1}^{k=s} U^2_{\mathbf{Y}|\mathbf{u}^k} \text{ pour le second.}$$

Elles fournissent des mesures globales de la variance d'un groupe qui peut être expliquée par l'autre. La redondance totale est l'expression de l'inter-relation des groupes; elle n'est pas contradictoire avec l'expression des coefficients de corrélation canonique qui mesurent l'intensité de la relation linéaire entre des composantes de chacun des groupes. En général, c'est une relation asymétrique car $V^2_{\mathbf{X}|\mathbf{v}^1,\ldots,\mathbf{v}^s} \neq U^2_{\mathbf{Y}|\mathbf{u}^1,\ldots,\mathbf{u}^s}$.

Si nous nous intéressons aux redondances des variables BLE, les deux premières redondances valent respectivement 0.221 et 0.120, alors que la redondance totale $U^2_{\mathbf{Y}|\mathbf{u}^1,\ldots,\mathbf{u}^9} = 0.533$, ce qui signifie que plus de la moitié de la variabilité de l'ensemble du BLE est prise en compte par les variables SOL et un peu plus de la moitié par les six premières variables canoniques $U^2_{\mathbf{Y}|\mathbf{u}^1,\ldots,\mathbf{u}^6} = 0.504$.

Si on associe ces indices au test de dimension précédent, on peut ne conserver que $t(< s)$ variables canoniques; il est alors possible d'étudier la distribution d'échantillonnage de la redondance totale. C'est ce qu'a fait Miller [17], cité par [10].

L'analyse de la redondance peut donc être une autre façon d'analyser la relation entre deux groupes de variables. Les résultats sont généralement différents; toutefois quand il y a un seul couple dominant, les résultats pour le premier couple sont très voisins. L'ordre des couples peut être modifié quand les valeurs propres d'une matrice intra-groupe, par exemple $\mathbf{R}_{22}$, sont proches. On pourra trouver des résultats sur la comparaison des deux approches dans [6].

5.3. Communautés des variables

Si maintenant nous regardons du côté de la variance des variables observées, nous pouvons le faire soit par rapport aux variables du même groupe soit par rapport aux variables de l'autre. La première peut nous renseigner sur la structure interne de chaque groupe dans l'espace des axes canoniques de dimension s si on conserve tous les axes, ou t si on se limite à la dimension de la représentation. Mais c'est la seconde qui peut s'avérer la plus utile; si on calcule la somme des carrés des coefficients de corrélation de la $j^{ème}$ variable du premier groupe avec les t premiers axes canoniques du second :

$$R^2_{\mathbf{x}^j,t} = \sum_{k=1}^{k=t} x^{(v)2}_{jk}; \forall t = 1,\ldots,s$$

ou la même quantité pour la $j^{ème}$ variable du second groupe avec les t premiers axes canoniques du premier :

$$R^2_{\mathbf{y}^j,t} = \sum_{k=1}^{k=t} y_{jk}^{(u)2}; \forall t = 1, \ldots, s$$

ces quantités, appelées *communautés*, sont de simples *coefficients de détermination* de la régression linéaire multiple de la variable j du groupe considéré sur les t premiers axes canoniques. Selon que le problème étudié est symétrique ou non, l'intérêt peut se porter sur les deux ensembles de communautés ou sur un seul des deux. Ce sera le cas de notre étude où nous cherchons à expliquer BLE ($\mathbf{Y}$) par SOL ($\mathbf{X}$), l'inverse n'ayant pas grand sens. Si $t = s = q$ (le nombre de variables du second groupe est inférieur ou égal à celui du premier et on conserve tous les axes canoniques), $R^2_{\mathbf{y}^j,s}$ est le coefficient de détermination de la régression de la $j^{ème}$ variable du second groupe sur l'ensemble des axes canoniques, mais aussi sur l'ensemble des variables observées du premier groupe. $R^2_{\mathbf{y}^j,s}$ donne donc la valeur maximale du coefficient de détermination que l'on peut obtenir avec le modèle complet à p variables; sur les s variables canoniques nous avons le même résultat global avec un nombre plus limité de régresseurs non corrélés (si $q < p$ naturellement). C'est une autre forme de *régression orthogonale* analogue à une régression sur les composantes principales du premier groupe, mais dans ce cas, en général, il faudrait conserver les p composantes principales pour obtenir le même résultat : donc les variables canoniques *concentrent toute l'information utile* pour expliquer les variables du second groupe par celles du premier.

Certaines variables canoniques peuvent faiblement contribuer à la valeur de $R^2_{\mathbf{y}^j,s}$, elles peuvent être ignorées. Par contre celles dont la contribution est importante, généralement peu nombreuses, doivent concentrer toute notre attention. Le graphe des observations d'une quelconque variable du second groupe en fonction de ces seules variables canoniques peut donner un éclairage sur les facteurs qui ont une grande influence sur la variable.

5.4. Résultats complémentaires

Munis de ces quelques indices supplémentaires, nous pouvons compléter les analyses précédentes. Étant donné la dissymétrie de l'analyse, nous nous placerons surtout dans l'optique naturelle de la prédiction du BLE par le SOL. Le plus simple consiste à regarder de nouveau les valeurs du tableau 7 légèrement transformées dans le tableau 8 pour faire apparaître le pourcentage de chaque variable canonique SOL dans le calcul de $R^2_{\mathbf{y}^j,9}$. La dernière colonne du tableau 7 (ou du tableau 8) nous fournit le maximum du coefficient de détermination que nous pouvons obtenir avec les 24 variables SOL : pour ZnB, MnB, MgB, CdB et NiB il est supérieur à $50\,\%$; pour CrB, CuB, FeB et PbB il ne dépasse pas cette valeur.

Pour les cinq premières variables *ETM* bien explicables par les variables SOL, pour dépasser $75\,\%$ ou presque du R^2 il faut une seule composante pour ZnB et MgB, mais il en faut deux pour CdB, MnB et NiB. Ce résultat nous permet de faire des graphiques des variables BLE en fonction des variables canoniques SOL

qui les «expliquent» le mieux. Dans les six *ETM* BLE que nous avons sélectionnés (*cf.* figure 8) ceux qui concernent ZnB, MgB et MnB sont les plus clairs : la liaison des deux premiers avec $\mathbf{u}^1$ est nette, comme celle du troisième avec $\mathbf{u}^2$. Ceci signifie qu'un modèle de régression possible de ZnB ou de MgB doit sélectionner des variables dans l'ensemble de celles qui permettent d'interpréter $\mathbf{u}^1$ c'est-à-dire : $\{SG, SF, ZnN, ZnD, MnS, PbS\}$ / $\{LG, LF, pH\}$. Pour ZnB il sera inutile d'aller en chercher d'autres puisque $90\,\%$ de ce qui peut servir à la prédire est contenu dans $\mathbf{u}^1$. Par contre, pour MgB il n'y en a que $74.4\,\%$, une partie encore importante ($19.2\,\%$) se trouve dans $\mathbf{u}^3$ donc dans les variables $\{CrS, CuS, FeS, MnS, NiS\}$ / $\{ZnD, ZnN\}$. Pour MnB il faudra regarder les variables de $\mathbf{u}^1$ et $\mathbf{u}^2$. L'*ACC* nous fournit donc des limites possibles pour les *ETM* de BLE individuels et des choix de variables SOL potentiellement bons prédicteurs.

TABLEAU 8
$\%$ du $R^2_{\mathbf{y}^j,9}$ des variables BLE observées pour chaque axe canonique

Blé	u^1	u^2	u^3	u^4	u^5	u^6	u^7	u^8	u^9	$R^2_{Y_j,9}$
CdB	5.2	46.0	14.2	30.5	2.1	0.1	1.2	0.0	0.7	0.588
CrB	30.2	9.8	18.7	1.3	20.5	3.8	0.2	8.8	6.7	0.338
CuB	28.6	4.5	17.3	36.3	0.6	0.6	10.2	1.4	0.7	0.484
FeB	3.5	22.9	8.4	0.3	27.5	33.8	0.8	0.8	2.2	0.377
MgB	74.4	0.1	19.2	1.6	3.0	0.3	0.3	0.5	0.7	0.672
MnB	33.7	60.1	0.6	4.3	0.0	0.6	0.1	0.4	0.0	0.695
NiB	53.8	21.5	12.2	0.0	0.5	0.0	5.5	6.0	0.4	0.504
PbB	9.7	35.8	1.7	0.2	21.6	22.3	5.2	3.0	0.5	0.404
ZnB	90.0	0.0	4.4	0.6	3.3	0.8	0.7	0.0	0.0	0.732

Naturellement il ne viendrait à l'idée d'aucun pédologue de s'intéresser à la prédiction des variables SOL par les variables BLE. Notons, toutefois que le pH a un statut assez particulier : pour la plupart des éléments traces métalliques (et notamment le cadmium), la phyto-disponibilité croît très rapidement (exponentiellement) à mesure que le pH baisse. Dans l'espace des variables SOL pour lequel son indice est $j = 9$, c'est la variable la mieux représentée ($\sum_{k=1}^{k=s} x_{jk}^{(u)2} = 0.784$). Si on regarde son coefficient de détermination $R^2_{\mathbf{x}^j,9}$ de la régression en fonction des variables BLE (dont nous avons dit qu'elle n'a pas de sens pour une prédiction), c'est lui qui a la valeur maximale (0.538). Peut-on, à la constatation de ces deux résultats, faire la conjecture que son rôle est essentiel? Nous n'avons pas d'argumentation scientifique sérieuse pour l'affirmer; néanmoins ceci confirmerait que, même s'il n'apparaît pas comme variable essentielle dans l'interprétation des variables canoniques (et ultérieurement dans des modèles de régression de chaque variable BLE), il est toujours là comme un facteur latent...

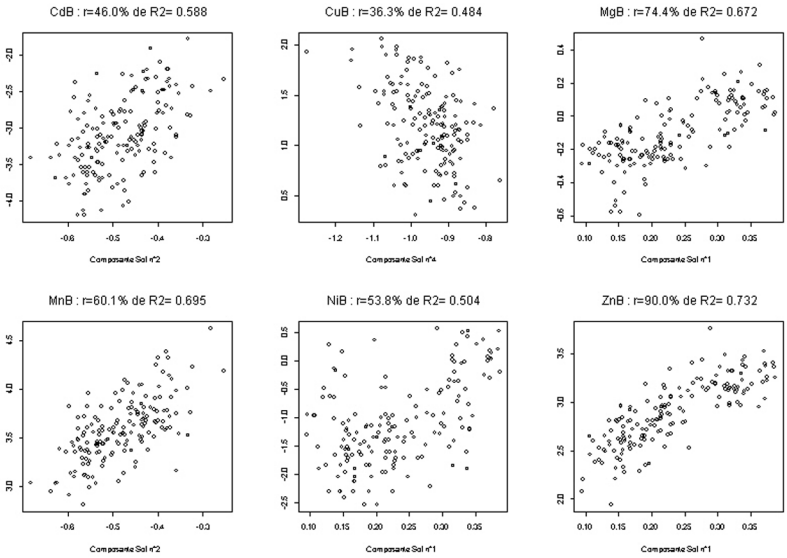

FIGURE 8
Sélections de relations ETM-Variables canoniques SOL

5.5. Ré-échantillonnage

On peut naturellement compléter ces calculs par les méthodes classiques de ré-échantillonnage, *jackknife* et *bootstrap [9]*. Les résultats sur les coefficients de corrélation canonique r_k (Tableau 9) montrent que les estimations que nous avons obtenues sont biaisées. Les graphiques jackknife (*cf.* figure 9) montrent qu'il n'existe pas d'observations particulièrement influentes sur l'estimation. Les graphiques bootstrap (*cf.* figure 10) donnent les images de distributions assez symétriques.

5.6. Vérifications, prédiction

Tous les résultats ont été complétés par des analyses dont nous ne donnerons pas les résultats ici. En particulier, nous avons supprimé les variables CEC, $CaCO_3$, CdS, CrS et NiS pour le SOL et de PbB pour le BLE, soit $p = 19$ et $q = 8$. Les nouveaux résultats sont très proches des précédents, la dimension de l'espace de représentation diminue de 1, soit $k = 5$. Les valeurs des coefficients de corrélation canonique sont : 0.8839, 0.8201, 0.7819, 0.6772 et 0.5821. L'interprétation des variables canoniques n'est pas modifiée de façon sensible. Nous pouvons donc accepter les premiers résultats sans arrière pensée.

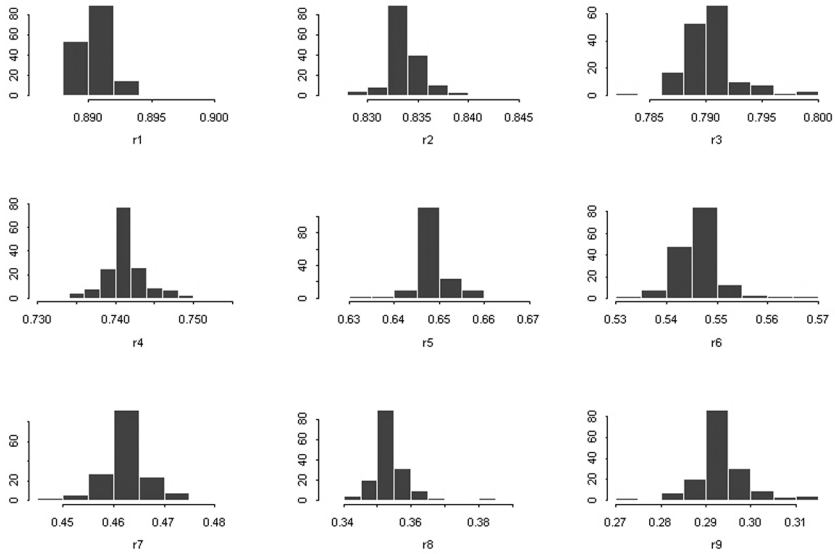

FIGURE 9
Distribution Jackknife des $r_k, k = 1 \ldots 9$

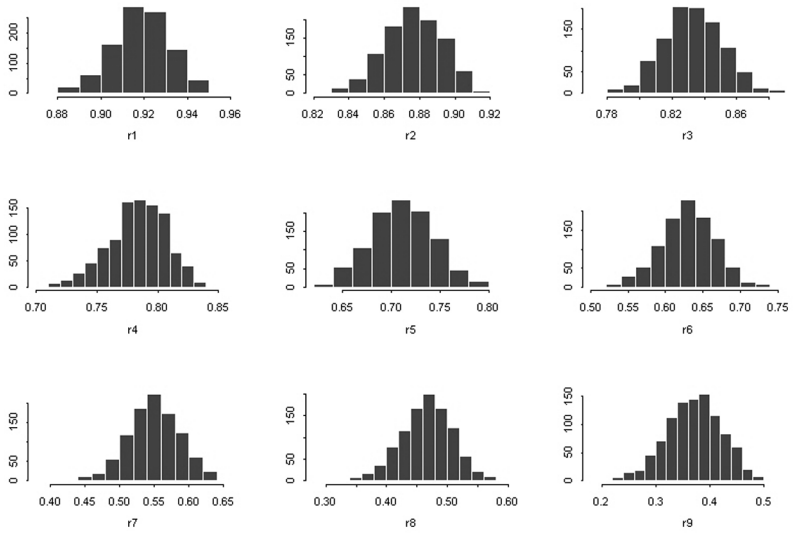

FIGURE 10
Distribution Bootstrap des $r_k, k = 1 \ldots 9$

TABLEAU 9
Estimations jackknife et bootstrap de r_k

		jackknife		bootstrap	
k	observée	biais	estimation	biais	estimation
1	0.890	0.034	0.856	0.028	0.862
2	0.833	0.059	0.774	0.043	0.791
3	0.790	0.054	0.736	0.043	0.747
4	0.741	0.052	0.689	0.042	0.699
5	0.648	0.076	0.572	0.063	0.585
6	0.546	0.106	0.440	0.084	0.462
7	0.462	0.094	0.367	0.088	0.374
8	0.353	0.244	0.109	0.114	0.239
9	0.293	0.083	0.210	0.078	0.216

De plus, les interprétations pédo-agronomiques qui ont été faites sont cohérentes avec les connaissances préalables des agronomes; elles leur apportent une vision globale et leur permettent d'envisager d'autres pistes de travail; elles les incitent à continuer et à approfondir cette démarche de prédiction de la composition de produits végétaux à partir de données analytiques du sol. Elles complètent des résultats partiels précédemment obtenus pour le cadmium particulièrement bien prédit [3], [4].

Comme nous l'avons dit en proposant d'autres présentations de l'*ACC*, celle d'un modèle linéaire multidimensionnel se prête bien à la prédiction de q valeurs des *ETM* de grain de blé pour de nouvelles valeurs analytiques possibles $\mathbf{x}_0$ du sol. Bien que nous n'ayons pas réalisé d'étude systématique de ces prédictions, les résultats permettent de voir quelles valeurs des p composantes de $\mathbf{x}_0$ sont susceptibles d'entraîner des valeurs «hors limite» du blé. En pratique, il est préférable de passer par la variable canonique $\mathbf{u}_0$ (qui n'a que s composantes au lieu de p) puis $\mathbf{v}_0$ et enfin $\mathbf{y}_0$. Le passage par $\mathbf{u}_0$ et $\mathbf{v}_0$ permet de visualiser la position de ces échantillons sur les graphiques tels que ceux de la figure 3.

5.7. Formulaires

Le tableau 10 contient les éléments principaux d'une ACC tandis que le tableau 11 fournit un résumé des outils interprétatifs d'une ACC.

TABLEAU 10
Eléments principaux issus d'une ACC

Eléments principaux	Définition	Relation
Obs. centrées réduites	$X_{n\times p} = \left[x^1 \quad \dots \quad x^p\right]$ $Y_{n\times q} = \left[y^1 \quad \dots \quad y^q\right]$	
corrélation initiale$\in M_{(p+q)\times(p+q)}$	$R = \frac{1}{n}\begin{bmatrix} X^TX & X^TY \\ Y^TX & Y^TY \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$	
Facteurs canoniques $\begin{cases} A_{p\times s} = \left[a^1 \quad \dots \quad a^k \quad \dots \quad a^s\right] \\ B_{q\times s} = \left[b^1 \quad \dots \quad b^l \quad \dots \quad b^s\right] \end{cases}$	$\begin{cases} R_{12}R_{22}^{-1}R_{21}a^k = \gamma_k^2 R_{11}a^k \\ R_{21}R_{11}^{-1}R_{12}b^k = \gamma_k^2 R_{22}b^k \end{cases}$	$\begin{cases} R_{22}^{-1}R_{21}a^k = \gamma_k b^k \\ R_{11}^{-1}R_{12}b^k = \gamma_k a^k \end{cases}$
Variables canoniques $\begin{cases} U_{n\times s} = X_{n\times p}A_{p\times s} \\ V_{n\times s} = Y_{n\times q}B_{q\times s} \end{cases}$	$\begin{cases} \underbrace{u^k}_{n\times 1} = X\underbrace{a^k}_{p\times 1} \\ \underbrace{v^k}_{n\times 1} = Y\underbrace{b^k}_{q\times 1} \\ k = 1,\dots s = \min(p,q) \end{cases}$	$cor\left(u^k, v^k\right) = r_k = \sqrt{\gamma_k^2}$

TABLEAU 11
Outils interprétatifs d'une ACC

Eléments principaux	Définitions
Corrélation initiale	$R = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$
corrélations canoniques	$\Gamma_{p\times q} = \left\{ \begin{bmatrix} diag(r_k) & \overbrace{0}^{s\times(q-s)} \\ \underbrace{0}_{(p-s)\times s} & \underbrace{0}_{(p-s)\times(q-s)} \end{bmatrix} \right\}, \quad 1 \geq r_1 \geq \dots \geq r_s \geq 0$
Corrélations intra-groupe	$\begin{cases} X^{(u)} = cor(X,U) = R_{11}A \in M_{p\times s} \\ Y^{(v)} = cor(Y,V) = R_{22}B \in M_{q\times s} \end{cases}$
Variance extraite pour la k var canonique	$\begin{cases} U_k^2 = \sum_{j=1}^{j=p} x_{jk}^{(u)2}/p \text{ pour le premier groupe} \\ V_k^2 = \sum_{j=1}^{j=q} y_{jk}^{(v)2}/q \text{ pour le second.} \end{cases}$
Corrélations inter-groupe	$\begin{cases} X^{(v)} = cor(X,V) = R_{11}A\ diag(r_k) \in M_{p\times s} \\ Y^{(u)} = cor(Y,U) = R_{22}B\ diag(r_k) \in M_{q\times s} \end{cases}$
Redondance ou variance expliquée	$\begin{cases} V_{X/v^k}^2 = r_k^2 U_k^2 = \sum_{j=1}^{j=p} x_{jk}^{(v)2}/p \text{ pour le premier groupe} \\ U_{Y/u^k}^2 = r_k^2 V_k^2 = \sum_{j=1}^{j=q} y_{jk}^{(v)2}/q \text{ pour le second.} \end{cases}$
Communauté des variables	$\begin{cases} R_{x^j,t}^2 = \sum_{k=1}^{k=t} x_{jk}^{(v)2}; \forall t = 1,\dots,s \\ R_{y^j,t}^2 = \sum_{k=1}^{k=t} y_{jk}^{(u)2}; \forall t = 1,\dots,s \end{cases}$

6. Conclusions

Au terme de cette présentation, pouvons nous dire que le but du programme GESSOL cité en début d'article a été atteint? Rappelons le : «*bâtir des modèles permettant de détecter les cas de concentrations excessives en ETM dans les grains de blé à partir de données pertinentes acquises sur des échantillons de sol*». Faisant fi de toute modestie, nous serions assez tentés de répondre par l'affirmative, ... avec beaucoup de prudence! Certes, nous avons obtenu des modèles utilisables à partir de données plus ou moins facilement mesurables; mais le chemin pour y parvenir est long et tortueux. Nous avons vu que pour choisir un modèle, il ne suffit pas de calculer un R^2! Ceux qui pensent que la pratique statistique consiste à avoir un bon logiciel de calcul commettent une erreur fondamentale.

Dans la mesure où l'*ACC* est une forme de régression multiple, toutes les précautions nécessaires pour faire une interprétation correcte peuvent être reprises. Il faut donc examiner les aspects suivants :

– Bien sûr, et avant tout, le corpus de données, l'échantillon, doit être représentatif d'une population plus large si on veut étendre les résultats. En outre, pour diverses raisons listées dans les références [3] et [4] on aurait pu ne pas trouver de lien entre variables SOL et variables BLE. Or, on en trouve! Et certaines particulièrement bonnes. Cela peut sans doute paraître évident. Mais le choix des « bonnes » variables est nécessaire si l'on veut voir apparaître ensuite des relations; ici ce choix judicieux concerne tout particulièrement le manganèse total et les extractions partielles par le $DTPA$ et le NH_4NO_3.

– Les données suspectes doivent être éliminées après un examen approfondi des distributions de chaque variable et de l'examen des premiers graphiques $\{\mathbf{u}^1, \mathbf{v}^1\}$, $\{\mathbf{u}^2, \mathbf{v}^2\}$, ...

– Les coefficients de corrélation sont des indices souvent peu fiables à cause de la présence d'observations perturbatrices; il peut donc s'avérer utile de les remplacer par des estimations robustes [24].

– Si les coefficients des variables ($\mathbf{A}$ ou $\mathbf{B}$) et les coefficients de corrélation avec les variables observées ($\mathbf{X}^{(u)}$ ou $\mathbf{Y}^{(v)}$) sont très différents il sera vraisemblablement utile de ne conserver qu'une variable parmi celles qui sont très corrélées entre elles et de recommencer l'analyse sur un nombre réduit de variables. Il faut aussi noter que si l'interprétation du premier couple $\{\mathbf{u}^1, \mathbf{v}^1\}$ est assez facile, celle des couples suivants peut l'être moins, car il faut intégrer dans l'interprétation que le second couple $\{\mathbf{u}^2, \mathbf{v}^2\}$, et les suivants, sont orthogonaux (c'est-à-dire non corrélés) avec les précédents.

– Utiliser les résultats de l'*ACC* pour choisir ensuite les modèles de régression les plus appropriés. On peut naturellement dans ces études plus fines étudier les influences de certaines observations en faisant des *régressions robustes* en utilisant les procédures issues des travaux de [26] proposées dans [25] (pp.171-174).

– Les tests statistiques n'ont de valeur que si la Normalité est approximativement respectée.

– Les méthodes de ré-échantillonnage doivent être utilisées, ne serait-ce que pour contrôler que les résultats ne sont pas différents des résultats obtenus par le calcul classique.

Avec toutes ces remarques en tête, l'interprétation d'une *ACC* est de même nature que celle d'une régression ou d'une analyse en composantes principales. Enfin, quelles que soient les difficultés qu'elle peut soulever, dans le domaine des données environnementales, son objectif nous paraît tout à fait adapté à celui des études qui y sont menées.

Références

[1] AFIFI F., CLARK V.A. and MAY S. (2004), *Computer-Aided Multivariate Analysis*, 4th. ed. Chapman & Hall, Londres.

[2] ANDERSON T.W. (1958), *An introduction to multivariate analysis*, Wiley, New York.

[3] BAIZE D. et TOMASSONE R. (2003), Modélisation empirique du transfert du cadmium et du zinc des sols vers les grains de blé tendre, *Étude et Gestion des Sols*, **4**, 219-238.

[4] BAIZE D. et TOMASSONE R. (2005), Prédiction de la teneur en cadmium du grain de blé tendre à partir de mesures sur échantillons de sols. *Journées techniques «Transfert des polluants des sols vers les végétaux cultivés et les animaux d'élevage, Outils pour l'évaluation des risques sanitaires»*, ADEME Paris, 1er février 2005, pp. 5-14.

[5] BRILLINGER (1975), *Time series : data analysis and theory*, Holt, Rinehart and Winston, New York.

[6] BUZAS T.E., FORNELL C. and RHEE B-D. (1989), Conditions under which canonical correlation and redundancy maximization produce identical results, *Biometrika*, **76** (3), 618-621.

[7] CHESSEL D., LEBRETON J.-D. and YOCCOZ N. (1987), Propriétés de l'analyse canonique des correspondances; une illustration en hydrobiologie. *Rev. Statistique Appliquée*, **XXXV** (4), 55-72.

[8] COURBE C., BAIZE D., SAPPIN-DIDIER V. et MENCH M. (2002), Impact de boues d'épuration anormalement riches en cadmium sur des sols agricoles du Limousin, Actes des $7^{ème}$ JNES, Orléans, pp. 15-16.

[9] DAVISON A.C. and HINKLEY D.V. (2003), *Bootstrap Methods and their Application*, Cambridge University Press, Cambridge.

[10] GITTINS G. (1980), *Canonical Analysis. A review with Applications in Ecology*, Springer Verlag, Berlin.

[11] HOTELLING H. (1936), Relations between two sets of variables. *Biometrika*, **28**, 321-377.

[12] JAMES M. (1979), The generalized inverse form of canonical correlation, *Communications in Statistics-Theory and Methods*, **A8**, 561-568.

[13] JOBSON J.D. (1992), *Applied Multivariate Data Analysis*. Volume 2 : Categorical and Multivariate Methods, Springer Verlag, New York.

[14] LAWLEY D.N. (1959), Tests of significance in canonical analysis, *Biometrika*, **46**, 59-66.

[15] LAZRAQ A., CLEROUX R. et KIERS H.A.L. (1992), Mesures de liaison vectorielle et généralisation de l'analyse canonique, *Rev. Statistique Appliquée*, **XL** (1), 22-35.

[16] LEBART L. FÉNELON J.-P. (1975), *Statistique et informatique appliquées*, Dunod, Paris.

[17] MILLER J.K. (1975), The sampling distribution and a test for the significance of the bimultivariate redondancy : a Monte Carlo study, *Multivariate Behavioral Research*, **10**, 233-244.

[18] MORRISON D.F. (1976), *Multivariate statistical methods*, 2nd ed. McGraw-Hill, New York.

[19] PINET C., LECOMTE J., VIMONT V. et AUBURTIN G. (2003), *Teneurs des plantes à vocation agronomique en éléments traces suite à l'épandage de déchets organiques*. ADEME, Angers.

[20] PONTIER J. et PERNIN M.O. (1989), Relations entre l'Analyse Canonique Complète et la méthode LONGI, *Rev. Statistique Appliquée*, **XXXVII** (4), 67-82.

[21] PONTIER J. et NORMAND M. (1992), A propos de généralisation de l'analyse canonique, *Rev. Statistique Appliquée*, **XL** (1), 57-75.

[22] SAPORTA G. (1990), *Probabilités, analyse des données et statistique*, Technip, Paris.

[23] TOMASSONE R. (2002), Epuration des boues et enquête publique : l'expertise citoyenne est-elle un leurre ? *Natures, Sciences, Sociétés*. **10** (3), 27-36.

[24] TUKEY J.W. (1969), Analyzing data : sanctification or detective work ? *American Psychologist*, **24**, 83-91.

[25] VENABLES W.N. and RIPLEY B.D. (1999), *Modern Applied Statistics with S-PLUS*, Springer, New York.

[26] YOHAI V., STAHEL W.A. and ZAMAR R.H. (1991), A procedure for robust estimation and inference in linear regression, in *Directions in Robust Statistics and Diagnostics, Part II*, W.A. Stahel and S.W. Weisberg, ed. Springer, New York.