ÉTUDE COMPARATIVE DE LA MÉTHODE DE RÉGRESSION SUR LES FACTEURS D'UNE ANALYSE DES CORRESPONDANCES

ÉTUDE COMPARATIVE DE LA MÉTHODE DE RÉGRESSION SUR LES FACTEURS D'UNE ANALYSE DES CORRESPONDANCES

par

Jean-Cléophas Ondo
Marius Lachance

Institut National de la Recherche Scientifique, INRS-Eau
2800 rue Einstein, Case postale 7500, SAINTE-FOY (Québec) G1V 4C7

Rapport de recherche No R-553

Novembre 1999

ISBN 2-89146-329-3

Avant-propos

Dans ce travail, nous voulons souligner l'importance de la méthode de régression sur les facteurs d'une analyse des correspondances comme une alternative de la méthode des moindres carrés ordinaires. L'intérêt de cette méthode est qu'elle vise à atténuer les inconvénients de la multicolinéarité importante qui se manifeste dans beaucoup de problèmes de régression multiple.

Malgré la démarche entreprise dans ce travail, coïncidant pratiquement avec les développements théoriques fondant la légitimité de cette méthode, elle demeure aujourd'hui un peu mystérieuse pour l'utilisateur. Plusieurs raisons pourraient expliquer cet état de fait. D'une part, la mise en œuvre des calculs réclame l'intervention directe de l'utilisateur, et les aspects informatiques auxquels il est confronté en premier lieu, ont longtemps estompé la véritable nature statistique du problème. D'autre part, la théorie statistique sous-jacente n'est pas très simple, et tous les problèmes concrets qui relèvent de cette méthode ne sont pas encore résolus.

En écrivant ce rapport, nous voulons respecter simultanément deux points de vue différents :

- le premier est la souplesse d'utilisation de la méthode que nous présentons ;

- le second résulte d'une réflexion personnelle des auteurs : il est vain d'expliquer des méthodes perfectionnées sans mettre à la disposition du lecteur les outils permettant de les utiliser : en conséquence, un exemple d'application est présenté au chapitre 5, les différents résultats que nous commentons sont obtenus à l'aide des logiciels ADDAD et STATISTICA.

Notre objectif est donc de rendre accessibles les notions utiles à la compréhension de cette méthode, et en les illustrant par un exemple d'application. Nous espérons ainsi montrer qu'il n'y a en fait aucun mystère en la matière.

Table des matières

Liste des tableaux

Liste des figures

1. INTRODUCTION

L'analyse des résultats d'une expérience relève très souvent de l'emploi d'un modèle de régression lorsque les observations qui en sont issues peuvent être représentées, chacune, comme la somme d'un terme systématique, dépendant de la valeur prise par une ou plusieurs autres variables, et de la réalisation d'une variable aléatoire. L'objectif est en général l'étude de la relation entre la variation d'une variable appelée variable réponse ou dépendante et d'une ou plusieurs autres variables appelées variables explicatives. Une telle relation est établie en estimant les paramètres associés aux variables explicatives.

Il arrive fréquemment dans la pratique que les données que nous utilisons dans une telle analyse de régression ne nous fournissent pas des réponses décisives aux questions que nous nous posons. Nous envisageons ici le cas où les erreurs des estimations des paramètres de régression sont importantes ou encore lorsque les valeurs observées de la statistique t de Student sont très petites. Ainsi, dans le premier cas, une difficulté apparaît toutefois : les intervalles de confiance construits sur ces paramètres d'intérêt sont donc trop grands. D'une façon générale, ces différentes situations apparaissent lorsque les variables explicatives ont une faible variance / ou forment un système de vecteur «presque» liés. Et, qu'alors la matrice $X'X$ (X étant la matrice des variables explicatives) a un certain nombre de valeurs propres très petites, qui pèsent numériquement sur son inversion. Dans cette dernière situation, on dit qu'on est en présence d'un cas de *multicolinéarité* (ou de *colinéarité* ou encore que la matrice X des variables explicatives est *mal conditionnée*). Ce terme a été pour la première fois introduit par *Ragnar (1934)*. Cette multicolinéarité peut engendrer de graves instabilités sur certaines estimations des paramètres du premier ordre obtenus par la méthode *des moindres carrés ordinaires (MCO)*. Dès lors qu'on ne peut

ignorer les conséquences fâcheuses engendrées par la présence de multicolinéarité, il convient d'analyser et de comprendre ses conséquences dans un modèle de régression multiple.

Dans l'étude de la liaison entre une variable expliquée y et K variables explicatives $X_1,\ldots,X_K$, la méthode des MCO est certainement le procédé le plus souvent retenu quand il s'agit d'ajuster une équation de régression linéaire multiple sur les données $(y, X_1,\ldots,X_K)$ constituant un échantillon. Ce procédé implique de minimiser la somme des carrés des écarts résiduelle. Des raisons à la fois pratiques et théoriques justifient le recours systématique à ce critère d'ajustement. En effet, l'estimation des coefficients de régression par la méthode des MCO peut être réalisée sans difficulté par n'importe quel logiciel statistique. Aussi, les estimateurs des MCO jouissent, sous certaines conditions d'applications, d'un ensemble de propriétés intéressantes.

L'une des difficultés majeures avec l'estimateur usuel des MCO est le problème de la multicolinéarité. Comme nous l'avons déjà mentionné, ce problème se produit lorsque deux variables explicatives ou davantage sont fortement corrélées ; c'est-à-dire, lorsque l'une d'entre elles varie, l'autre a une forte tendance à varier aussi. Dans pareille circonstance, il est alors difficile, sinon impossible, d'isoler l'effet particulier de chacune de ces variables sur la variable dépendante. Aussi, les coefficients MCO estimés peuvent être sans signification statistique (ou même être affectés d'un signe erroné) alors que le coefficient de détermination R^2 atteint une valeur «élevée».

La multicolinéarité peut donc engendrer de graves instabilités sur certaines estimations des paramètres du premier ordre obtenus par la méthode des MCO. Ces instabilités résultent d'une part de difficultés purement numériques liées à l'inversion d'une matrice presque singulière. On peut de ce fait se heurter aux limites de l'algorithme de calcul implanté dans le logiciel statistique. D'autre part, certains estimateurs peuvent être caractérisés par une forte variabilité. En effet, la matrice de variance covariance de l'estimateur des MCO est

proportionnelle à l'inverse de la matrice $X'X$. Une partie de ses éléments peuvent donc être très grands. La multicolinéarité s'apparente donc à un problème numérique doublé d'un problème statistique. Il faut noter aussi que dans un modèle de régression linéaire ordinaire affecté de multicolinéarité, l'estimateur des MCO demeure sans biais linéaire optimal. Il a cependant l'inconvénient d'être peu robuste au voisinage de la multicolinéarité stricte.

Pour surmonter le problème de la multicolinéarité dans un modèle de régression multiple, diverses approches ont été proposées dans la littérature. Une possibilité est la réduction de l'intensité de la colinéarité par l'élimination d'une ou de plusieurs variables explicatives. C'est-à-dire d'essayer de choisir avec sagesse les variables explicatives qui ont aussi peu d'inter-corrélation que possible. Il se pose alors le problème du choix des variables à faire figurer dans l'équation de régression linéaire multiple. Une autre approche consiste à faire appel à des techniques de régression spécialement mise au point pour atténuer les effets de la multicolinéarité. Ces techniques de régression sont basées sur le calcul de nouvelles variables explicatives : il s'agit entre autres de la régression en fonction des composantes principales, de la régression proposée par *Webster*, *Gunst* et *Mason (1974)* et de la régression par les moindres carrés partiels (*Martens et Naes, 1989*).

Le but de ce travail est de présenter une autre méthode de régression particulièrement mise au point pour atténuer aussi les effets de la multicolinéarité. Cette méthode est basée sur l'utilisation des facteurs d'une analyse des correspondances comme nouvelles variables explicatives : il s'agit de la méthode de régression basée sur les facteurs à partir des éléments supplémentaires d'une analyse des correspondances. Nous sommes donc intéressés à faire une étude comparative de cette méthode.

Puisque le sujet que nous traitons est vaste, nous ne pouvons faire justice à toutes les dimensions de la question. Nous abordons tout d'abord le sujet en rappelant quelques notions de régression multiple classique en accord avec les problèmes liés à la multicolinéarité. Puis, nous décrivons la méthode de régression sur les facteurs d'une

analyse des correspondances. Par la suite, nous exposons la méthode de régression basée sur les facteurs à partir des éléments supplémentaires d'une analyse des correspondances. Nous insistons notamment sur les avantages, limites et portée de cette méthode par rapport à la régression traditionnelle. Aussi, nous survolons la revue de littérature des développements et travaux portant sur cette méthode de 1980 jusqu'à aujourd'hui. Finalement, nous envisagerons la comparaison de cette méthode avec l'approche traditionnelle à partir d'un exemple d'application sur un ensemble de données. Cet exemple concerne un ensemble de données tirées à partir de la base de données météorologiques de la province de l'Ontario. Il s'agit ici d'expliquer une variable hydrologique en fonction d'un certain nombre de variables physiographiques.

2. NOTIONS DE RÉGRESSION MULTIPLE AU SENS DES MOINDRES CARRÉS

Dans ce chapitre, il s'agira essentiellement de situer la régression multiple par rapport au problème de multicolinéarité. Notre objectif ici n'est pas de présenter les notions de régression de façon exhaustive. Il existe un large éventail d'ouvrages concernant la régression linéaire. Le lecteur désireux d'obtenir des informations complémentaires consultera les ouvrages de *Draper et Smith (1981)*, de *Theil (1978)* et de *Weisberg (1985)*.

2.1. Le cadre de référence

On se place dans le modèle de régression multiple ordinaire à K variables explicatives et N observations :

$$\underset{(N,1)}{y} = \underset{(N,K)(K,1)}{X\beta} + \underset{(N,1)}{\varepsilon} \qquad (2.1)$$

où y est le vecteur de la variable dépendante, X est la matrice des variables explicatives, β est le vecteur des paramètres du premier ordre, et ε est le vecteur des perturbations aléatoires. On considère en outre que les perturbations aléatoires relatives à des individus différents sont des réalisations indépendantes d'une même variable aléatoire normale de moyenne nulle et de matrice de variances-covariances :

$$V(\varepsilon) = \sigma^2 I_N .$$

En pratique, l'objectif poursuivi est d'estimer β à partir des données observées sur N individus choisis de manière aléatoire et simple et pour lesquels le modèle spécifié à

l'équation 2.1, est applicable. y étant le vecteur des valeurs observées de la variable à expliquer et X la matrice des variables explicatives, l'estimation au sens des moindres carrés ordinaires du vecteur β consiste à minimiser la quantité $\varepsilon'\varepsilon$. La caractéristique essentielle de l'estimateur des moindres carrés ordinaires est d'être l'estimateur linéaire non biaisé de variance minimum. On peut montrer que les formules relatives au calcul de

$$\hat{\beta},\ V(\hat{\beta}) \text{ et } E(\hat{\beta}-\beta)'(\hat{\beta}-\beta)$$

sont données par :

$$\hat{\beta} = (X'X)^{-1}X'y \tag{2.2}$$

$$V(\hat{\beta}) = \sigma^2(X'X)^{-1} \quad \Leftrightarrow \quad V(\hat{\beta}) = \sigma^2 \sum_{i=1}^{K} \lambda_i^{-1} > \sigma^2 \lambda_{k0}^{-1} \tag{2.3}$$

$$E\left[(\hat{\beta}-\beta)'(\hat{\beta}-\beta)\right] = \sigma^2 tr(X'X)^{-1} \tag{2.4}$$

où, $tr(X'X)^{-1}$ représente la trace de la matrice $(X'X)^{-1}$. Ces formules font intervenir l'inverse de la matrice $X'X$ (ou encore la matrice de corrélation des K variables explicatives). Elles ne peuvent donc être utilisées que si la matrice $X'X$ est non singulière, c'est-à-dire s'il n'existe pas de relations linéaires entre les variables explicatives.

2.2. La multicolinéarité et la stabilité des coefficients de régression

Lorsque deux variables explicatives sont fortement liées, il est difficile de «dissocier» leurs effets séparés sur la variable réponse y. Lorsque l'une d'entre elles augmente, l'autre augmente en même temps. À quel accroissement attribuer alors l'accroissement de la

variable réponse y ? C'est bien difficile à dire - et c'est précisément dans ce type de situation qu'apparaît le problème de la multicolinéarité.

Autant la multicolinéarité exacte des mathématiciens est une notion clairement établie, autant la multicolinéarité approchée est délicate à concevoir et à définir. Sa définition même dans le simple cadre du modèle de régression multiple spécifié à l'équation 2.1 traduit cette difficulté : une situation de multicolinéarité (au sens de multicolinéarité approchée) apparaît lorsque les vecteurs colonnes de la matrice des variables explicatives sont «presque» colinéaires. Ce faisant, la multicolinéarité exacte implique que la matrice $X'X$ est singulière, c'est-à-dire que la plus petite valeur propre, λ_{k0}, est nulle $\left(\lambda_{k0} = 0\right)$. La multicolinéarité approchée signifie que la matrice $X'X$ est presque singulière et la plus petite valeur propre, λ_{k0}, est proche de zéro $\left(\lambda_{k0} \to 0\right)$.

Dans le cas de la multicolinéarité exacte, le vecteur $\hat{\beta}$ est indéterminé, une infinité de vecteurs différents conduisant à la même valeur minimum de $\varepsilon'\varepsilon$. La multicolinéarité approchée cause des problèmes de précision numérique lors du calcul des coefficients. De plus, elle conduit à des variances des coefficients et donc aussi à un carré moyen de l'erreur de $\hat{\beta}$ importants.

En somme, pour un modèle de régression multiple, un paramètre est généralement interprété comme étant la variation de la variable dépendante pour l'accroissement d'une unité de la variable explicative considérée, toutes les autres variables étant fixées. Or cette interprétation n'est plus appropriée lorsque les variables explicatives sont fortement corrélées entre elles. La présence d'une forte corrélation entre deux ou plusieurs variables explicatives est appelée multicolinéarité. Elle se présente souvent lorsqu'un grand nombre de variables explicatives sont incluses au modèle et que certaines d'entre elles fournissent une information similaire. Cela entraîne les problèmes suivants :

- La variance des estimations, $\hat{\beta}_i \quad (i = 1,\ldots,K)$, des paramètres a tendance à être gonflée : pour la multicolinéarité approchée, $\lambda_{k_0} \to 0$ et l'équation 2.3 implique que $V(\hat{\beta}) \to \infty$, c'est-à-dire que $\hat{\beta}$ est sujet à une très grande variance. Dans cette situation, l'information fournie sur ces paramètres est donc peu précise.

- Le fait d'ajouter ou de retrancher une variable explicative du modèle de régression spécifié à l'équation 2.1 peut modifier de façon importante l'estimation des paramètres.

- L'estimation d'un paramètre peut être de signe opposé à celui prévu.

- Les paramètres associés aux variables fortement corrélées peuvent être non significatifs en dépit d'une forte relation entre la variable dépendante et celles-ci.

- La variance des estimations des valeurs prises par la variable dépendante pour certaines valeurs des variables explicatives a aussi tendance à être grande, particulièrement lorsque cette combinaison de valeurs ne fait pas partie de l'échantillon.

La détection de la présence de multicolinéarité est très délicate. En effet, celle-ci ne peut être appréhendée comme un problème d'inférence statistique si elle relève des données observées et non des variables aléatoires à la base de leur constitution (voir par *exemple Maddala (1977)* ou *Erkel-Rousse (1994 et 1994/95)*). Plusieurs auteurs ont proposé des indices de détection de la multicolinéarité. Une documentation appropriée est donnée par *Farrar* et *Glauber (1967)*, *Gunst* et *Mason (1977)*, *Mason*, *Gunst* et *Webster (1975)*, *Silvey (1969)*. Les indicateurs *BKW* de *Belsley*, *Kuh* et *Welsch (1980)* demeurent cependant parmi les plus riches, les plus célèbres et les plus accessibles. Malgré cela, il n'en demeure pas moins que certains auteurs leur ont depuis 1980 opposé des concurrents, comme les «***variance inflation factors***» ou VIF et leurs inverses appelés «***tolérance***» ou TOL.

Dans la régression usuelle, pour réduire les inconvénients liés à la multicolinéarité sans remettre en cause le principe des moindres carrés, on préfère parfois supprimer une ou

plusieurs variables explicatives. On se trouve alors confronté au problème du choix des variables explicatives qui doivent figurer dans le modèle de régression multiple.

Dans la littérature, plusieurs méthodes de sélection de variables sont connues. La plupart d'entre elles peuvent être regroupées en deux catégories :

- les méthodes de recherche exhaustives ;
- les algorithmes de sélection systématique.

La première catégorie de ces méthodes est basée sur l'inspection de sous-ensembles de variables explicatives possibles par rapport à un certain critère. La deuxième catégorie quant à elle est basée sur des algorithmes tels que les régressions d'inclusion et d'exclusion «pas à pas» ("stepwise"), la régression d'exclusion «pas à pas» ("backward") et la régression d'inclusion «pas à pas» ("forward"). Bien que ces méthodes produisent souvent de bons sous-ensembles de variables explicatives, elles présentent aussi quelques faiblesses. En effet, les procédures de recherches exhaustives demandent des calculs considérables et deviennent par le fait même très coûteuses ou même difficilement réalisables dans un grand nombre de problèmes. Les algorithmes de sélection, bien que leur temps de calcul soit efficace, parfois, ils ne parviennent pas à détecter le meilleur sous-ensemble de variables explicatives. La question du choix du meilleur sous-ensemble de variables explicatives reste donc poser. En effet, comme nous venons de le voir, aucune des méthodes proposées ne nous permet, de façon formelle, de faire le bon choix. Pour une étude plus complète des méthodes de sélection de variables, le lecteur est référé à *Miller (1990)*. De plus, une synthèse bibliographique a été faite par *Hocking (1976)* et *Thompson (1978a, 1978b)*.

D'autres méthodes permettent quelquefois de surmonter ou du moins de réduire l'inconvénient de la présence de la multicolinéarité dans un modèle de régression multiple. Il s'agit des méthodes de régression spécialement mise au point pour atténuer les effets de la multicolinéarité. Ces méthodes de régression consistent à calculer de nouvelles variables

explicatives, à effectuer la régression de la variable expliquée par ces nouvelles variables. Nous allons examiner l'une d'entre elle dans la suite de ce rapport.

3. MÉTHODE DE RÉGRESSION SUR LES FACTEURS D'UNE ANALYSE DES CORRESPONDANCES

La méthode de régression sur les facteurs d'une analyse des correspondances est fondée sur le même principe que la méthode de *régression en fonction des composantes principales*, appelée aussi *régression orthogonalisée* : elle repose sur l'utilisation, comme variables explicatives, des valeurs des composantes principales calculées à partir de la matrice X des variables explicatives du modèle de régression indiqué à l'équation 2.1.

Dans la régression sur les facteurs d'une analyse des correspondances, au lieu de prendre les composantes principales comme variables explicatives, on utilise plutôt les valeurs des facteurs obtenus *après analyse des correspondances multiples* (ACM). Le vecteur $\hat{\beta}$ ainsi obtenu par l'intermédiaire des facteurs d'une ACM est théoriquement identique au vecteur $\hat{\beta}$ obtenu par les moindres carrés ordinaires. L'intérêt du passage par les facteurs d'une ACM peut être numérique. En effet, comme nous l'avons déjà mentionné au début de ce rapport, on évite l'inversion d'une matrice qui, dans certains cas, est quasi singulière. Il est également théorique, car, il permet, notamment, d'expliciter la variance des coefficients de régression β_i afin de mettre en évidence l'incidence de la présence de multicolinéarité. Il convient maintenant de présenter le principe sous-jacent à la méthode de régression sur les facteurs d'une analyse des correspondances. Avant cela, rappelons quelques définitions et compléments sur l'analyse des correspondances multiples et dont la connaissance représente l'essentiel de ce qui est nécessaire pour aborder la suite.

3.1 Rappels sur l'analyse des correspondances multiples

L'analyse des correspondances multiples permet d'étudier une population de I individus décrits par J variables qualitatives. Cette étude met en jeu trois familles d'objets : individus, variables et modalités. L'unicité du tableau qui est à l'origine est réalisée en articulant les interprétations autour de la typologie des modalités. En effet, cette typologie permet d'étudier l'association mutuelle entre les modalités c'est-à-dire les liaisons entre les couples de variables. Elle permet aussi d'aborder celle des individus en examinant le comportement moyen de classes d'individus. Les objectifs indiqués dans l'étude des variables et des individus s'expriment aussi en grande partie à l'aide des modalités. Le principe de l'ACM est en fait une analyse des correspondances simple sur un tableau de données composé d'une population de I individus décrits par J variables qualitatives : les lignes représentent les individus, les colonnes représentent les variables. Ainsi, à l'intersection de la ligne i et de la colonne j, se trouve la valeur x_{ij} (ou encore le codage condensé) de l'individu i pour la variable j. On réalise ainsi une *analyse factorielle des correspondances (AFC)* sur ce tableau. Toutefois, comme les valeurs x_{ij} sont des codifications qui ne possèdent pas de propriétés numériques, il n'est donc pas possible de traiter directement ce tableau de données par AFC.

Une autre façon de présenter ces mêmes données est de construire un Tableau Disjonctif Complet (TDC) : à l'intersection de la ligne i et de la colonne k on trouve x_{ik} qui vaut 1 ou 0 selon que l'individu i possède la modalité k ou non. Le développement de l'AFC sur ce tableau est sans intérêt pratique immédiat. En revanche, pour généraliser l'AFC à l'étude des croisements entre plus de deux variables, on peut construire un tableau contenant l'ensemble des tableaux de contingence entre les variables prises deux à deux. Le Tableau de BURT ainsi construit n'est pas exactement un tableau de contingence, mais une juxtaposition de tels tableaux ; chaque individu y apparaît J^2 fois, car on croise l'ensemble des modalités de toutes les variables avec lui-même. La problématique de l'ACM peut donc être considérée comme une généralisation de l'AFC. Par ailleurs, si l'on veut obtenir de nouvelles variables explicatives pour atténuer l'effet de la multicolinéarité

dans un modèle de régression multiple, il faut d'abord faire l'analyse des correspondances multiples. Pour cela, il convient tout d'abord de faire quelques rappels sur le principe de l'analyse factorielle des correspondances.

3.1.1. Principe de l'analyse factorielle des correspondances

L'analyse factorielle des correspondances (AFC) est une technique récemment mise au point par l'équipe de recherche du professeur *J. P. BENZECRI* au début des années 60, à l'université Paris IV. Elle a été conçue pour étudier des tableaux appelés couramment tableaux de contingence (ou tableaux croisés). Il s'agit de tableaux d'effectifs obtenus en croisant les modalités de deux variables qualitatives définies sur une même population de n individus. De ce fait, les données en AFC apparaissent habituellement sous forme de fréquences dans un tableau à double classifications. Aucune des classifications ne joue de rôle particulier, comme c'est le cas en analyse des composantes principales (ACP) où l'une est faite des sujets et l'autre des variables sur lesquelles les sujets sont évalués. D'ailleurs, l'extraction des facteurs se fera successivement sur les classifications lignes et colonnes. En somme, les objectifs de l'AFC peuvent s'exprimer de manière tout à fait analogue à ceux de l'ACP : on cherche à réduire la dimension des données en conservant le plus d'information possible.

En générale, le tableau brut de l'AFC présenté précédemment n'est pas analysé directement. En effet, les comparaisons de lignes (qui peuvent donner lieu au tracé de profils) ne sont instructives que si le total de la i ème ligne est le même pour chaque ligne. C'est pourquoi il y a intérêt dans une première opération à transformer le tableau des données. Pour cela, on divise chaque terme (ou fréquence) f_{ij} de la ligne i (respectivement colonne j) par la marge $f_{i.}$ (respectivement $f_{.j}$) de cette ligne i (respectivement colonne j). La nouvelle ligne (respectivement colonne) est appelée *profil ligne* (respectivement *profil colonne*). La figure 3.1 montre une illustration du tableau final. Cette transformation découle de l'objectif qui vise à étudier la liaison entre les deux variables au travers de l'écart entre les pourcentages en lignes.

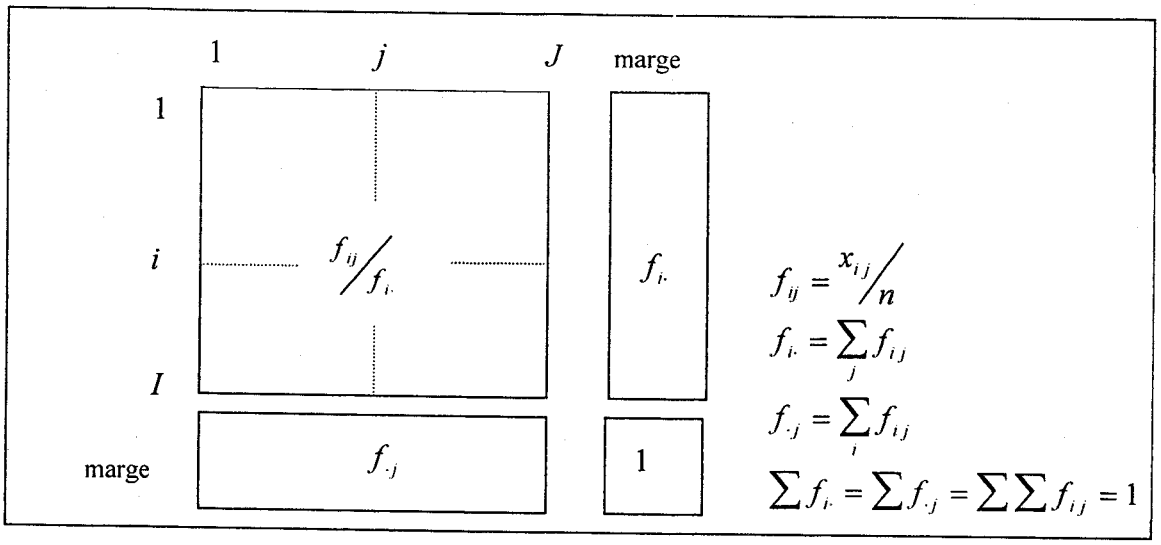


Figure 3.1 : La figure illustrative des profils ligne et colonne ainsi que ses marges

La ressemblance entre deux lignes ou entre deux colonnes peut s'évaluer objectivement en introduisant une distance entre leurs profils. Cette distance est connue sous le nom de distance du χ^2 :

$$d\chi^2(\textit{profil ligne } i, \textit{profil ligne } l) = \sum_j \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{lj}}{f_{l.}} \right)^2$$

$$d\chi^2(\textit{profil colonne } j, \textit{profil colonne } k) = \sum_j \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ik}}{f_{.k}} \right)^2$$

Cette distance du χ^2 jouit d'une propriété fondamentale appelée ***équivalence distributionnelle***. Comme les ensembles d'indices I et J jouent un rôle symétrique et l'on n'a pas à distinguer entre «individus » et «variables », on peut donc définir deux nuages : N_I, nuage des profils ligne munis des poids $f_{i.}$ $(1 \le i \le I)$, et N_J, nuage des profils colonne munis des poids $f_{.j}$ $(1 \le j \le J)$. On peut ainsi analyser le nuage N_I dans R^J qui

va nous intéresser. Les résultats obtenus se transposent alors sans difficulté à celle du nuage N_J dans R^I.

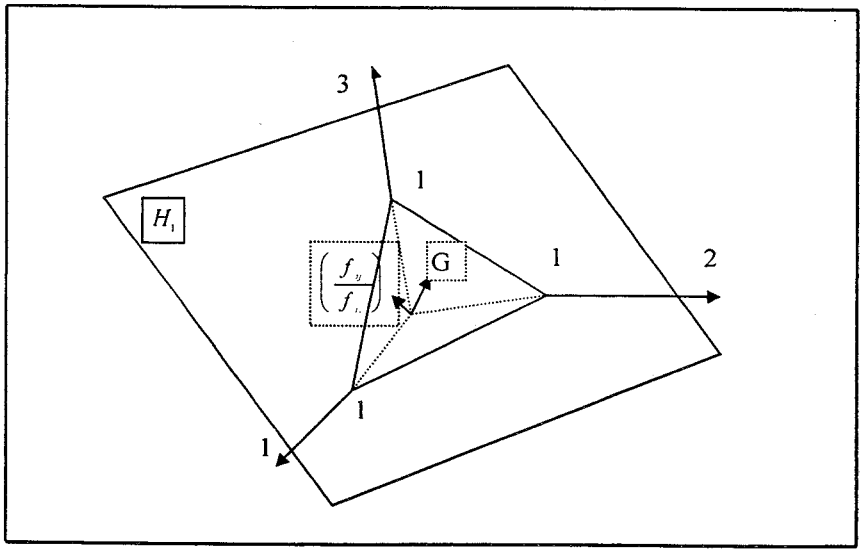
À présent, considérons donc le nuage N_I dans R^J. Soit G son centre de gravité : sa j ème composante s'écrit :

$$g_j = \sum_{i=1}^{n} f_{i.} \left(\frac{f_{ij}}{f_{i.}} \right)$$

Les $\left(\frac{f_{ij}}{f_{i.}} \right)_{i=1,\ldots,I}$ et G sont contenus dans une partie d'un *hyperplan afine*, noté H_1 de R^J tel que :

$$\forall \left(\frac{f_{ij}}{f_{i.}} \right) \in N_I, \quad \sum_{j}^{J} \frac{f_{ij}}{f_{i.}} = 1 \quad \text{et} \quad \sum_{j}^{J} g_j = 1.$$

Si $J = 3$ par exemple, on peut représenter cette propriété par la figure ci-dessous :



Par ailleurs, en définissant les matrices suivantes

$$\underset{(J,J)}{Q_I} = \begin{pmatrix} 1/f_{.1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/f_{.J} \end{pmatrix}$$: la métrique associée à une distance diagonale ;

$$\underset{(I,J)}{X} = \begin{pmatrix} f_{11}/f_{1.} & \cdots & f_{1J}/f_{I.} \\ \vdots & \ddots & \vdots \\ f_{I1}/f_{1.} & \cdots & f_{IJ}/f_{I.} \end{pmatrix}$$: la matrice de données ;

$$\underset{(I,I)}{D_I} = \begin{pmatrix} 1/f_{1.} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/f_{I.} \end{pmatrix}$$: la matrice diagonale des poids

on pose alors $\underset{(J,J)}{V} = (XQ_I)' D_I (XQ_I)$ comme étant la matrice d'inertie du nuage N_I par rapport à l'origine.

Contrairement à ce qui est fait en analyse en composantes principales, où les colonnes sont centrées, ici on va effectuer l'analyse d'une matrice non centrée. Ceci permettra d'utiliser les résultats de cette analyse pour effectuer l'analyse du nuage N_J, pratiquement sans calcul supplémentaire et pour obtenir une représentation simultanée de N_I et N_J. Pour la matrice d'inertie du nuage N_I par rapport à l'origine, V, on peut montrer que :

- les valeurs propres, notées $\lambda_1,\ldots,\lambda_J$, sont comprises entre 0 et 1. Par ordre décroissant, on a : $\lambda_1 = 1 \geq \lambda_2 \geq \cdots \lambda_J \geq 0$;

- G, vecteur centre de gravité, est le vecteur associé à la valeur propre $\lambda_1 = 1$;
- le deuxième axe principal Δu_2 correspond à λ_2, etc.
- soient $u_1 = G,\ u_2,\ldots,u_\alpha,\ldots u_q$, les q premiers vecteurs propres. Comme en ACP, on appelle facteurs les coordonnées des projections des I points du nuage N_I sur l'axe u. C'est donc dire qu'un facteur est un vecteur de dimension I. Le α ème facteur principal calculé dans R^J est noté :

$$F_\alpha = XQ_I u_\alpha$$

De façon analogue à l'espace R^J, on définit les choses suivantes dans l'autre espace :

a) la matrice associée à la forme quadratique

$$\underset{(I,I)}{Q_J} = \begin{pmatrix} 1/f_{1.} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/f_{I.} \end{pmatrix};$$

b) la matrice $\underset{(n,J)}{Y}$ contient en colonne les éléments de N_J ;

c) la matrice des poids $\underset{(J,J)}{D_J} = Q_I^{-1}$;

d) la matrice d'inertie par rapport à l'origine est :

$$\underset{(I,I)}{W} = (YQ_J)D_J(YQ_J)'$$

Les axes principaux ont alors comme vecteurs unitaires les vecteurs propres $v^2,\ldots,v^l,\ldots,v^q \in R^I$ associés aux valeurs propres $\mu_2 \geq \cdots \geq \mu_q$ (on exclut le centre de

gravité de N_J associé à la valeur propre 1). En notant G_α, le α ème facteur principal dans R^I, on montre alors les propriétés suivantes :

- les valeurs propres $\mu_1 = 1, \quad \mu_2, \ldots, \mu_q$ associées aux vecteurs propres de W sont égales respectivement aux valeurs propres $\lambda_1 = 1, \quad \lambda_2, \ldots, \lambda_q$ de V ;

- pour les λ_α positifs, on a les relations :

$$\forall\, \alpha \in \{2, \ldots, q\} \quad G_\alpha(j) = \frac{1}{\sqrt{\lambda_\alpha}} X F_\alpha(i) \quad \text{et} \quad F_\alpha(i) = \frac{1}{\lambda_\alpha} Y' G_\alpha(j)$$

Ces deux propriétés, qui expriment les résultats de l'analyse d'un nuage en fonction des résultats de l'analyse de l'autre nuage, conduisent à une économie de calcul. Mais surtout, elles donnent un sens à une représentation simultanée des lignes et des colonnes.

Les praticiens de l'AFC savent bien que les directions indiquées par les derniers vecteurs propres sont déterminées par les particularités de certains individus de la population. En conséquence, les indices d'aide à l'interprétation (qualité de représentation d'un élément par un axe ou un plan et contribution d'un élément à l'inertie d'un axe) définis en ACP sont applicables et valables pour un nuage quelconque. Notons cependant que si en ACP, en général, les poids de tous les éléments sont égaux, ce n'est pas le cas en AFC ; or ces poids interviennent dans la contribution d'un point à l'inertie d'un axe. Comme l'ACP, l'AFC utilise aussi presque systématiquement la technique des éléments supplémentaires. Cette technique consiste à projeter sur les axes factoriels des profils ligne et colonne qui n'interviennent pas dans le calcul de ces axes. Ces éléments supplémentaires servent très souvent, eux aussi, d'aides à l'interprétation.

En conclusion, bien que l'AFC soit applicable à une matrice traditionnelle de cas mesurés sur un certain nombre de variables, cette technique a été conçue surtout pour une matrice de fréquences. Elle a pour but de mettre en évidence les relations et rapprochements qui

peuvent exister entre les critères de classification des lignes et des colonnes. Ces lignes et colonnes jouent donc des rôles parfaitement symétriques. L'AFC met ainsi en places des droites Δ_α dans l'espace euclidien où est situé le nuage N_I. Ces droites passent par le centre de gravité du nuage N_I :

- Δ_1 est l'axe principal d'allongement du nuage ;
- Δ_2 est, parmi les droites perpendiculaires à Δ_1, celle sur laquelle le nuage se projette avec la plus grande dispersion (c'est-à-dire celle le long de laquelle l'inertie du nuage est la plus grande) ;
- Δ_3 est, parmi les droites perpendiculaires à Δ_1 et Δ_2, celle sur laquelle le nuage se projette avec la plus grande dispersion ; etc.

Le système des droites Δ_α, deux à deux perpendiculaires, est alors utilisé comme système d'axes de coordonnées. Ces coordonnées seront appelées les facteurs. La i ème cordonnée du α ème facteur principal calculé dans R^J est notée $F_\alpha(i)$. De cette façon, le principe de l'AFC consiste alors à mettre en évidence les rapprochements qui existent entre les composantes principales de classification des colonnes et des lignes. Ces rapprochements se font souvent dans un espace où les composantes les plus importantes sont illustrées par rapport aux axes orthogonaux Δ_α qui sont les mêmes pour les composantes lignes et colonnes. En réalité, on ne retient que les deux ou trois premières composantes qu'on projette dans le même espace des facteurs. Ainsi, l'interprétation à donner à une telle illustration est la suivante : les lignes qui donnent lieu à des projections rapprochées et voisines de celles d'une colonne signifient une ressemblance entre les deux modes de classification pour les classes en question. Enfin, l'AFC ne prétend donc à aucune généralisation et inférence statistique : elle n'est utilisée que dans un but descriptif.

L'exposé présenté ci-dessus n'a pas la prétention d'avoir décrit l'analyse factorielle des correspondances de façon exhaustive. Les informations complémentaires peuvent être trouvées dans les ouvrages de *Benzécri (1982)*, *Lebart et al. (1982)*, *Greenacre (1984)* et *Escofier et Pagès (1988).*

3.2. Généralités sur la méthode de régression sur les facteurs d'une analyse des correspondances

Soient y, $x_1, x_2, \ldots, x_k$, l'ensemble des variables. y, la variable dépendante, étant à expliquer en fonction des variables explicatives $x_1, x_2, \ldots, x_k$. Nous supposons par ailleurs que ces $k+1$ variables ont été rendues qualitatives par un découpage préalable en classes. Désignons par J_q l'ensemble des modalités de la q ème variable $(0 \le q \le k)$. On posera :

$K = \{0,1,2,\ldots,k\}$	ensemble des variables ;
$K_e = \{1,2,\ldots,k\}$	ensemble des variables explicatives ;
$J = \cup\{J_q / q \in K\}$	ensemble des modalités de toutes les variables ;
$J_e = \cup\{J_q / q \in K_e\}$	ensemble des modalités des variables explicatives ;
$I = \{1,2,\ldots,n\}$	ensemble des n individus (ou observations) pour lesquels on a mesuré les variables y, $x_1, x_2, \ldots, x_k$.

I désignant l'ensemble des n observations, on considère alors k_{IJ}, le tableau initial des données, qui est un tableau disjonctif complet (TDC) :

$$\forall i \in I, \quad \forall j \in J_q \subset J : \quad k_{IJ}(i,j) = \begin{cases} 1 & \text{si } i \text{ a adopté la modalité } j \text{ de } J_q \\ 0 & \text{sinon} \end{cases}$$

D'une façon générale, la méthode de régression sur les facteurs d'une analyse des correspondances revient à effectuer les étapes suivantes :

1) après la division en tranches des variables x et y, construire le sous tableau de BURT, $B_{J_e J_0}$, qui assemble le tableau de contingence croisant les modalités de toute variable x_i $(1 \le i \le k)$ avec les modalités de la variable y ;

2) effectuer l'AFC du tableau $B_{J_eJ_0}$. On désignera par $\left(\varphi_\alpha^{Kx}, \varphi_\alpha^{Ky}\right)$ le α ème couple de facteurs associés de cette analyse et par λ_α la valeur propre correspondante ;

3) rajouter le tableau disjonctif complet, k_{IJ_e}, associé aux modalités des variables $x_i (1 \le i \le k)$ en supplémentaire de $B_{J_eJ_0}$. Soit $F_\alpha(i)$ le α ème facteur pour individus supplémentaires ;

4) effectuer la régression usuelle de y (avant découpage en classes) sur les F_α ;

Remarquons tout d'abord que le sous tableau de BURT $B_{J_eJ_0}$ représente le tableau de régression ou encore de dépendance de la variable dépendante y avec les variables explicatives $x_1, x_2, \ldots, x_k$. Ainsi, c'est donc en effectuant l'AFC de ce tableau qu'on étudiera la liaison de la variable dépendante avec les k variables explicatives. Aussi, la mise en supplémentaire du tableau k_{IJ_e} permet de faire sur les facteurs associés aux n individus supplémentaires une régression usuelle. Par ailleurs, remarquons aussi qu'au lieu de faire la régression sur les facteurs F_α, on aurait pu faire la régression sur les facteurs issus de l'AFC du tableau disjonctif complet k_{IJ} (en mettant $B_{J_eJ_0}$ en supplémentaire pour visualiser les liaisons entre les x_i $(1 \le i \le k)$ et la variable y).

3.2.1. Lien théorique entre la méthode de régression traditionnelle et la régression sur les facteurs

Dans le présent paragraphe, nous allons étayer notre propos en considérant le modèle de régression classique suivant :

$$Y = \beta_0 1 + X\beta + \varepsilon \qquad (3.1)$$

où la matrice de données est : $X = \begin{bmatrix} x_{11},\ldots,x_{k1} \\ \\ x_{1n},\ldots,x_{kn} \end{bmatrix} = [x_1,\ldots,x_k]$

Comme nous l'avons déjà dit précédemment, l'AFC sur X permet de réexprimer X avec un petit nombre de nouvelles variables, Z, appelées facteurs, qui sont des combinaisons linéaires des Z. Ces variables Z capturent autant que possible la variation dans les X. Puisque les vecteurs propres v_k de la matrice $X'X$ sont orthogonaux et unitaires, nous avons :

$$V = [v_1,\ldots,v_k] \quad \Rightarrow \quad VV' = I \text{, matrice unitaire.}$$

En considérant $\lambda_1,\ldots,\lambda_q$ les valeurs propres associées, l'équation 3.1 peut encore s'écrire :

$$Y = \beta_0 1 + XVV'\beta + \varepsilon,$$

ou encore,

$$Y = \beta_0 1 + Z\alpha + \varepsilon \qquad (3.2)$$

où $Z = XV$ est une matrice de dimension $(n \times k)$ et $\alpha = V'\beta$ un vecteur de dimension $(k \times 1)$. Les colonnes de Z peuvent alors être vues comme étant des lectures sur k nouvelles variables, les facteurs (ce sont des combinaisons linéaires des x). Il est facile de voir que ces facteurs sont orthogonaux l'un à l'autre. Nous avons :

$$\begin{aligned} Z'Z &= (XV)'(XV) \\ &= V'X'XV \\ &= \text{diag}(\lambda_1,\ldots,\lambda_k) \end{aligned}$$

Ainsi, si on effectue la régression classique sur les nouvelles variables Z via l'équation 3.2, les variances des nouveaux coefficients sont telles que :

$$\frac{Var\left(\hat{\alpha}_j\right)}{\sigma^2} = \frac{1}{\lambda_j} \qquad \left(j = 1,\ldots,k\right).$$

On remarque donc que si la multicolinéarité est sévère, il y aura au moins une valeur propre proche de zéro. L'élimination du facteur associé à cette valeur propre, pourrait réduire substantiellement la variance totale dans le modèle et donc produire une amélioration appréciable de l'équation de prédiction.

Soit Z^* la matrice dénotant les r premiers facteurs de Z qui expliquent un pourcentage de variabilité acceptable dans l'analyse factorielle des correspondances. En ne retenant que les r premières colonnes de V, la matrice V^* s'exprime alors :

$$Z^* = XV^* \qquad (3.3)$$

En régressant y sur Z^*, on note les coefficients de régression par β^* : $\hat{y} = Z^*\hat{\beta}^*$. En substituant l'équation 3.3, dans cette dernière expression nous obtenons :

$$\hat{y} = XV^*\hat{\beta}^* \qquad (3.4)$$

Ainsi, la combinaison de l'expression $\hat{y} = X\hat{\beta}$ avec celle obtenue à l'équation 3.4, nous donne :

$$\hat{\beta} = V^*\hat{\beta}^* \qquad (3.5)$$

Cette dernière équation montre bien le lien existant entre les coefficients estimés de la régression classique et ceux obtenus à l'aide de la méthode de régression sur les facteurs d'une analyse des correspondances.

En résumé, lorsque la matrice $X'X$ est «presque» singulière, son inversion est délicate, et les variances des coefficients β_i sont très élevées. Ceci se produit si les variables $x_j \quad (j = 1,\ldots,k)$ sont voisines de la multicolinéarité. Nous pouvons considérablement simplifier les calculs en ne retenant que les vecteurs propres et valeurs propres relatifs aux facteurs qui peuvent être considérés comme suffisants pour la reconstitution approchée du tableau de données X. Par la même occasion, nous améliorons la stabilité de l'estimateur. Celui-ci devient donc moins sensible aux particularités de l'échantillon de variables et d'individus retenus pour constituer ce tableau de données.

4. MÉTHODE DE RÉGRESSION BASÉE SUR LES FACTEURS À PARTIR DES ÉLÉMENTS SUPPLÉMENTAIRES D'UNE ANALYSE DES CORRESPONDANCES

Dans la section 3.2, nous avons traité les généralités sur la méthode de régression sur les facteurs d'une analyse des correspondances. Nous avons noté qu'il était essentiellement important d'effectuer l'analyse factorielle des correspondances sur le tableau de BURT $B_{J_eJ_0}$, dans la mesure où cette façon de procéder nous permettait d'étudier la liaison de la variable dépendante y avec les variables explicatives $X_i(1 \le i \le K)$. Nous abordons maintenant le sujet principal en revenant sur cette excellente propriété et en mettant en supplémentaire le tableau disjonctif complet k_{IJ_0}. Cette mise en supplémentaire permet de faire sur les facteurs associés aux I individus supplémentaires une régression usuelle : c'est ce que l'on appelle la méthode de régression basée sur les facteurs à partir des éléments supplémentaires d'une analyse des correspondances. Cette méthode est une variante de la méthode de régression sur les facteurs d'une analyse des correspondances. C'est elle que nous allons examiner et appliquer dans la suite de ce rapport.

4.1. Description de la méthode

La démarche de la méthode de régression basée sur les facteurs à partir des éléments supplémentaires d'une analyse des correspondances correspond aux différentes étapes suivantes :

Étape_1 : Après la division en classes des variables x et y, construisez le tableau disjonctif complet, k_{IJ} (tableau 4.1).

Étape_2 : Construisez le tableau de BURT, B_{JJ}, associé à k_{IJ}, obtenu en croisant l'ensemble des modalités de toutes les variables avec lui-même (tableau 4.2).

Étape_3 : Obtenez le tableau $B_{J_e J_0}$, qui est le sous tableau de $BURT(B_{JJ})$, croisant J_0 ensemble des modalités de la variable à expliquer avec J_e ensemble de toutes les modalités des variables explicatives.

Étape_4 : Faites une AFC sur le tableau $B_{J_e J_0}$ et mettez en observations supplémentaires le tableau k_{IJ_0} (tableau 4.3).

Étape_5 : Faites la régression multiple pour trouver la relation suivante :

$$y_i = \alpha_1 F_1(i) + \alpha_2 F_2(i) + \ldots + \alpha_{k_0} F_{k_0}(i) + \varepsilon_i \qquad (3.6)$$

où, la variable y est obtenue de votre tableau initiale de données et k_0 désigne l'ensemble des facteurs considérés dans le modèle.

Étape_6 : Discutez et interprétez les résultats de votre régression.

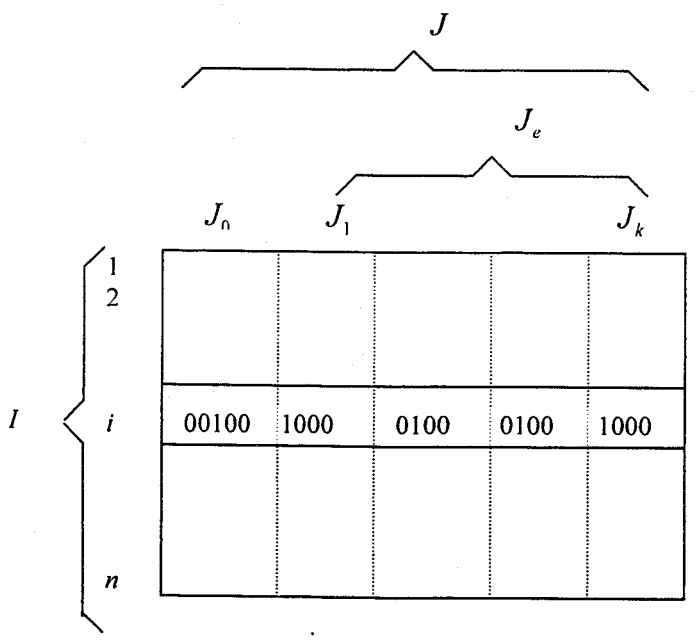


Tableau 4.1 : Illustration d'un tableau disjonctif complet

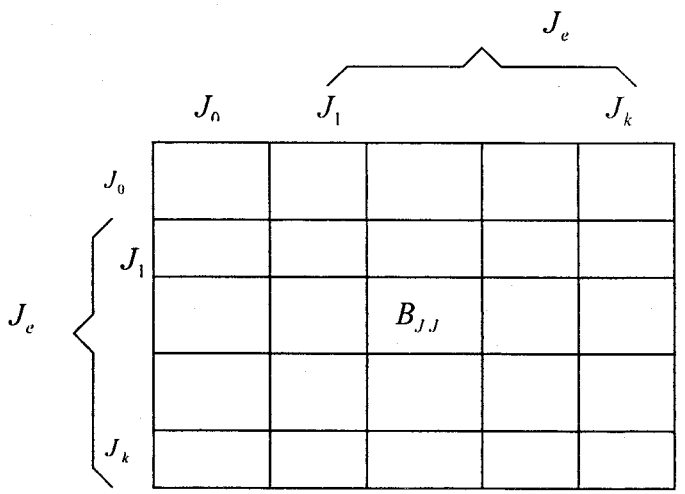


Tableau 4.2 : Illustration d'un tableau de BURT

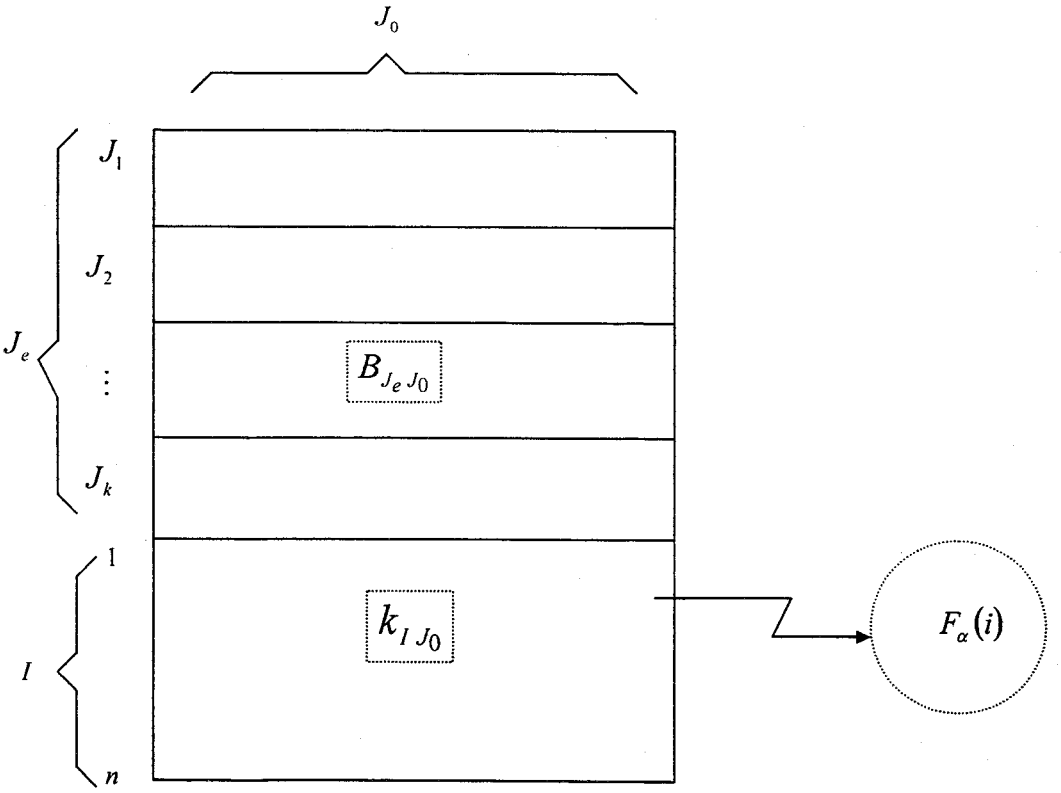


Tableau 4.3 : Illustration du tableau final s'apprêtant à l'AFC

4.2. Avantages de la méthode

Rappelons pour commencer que l'analyse factorielle des correspondances sur le sous tableau $B_{J_e J_0}$ conserve les propriétés du tableau de BURT. En effet, on note que :

- sur un axe factoriel, le centre de gravité des j points décrivant J_q est identique au centre de gravité de l'ensemble des points du nuage ;
- le centre de gravité des j' points décrivant $J_{q'}$ est identique à l'origine.

L'avantage de découper les variables en classes puis de faire la régression sur les facteurs issus de l'AFC du tableau $B_{J_e J_0}$ avec $k_{I J_0}$ en supplémentaire réside dans les points suivants :

- la liaison de la variable réponse, y, avec les variables explicatives, $x_1, x_2, \ldots, x_k$, est mieux visualisée par la représentation simultanée de ces deux ensembles de variables fournie par l'AFC du tableau $B_{J_e J_0}$ avec $k_{I J_0}$ en supplémentaire ;
- la méthodologie précédente se transpose facilement au cas où nous ayons à faire la régression sur les variables explicatives qualitatives ou le cas d'un mélange de variables qualitatives et quantitatives ;
- on peut aussi adapter cette méthodologie dans la situation où la variable à expliquer est qualitative et dans la situation où nous avons plusieurs variables à expliquer ;
- comme nous l'avons déjà mentionné, la méthodologie précédente permet aussi de s'affranchir du problème de la multicolinéarité dans une régression multiple afin d'éliminer le bruit dû aux fluctuations d'échantillonnage ;

- le nombre de facteurs à conserver dans le modèle de régression spécifié à l'équation 3.6 peut facilement être déduit à partir du critère du R^2 maximal. Les facteurs étant orthogonaux, la valeur de R^2 atteint rapidement un palier avec l'augmentation du nombre de facteurs .

En conséquence, comparativement à la régression usuelle, la méthode de régression sur les facteurs à partir des éléments supplémentaires d'une analyse des correspondances présente plusieurs avantages. De façon concrète, elle permet :

- de s'affranchir du problème de la multicolinéarité, permettant par la même occasion l'amélioration de la stabilité de l'estimateur des coefficients de régression qui devient moins sensible ;

- de faire la régression sur les variables qualitatives ;

- de choisir plus facilement un sous ensemble de variables explicatives à l'aide du critère R^2 maximal.

4.3. Limites et portée de la méthode

Dans notre propos, nous avons indiqué que l'avantage d'utiliser la méthode de régression basée sur les facteurs à partir des éléments supplémentaires d'une analyse des correspondances est d'optimiser dans un certain sens la régression lorsque les variables explicatives sont liées entre elles. Par contre, certaines limites et la portée de cette méthode sont à prendre en considération. Parmi celles-ci, nous dénombrons ce qui suit :

- lorsque les variables explicatives du modèle de régression exprimé à l'équation 2.1 ne présentent pas des phénomènes de multicolinéarité, l'estimation par les moindres carrés ordinaires est préférable à l'estimation obtenue à l'aide de la méthode de régression basée sur les facteurs à partir des éléments

supplémentaires d'une analyse des correspondances. En effet, la méthode des MCO demande un temps de calcul moins fastidieux ;

- malgré le mérite que l'on peut accorder à la méthode de régression basée sur les facteurs à partir des éléments supplémentaires d'une analyse des correspondances, il faut tout de même souligner que les résultats de la régression sont biaisés puisqu'on prend pour variables explicatives, des facteurs qui sont des variables dépendant des liaisons entre y et les x_j. À ce sujet, *Cazes (1997)* suggère d'utiliser un échantillon test pour valider les résultats obtenus. Pour illustrer ce propos, supposons que nous considérons la procédure des facteurs d'une AFC avec r facteurs à éliminer (ceux ayant une valeur propre proche de zéro) et s facteurs retenus $(s+r=k)$. De plus, considérons la matrice V des vecteurs propres normalisés de $X'X$ partitionnée comme suit :

$$V = \left[V_r \vdots V_s \right].$$

Et, similairement, considérons la matrice diagonale des valeurs propres de $X'X$, Λ, comme :

$$\Lambda = \begin{bmatrix} \Lambda_r & 0 \\ 0 & \Lambda_s \end{bmatrix},$$

où, Λ_r et Λ_s sont des matrices diagonales, avec Λ_r contenant les valeurs propres associées aux facteurs éliminés. Ainsi, nous avons :

$$\begin{aligned} Z'Z &= (XV)'(XV) \\ &= V'X'XV \\ &= \text{diag}(\lambda_1,\ldots,\lambda_k) \end{aligned}$$

À partir du modèle de régression exprimé à l'équation 3.2, nous obtenons,

$$\hat{\alpha} = (Z'Z)^{-1} X'Y = \Lambda^{-1} V'X'Y \quad \Rightarrow \quad \hat{\alpha}_s = \Lambda_s^{-1} V_s' X'Y$$

On peut montrer facilement que $\hat{\alpha}_s$ est un estimateur sans biais de α_s.

Soit $\hat{\beta}_F$ l'estimateur de β obtenu par les MCO. À partir de l'équation 3.5, nous avons :

$$\hat{\beta}_F = \begin{bmatrix} \hat{\beta}_{1,F} \\ \vdots \\ \hat{\beta}_{k,F} \end{bmatrix} = \begin{bmatrix} v_1 & v_2 \cdots v_{k-r} \end{bmatrix} \begin{bmatrix} \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_{k-r} \end{bmatrix} = V_s \hat{\alpha}_s$$

(en effet, l'élimination des facteurs n'implique pas nécessairement l'élimination des *k* variables explicatives).

Donc,

$$E\left(\hat{\beta}_F\right) = V_s \alpha_s = V_s V_s' \beta$$

or, $VV' = I = V_r V_r' + V_s V_s'$, d'où,

$$\begin{aligned} E\left(\hat{\beta}_F\right) &= \left[I - V_r V_r' \beta \right] \beta \\ &= \beta - V_r V_r' \beta \\ &= \beta - V_r \alpha_r \end{aligned}$$

En somme, nous voyons que les estimateurs des *k* coefficients de régression sont biaisés par la quantité $V_r \alpha_r$, avec α_r étant le vecteur des facteurs qui ont été enlevés ;

- dans la méthode de régression basée sur les facteurs à partir des éléments supplémentaires d'une analyse des correspondances, certains facteurs de faible variance peuvent être éliminés comme on vient de le voir. Or, ces facteurs peuvent être corrélés de façon non négligeable à la variable à expliquer, ici y. Ce qui constitue aussi une certaine limitation de cette méthode.

En résumé, cette méthode ne permet pas de discriminer le meilleur ensemble de variables explicatives retenu. Elle est intéressante, car elle permet d'améliorer la prédiction avec les variables dont on dispose.

4.4. Revue de littérature des développements et travaux portant sur cette méthode de 1980 jusqu'à aujourd'hui

La méthode de régression sur les facteurs d'une analyse des correspondances représente une autre technique de régression spécialement mise au point pour atténuer les effets de la multicolinéarité. Son principe est basé sur celui de la régression en fonction des composantes principales d'une ACP : au lieu d'utiliser les composantes principales d'une ACP comme nouvelles variables explicatives, on utilise plutôt les facteurs d'une analyse des correspondances. Cette méthode de régression en fonction des composantes principales a été originellement proposée *par Kendall (1957)* et elle a été depuis considérée par *Massy (1965)*, *Daling et Tamura (1970)*, *Basilevsky (1968)* et d'autres. Des exemples d'application sont donnés par *Jolliffe (1982)* et *Shahar et Gonzalo (1994)*. Aussi, *Coxe (1982, 1984)*, *Jolliffe (1986)* et *Jackson (1991)* élaborent des discussions intéressantes sur cette méthode. Les références citées précédemment peuvent toujours contribuer à éclairer le lecteur dans la compréhension de la méthode de régression sur les facteurs d'une analyse des correspondances. Dans la période antérieure à 1980, *Cazes* a apporté une grande contribution sur la méthode de régression basée sur les facteurs à partir des éléments supplémentaires d'une analyse des correspondances. Dans une publication récente *(Cazes, 1997)*, il a adapté la régression PLS à la régression après analyse des correspondances multiples. Ce procédé permet d'obtenir des composantes principales non corrélées sur

lesquelles on peut effectuer la régression. D'autres auteurs ont aussi publié sur la méthode de régression sur les facteurs d'une analyse des correspondances. À ce sujet, on peut noter l'article de *Shahid (1982)* qui présente un exemple d'application de cette méthode dans l'estimateur des paléoclimats d'après l'écologie des foraminifères. Pour la référence théorique, outre les recommandations sur la méthode de régression en fonction des composantes principales, le lecteur peut aussi se référer à *Basilevsky (1981)*.

5. EXEMPLE D'APPLICATION

5.1. Présentation des données et position du problème

5.1.1. Les données

Les données analysées dans ce chapitre ont été tirées à partir de la base de données météorologiques de la province de l'Ontario. Elles concernent un ensemble I de 106 stations météorologiques (les individus) caractérisées par 10 variables :

AIRE :	la superficie du bassin versant
LCP :	la longueur du cours d'eau principal
PCP :	la pente du cours d'eau principal
SLM :	la superficie du bassin versant contrôlée par les lacs et les marais
LAT :	la latitude
LONG :	la longitude
DTY5 :	la médiane de la moyenne des températures moyennes journalières, 5 jours précédant la crue
YPR5 :	moyenne de la précipitation totale, 5 jours précédant la crue
YNS5 :	moyenne de la neige au sol, 5 jours précédant la crue
PTMA :	précipitation totale moyenne annuelle.

De cet ensemble, nous avons les catégories de variables suivantes :

- **les variables physiographiques** :
  1. la superficie du bassin versant (AIRE) ;
  2. la longueur du cours d'eau principal (LCP) ;
  3. la pente du cours d'eau principal (PCP) ;
  4. la superficie du bassin versant contrôlée par les lacs et les marais (SLM) ;
  5. la latitude (LAT) ;
  6. la longitude (LONG) ;

- **les variables hydrologiques** :
  1. la médiane de la moyenne des températures moyennes journalières, 5 jours précédant la crue (DTY5) ;
  2. la moyenne de la précipitation totale, 5 jours précédant la crue (YPR5) ;
  3. la moyenne de la neige au sol, 5 jours précédant la crue (YNS5) ;
  4. la précipitation totale moyenne annuelle (PTMA) .

Chacune de ces variables hydrologiques peut logiquement jouer le rôle de la variable dépendante qui sera expliquée par les variables physiographiques (variables explicatives).

5.1.2. Choix de la variable réponse et position du problème

Pour choisir la variable réponse appropriée, nous allons analyser les relations linéaires qui existent entre les variables. Le tableau A1 de l'annexe A montre que :

- les quatre variables dépendantes probables (variables hydrologiques) sont corrélées significativement ;

- les variables explicatives (variables physiographiques) sont plus corrélées positivement avec la variable dépendante DTY5 et négativement avec la variable PTMA. On mentionne cependant que seule la variable explicative LCP est faiblement reliée aux quatre variables hydrologiques.

Les graphiques illustrant les relations entre les variables prises deux à deux (figure A1 de l'annexe A) fournissent ici la même information que les coefficients de corrélations précédents. Il n'est cependant pas inutile de faire remarquer que seul l'histogramme de la variable dépendante YPR5 a une forme proche de celle d'une loi normale. Ce fait combiné aux observations précédentes nous amène à retenir la variable YPR5 comme étant la variable dépendante à expliquer dans la suite de cette étude. Toutefois, le fait de ne pas avoir choisi une autre variable hydrologique comme variable dépendante à être expliquée ne saurait compromettre notre propos. En effet, hormis les critères en faveur du choix de la variable YPR5, nous avons observé que les quatre variables hydrologiques sont corrélées significativement et, en conséquence, elles sont redondantes : l'information apportée par l'une est déjà comprise dans l'autre. La rétention de la variable YPR5 est surtout bénéfique parce qu'elle a des propriétés prérequises à l'application d'un modèle de régression linéaire : son histogramme ressemble beaucoup à celui d'une loi normale. La variable à expliquer dans cette étude est donc la moyenne de la précipitation totale, 5 jours précédant la crue (YPR5). Les six variables explicatives sont les variables physiographiques (AIRE, LCP, PCP, SLM, LAT, LONG) mesurées sur les 106 stations météorologiques.

5.2. Conception d'un modèle de régression multiple

Dans cette section, nous voulons construire un modèle de régression inspiré des règles de *Neter et al. (1985)*. La variable dépendante à expliquer est YPR5, les variables explicatives sont : AIRE, LCP, PCP, SLM, LAT, et LONG.

5.2.1.Analyse exploratoire des données

Avant de modéliser la relation existant entre la variable dépendante et les six variables explicatives, il est souhaitable de faire une analyse exploratoire des données de manière à nous familiariser avec celles-ci. Cette analyse devrait nous permettre de découvrir les relations existant entre les variables à l'étude et d'anticiper les difficultés qui pourraient survenir dans l'ajustement du modèle de régression (exemple : données extrêmes, regroupements d'observations, relations non linéaires etc.). Examinons d'abord la relation existant entre la variable dépendante et les six variables explicatives, à l'aide des

graphiques et de la matrice de corrélations. La figure A2 de l'annexe A nous permet d'observer ce qui suit :

- les graphiques montrent que la variable dépendante est fortement liée avec les variables explicatives LAT et LONG ;
- la relation entre la variable dépendante (YPR5) et les autres variables explicatives (SLM, PCP, LCP, AIRE) n'est pas apparente ;
- trois stations ont une superficie (AIRE) élevée sans que leur valeur prise par la variable dépendante (YPR5) le soit ;
- une station a une valeur de PCP élevée sans que sa valeur sur la variable dépendante (YPR5) le soit ;
- deux stations ont une valeur de SLM élevée sans que leur valeur sur la variable dépendante (YPR5) le soit.

On peut aussi évaluer le degré de liaison entre deux variables à l'aide du coefficient de corrélation linéaire. Ce faisant, l'interprétation de la matrice de corrélations (tableau A2 de l'annexe A) nous permet de faire les observations suivantes :

- les coefficients de corrélation fournissent une information similaire à celle obtenue précédemment ;
- la corrélation la plus élevée relie les variables explicatives SLM et AIRE ($r = 0.96$) ;
- la variable dépendante YPR5 est plus corrélée avec la variable latitude (LAT) ($r = -0.65$) ;

À première vue, la variable dépendante (YPR5) semble être reliée fortement à la variable LAT et faiblement aux variables AIRE, PCP, SLM, et LONG. Or ces cinq variables explicatives sont reliées entre elles. On peut donc se poser les questions suivantes :

- *Une fois que l'on a tenu compte de la latitude (LAT) pour expliquer les différences observées dans la moyenne de la précipitation totale, 5 jours précédant la crue (YPR5), qui est la variable dépendante, demeure-t-il pertinent de considérer aussi les autres variables explicatives, à savoir, la superficie du bassin versant (AIRE), la pente du cours d'eau principal (PCP), la superficie du bassin versant contrôlée par les lacs et les marais (SLM) et la longitude (LONG) ?*

- *D'autre part, la longueur du cours d'eau principal (LCP), qui n'est pas significativement reliée à la variable dépendante YPR5, permet-elle d'expliquer une part des différences observées, une fois les cinq autres variables explicatives considérée ?*

Les graphiques illustrant les relations entre les variables prises deux à deux ne permettent pas de répondre à ces questions. La modélisation de la relation globale entre la variable dépendante YPR5 et les variables explicatives peut toutefois permettre de mieux discerner comment ces variables physiographiques, considérées dans leur ensemble, permettent d'expliquer les variations observées de la variable YPR5. De plus, nous prendrons aussi en compte l'influence des points extrêmes observés dans les graphiques plus loin dans l'étude.

5.2.2. Ajustement d'un premier modèle et analyse sommaire des résidus

5.2.2.1. Modèle de régression

Ajustons d'abord un modèle de régression multiple incluant les six variables explicatives :

$$\text{YPR5} = \beta_0 + \beta_1 \text{AIRE} + \beta_2 \text{LCP} + \beta_3 \text{PCP} + \beta_4 \text{SLM} + \beta_5 \text{LAT} + \beta_6 \text{LONG} + \varepsilon$$

Les résultats obtenus avec ce modèle de régression sont présentés au tableau A3 de l'annexe A. Nous observons que l'estimé du paramètre β_6 est positif, contrairement à ce qui était prévisible. En effet, on avait remarqué au tableau A2 et à la figure A2 de l'annexe A que la variable dépendante YPR5 et la variable explicative LONG étaient reliées négativement. Au seuil de signification 5%, seuls les coefficients $\beta_0, \beta_2, \beta_5$ et β_6 sont significatifs. Avant d'aller plus loin dans l'interprétation de ces résultats, attardons-nous d'abord à analyser les résidus de ce premier modèle.

5.2.2.2. Analyse des résidus

Notons que nous ne cherchons pas ici à justifier rigoureusement que les résidus de notre modèle de régression sont indépendants et équidistribuées : les décisions seront prises selon des critères qualitatifs et non quantitatifs. C'est pourquoi nous nous bornons à la présentation de diagnostics fondés sur des graphiques, et non sur des tests dont nous ne serions pas capables d'évaluer les performances. Même alors, la circonspection est de rigueur : c'est ce que l'on constate en étudiant de façon un peu fine les résidus, comme cela est indiqué dans l'ouvrage de *Draper et Smith (1981).* Ainsi, à partir des graphiques des résidus présentés aux figures A3 et A4 de l'annexe A, nous pouvons tirer les renseignements suivants :

- sur la figure A3, les résidus sont bien repartis de part et d'autres de la droite de Henry. Le graphique de la figure A4 montre une bonne répartition des résidus. Ainsi, à l'évidence, les hypothèses d'indépendance, de normalité et d'homogénéité de la variance des résidus semblent satisfaisantes ;

- l'observation des figures A3 et A4 révèle aussi que deux stations présentent de grands résidus (en valeur absolue).

5.2.3. La multicolinéarité

La détection de la présence de multicolinéarité est facilitée par le calcul des "**tolérance**" où **TOL**. Nous rappelons que les VIF (ou "**variance inflation factors**") sont des facteurs qui

indiquent de combien de fois la variance de chacun des paramètres est gonflée par la présence de multicolinéarité. Pour la k ème variable explicative, le VIF se calcule comme suit :

$$\mathrm{VIF_k} = \frac{1}{1 - R_k^2},$$

où R_k^2 est le coefficient de détermination pour la régression de la k ème variable explicative sur les autres variables. Le dénominateur du VIF est égal à la "**tolérance**" ou TOL et prend une valeur entre 0 et 1, une petite valeur étant signe de multicolinéarité.

Pour notre modèle de régression, les valeurs des "**tolérance**" ou TOL sont présentées dans le tableau A4 de l'annexe A. À la lumière des résultats obtenus, l'examen des tolérances indique la présence de multicolinéarité. En effet, les variables explicatives AIRE (0.032), PCP (0.22), SLM (0.048), LAT (0.28) et LONG (0.39) ont de petites valeurs de tolérance (TOL). Ces résultats révèlent aussi que, pour chacune d'elles, la valeur du VIF correspondante est très élevée. Nous sommes donc en droit de soupçonner la présence de multicolinéarité dans cet ensemble de variables explicatives retenues. Cette présence de multicolinéarité est probablement causée par la forte corrélation observée entre ces cinq variables explicatives (tableau A2 de l'annexe A). Nous sommes donc incités à la prudence au sujet de l'interprétation des paramètres estimés du modèle de régression retenu.

5.2.4. Ajustement final du modèle

5.2.4.1. Sélection du modèle

Dans cette partie de l'étude, il est nécessaire d'effectuer un dernier ajustement de manière à ne conserver dans le modèle que des variables explicatives importantes. À cet effet, puisque nous avons observé au tableau A3 de l'annexe A que les paramètres associés aux variables AIRE, PCP, et SLM ne sont pas significativement non nuls au seuil de signification 5%, ces variables explicatives sont retranchées du modèle pour l'ajustement final. Notre modèle de régression est donc le suivant :

$$YPR5 = \beta_0 + \beta_1 LCP + \beta_2 LAT + \beta_3 LONG + \varepsilon$$

Le tableau A5 de l'annexe A présente les résultats obtenus à l'aide de ce modèle de régression. Il est bien évident qu'au seuil de signification 5%, la relation entre la moyenne de la précipitation totale, 5 jours précédant la crue (YPR5) et les variables explicatives (la longueur du cours d'eau principal (LCP), la latitude (LAT) et la longitude (LONG)) peut être approximée par l'équation de régression

$$YPR5 = 74.26 - 0.94\ LCP - 2.11\ LAT + 0.56\ LONG$$

Pour revenir à notre préoccupation, il n'est pas inutile de rappeler que lorsqu'un grand nombre de variables explicatives sont en cause et que l'ajustement du modèle complet entraîne des problèmes de multicolinéarité, ce qui est le cas ici, il peut être utile d'avoir recours à un processus de sélection automatique du modèle. Les processus de sélection automatiques sont fonction de divers critères d'optimisation. L'approche que nous avons retenue dans cette étude est celle de la méthode *STEPWISE*.

La méthode STEPWISE encore appelée méthode de régression d'inclusion d'exclusion «pas à pas» permet de construire un modèle en incluant dans celui-ci une seule variable à la fois. La première variable à entrer dans le modèle est celle étant corrélée la plus fortement avec la variable dépendante. Elle est retenue dans le modèle seulement si le paramètre correspondant est significatif au seuil d'entrée choisie. De même, la seconde variable à entrer dans le modèle est celle étant la plus fortement reliée avec la variable dépendante. Une fois que celle-ci a été ajustée pour l'effet de la première variable incluse dans le modèle. Autrement dit, la seconde variable choisie est celle dont le seuil observé pour le test sur son paramètre est le plus petit, tout en étant sous le seuil d'entrée. Cette procédure se poursuit jusqu'à ce que toutes les variables non incluses au modèle aient des paramètres non significatifs.

Le tableau A6 de l'annexe A présente les résultats obtenus en utilisant les seuils d'entrée et de sortie par défaut du logiciel STATISTICA. Il appert que les résultats fournis par la méthode STEPWISE concordent avec le modèle choisi précédemment. La première variable sélectionnée est LAT avec un $R^2 = 0.42$, ensuite on a les variables LONG et LCP qui ont été choisies dans l'ordre.

5.2.4.2.Validation du modèle de régression retenu

5.2.4.2.1. Détection d'observations influentes

Les statistiques d'influence sont des outils permettant d'identifier les données qui ont eu une influence importante sur l'estimation des paramètres du modèle. En présence de telles données, le modèle ajusté ne reflète pas la tendance de l'ensemble des observations, mais plutôt l'influence de quelques observations isolées. De nombreuses statistiques peuvent être calculées pour quantifier l'influence des observations. Dans cette étude, nous avons retenu deux statistiques d'influence parmi les plus répandues : le *résidu* (*Studentisé*) et la *distance de Cook*. En effet, l'analyse des résidus standards aura tendance à sous estimer l'influence des observations dont les valeurs prises par les variables explicatives diffèrent beaucoup de valeurs moyennes. Comme ces observations sont éloignées de l'ensemble quant aux valeurs des variables explicatives, elles peuvent avoir un effet de levier important sur l'estimation des paramètres. Or, les résidus Studentisés permettent de corriger cette situation d'où leur sollicitation dans l'étude des observations influentes. Toutefois, il convient de mentionner que le résidu Studentisé peut comme le résidu standard, être petit même pour les observations très influentes. Ce qui nous autorise à étudier d'autres statistiques d'influence comme la distance de Cook. Cette distance mesure l'écart entre les valeurs prédites pour le modèle ajusté à l'ensemble des observations et les valeurs prédites pour le modèle ajusté sans la i ème observation. Une valeur élevée de cette distance, signifie que l'observation en question a exercé une influence importante sur l'ajustement du modèle.

Les résultats obtenus avec le logiciel STATISTICA sont rapportés au tableau A7 de l'annexe A. Nous avons l'interprétation suivante :

- les stations S31, S35, S45, S60, S98 et S105 ont donné lieu à de grands résidus Studentisés ;

- les stations S98 , S52 et S45 ont, dans l'ordre, une distance de Cook élevée.

À la lumière de ces résultats, les stations S98 et S45 qui présentent chacune un grand résidu Studentisé et une grande distance de Cook, ont peut-être eu une influence importante sur l'estimation des paramètres du modèle de régression retenu. Afin de vérifier si nous avons raison de craindre cette influence sur l'ajustement final de notre modèle de régression, nous avons refait la régression en omettant ces deux stations dans les données. Les résultats obtenus sont présentés au tableau A8 de l'annexe A. En comparaison avec les résultats obtenus au tableau A5, nous remarquons que le retrait des stations S98 et S45 n'a à peu près pas modifié l'estimation des paramètres du modèle de régression retenu. Nous nous accordons en conséquence, jusqu'à ce point, que ce modèle reste acceptable. Mais, avant de conclure, examinons les hypothèses qui gouvernent le modèle de régression retenu.

5.2.4.2.2. Validation des hypothèses du modèle

Nous cherchons à vérifier les hypothèses standards du modèle de régression retenu, à savoir, l'homogénéité de la variance et la normalité des résidus. Bien entendu, nous supposons que l'échantillonnage des données de cette étude a été réalisé de manière à satisfaire l'hypothèse d'indépendance des observations. Dans l'étape d'analyse des résidus (paragraphe 5.2.2.2) nous avons vu, dans la figure A4 de l'annexe A, que les résidus sont assez homogènes autour de la valeur zéro. De plus, aucun pattern particulier ne semble se dessiner dans cette figure. Nous admettons donc que l'hypothèse d'homogénéité est satisfaite. Dans la figure A3 de l'annexe A, les observations se répartissent de part et d'autres de la droite d'Henry. L'hypothèse de normalité des résidus est aussi satisfaite. En

résumé, les hypothèses standards qui gouvernent notre modèle de régression semblent visuellement confirmées.

5.2.5. Conclusion et interprétation des résultats du modèle final

Le choix dirigé du modèle et la méthode de sélection automatique utilisée (STEPWISE) ont fourni le même modèle final à trois variables explicatives : LAT, LONG et LCP. La relation entre la moyenne de la précipitation totale, 5 jours précédant la crue (YPR5) et les variables explicatives (la longueur du cours d'eau principal (LCP), la latitude (LAT) et la longitude (LONG)) peut ainsi être exprimée par l'équation de régression

$$YPR5 = 74.26 - 0.94\ LCP - 2.11\ LAT + 0.56\ LONG$$

C'est donc dire qu'une fois la latitude (LAT) et la longitude (LONG) des stations météorologiques considérées, la longueur du cours d'eau principal (LCP) est la variable permettant d'étudier l'évolution de la moyenne de la précipitation totale, 5 jours précédant la crue (YPR5). Notons aussi que 56% de la variabilité de la variable dépendante YPR5 est expliquée par ces trois variables explicatives. Une part importante de la variabilité demeure donc inexpliquée ; d'autres variables explicatives seraient nécessaires pour mieux expliquer les variations de la moyenne de la précipitation totale, 5 jours précédant la crue (YPR5). Les différentes étapes de validation n'ont pas permis de remettre en cause les hypothèses sous-jacentes au modèle. D'autre part, nous avons observé, dans ce modèle de régression, que la multicolinéarité était présente dans l'ensemble des variables explicatives retenues. En conséquence, nous devons user d'un luxe de précaution dans l'interprétation des paramètres estimés de ce modèle de régression. En effet, comme nous l'avons déjà mentionné dans ce rapport, la multicolinéarité survient lorsqu'un nombre de variables explicatives fournissent une information similaire. Sa présence peut entraîner beaucoup de problèmes parmi lesquels nous rappelons :

- le fait d'ajouter ou de retrancher une variable explicative du modèle peut modifier de façon importante l'estimation des paramètres ;

- l'estimation d'un paramètre peut être de signe opposé à celui prévu ;

- la variance des estimations des paramètres a tendance à être gonflée, l'information fournie sur ces paramètres est donc peu précise.

5.3. Construction d'un modèle de régression sur les facteurs d'une AFC

Dans cette section, nous cherchons à surmonter le problème de multicolinéarité rencontré dans la première partie de l'exemple d'application. Pour cela, nous allons élaborer un modèle de régression sur les facteurs d'une analyse factorielle des correspondances.

5.3.1. Les données et les analyses factorielles effectuées

5.3.1.1. Obtention du sous tableau de BURT $B_{J_e J_0}$

Nous rappelons que nos données concernent un ensemble I de 106 stations météorologiques caractérisées par un ensemble de 7 variables quantitatives :

- six variables physiographiques : la superficie du bassin versant (AIRE), la longueur du cours d'eau principal (LCP), la pente du cours d'eau principal (PCP), la superficie du bassin versant contrôlée par les lacs et les marais (SLM), la latitude (LAT) et la longitude (LONG) ;

- une variable hydrologique : la moyenne de la précipitation totale, 5 jours précédant la crue (YPR5).

Il est bien entendu que cet ensemble de variables est le résultat des conclusions obtenues dans la première partie de cet exemple d'application. Aussi, les six variables physiographiques sont les variables indépendantes (variables explicatives), la variable

hydrologique est quant à elle la variable dépendante. YPR5 est donc la variable que l'on désire expliquer en fonction des variables physiographiques.

Dans une première étape, nous avons découpé en classes toutes les variables à l'étude (voir tableau B1 de l'annexe B) : les variables explicatives sont codées 0 ou 1 selon trois modalités et la variable à expliquer est codée 0 ou 1 en huit modalités ; soit, 26 variables logiques, 8 caractérisant YPR5 et notées YP1, YP2, YP4, ... , YP8 et 18 caractérisant les variables explicatives. Puis, nous avons obtenu un tableau disjonctif complet (voir tableau B2 de l'annexe B) $106 \times 26 : k_{IJ}$, où $J = J_e \cup J_0$, avec : J_e, ensemble des modalités des variables physiographiques $\left(\text{Card } J_e = 18\right)$ et J_0, ensemble des modalités de la variable YPR5 $\left(\text{Card } J_0 = 8\right)$. Et, finalement, nous obtenons le sous tableau de BURT, $B_{J_e J_0}$, croisant les variables J_e en ligne et J_0 en colonne. Il est présenté au tableau B3 de l'annexe B.

5.3.1.2. Analyse factorielle des correspondances sur le sous tableau de BURT $B_{J_e J_0}$

Dans la présente analyse, on a été amené à mettre en observations supplémentaires, le tableau disjonctif complet $k_{I J_0}$. Nous possédons donc pour les points ligne 124 individus et 8 individus pour les points colonne. Les résultats de l'AFC sont consignés au tableau B4 de l'annexe B. Le premier facteur (on ne tient pas compte de la première valeur propre) rend compte à lui tout seul de 70% de l'inertie totale, tandis que le deuxième facteur correspond à 18% de cette inertie. Le plan 1-2 résume donc 88% de l'inertie totale. Ainsi, les deux premiers facteurs restituent l'information contenue dans les données à 88 %. Comme ce pourcentage d'inertie est acceptable, nous retenons donc les deux premiers axes factoriels pour analyser les résultats de l'AFC. Toutefois, nous nous réservons le droit d'utiliser les autres axes factoriels si nous le jugeons nécessaire.

En procédant notre analyse dans le nuage des points ligne (espace R^8), nous pouvons observer ce qui suit

- ***Qualité de la représentation*** *(QLT)* : pour les trois premiers axes factoriels, tous les individus (les modalités des variables explicatives) sont bien représentés. Leur QLT est proche de 1.

- ***Poids*** : tous les points ligne ont un poids comparable. Nous mentionnons cependant que la modalité faible de la variable LONG (LO1) a le plus grand poids (0.0074) .

- ***L'inertie (INR)*** : les points ligne ont aussi une inertie comparable sur la formation des trois premiers axes factoriels. On remarque toutefois que les modalités fortes des variables AIRE (AI3) PCP (PC3) et LAT (LA3) ainsi que la deuxième modalité de la variable LONG (LO2) ont le plus contribué à la formation de ces axes.

- ***Contribution relative à l'inertie de l'axe (CTR) et contribution de l'axe à l'élément (COR)*** : les modalités fortes des variables AIRE (AI3), PCP (PC3), LAT (LA3) et LONG (LO3) sont celles qui ont le plus contribué à la formation du premier axe factoriel dans le nuage des points ligne (d'après les valeurs de CTR au tableau B4 de l'annexe B) ; elles y sont aussi mieux représentées (d'après les valeurs de COR associées au tableau B4 de l'annexe B). Par contre, les modalités LC2, LC3, LO1 et LO2 qui sont mal représentées sur le premier axe factoriel, jouent maintenant un rôle essentiel sur le deuxième axe factoriel. En effet, ce sont elles qui ont le plus contribué à sa formation. De plus, elles y sont aussi mieux représentées.

En regardant les coordonnées sur les deux premiers axes factoriels, nous observons que le premier axe factoriel est essentiellement décrit par les modalités fortes des variables AIRE (AI3), PCP (PC3) , LAT (LA3) et LONG (LO3). Par contre, le deuxième axe factoriel oppose essentiellement les modalités LC2 et LO2 aux modalités LO1 et LC3.

En procédant comme précédemment, l'analyse du nuage des points colonne (espace R^{124}) donne lieu à l'interprétation suivante :

- ***Qualité de la représentation (QLT)*** : tous les points colonne, excepté la modalité YPR8, sont bien représentés sur les trois premiers axes factoriels.

- ***Poids*** : les modalités YP1, YP2, YP4, YP5, YP6 et YP7 ont les poids les plus élevés.

- ***L'inertie (INR)*** : c'est la modalité YP1 (0.480) qui a le plus contribué à la formation des trois premiers axes factoriels ; suivie dans l'ordre des modalités YP6 et YP7.

- ***Contribution relative à l'inertie de l'axe (CTR) et contribution de l'axe à l'élément (COR)*** : les modalités YP1 et YP6 sont celles qui ont le plus contribué à la formation du premier axe factoriel dans le nuage des points colonne. Cependant, dans la formation du deuxième axe factoriel, ce sont les modalités YP2, YP4 et YP7 qui ont le mieux joué ce rôle. Leur qualité de représentation est aussi la meilleure pour chacun de ces axes.

Dans le nuage des points colonnes, l'analyse des coordonnées sur les deux premiers axes factoriels révèle que le premier axe factoriel oppose principalement la modalité YP6 à la modalité YP1. Le deuxième axe factoriel oppose quant à lui la modalité YP7 aux modalités YP2 et YP4.

La figure B1 de l'annexe B illustre la représentation des modalités des variables explicatives dans le repère formé par les deux premiers facteurs retenus. Ce graphique présente des propriétés remarquables :

- les modalités des variables explicatives se suivent dans l'ordre ;

- le parcours de la variable LCP est inverse de celui des autres variables explicatives. Il existe donc une corrélation négative de LCP avec les autres variables explicatives. Ce fait vient confirmer l'observation faite au tableau A2 de l'annexe A ;

- on note le parcours particulier pour la variable LAT.

La figure B2 de l'annexe B illustre une représentation simultanée de deux espaces dans le repère formé par les deux premiers facteurs retenus. Elle nous permet de tirer les renseignements suivants :

- sur le premier axe factoriel, les modalités de la variable dépendante YPR5 se suivent dans l'ordre à part une légère inversion entre les modalités YP3 et YP4 ;

- les modalités intermédiaires des variables explicatives semblent s'associer aux modalités fortes (YP6, YP7 et YP8) de la variable dépendante YPR5.

En résumé, l'analyse factorielle des correspondances du sous tableau de BURT $B_{J_e J_0}$ des modalités des variables explicatives avec celles de la variable à expliquer, ici la moyenne de la précipitation totale, 5 jours précédant la crue (YPR5), a permis de mieux apprécier les liaisons entre la variable dépendante et les variables explicatives. Ainsi, grâce au découpages en classes, les modalités intermédiaires des variables explicatives semblent s'associer aux modalités fortes (YP6, YP7 et YP8) de la variable à expliquer. Aussi, à partir de la figure B1 de l'annexe B, les modalités fortes du sous-ensemble des variables explicatives AIRE, SLM, PCP et LAT (AI3, SL3, PC3 et LA3) sont fortement corrélées avec le premier axe factoriel ; on note toutefois que leurs modalités faibles et intermédiaires sont, dans la majorité des cas, proches de l'origine des axes.
D'autre part, nous avons conclu, dans la première partie de cet exemple d'application, à la présence de multicolinéarité dans ce même sous-ensemble de variables explicatives. Dès

lors, on peut donc penser que l'application d'un modèle de régression sur les facteurs associés aux 106 individus supplémentaires, aura pour effet d'atténuer les effets de la présence de multicolinéarité. Cela permettra, à notre avis, d'améliorer grandement la prédiction de la variable dépendante YPR5.

5.3.2. Régression sur les facteurs associés aux 106 individus supplémentaires de l'AFC

Nous possédons les variables suivantes : YPR5, F_1, $F_2,\ldots,F_\alpha,\ldots,F_{k_0}$, où F_α est le α ème facteur associé aux 106 individus supplémentaires de l'AFC effectuée précédemment. Le modèle de régression que nous voulons ajuster est le suivant :

$$\text{YPR5}(i)=\sum_{\alpha=1}^{k_0} a_i F_\alpha(i)+\varepsilon(i), \qquad i=1,\ldots,106 \quad \text{et } k_0=5$$

5.3.2.1. Analyse exploratoire des données

Nous souhaitons découvrir les relations entre les variables du modèle de régression ci-haut mentionné. D'après le tableau A9 de l'annexe A, nous observons que le premier facteur (F_1) est celui qui est le plus corrélé avec la variable dépendante YPR5 ($r=-0.84$). La relation entre ces deux variables est probablement due au fait que, d'après l'analyse factorielle des correspondances, F_1 est celui qui a le plus résumé l'information comprise dans les données. On remarque par ailleurs que la corrélation entre les variables explicatives (les facteurs) est presque négligeable : la plus grande, qui est égale à 0.58, est obtenue entre le deuxième facteur (F_2) et le quatrième facteur (F_4). On peut donc s'attendre à ce que le problème de multicolinéarité ne soit pas important dans ce cas-ci.

5.3.2.2. Ajustement d'un premier modèle de régression

Le modèle de régression linéaire multiple que nous voulons ajuster a la forme suivante :

$$\text{YPR5}\,(i)=\sum_{\alpha=1}^{5} a_i F_\alpha(i)+\varepsilon(i), \qquad i=1,\ldots,106$$

Les résultats obtenus à l'aide de ce modèle de régression sont consignés au tableau A10 de l'annexe A. L'analyse de ce tableau permet d'éprouver l'hypothèse que tous les coefficients du modèle de régression sont non nuls conjointement au seuil de signification de 5%. Le premier modèle d'ajustement retenu est donc celui qui contient les 5 premiers facteurs. Avant d'en venir à cette conclusion, appliquons d'abord une méthode de sélection automatique

5.3.2.3. Ajustement final du modèle de régression

Dans cet ajustement final, nous nous sommes inspirés des résultats de la sélection automatique d'un modèle par la méthode du R^2 *maximal*. Cette méthode consiste à déterminer quel sous-ensemble des variables explicatives est le meilleur prédicteur de la variable dépendante par régression linéaire. Cette méthode a été préférée à la méthode du STEPWISE utilisée à la première partie de cet exemple d'application, à cause de sa spécificité. En effet, n'oublions pas que cette deuxième partie de l'étude a pour objectif de réduire les effets de la multicolinéarité rencontré dans le modèle de régression usuelle appliqué à cette première partie. Comme le VIF nous permet de voir si une variable est sujet à la manifestation de multicolinéarité, on pourrait donc utiliser sa valeur pour discriminer les variables susceptibles de favoriser sa présence dans le sous-ensemble de variables que nous souhaitons retenir. Car, si toutes les variables sont orthogonales alors tous les VIF sont égaux à 1 mais, si une sévère multicolinéarité existe, alors le VIF sera très grand pour ces variables. Ainsi, comme $VIF = 1/(1-R^2)$, si VIF est supérieur à 10, alors R^2 sera supérieur à 0.90, d'où la rétention de cette méthode de sélection. Les résultats de la régression par étapes sont présentés au tableau A11 de l'annexe A. Dans la figure A5 de l'annexe A, nous avons représenté la variation du R^2 maximal. Il s'ensuit que le palier est vite atteint après considération du premier et du deuxième facteur comme variables explicatives du modèle de régression. Ce faisant, l'analyse des résultats de la méthode du R^2 maximal favorise le choix des deux premiers facteurs associés aux 106 individus supplémentaires de l'AFC comme variables explicatives. En effet, le R^2 incluant ces deux variables est de 0.91, comparativement au modèle complet à cinq variables explicatives qui

est de 0.94. D'autre part, le R^2 ajusté est comparable pour ces deux modèles. Ainsi, la relation entre la moyenne de la précipitation totale, 5 jours précédant la crue (YPR5) et les deux premiers facteurs de l'AFC peut être approximée par l'équation de régression

$$YPR5 = 19.83 - 5.08 F_1 - 2.37 F_2$$

5.3.2.4. Discussions des résultats du modèle final

Nous allons surtout discuter du problème de multicolinéarité et des hypothèses du modèle de régression.

5.3.2.4.1. Discussion de la présence de multicolinéarité

Pour étudier la multicolinéarité, nous nous sommes aussi servis de la valeur de TOL comme dans la première partie de cet exemple d'application. Les résultats obtenus au tableau A12 de l'annexe A montrent que les valeurs de TOL sont très élevées pour les 5 variables explicatives (les 5 premiers facteurs associés aux 106 individus supplémentaires de l'AFC). Cela veut donc dire que la multicolinéarité est absente dans le sous-ensemble de variables explicatives retenus. Ce fait vient ainsi confirmer l'observation faite sur le tableau A9 de l'annexe A, à savoir que la corrélation entre les variables explicatives (les facteurs) est presque négligeable. En conséquence, ce résultat nous permet d'interpréter avec beaucoup de confiance les estimés des paramètres du modèle de régression retenu. C'est donc un avantage d'utiliser ce genre de régression lorsque nous sommes confrontés à la présence de multicolinéarité dans une régression classique.

5.3.2.4.2.Discussion de la validation du modèle de régression retenu

Dans la validation du modèle retenu, nous avons vérifié en premier s'il y a présence d'observations qui auraient pu exercer une influence indue sur l'estimation des paramètres du modèle. Ensuite, nous avons examiné la vérification des hypothèses du modèle de régression. Il ressort de cette analyse, les observations suivante :

- **observations influentes** : en procédant comme dans la première partie de cet exemple d'application, nous avons observé qu'aucune observation n'a influencé sérieusement l'estimation des paramètres du modèle de régression retenu ;

- **vérification des hypothèses du modèle** : nous considérons, comme dans la première partie de cet exemple d'application, que l'hypothèse d'indépendance est satisfaite. L'examen de la figure A6 de l'annexe A montre que, excepté quelques points qui se détachent dans les deux extrémités du graphique, pour la grande majorité, ils sont bien regroupés autour de la valeur zéro. Nous admettons donc que l'hypothèse d'homogénéité de la variance est satisfaite. Il en est de même aussi avec l'hypothèse de normalité des résidus. En effet, sur la figure A7 de l'annexe A, malgré le fait que la loi de la distribution a des queues lourdes (voir les observations extrêmes aux deux extrémités), nous observons tout de même que les résidus se répartissent de part et d'autres de la droite de Henry. L'hypothèse de normalité est donc valide.

En somme, le modèle de régression retenu ne comporte pas d'observations influentes qui auraient joué un rôle dans l'estimation de ses paramètres. Cela est particulièrement dû au découpage en classes. De plus, les erreurs de ce modèle sont indépendantes, homogènes et normalement distribuées.

5.3.3. Conclusion et interprétation des résultats du modèle final

L'analyse factorielle des correspondances du tableau de BURT $B_{J_eJ_0}$ des modalités des variables explicatives (variables physiographiques) avec celles de la variable à expliquer YPR5, ici la moyenne de la précipitation totale, 5 jours précédant la crue, a permis de bien voir les liaisons entre cette variable hydrologique et les différentes variables physiographiques. La régression sur les facteurs associés aux 106 individus supplémentaires (stations météorologiques) de l'AFC a permis quant à elle d'estimer la valeur de YPR5, connaissant les variables explicatives. Aussi, elle permet de nous rassurer sur la précision de cette estimation, à cause de l'absence de multicolinéarité dans le sous-ensemble de variables explicatives retenus. Dans l'espace des cinq premiers facteurs issus de cette analyse factorielle des correspondances, la relation entre la moyenne de la

précipitation totale, 5 jours précédant la crue (YPR5) et les deux premiers facteurs peut être approximée par l'équation de régression :

$$YPR5 = 19.83 - 5.08 F_1 - 2.37 F_2$$

C'est donc dire qu'une fois connu le premier facteur associé aux 106 individus supplémentaires de l'analyse factorielle des correspondances, le deuxième facteur issu de cette même analyse est la variable explicative permettant d'étudier l'évolution de la moyenne de la précipitation totale, 5 jours précédant la crue (YPR5). 91% de la variabilité de la variable dépendante YPR5 est expliquée par ces deux variables explicatives. Une part peu importante de la variabilité demeure donc inexpliquée. Les différentes étapes de validation n'ont pas permis de remettre en cause les hypothèses sous-jacentes au modèle.

5.4. Conclusion

L'intérêt de la régression sur les facteurs d'une analyse factorielle des correspondances par rapport à la régression usuelle réside, comme nous l'avons déjà mentionné dans le fait qu'elle fournit une meilleure précision pour chaque estimation de YPR5 (i), i = 1, ... ,106. D'un point de vue pratique, cette propriété est fort intéressante, surtout quand l'ensemble des variables explicatives retenues pour expliquer la variable réponse présente des problèmes de multicolinéarité (ce qui est le cas dans cet exemple d'application).

Pour se rendre compte de la qualité globale de la régression sur les facteurs d'une analyse factorielle des correspondances, on peut calculer sur l'échantillon le coefficient de corrélation (R) entre la variable dépendante et sa valeur prédite. Ce coefficient est l'analogue du coefficient de corrélation multiple dans la régression usuelle. On pourra noter aussi la valeur comparable du coefficient de détermination R^2. En effet, un modèle de régression qui s'ajuste bien donnera lieu à une somme des carrés de l'erreur petite par rapport à la somme des carrés totale. Une mesure sommaire d'évaluation de cet ajustement est le R^2. Il prend une valeur entre 0 et 1 et quantifie la proportion de la variabilité de la variable dépendante qui est expliquée par le modèle. Dans notre exemple d'application, les

valeurs de R et de R^2, pour la régression usuelle et la régression sur les facteurs de l'AFC, sont consignées dans le tableau 5.1 qui suit. Dans ce tableau, on voit que la régression sur les facteurs d'une analyse factorielle des correspondances donne des résultats meilleurs que la régression usuelle : les valeurs de R et R^2 sont de loin supérieures à celles de la régression usuelle.

Les principaux désavantages de la régression sur les facteurs d'une analyse factorielle des correspondances par rapport à la régression usuelle sont les suivants :

- elle demande beaucoup d'étapes de calculs comparativement à la régression usuelle. En effet, il faut d'abord faire une analyse factorielle des correspondances avant de faire la régression usuelle sur les facteurs retenus ;

- elle ne permet pas de discriminer le meilleur ensemble de variables explicatives retenus : sa méthodologie consiste d'abord à ne retenir que les premiers facteurs de l'AFC qui expliquent le plus de variabilité dans les données et ensuite appliquer la régression usuelle. Il est donc probable que l'on puisse négliger, dans l'étape de l'AFC, certains facteurs qui auraient peut-être mieux expliqué la variable dépendante, ici YPR5. C'est donc un défaut imputable à cette méthode de régression. On note cependant que, dans notre exemple d'application, nous pouvons nous considérer être à l'abri de cette lacune, car la valeur de R^2 obtenue avec cette méthode est suffisamment élevée (voir le tableau 5.1).

TYPE de RÉGRESSION	**MODÈLE D'AJUSTEMENT**	R	R^2
Usuelle	YPR5 = 74.26 - 0.94 LCP -2.11 LAT + 0.56 LONG	0.75	0.56
Sur les Facteurs d'une AFC	YPR5 $= 19.83 - 5.08 F_1 - 2.37 F_2$	0.95	0.91

Tableau 5.1 : Comparaison des deux régressions

6. CONCLUSION GÉNÉRALE

Dans ce travail, nous avons présenté une alternative à l'estimation classique des coefficients de régression par la méthode des moindres carrés ordinaires. Celle-ci vise à atténuer les inconvénients de la multicolinéarité importante qui se manifeste dans beaucoup de problèmes de régression multiple. Cette multicolinéarité se présente souvent lorsqu'un grand nombre de variables explicatives sont incluses au modèle et que certaines d'entre elles fournissent une information similaire. D'une manière générale, l'estimation par les moindres carrés ordinaires obtenue dans ces conditions a notamment le défaut de conduire à des estimateurs dont les variances sont très grandes. L'estimation des coefficients de régression par la méthode de régression classique est plus «fidèle» lorsque les variables explicatives ne présentent pas des phénomènes de multicolinéarité. Mais cela ne suffit pas à la préférer si l'on recherche une estimation des coefficients de régression qui soit le plus stable possible sur la population. En l'occurrence, lorsque la multicolinéarité est présente dans l'ensemble des variables explicatives, et que celle-ci est indépendante de l'interprétation des variables et des erreurs de mesures, ce serait alors la méthode de régression sur les facteurs d'une analyse des correspondances qui devrait être préférée.

L'idée sous-jacente à l'utilisation de la méthode présentée est la suivante. On s'efforce d'extraire de X l'information utile qui est vraisemblablement contenue dans les premiers facteurs et on néglige les fluctuations «aléatoires» ou le «bruit» qui est contenu dans les derniers facteurs. Il est donc probable que l'on puisse négliger certains facteurs qui auraient peut-être mieux expliqué la variable dépendante. La suppression automatique des facteurs associés à de faibles valeurs propres est un défaut imputable à cette méthode. D'ailleurs, *Greenberg (1975)* signale que la prise en considération des facteurs correspondant à de

faibles valeurs propres augmente la variance des coefficients de régression. Par contre elle permet de diminuer le biais, si ces facteurs sont corrélés avec la variable dépendante. Dans certaines situations cependant, la prise en considération des premiers facteurs uniquement peut se justifier si on considère, a priori, que les derniers facteurs prennent en compte des fluctuations aléatoires dans les variables explicatives.

L'exemple d'application traité au chapitre 5 est destiné à illustrer les principes de la méthode de régression sur les facteurs d'une analyse des correspondances que nous avons exposés tout au long de ce travail. Il présente des phénomènes de multicolinéarité accentué et il est normal que cette méthode alternative soit supérieure à la régression usuelle. Effectivement, pour se rendre compte de la qualité globale de la régression sur les facteurs d'une analyse factorielle des correspondances, nous avons calculé sur l'échantillon le coefficient de corrélation (R) entre la variable dépendante YPR5 et sa valeur prédite. Puis, nous avons évalué la valeur comparable du coefficient de détermination R^2. Dans le tableau 5.1, il appert que la régression sur les facteurs d'une analyse factorielle des correspondances donne des résultats meilleurs que la régression usuelle : les valeurs de R et de R^2 sont supérieures à celles de la régression usuelle. On notera cependant que, dans cet exemple d'application, la valeur de R^2 obtenue, en ne retenant que les deux premiers facteurs, dans le modèle de régression est suffisamment élevée. Nous n'avons donc pas à nous inquiéter de l'information qu'auraient apporter des facteurs correspondant à de faibles valeurs propres.

Enfin, on remarquera aussi que la méthode alternative qui a été décrite dans ce travail n'est pas la seule solution possible faisant appel à des techniques de régression spécialement mises au point pour atténuer les effets de la multicolinéarité. En effet, nous avons signalé la méthode de *Webster, Gunst et Mason (Webster et al., 1974)* qui est la régression par l'analyse des valeurs latentes et la méthode de régression par les moindres carrés partiels (*Martens et Naes, 1989*). On peut élargir cette liste à la régression "ridge" (*Chatterjee and Price (1977)*, pp. 181-192, *Draper and Smith (1981)*, pp. 313-325, *Neter et al. (1985)*, pp. 393-400) et à la régression PLS (*Cazes, 1997*). Effectivement, dans le cas de la régression

"ridge", comme les estimateurs des moindres carrés ne sont pas biaisés, cela signifie que si l'on répétait l'échantillonnage un grand nombre de fois, l'estimation des paramètres serait en moyenne égale à la vraie valeur des paramètres. Or en présence de multicolinéarité, la variance de ces estimateurs a tendance à être grande. La régression "ridge" est une méthode permettant de corriger cette situation en fournissant des estimateurs biaisés des paramètres mais dont la variance est plus petite. Pour la régression PLS, il est avantageux de raisonner sur les composantes PLS et non sur les facteurs d'une analyse des correspondances pour effectuer la régression. En effet, la régression PLS permet d'avoir, comme nous l'avons déjà dit, des composantes explicatives dépendant de la liaison entre la variable dépendante et les variables explicatives. Ce qui permet d'optimiser dans un certain sens la régression, si on ne garde que les premières composantes PLS. Cette façon de procéder est contraire à la régression sur les facteurs d'une analyse des correspondances où certains facteurs de faible variance peuvent être éliminés alors qu'ils peuvent être corrélés de façon non négligeable à la variable à expliquer (la variable dépendante). Par ailleurs, les résultats obtenus avec la régression PLS sont biaisés, puisqu'on utilise pour expliquer la variable dépendante des composantes qui sont déjà fonction des liaisons entre la variable dépendante et les variables explicatives. Des informations relatives à ces méthodes et des exemples d'application sont donnés par *Cazes (1997)* et par *Shahid (1982)*.

En somme, la méthode de régression sur les facteurs d'une analyse factorielle des correspondances ne permet pas de discriminer le meilleur ensemble de variables explicatives retenus. Elle est intéressante, car elle permet d'améliorer la prédiction de la variable dépendante avec les variables explicatives dont on dispose.

7. REVUE BIBLIOGRAPHIQUE

Basilevsky, A. (1968). Factor Analysis as a Variable Aggregator. *Technical Report*, Dept. of Economics, University of Southampton, England.

Basilevsky, A. (1981). Factor Analysis Regression. *The Canadian Journal of Statistics*, Vol. 9. No 1, pp. 109-117.

Belsley, D. A., Kuh E., Welsch R. E. (1980). Regression Diagnostics : Identifying Influential data and Sources of Collinearity, *Wiley* Ed.

Benzécri, J. P. (1982). L'Analyse des Données, Tome 2 : L'Analyse des correspondances, *Dunod*, 4^{e} édition, 620pp.

Cazes, P. (1997). Adaptation de la Régression PLS au cas de la Régression après des Correspondances Multiples. *Rev. Statist. Appli.*, **XLV** (2), pp. 89-99.

Chatterjee, S. and Price B. (1977). Regression Analysis by Example, New York, *Wiley*.

Coxe, K. L. (1982). Selection Rules for Principal Component Regression, *ASA Proceed. Bus & Econ. Section*, pp. 22-227.

Coxe, K. L. (1984). Multicollinearity, Principal Component Regression and Selection Rules for these Components. *ASA Proceed. Bus & Econ. Section*, pp. 22-227.

Daling, J. R., and Tamura, H. (1970). Use of Orthogonal Factors for Selection of Variables in a Regression Equation- An illustration. *Appl. Statist.*, 19, pp 206-268.

Draper, N. R. and Smith H. (1981). Applied Regression Analysis. New York, *Wiley*, 709 pp.

Erkel-Rousse, H. (1994/95). Multicolinéarité dans le Modèle Linéaire Ordinaire : Définition, Détection, Proposition de Solutions, in «Introduction à l'Économétrie du Modèle Linéaire», *Polycopié ENSAE*, pp. 177-252.

Erkel-Rousse, H. (1994). Multicolinéarité dans le Modèle Linéaire Ordinaire : Quelques Éléments pour un Usage Averti des Indicateurs de Belsley, Kuh et Welsh, Projet de Document de *Travail de la Collection Méthodologique de l'INSEE*, 21 avril 1994.

Escofier, B. et Pagès J. (1988). Analyses Factorielles Simples et Multiples. *Dunod*, Paris.

Farrar, D. E. et Glauber R. R. (1967). Multicollinearity in Regression Analysis : the Problem Revisited, *Review of Economics and Statistics*, Vol. 49, pp. 92-107.

Greenacre, M. J. (1984). Theory and Application of Correspondence Analysis. *Academic Press*, London, 364 pp.

Greenberg, E. (1975). Minimum Variance Properties of Principal Component Regression. *J. Amer. Stat. Assoc.* 70, pp. 194-197.

Hocking, R. R. (1976). The Analysis and Selection of Variables in Linear Regression, *Biometrics* 32, pp. 1-49.

Jackson, J. E. (1991). A User's Guide to Principal Components, *Wiley*, New York.

Jolliffe, I. T. (1982). A Note on the Use of Principal Components in Regression, *Appl. Statist., 31, pp. 300-303.*

Jolliffe, I. T. (1986). Principal Component Analysis, *Springer-Verlag*, New York.

Kendall, M. G. (1957). A Course in Multivariate Analysis, *Griffin*, London.

Lebart, L., Morineau A. et Fénelon J. P. (1982). Traitement des Données Statistiques : Méthodes et Programmes. *Dunod*, Paris, 2^{e} édition.

Maddala, G.S. (1977). Econometrics, *Mc Graw-Hill* Ed.

Martens H., Naes T. (1989). Multivariate Calibration. New York, *Wiley*, 419 pp.

Mason, R. L., Gunst R. F., Webster J. T. (1975). Regression Analysis and Problem of Multicollinearity, *Communication in Statistics*, 4 (3), pp. 277-292.

Massy, W. F. (1965). Principal Component Regression in Exploratory Statistical Research. *J. Amer. Statist. Assoc.*, 60, pp. 234-256.

Miller, A. J. (1990). Subset Selection in Regression, *Chapman and Hall*, London, 229pp.

Neter, J., Wasserman W. and Kutner M. (1985). Applied Linear Statistical Models, 2nd ed., *Richard D. Irwin*, Inc.

Ragnar Frisch (1934). Statistical Confluence Analysis by Means of Complete Regression Systems. Publication 5 (Oslo ; University Institute of Economics).

Shahar, B. et Gonzalo R. M. (1994). Variable Selection in Regression Models Using Principal Componets. *Commun., Statist.-Theory Meth.*, 23 (1), pp 197-213.

Shahid, A. A. (1982). Application de la Régression d'après un Tableau de Correspondance : l'Estimateur des Paléoclimats d'après l'Écologie des Foraminifères. *Cahiers de l'Analyse des Données*, 7 (1) : pp.93-111.

Silvey, S. D. (1969). Multicollinearity and Imprecise, Estimations, *Journal of Royal Statistical Society, Series B*, Vol. 31, pp.539-552.

Stewart, G. W. (1987). Collinearity and Least Squares Regression, *Statistical Science* Vol. 2, No 1, pp. 68-84, suivi des Commentaires de D. A. Belsley pp. 86-91.

Theil, H. (1971). Principles of Econometrics. New York, *Wiley*, 736p.

Thompson, M. (1978a). Selection of Variables in Multipler Regression. Part I : a Review and Evaluation. *Int. Stat. Rev.* 46 pp. 1-19.

Thompson, M. (1978a). Selection of Variables in Multipler Regression. Part II : Chosen Procedures, Computation and Exemples. *Int. Stat. Rev.* 46, pp. 129-146.

Webester, J. T., Gunst R. F., Mason R. L. (1974). Latent Root Regression Analysis. *Technometrics* 16, pp. 513-522

Weisberg, S. (1985). Applied Linear Regression. New York. *Wiley*, 324p.

ANNEXE A : Résultats obtenus avec le logiciel STATISTICA

Correlations (lah99don.sta)

Marked correlations are significant at p < .05000

N=106 (Casewise deletion of missing data)

	AIRE	*LCP*	*PCP*	*SLM*	*LAT*	*LONG*	*DTY5*	*YPR5*	*YNS5*	*PTMA*
AIRE	1.00	***-.28***	***.79***	***.96***	***.61***	***.39***	***.38***	***-.39***	***-.27***	***-.47***
LCP	***-.28***	1.00	***-.41***	***-.25***	***-.40***	***-.39***	***-.46***	.02	.10	.19
PCP	***.79***	***-.41***	1.00	***.68***	***.69***	***.51***	***.52***	***-.42***	***-.37***	***-.50***
SLM	***.96***	***-.25***	***.68***	1.00	***.54***	***.37***	***.38***	***-.33***	***-.24***	***-.43***
LAT	***.61***	***-.40***	***.69***	***.54***	1.00	***.75***	***.76***	***-.65***	***-.45***	***-.77***
LONG	***.39***	***-.39***	***.51***	***.37***	***.75***	1.00	***.73***	***-.30***	***-.39***	***-.65***
DTY5	***.38***	***-.46***	***.52***	***.38***	***.76***	***.73***	1.00	***-.39***	***-.49***	***-.54***
YPR5	***-.39***	.02	***-.42***	***-.33***	***-.65***	***-.30***	***-.39***	1.00	***.23***	***.59***
YNS5	***-.27***	.10	***-.37***	***-.24***	***-.45***	***-.39***	***-.49***	***.23***	1.00	***.40***
PTMA	***-.47***	.19	***-.50***	***-.43***	***-.77***	***-.65***	***-.54***	***.59***	***.40***	1.00

Tableau A1 : Matrice de corrélations des 10 variables

Correlations (lah99do2.sta)

Marked correlations are significant at p < .05000

N=106 (Casewise deletion of missing data)

	AIRE	*LCP*	*PCP*	*SLM*	*LAT*	*LONG*	*YPR5*
AIRE	1.00	***-.28***	***.79***	***.96***	***.61***	***.39***	***-.39***
LCP	***-.28***	1.00	***-.41***	***-.25***	***-.40***	***-.39***	.02
PCP	***.79***	***-.41***	1.00	***.68***	***.69***	***.51***	***-.42***
SLM	***.96***	***-.25***	***.68***	1.00	***.54***	***.37***	***-.33***
LAT	***.61***	***-.40***	***.69***	***.54***	1.00	***.75***	***-.65***
LONG	***.39***	***-.39***	***.51***	***.37***	***.75***	1.00	***-.30***
YPR5	***-.39***	.02	***-.42***	***-.33***	***-.65***	***-.30***	1.00

Tableau A2 : Matrice de corrélations des sept variables

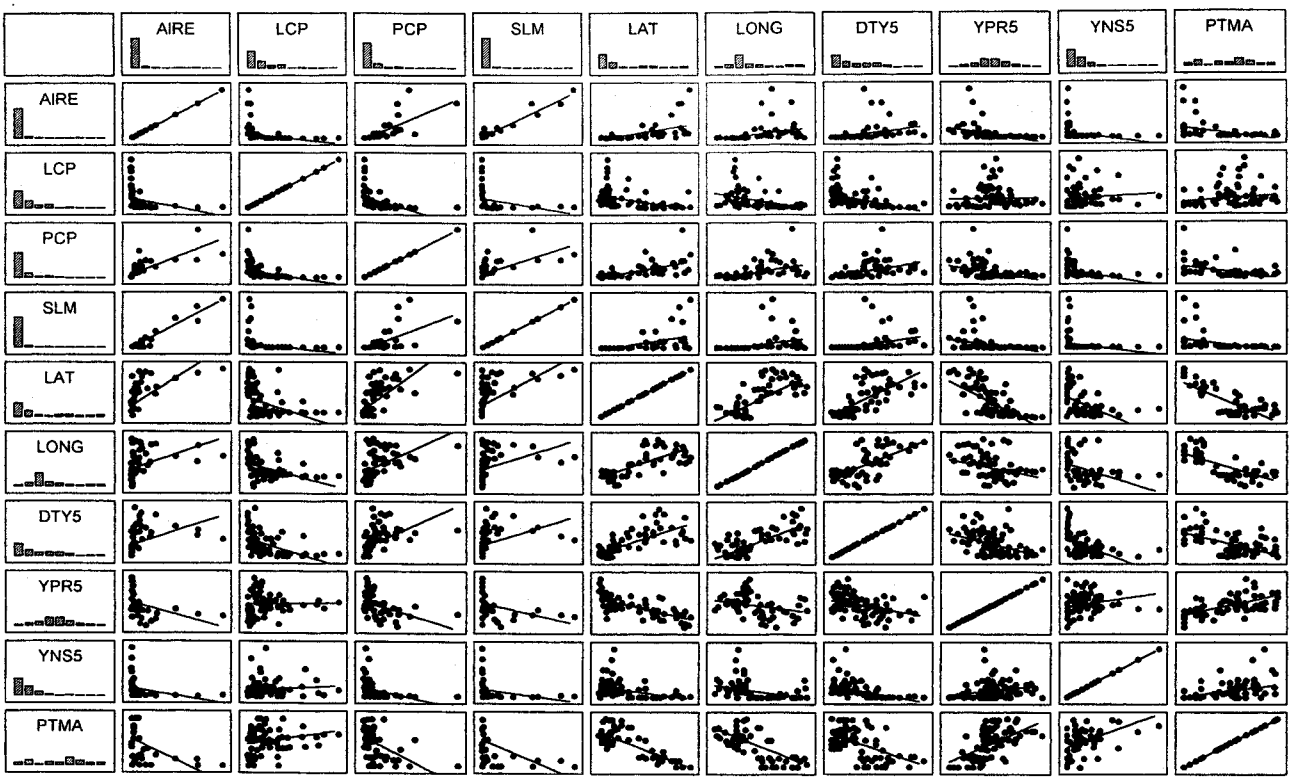


Figure A1 : Relations entre les 10 variables

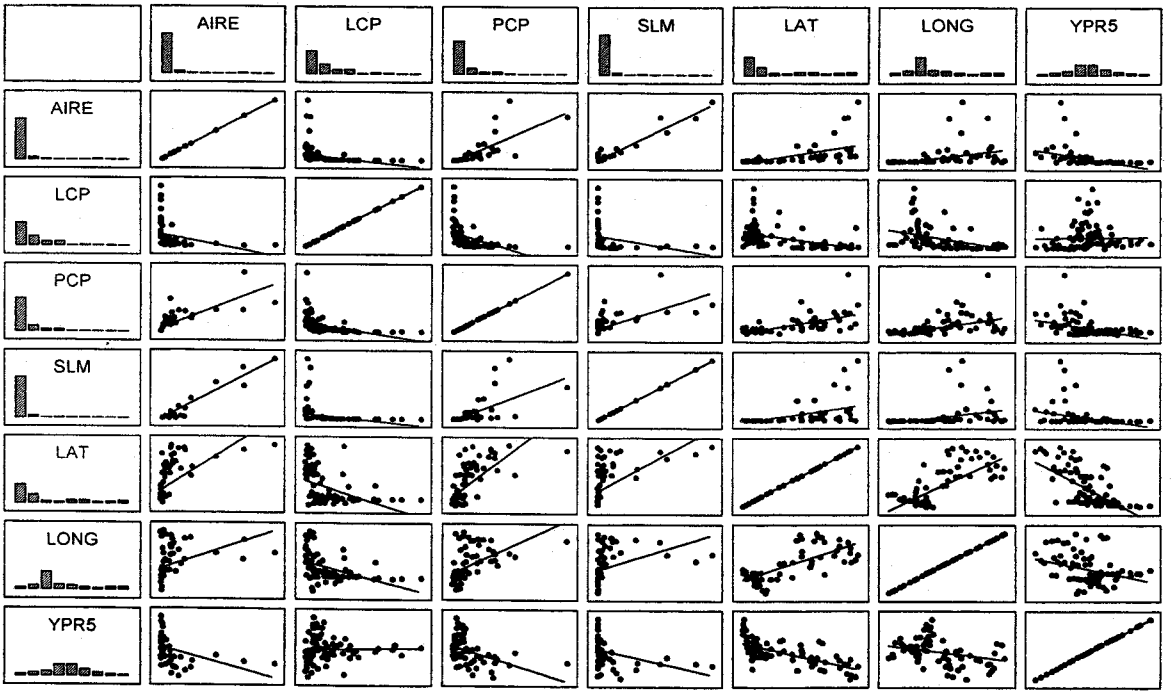


Figure A2 : Relations entre les 7 variables

Analysis of Variance (lah99do2.sta)

	Sums of Squares	df	Mean Squares	F	p-level
Regress.	2905.958	6	484.3263	20.97219	.000000
Residual	2286.281	99	23.0937		
Total	5192.238				

Regression Summary for Dependent Variable: YPR5

R= .74811326 R^2= .55967345 Adjusted R^2= .53298699

F(6,99)=20.972 p<.00000 Std.Error of estimate: 4.8056

	BETA	St. Err. of BETA	B	St. Err. of B	t(99)	p-level
Intercp			74.57591	10.08328	7.39600	.000000
AIRE	.18394	.370760	.00017	.00035	.49612	.620906
LCP	-.25101	.075924	-.97027	.29348	-3.30610	.001319
PCP	-.08084	.143494	-.00479	.00850	-.56334	.574480
SLM	-.09089	.304495	-.00010	.00035	-.29851	.765940
LAT	-1.06658	.125773	-2.15491	.25411	-8.48021	.000000
LONG	.40615	.107220	.58256	.15379	3.78798	.000261

Tableau A3 : Résultats de la régression de YPR5 vs (AIRE, LCP, PCP, SLM, LAT, LONG)

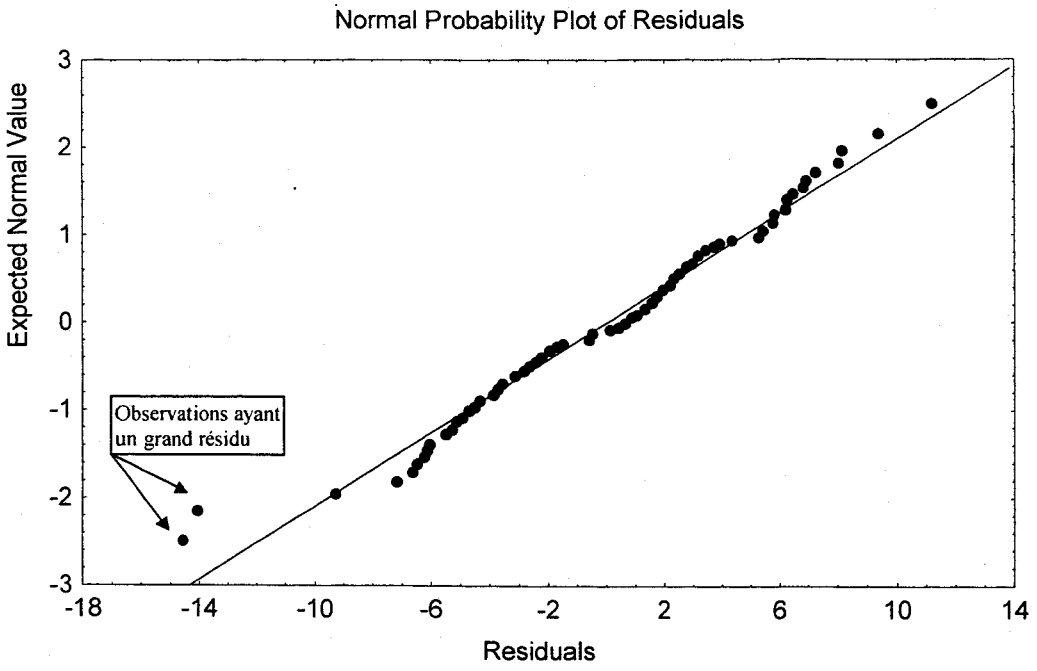


Figure A3 : Graphique de la droite de Henry dans la régression de YPR5 vs (AIRE, LCP, PCP, SLM, LAT, LONG)

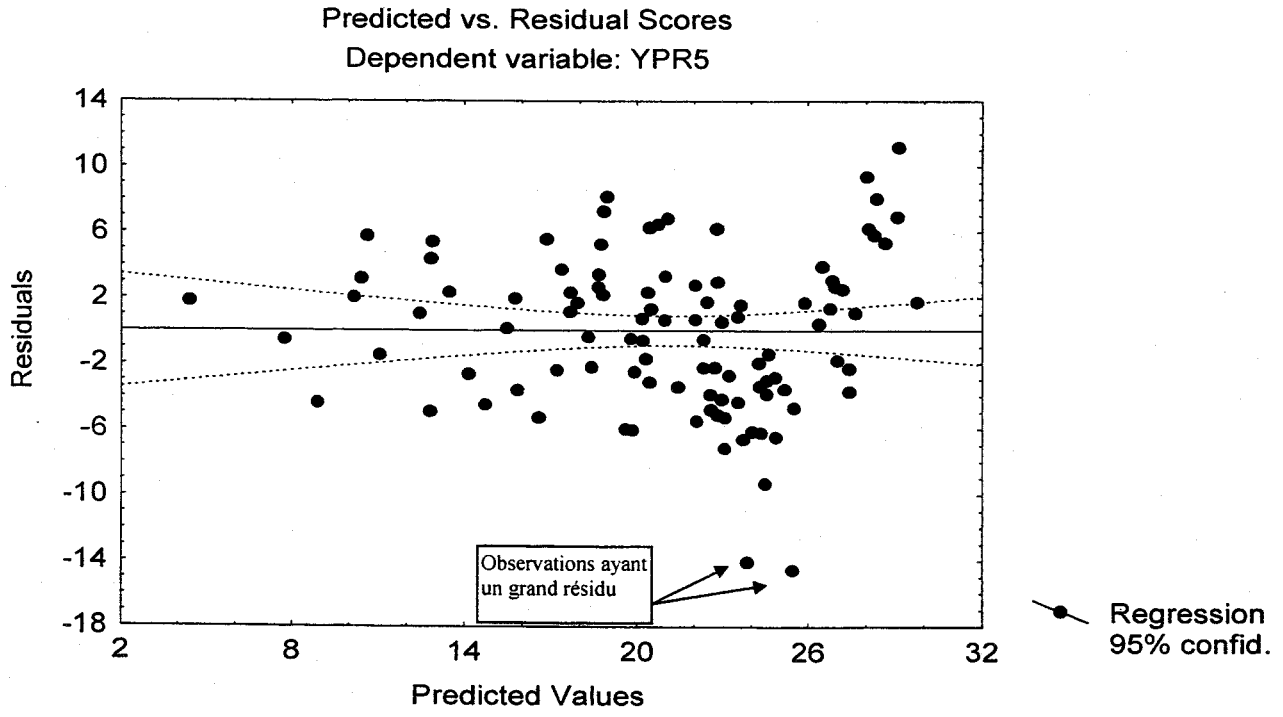


Figure A4 : Graphique des résidus contre les valeurs prédites dans la régression de YPR5 vs (AIRE, LCP, PCP,SLM, LAT, LONG)

Redundancy of Independent Variables (lah99do2.sta)

R-square column contains R-square of respective variable with all other independent variables

	Tolerance.	R-square	Partial Cor.	Semipart Cor.
AIRE	.032356	.967644	.049801	.033087
LCP	.771590	.228410	-.315324	-.220488
PCP	.216009	.783991	-.056527	-.037570
SLM	.047971	.952029	-.029988	-.019908
LAT	.281166	.718834	-.648661	-.565557
LONG	.386893	.613107	.355795	.252626

Tableau A4 : Résultats des valeurs de TOL (YPR5 vs (AIRE,..., LONG))

Analysis of Variance (lah99do2.sta)

	Sums of Squares	df	Mean Squares	F	p-level
Regress.	2892.540	3	964.1799	42.76488	.000000
Residual	2299.699	102	22.5461		
Total	5192.238				

Regression Summary for Dependent Variable: YPR5

R= .74638405 R^2= .55708915 Adjusted R^2= .54406236

F(3,102)=42.765 p<.00000 Std.Error of estimate: 4.7483

	BETA	St. Err. of BETA	B	St. Err. of B	t(102)	p-level
Intercp			74.26333	8.782204	8.4561	.000000
LCP	-.24375	.072858	-.94220	.281629	-3.3455	.001150
LAT	-1.04235	.101708	-2.10595	.205489	-10.2485	.000000
LONG	.38804	.101295	.55658	.145294	3.8308	.000221

Tableau A5 : Résultats de la régression de YPR5 vs (LCP, LAT, LONG)

Étape 1

Regression Summary for Dependent Variable: YPR5

R= .65183966 R^2= .42489494 Adjusted R^2= .41936508

F(1,104)=76.837 p<.00000 Std.Error of estimate: 5.3584

	BETA	St. Err. of BETA	B	St. Err. of B	t(104)	p-level
Intercp			82.10531	6.968078	11.78306	.000000
LAT	-.651840	.074363	-1.31696	.150242	-8.76564	.000000

Étape 2

Regression Summary for Dependent Variable: YPR5

R= .71308352 R^2= .50848811 Adjusted R^2= .49894419

F(2,103)=53.279 p<.00000 Std.Error of estimate: 4.9777

	BETA	St. Err. of BETA	B	St. Err. of B	t(103)	p-level
Intercp			60.96823	8.209992	7.42610	.000000
LAT	-.982616	.104966	-1.98526	.212071	-9.36128	.000000
LONG	.439325	.104966	.63015	.150559	4.18540	.000060

Étape 3

Regression Summary for Dependent Variable: YPR5

R= .74638405 R^2= .55708915 Adjusted R^2= .54406236

F(3,102)=42.765 p<.00000 Std.Error of estimate: 4.7483

	BETA	St. Err. of BETA	B	St. Err. of B	t(102)	p-level
Intercp			74.26333	8.782204	8.4561	.000000
LAT	-1.04235	.101708	-2.10595	.205489	-10.2485	.000000
LONG	.38804	.101295	.55658	.145294	3.8308	.000221
LCP	-.24375	.072858	-.94220	.281629	-3.3455	.001150

Tableau A6 : Résultats obtenus avec la méthode STEPWISE dans la régression de YPR5 vs (AIRE, LCP, PCP,SLM, LAT, LONG)

Deleted residuals (lah99do2.sta)	Standard	Deleted	Cook's
Dependent variable: YPR5	Residual	Residual	Distance
S1	0,22433411	1,11155891	0,00057141
S2	0,66557544	3,21445823	0,00192924
S3	0,48015136	2,31621885	0,00093308
S4	0,02403251	0,11632725	2,8563E-06
S5	-0,62391621	-2,99817705	0,00118533
S6	0,09541424	0,45820171	2,6162E-05
S7	1,40837002	6,75984478	0,0054361
S8	0,4879888	2,50054288	0,00508626
S9	1,37052608	6,61489868	0,00786822
S10	0,75412071	3,68562365	0,00428519
S11	-0,3812488	-1,83952403	0,00059666
S12	0,47595364	2,2979455	0,00096799
S13	-0,58366734	-2,82429457	0,00165618
S14	0,188107	0,90559459	0,00012463
S15	-1,09411645	-5,28116703	0,00503654
S16	-1,32303262	-6,45153666	0,01211992
S17	-1,39241672	-6,7007246	0,0066242
S18	-1,00400174	-4,83011913	0,00336604
S19	-0,66339546	-3,20401025	0,00191954
S20	-0,69574946	-3,34827852	0,00165856
S21	-0,59909838	-2,88937306	0,00143188
S22	-0,71427393	-3,44159579	0,00190925
S23	-0,40618753	-1,95376074	0,00054318
S24	0,11978256	0,58194405	8,5075E-05
S25	0,08219852	0,39874312	3,7327E-05
S26	0,84306085	4,08264637	0,00360196
S27	1,21862173	5,9684577	0,01205237
S28	0,40129974	1,95834792	0,00114804
S29	-0,46452308	-2,25838995	0,00131994
S30	-0,73492259	-3,5764811	0,00344507
S31	-3,04686379	-14,6957817	0,03722668
S32	-0,3751483	-1,82431495	0,00087003
S33	0,6465978	3,13838744	0,00237217
S34	0,56544614	2,74466896	0,00181929
S35	2,00045204	9,74826527	0,02697776
S36	1,33930278	6,53476238	0,01270887
S37	1,25891471	6,1386838	0,01096015
S38	0,26260665	1,27832794	0,00044509
S39	0,66900349	3,24423456	0,00243272
S40	0,30626109	1,48604918	0,00052464
S41	1,72090912	8,41358852	0,02260006
S42	1,474177	7,25331259	0,02039021
S43	0,3744835	1,85633445	0,00160936
S44	1,11657465	5,48638296	0,01122927
S45	2,39749289	11,7901134	0,0531001

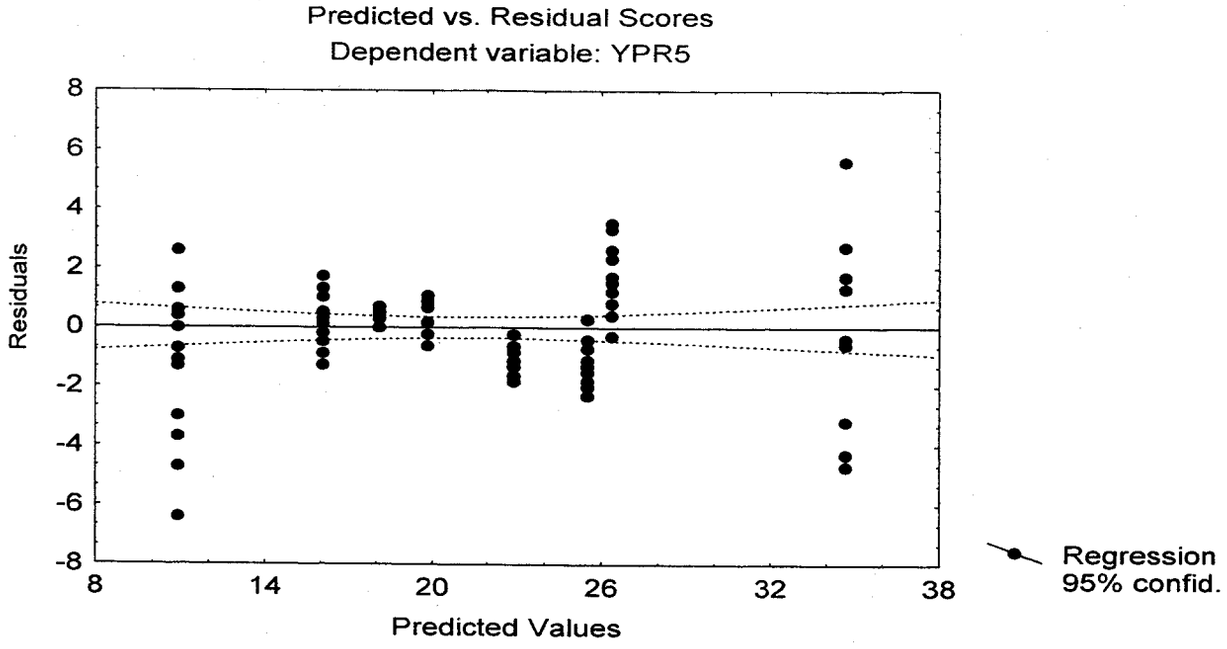


Figure A6 : Graphique de la droite de Henry dans la régression de YPR5 vs FACTEURS

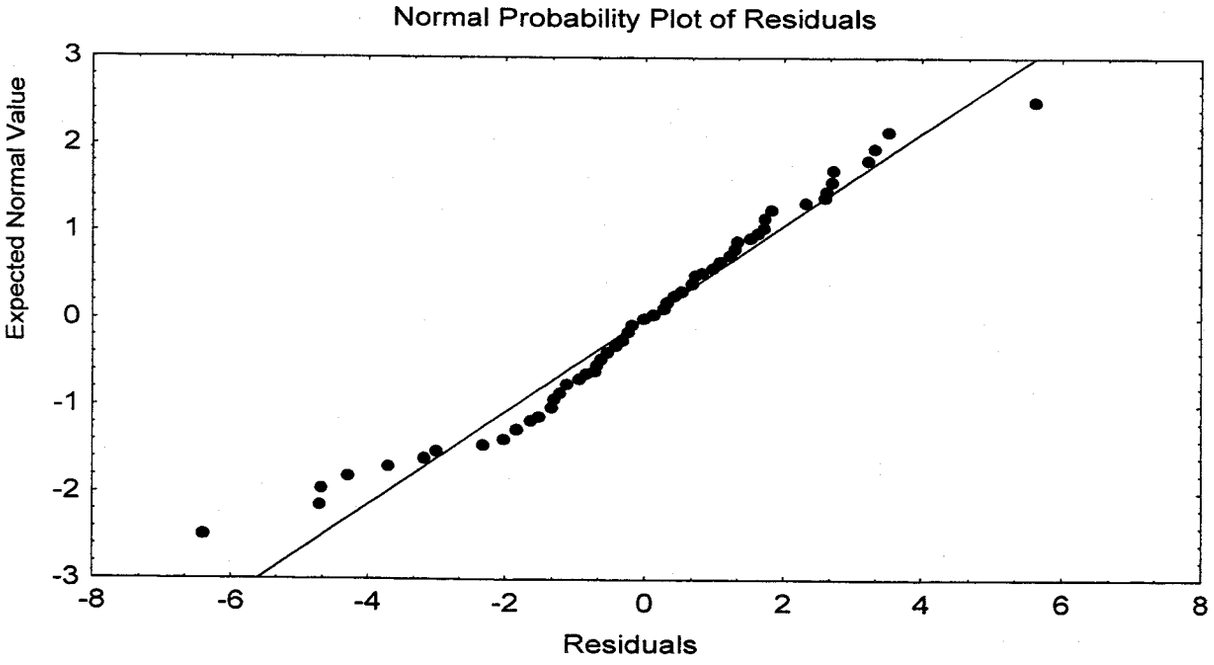


Figure A7 : Graphique des résidus contre les valeurs prédites dans la régression de YPR5 vs FACTEURS

ANNEXE B : Résultats obtenus avec le logiciel ADDAD

Programme Addad

```
$RUN RECOD2
$L080
$PRT=zzCREACM.SOR
$PAR=.
   TITRE ESSAI RECOD2 SUR MES DONNEES (87);
   PARAM NI1=106 NQ=7  NJ=33 MAXMOD=8 STIJ=1;
   RECOD NVF=7;
   FCODE 1-7;
   CLASSES 6*3 8 ;

   LISTE IDEN(1,4)
      AIRE(7,4,1)  LCP(14,4,2) PCP(19,4,2) SLM(27,4,1) LAT(34,4,2)
      LONG(40,4,2) YPR5(51,4,1);

   LISTE IDEN(1,4)
      AI1(5,2)   AI2(7,2)   AI3(9,2)
      LC1(11,2)  LC2(13,2)  LC3(15,2)
      PC1(17,2)  PC2(19,2)  PC3(21,2)
      SL1(23,2)  SL2(25,2)  SL3(27,2)
      LA1(29,2)  LA2(31,2)  LA3(33,2)
      LO1(35,2)  LO2(37,2)  LO3(39,2)
      YP1(41,2)  YP2(43,2)  YP3(45,2)
      YP4(47,2)  YP5(49,2)  YP6(51,2)
      YP7(53,2)  YP8(55,2)
      AIRE(65,6,3) LCP(71,6,3) PCP(77,6,3) SLM(83,6,3)LAT(89,6,3)
      LONG(95,6,3)YPR5(107,6,3);
$F11=Donndev1.dat
$F21=zzACM.DAT
$END
```

Résultats : VARIABLES RECODEES SUIVANT LE TYPE F

```
     L'ENTREE NUMERO   1 -AIRE-  DEVIENT EN SORTIE
     NUMERO   1 POUR LES BORNES    .10 -   2.00    36 OBSERVATIONS
     NUMERO   2 POUR LES BORNES   2.00 -  11.90    36 OBSERVATIONS
     NUMERO   3 POUR LES BORNES  11.90 - 500.00    34 OBSERVATIONS
     L'ENTREE NUMERO   2 -LCP -  DEVIENT EN SORTIE
     NUMERO   4 POUR LES BORNES    .00 -    .70    41 OBSERVATIONS
     NUMERO   5 POUR LES BORNES    .70 -   1.70    30 OBSERVATIONS
     NUMERO   6 POUR LES BORNES   1.70 -   9.40    35 OBSERVATIONS
     L'ENTREE NUMERO   3 -PCP -  DEVIENT EN SORTIE
     NUMERO   7 POUR LES BORNES    .05 -    .34    37 OBSERVATIONS
     NUMERO   8 POUR LES BORNES    .34 -    .84    34 OBSERVATIONS
     NUMERO   9 POUR LES BORNES    .84 -   8.47    35 OBSERVATIONS
     L'ENTREE NUMERO   4 -SLM -  DEVIENT EN SORTIE
     NUMERO  10 POUR LES BORNES    .00 -    .00    49 OBSERVATIONS
     NUMERO  11 POUR LES BORNES    .00 -   2.00    22 OBSERVATIONS
     NUMERO  12 POUR LES BORNES   2.00 - 400.00    35 OBSERVATIONS
     L'ENTREE NUMERO   5 -LAT -  DEVIENT EN SORTIE
     NUMERO  13 POUR LES BORNES  42.00 -  43.00    43 OBSERVATIONS
     NUMERO  14 POUR LES BORNES  43.00 -  46.00    29 OBSERVATIONS
     NUMERO  15 POUR LES BORNES  46.00 -  54.00    34 OBSERVATIONS
     L'ENTREE NUMERO   6 -LONG-  DEVIENT EN SORTIE
     NUMERO  16 POUR LES BORNES  74.00 -  79.00    42 OBSERVATIONS
     NUMERO  17 POUR LES BORNES  79.00 -  82.00    29 OBSERVATIONS
     NUMERO  18 POUR LES BORNES  82.00 -  94.00    35 OBSERVATIONS
     L'ENTREE NUMERO   7 -YPR5-  DEVIENT EN SORTIE
     NUMERO  19 POUR LES BORNES   4.00 -  13.00    16 OBSERVATIONS
     NUMERO  20 POUR LES BORNES  13.00 -  17.00    18 OBSERVATIONS
     NUMERO  21 POUR LES BORNES  17.00 -  18.00     8 OBSERVATIONS
     NUMERO  22 POUR LES BORNES  18.00 -  20.00    13 OBSERVATIONS
     NUMERO  23 POUR LES BORNES  20.00 -  22.00    14 OBSERVATIONS
     NUMERO  24 POUR LES BORNES  22.00 -  25.00    13 OBSERVATIONS
     NUMERO  25 POUR LES BORNES  25.00 -  29.00    13 OBSERVATIONS
     NUMERO  26 POUR LES BORNES  29.00 -  40.00    11 OBSERVATIONS
```

Tableau B1 : Découpage des variables en classes

S1	0	1	0	0	0	1	0	1	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	0	0	0
S2	0	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	0	0	1	0	0	0
S3	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	0	0	1	0	0	0
S4	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0
S5	0	0	1	0	1	0	0	0	1	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	0
S6	0	1	0	0	1	0	0	0	1	0	0	1	0	1	0	0	0	1	0	0	0	0	1	0	0	0
S7	0	1	0	0	1	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	1	0
S8	1	0	0	0	0	1	1	0	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0
S9	0	1	0	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0
S10	1	0	0	0	0	1	1	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0
S11	0	1	0	0	0	1	0	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	0	0	0	0
S12	1	0	0	0	0	1	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0
S13	0	0	1	0	1	0	0	0	1	1	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0
S14	1	0	0	0	0	1	1	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0
S15	1	0	0	0	0	1	1	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0
S16	0	1	0	1	0	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0
S17	0	1	0	0	0	1	0	1	0	0	0	1	0	1	0	1	0	0	0	1	0	0	0	0	0	0
S18	1	0	0	0	0	1	1	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0
S19	0	1	0	0	1	0	0	1	0	0	0	1	0	1	0	0	1	0	0	0	0	0	1	0	0	0
S20	1	0	0	0	1	0	1	0	0	0	1	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0
S21	0	0	1	0	1	0	0	0	1	1	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0
S22	0	0	1	0	1	0	0	1	0	0	0	1	0	1	0	0	1	0	0	0	0	0	1	0	0	0
S23	0	1	0	0	0	1	0	1	0	0	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0
S24	1	0	0	0	0	1	1	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0
S25	0	1	0	0	1	0	0	1	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0
S26	0	1	0	0	1	0	0	1	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1
S27	0	1	0	1	0	0	0	0	1	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1
S28	0	0	1	0	0	1	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0
S29	0	1	0	0	1	0	0	1	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0
S30	0	1	0	0	1	0	0	1	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0
S31	0	1	0	0	1	0	0	1	0	1	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0
S32	0	1	0	0	1	0	0	1	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0
S33	0	1	0	0	1	0	0	1	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0
S34	1	0	0	0	1	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0
S35	0	1	0	0	1	0	0	1	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1
S36	0	1	0	0	1	0	0	1	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1
S37	0	1	0	0	1	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1
S38	1	0	0	0	1	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0
S39	1	0	0	0	1	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1
S40	1	0	0	0	0	1	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0
S41	1	0	0	0	1	0	0	1	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1
S42	0	1	0	1	0	0	0	1	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1
S43	0	1	0	1	0	0	0	0	1	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1
S44	1	0	0	1	0	0	0	1	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1
S45	0	1	0	1	0	0	0	1	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1
S46	0	1	0	1	0	0	0	1	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0
S47	1	0	0	0	0	1	1	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0
S48	1	0	0	0	0	1	1	0	0	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0
S49	1	0	0	0	0	1	0	1	0	0	1	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0
S50	1	0	0	0	0	1	1	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0
S51	1	0	0	0	0	1	1	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0
S52	1	0	0	0	0	1	1	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0
S53	1	0	0	0	0	1	1	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0
S54	0	1	0	0	1	0	0	1	0	0	1	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0
S55	1	0	0	0	0	1	1	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0
S56	1	0	0	0	0	1	1	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0
S57	1	0	0	0	0	1	1	0	0	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0
S58	1	0	0	0	0	1	1	0	0	1	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0
S59	1	0	0	0	0	1	1	0	0	1	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0
S60	1	0	0	0	0	1	1	0	0	0	1	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0
S61	1	0	0	0	0	1	1	0	0	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0
S62	0	1	0	0	0	1	1	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0
S63	0	0	1	0	0	1	1	0	0	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0
S64	1	0	0	0	0	1	1	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0
S65	1	0	0	0	0	1	1	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0
S66	1	0	0	0	1	0	1	0	0	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0
S67	1	0	0	0	0	1	1	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0
S68	1	0	0	0	0	1	1	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0
S69	0	1	0	0	0	1	0	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	1	0	0
S70	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0
S71	1	0	0	0	1	0	1	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0
S72	1	0	0	0	1	0	1	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0
S73	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	1	0	0	0	0	0	1	0	0	0	0
S74	0	1	0	1	0	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0
S75	0	1	0	1	0	0	0	1	0	1	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0
S76	0	1	0	1	0	0	1	0	0	0	0	1	0	1	0	1	0	0	0	0	1	0	0	0	0	0
S77	0	1	0	1	0	0	0	1	0	1	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0
S78	0	1	0	1	0	0	0	1	0	1	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0
S79	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	0	0	0	0	0
S80	0	1	0	1	0	0	0	1	0	0	1	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0
S81	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	0	0	0	0	0
S82	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	0
S83	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	0
S84	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	0	0	0	0	0
S85	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	0	0	0	0	0
S86	0	0	1	0	0	1	0	0	1	1	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0
S87	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	0	0	0	0	0
S88	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	0	0	0
S89	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	0
S90	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	0	0	0	0	0
S91	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	0
S92	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	0
S93	0	0	1	1	0	0	0	0	1	0	1	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0
S94	0	0	1	1	0	0	0	0	1	0	1	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0
S95	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	0	0	0	0	0
S96	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	0
S97	0	0	1	0	1	0	0	0	1	1	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0
S98	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	0	0	0	0	0
S99	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0
S100	1	0	0	1	0	0	0	1	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0
S101	0	1	0	1	0	0	0	1	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0
S102	0	0	1	1	0	0	0	1	0	0	1	0	0	1	0	0	1	1	1	0	0	0	0	0	0	0
S103	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	0	0	0	0	0
S104	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0
S105	0	0	1	0	1	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0
S106	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	0

Tableau B2 : Le tableau disjonctif complet obtenu à l'aide du programme Addad

Programme Addad

```
$RUN TABACO
$L080
$PRT=ztabaco.sor
$PAR=.
   TITRE CONSTRUCTION D'UN TABLEAU DE BURT SUR NOS DONNÉES;
   PARAM NI=106 NJ=26 NJL=18 NJC=8 LECIJ=1 STCR=1  STIJ=1 ;
   OPTIONS IOUT=1;
   LIGNE 1-18 ;
   COLONNE 19-26;
   LISTE IDEN(1,4)
      AI1(5,2)   AI2(7,2)   AI3(9,2)
      LC1(11,2)  LC2(13,2)  LC3(15,2)
      PC1(17,2)  PC2(19,2)  PC3(21,2)
      SL1(23,2)  SL2(25,2)  SL3(27,2)
      LA1(29,2)  LA2(31,2)  LA3(33,2)
      LO1(35,2)  LO2(37,2)  LO3(39,2)
      YP1(41,2)  YP2(43,2)  YP3(45,2)
      YP4(47,2)  YP5(49,2)  YP6(51,2)
      YP7(53,2)  YP8(55,2);
$F11=ZZACM.DAT
$F21=ZZBURT.DAT
$F22=ZZAFC.DAT
$END
```

TABLEAU CROISE $B_{J_e J_0}$

	YP1	YP2	YP3	YP4	YP5	YP6	YP7	YP8
AI1	0	8	2	6	7	6	4	3
AI2	2	2	5	2	4	7	6	8
AI3	14	8	1	5	3	0	3	0
LC1	13	8	3	5	1	2	4	5
LC2	2	1	1	3	5	5	7	6
LC3	1	9	4	5	8	6	2	0
PC1	0	7	3	7	6	6	6	2
PC2	2	3	4	2	4	7	5	7
PC3	14	8	1	4	4	0	2	2
SL1	3	7	2	8	5	8	5	11
SL2	3	2	2	2	4	4	5	0
SL3	10	9	4	3	5	1	3	0
LA1	1	6	3	5	4	6	7	11
LA2	0	4	3	5	8	5	4	0
LA3	15	8	2	3	2	2	2	0
LO1	0	10	6	9	7	7	3	0
LO2	2	0	0	1	4	4	7	11
LO3	14	8	2	3	3	2	3	0

Tableau B3 : Obtention du tableau de BURT à l'aide du programme Addad

Programme Addad

```
$RUN ANCORR
$L080
$PRT=ZANCORR.SOR
$PAR=.
   TITRE AFC SUR NOS DONNEES;
   PARAM NI=124 NJ=8 NF=5 NI2=106 STFI=3;
   OPTIONS IMPFI=1 IMPFJ=1 NGR=5;
   GRAPHE X=1 Y=2 GI=1;
   GRAPHE X=1 Y=3 GI=1;
   GRAPHE X=1 Y=4 GI=1;
   GRAPHE X=1 Y=3 GI=1 GJ=1;
   GRAPHE X=1 Y=3 GI=3 GJ=1;
   LISTE IDEN(1,4)
     YP1(7,4) YP2(13,4) YP3(20,4)
     YP4(26,4) YP5(32,4) YP6(38,4)
     YP7(44,4) YP8(49,4);
$F11=ZBURACM.DAT
$F21=ZFACT.FAI
$F22=ZFACSUP.FAI
$END
```

LES VALEURS PROPRES VAL(1)= 1.00000

NUM	VAL PROPRE	POURC.	CUMUL	VARIAT.	HISTOGRAMME DES VALEURS PROPRES
2	.25389	70.07	70.075	*******	**********
3	.06425	17.73	87.809	52.340	**********
4	.02225	6.140	93.949	11.595	**********
5	.01319	3.640	97.589	2.500	******
6	.00509	1.405	98.994	2.235	**
7	.00321	.886	99.881	.5191	*
8	.00043	.119	100.00	.7671	*

COORDONNÉES DES POINTS LIGNE

	I1	QLT	POID	INR	1#F	COR	CTR	2#F	COR	CTR	3#F	COR	CTR
1	AI1	971	58	40	-395	626	36	233	217	49	-170	116	75
2	AI2	995	49	55	-436	470	371	-268	177	55	375	347	311
3	AI3	996	60	95	744	957	130	-5	0	0	-144	36	56
4	LC1	998	63	66	576	884	83	-37	4	1	76	15	16
5	LC2	991	42	49	-420	422	29	-457	499	137	-168	67	53
6	LC3	983	61	43	-305	367	22	352	489	118	37	5	4
7	PC1	985	61	37	-421	807	43	125	70	15	-112	58	35
8	PC2	998	47	38	-398	546	30	-189	123	26	296	303	187
9	PC3	997	58	100	773	958	136	23	1	0	-124	25	40
10	SL1	990	67	26	-254	453	17	77	42	6	-127	113	48
11	SL2	970	39	17	-221	310	7	-307	600	57	43	12	3
12	SL3	986	61	41	416	713	42	109	49	11	111	51	34
13	LA1	952	56	30	-377	748	32	-57	17	3	-12	1	0
14	LA2	953	51	45	-492	759	48	112	39	10	-83	21	16
15	LA3	999	60	102	776	968	141	-41	3	2	81	11	18
16	LO1	989	74	62	-364	431	38	402	526	185	71	16	17
17	LO2	993	32	80	-433	207	23	-805	714	320	-240	64	82
18	LO3	997	61	76	661	976	105	-66	10	4	39	3	4
			1000			1000			1000			1000	

COORDONNÉES DES POINTS COLONNES

	J1	QLT	POID	INR	1#F	COR	CTR	2#F	COR	CTR	3#F	COR	CTR
1	YP1	999	168	480	998	964	660	-185	33	90	30	1	7
2	YP2	942	189	73	194	268	28	306	665	275	-25	5	5
3	YP3	992	84	61	-224	193	17	160	97	33	417	665	656
4	YP4	939	137	49	-129	127	9	239	439	121	-166	211	169
5	YP5	997	147	72	-320	580	59	38	8	3	-110	68	80
6	YP6	1000	137	138	-563	870	171	-112	34	271	94	24	55
7	YP7	997	137	122	-318	312	55	-454	635	438	-68	14	28
8	YP8	743	1	5	-544	177	1	-839	420	13	-71	3	0
			1000			1000			1000			1000	

Tableau B4 : Obtention des résultats de l'AFC à l'aide du programme Addad

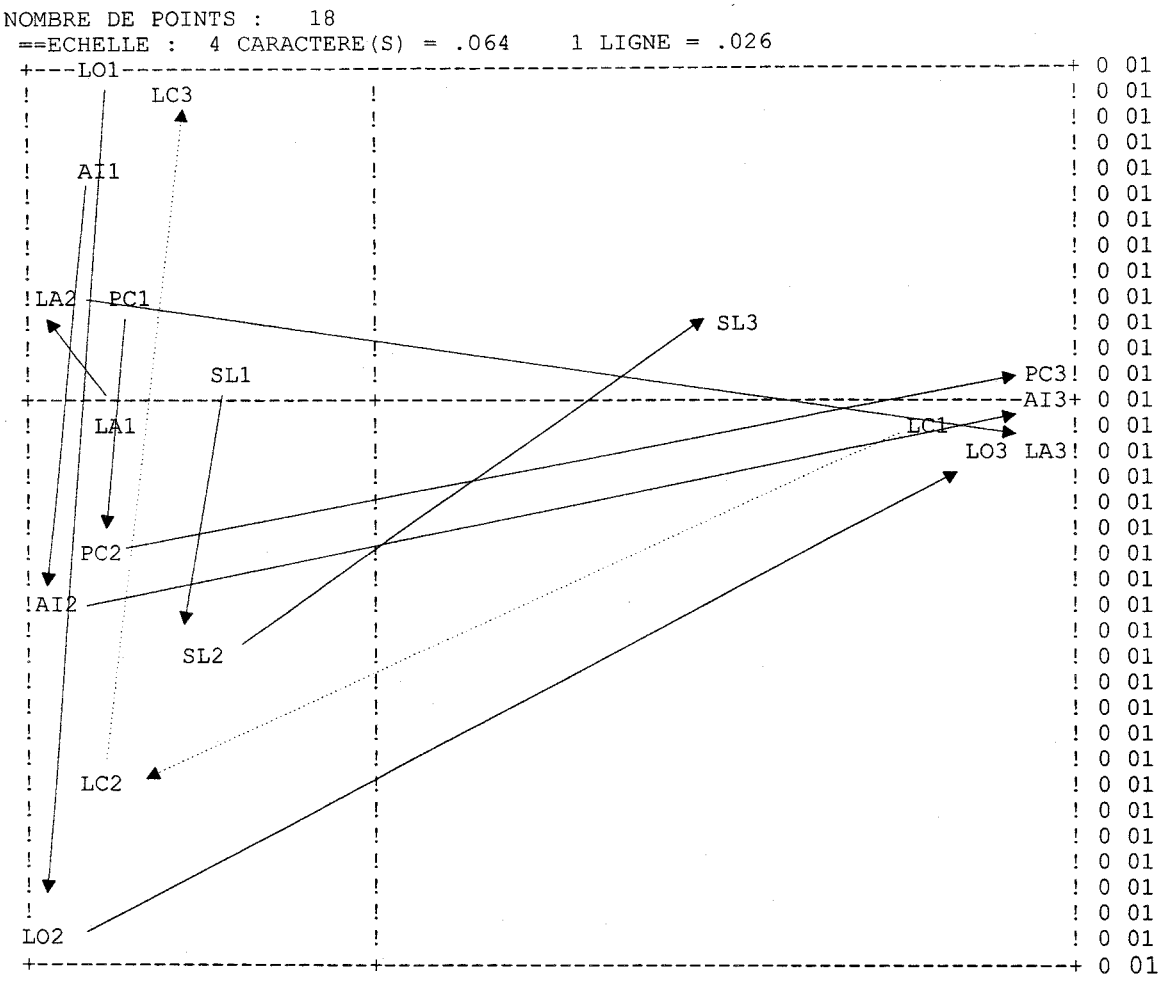


Figure B1 : Représentation des modalités des variables explicatives dans le repère formé par les deux premiers facteurs retenus (suite du programme inscrit au tableau B4)

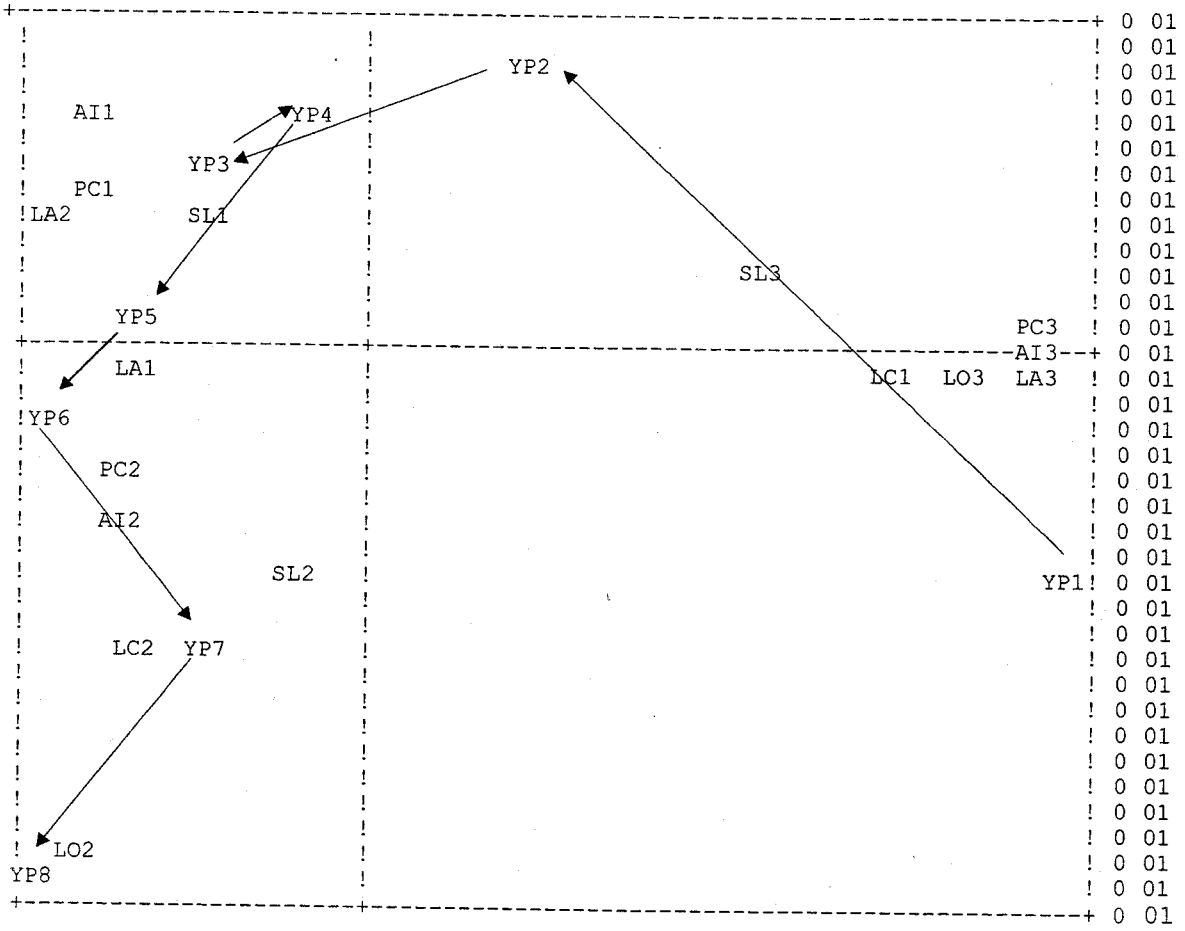


Figure B2 : Représentation simultanée des deux espaces dans le repère formé par les deux premiers facteurs retenus (suite du programme inscrit au tableau B4)