UNIVERSITE MONTPELLIER II

ORSTOM ENSAM

DEA DE BIOSTATISTIQUE
RAPPORT DE STAGE DE FIN D'ANNEE

PREVISION DES CRUES DE L'AMAZONE à MANAUS

par CATHERINE BEAL

Soutenu devant le jury composé de :

A. BERLINET
G. CARAUX
J.P. DAURES
R. SABATIER
G. VIGNAU

le 29 juin 1998

UNIVERSITE MONTPELLIER II

ORSTOM ENSAM

DEA DE BIOSTATISTIQUE
RAPPORT DE STAGE DE FIN D'ANNEE

PREVISION DES CRUES DE L'AMAZONE à MANAUS

par CATHERINE BEAL

Soutenu devant le jury composé de :

A. BERLINET
G. CARAUX
J.P. DAURES
R. SABATIER
G. VIGNAU

le 29 juin 1998

SOMMAIRE

Introduction

LE POURQUOI DE L'ETUDE

Le problème des crues de l'Amazone à Manaus

Les 2 500 000 habitants de Manaus, capitale de l'état Amazonas (Brésil), subissent annuellement les effets dévastateurs et prolongés des crues de l'Amazone. Ce fleuve constitue le plus important de la planète. Ses affluents les plus importants sont au nombre de trois : le Rio Negro, le Rio Madeira et le Rio Branco (voir carte, annexe 1).

Quelques chiffres

Le bassin de l'Amazone couvre plus de 6 000 000 de km^2, et s'étale sur six pays : le Brésil (68% du bassin), la Bolivie, l'Equateur, la Guyane, le Pérou et le Vénézuela. Avec un débit annuel moyen estimé à 209 000 m^3/s, il contribue à 20% des apports mondiaux d'eau douce aux océans. Son réseau hydrographique est particulièrement complexe. C'est donc un fleuve record, également en ce qui concerne sa longueur (6518 km), qui le place au deuxième rang mondial après le Nil (6671 km); la largeur de son lit principal (le Rio Solimoes), qui peut atteindre 4 ou 5 km; et sa profondeur, située aux alentours de 40 ou 50 m et pouvant atteindre jusqu'à 100 m.

Dénivelée du bassin

La pente, très faible sur tout le réseau hydrographique, accentue la gravité de la crue. A Iquitos, situé à environ 3 600 km de l'Atlantique, le niveau d'eau de la rivière est à environ 100 m au-dessus du niveau de la mer en saison sèche. Il n'est plus qu'à 15 m au confluent du Rio Negro (encore à 1 500 km de l'océan). La dénivelée moyenne est alors estimée à environ 1 cm/km. En saison des pluies, ce taux ne dépasse pas 2 cm/km.

Ce phénomène provoque, au niveau de la confluence du Rio Madeira, la formation d'un étang caractérisé par de violents remous. Ainsi, malgré la distance de ce point à Manaus (de l'ordre de 60 km), et malgré sa situation en aval de la ville, la crue du Madeira non seulement ralentit l'écoulement des eaux à partir de Manaus une fois la crue amorcée, mais participe à la montée des eaux.

La saison des pluies

Les précipitations se déplacent du sud vers le nord. Ainsi, le premier affluent affecté par la crue est le Madeira, en mars-avril. Le Rio Negro atteint quant à lui son maximum vers juin-juillet. Les contributions cumulées de tous les affluents provoquent alors une crue unique annuelle et étalée sur une période de trois ou quatre mois, d'avril à juillet, à l'aval de Manaus.

Une solution : la prévision

Devant la gravité de la situation, les autorités brésiliennes (Défense Civile) ont contacté le DNAEE (département national de l'eau et de l'énergie électrique du Brésil) et lui ont demandé d'élaborer un outil de prévision des crues à Manaus. Depuis plus de

quinze ans, sept cents postes pluviométriques et quatre cents stations limnimétriques ont été réparties sur le bassin. Les stations de télétransmission ARGOS, permettant le suivi en temps réel des niveaux d'eau des formateurs de l'Amazone, ont été installées en coopération avec l'ORSTOM en 1985.
Ne disposant pas de données assez nombreuses sur les caractéristiques hydrauliques du fleuve (débits, largeur du lit, ...) pour établir un modèle physique, une modélisation de type statistique est entreprise.

Objectif de l'étude
Est recherché un modèle de prévision du niveau d'eau à Manaus pour un délai de 30 jours. Le problème posé est donc l'établissement d'une relation expliquant le niveau d'eau à Manaus un jour J par les niveaux d'eau sur les stations en amont de Manaus à des dates antérieures à J-30.

LES DONNEES

Les variables
Une variable est caractérisée par une station et un décalage temporel par rapport au jour de prévision.
Introduire plusieurs décalages en temps par station sert à prendre en compte l'évolution de la crue. Toutefois dans le cadre de la prévision du niveau d'reau à Manaus se pose le problème de la sélection des variables : s'il relève du bon sens de considérer que les stations situées sur le Negro et sur le Solimoes risquent d'être les plus importantes dans le signal de crue à Manaus, le choix de décalages temporels pertinents s'avère bien plus délicat.
Il s'agit donc de mettre sur pied une technique de sélection de variables explicatives, efficace et adaptée aux circonstances présentes.

Nous rencontrerons trois difficultés majeures :

- le caractère doublement corrélé des variables,
- la présence de données manquantes,
- la taille des échantillons

Les corrélations entre variables sont à la fois spatiales et temporelles : spatiales car deux variables se rapportant à deux stations limnimétriques différentes peuvent être significativement corrélées (position sur le même affluent, soumises aux mêmes conditions de précipitation, ...), temporelles puisque les données sont des séries chronologiques au pas de temps journalier donc autocorrélées. La stationnarité des séries est mise en évidence par le tracé de leur autocorrélogramme (nous avons seulement tracé celui de Manaus, le phénomène sous-jacent régissant la crue étant supposé le même sur toutes les stations).

Les échantillons dont nous disposons comportent des données manquantes. Pour éviter de gaspiller de l'information, il faut donc que la procédure de modélisation en tienne compte.

La taille des échantillons constitue également un obstacle.
Nous disposons des niveaux d'eau sur trente-neuf des stations du bassin, pratiquement tous les jours depuis 15 ans. Ceci nous amènera à manipuler, pour un délai de 30 jours, des tableaux d'environ 2 000 lignes pour 90 colonnes. Nous verrons que la technique que nous allons utiliser pour sélectionner les variables est en général appréciée par la validation croisée. Ce procédé perd en efficacité lorsque les observations sont trop nombreuses par rapport au nombre de variables.

Travail antérieur sur le sujet et ses lacunes
La précédante étude, menée par une équipe d'hydrologues (H. Thepaut, 1994), a conduit à des modèles de type linéaire pour différents délais de prévision. Ces résultats ont été obtenus par régression pas-à-pas. Les modèles générés présentent de bonnes propriétés d'ajustement ainsi que de prévision.
Toutefois, les conditions de validité de la régression multiple n'ont pas été vérifiées (autocorrélations spatio-temporelles), et la présence de données manquantes a nécessité le développement d'une procédure spécifique afin d'utiliser au mieux toute l'information disponible pour l'identification du modèle.
Les hydrologues se sont donc adressés aux statisticiens, afin de disposer d'une méthode tenant compte de la présence de variables corrélées aussi bien spatialement que temporellement, et admettant directement des données manquantes dans l'algorithme de détermination du modèle.

L'étude initiale a cependant apporté des éléments que nous allons utiliser :

- Pour un délai de 30 jours, la considération de stations du Rio Madeira n'améliore pas de façon intéressante la qualité prévisionnelle du modèle. Nous n'en tiendrons pas compte dans notre recherche.
- Elle fournit un critère permettant de présélectionner un premier lot de variables. Cette idée sera présentée en début de deuxième partie.

Etapes de notre travail
Nous commencerons, dans une première partie, par présenter les techniques statistiques. Nous montrerons en particulier en quoi la méthode utilisée lors du travail de H. Thepaut (régression pas à pas) n'est théoriquement pas valide dans les circonstances de la prévision des crues à Manaus. Nous nous attacherons ensuite à exposer une autre technique de régression, dite régression Partial Least Square, qui a fait ses preuves, et plus compatible avec les circonstances présentes (fortes auti-corrélations). Nous proposerons un nouvel algorithme de sélection de variables. Ces procédés feront ensuite l'objet de quelques modifications pour permettre d'intégrer la présence des données manquantes.

Nous passerons enfin à l'établissement de modèles de prévision à 30 jours, d'abord en restreignant les échantillons de façon à éliminer les données manquantes, puis en exploitant la totalité de l'information.

1ère partie :

Exposition des techniques statistiques

A/ METHODE PARTIAL LEAST SQUARE

Nous exposerons tout d'abord les raisons de l'inapplicabilité de la régression multiple à notre problème, c'est-à-dire le cadre théorique idéal de validité de cette méthode. Ensuite viendra la présentation de la technique Partial Least Squares, qui exige des conditions d'application moins strictes. Nous montrerons dans un troisième temps comment utiliser cette procédure pour sélectionner des variables explicatives. Enfin, la dernière partie concernera l'analyse des résidus du modèle obtenu.

1-*Les limites de la régression multiple classique*

La régression multiple consiste à construire une relation présupposée linéaire entre une variable à expliquer Y quantitative et un ensemble de variables (dites explicatives, ou régresseurs), X^1, X^2,...X^p, également quantitatives. On dispose de n observations sur ces variables.

Notations :

X la matrice n x p des variables explicatives
Y la matrice n x q des réponses
n le nombre d'observations
p le nombre de variables explicatives
x^j_i valeur de la variable X^j sur l'observation i
y_i valeur de la variable Y sur l'observation i
ε_i erreur aléatoire sur l'observation i
u_i résidu sur l'observation i (estimation de ε_i sur un échantillon)
X' transposée de la matrice X
1 vecteur n x 1 constitué de 1

La relation recherchée s' écrit :

$$y_i = a_0 + a_1 x^1_i + \ldots + a_p x^p_i + \varepsilon_i \qquad (1)$$

On distinguera l'échantillon de calage (ou d'apprentissage) de l'échantillon test (ou de validation). Le premier servira à calculer les coefficients tandis que le second nous permettra de mesurer la qualité du modèle par le calcul des écarts quadratiques moyens (Nous écrirons Eqm). L'intérêt de cette distinction est de permettre d'apprécier la qualité prédictive du modèle final, par le calcul de l'écart quadratique moyen en test, et la qualité de l'ajustement, révélée par l'écart quadratique moyen sur l'échantillon de calage.

Les x^j_i sont supposés connus sans erreur.
Sans cette condition, une régression est toujours possible, mais conditionnellement aux valeurs prises par les régresseurs . On écrit dans ce cas :

$$E(Y/X) = a_0 + a_1 X^1 + \ldots + a_p X^p$$

Les hypothèses sur $(\varepsilon_i)_{i\leq n}$ doivent être mathématiquement claires et plausibles. Représentant la partie "non explicable" par le modèle, la partie "incontrôlable", elle ne devra manifester aucune structure. R. Tomassone (1992), expose trois conditions :

Les variables $(\varepsilon_i)_i$ doivent être

- *indépendantes*
- *de loi normale*
- *de même variance σ^2 (homoscédasticité)*

La première condition suppose qu' *aucune variable ne soit auto-corrélée.* Lorsqu'elle n'est pas vérifiée, on peut toujours apporter une amélioration en écrivant $\varepsilon_i = \alpha\, \varepsilon_{i-1} + \nu_i$, α étant un paramètre à estimer, et en considérant, jusqu'à vérification, que (ν_i) est un bruit blanc. Le modèle (1) devient alors :

$$y_i^* = a_0 + a_1 x^1{}_i^* + \ldots + a_p x^p{}_i^* + \nu_i,$$

où $y_i^* = y_i - \alpha\, y_{i-1}$ et
$x^j{}_i^* = x^j{}_i - \alpha\, x^j{}_{i-1}$

L'hypothèse de normalité des $(\varepsilon_i)_i$ sert à construire des tests et des intervalles de confiance pour les coefficients.

On désigne par $(u_i)_i$ la collection des résidus : $\forall i,\ u_i = y_i - y_{i\ estimé}$
Leur analyse, phase terminale et indispensable à la régression multiple, permet de valider ou d'infirmer le modèle. Elle consiste à étudier la variabilité des résidus réduits en fonction des divers facteurs connus susceptibles d'exercer une influence. A savoir : les régresseurs, l'estimation de Y, et les numéros des observations. Le modèle sera adopté si, sur chaque graphe, les résidus réduits sont répartis aléatoirement autour de 0 (indépendance) et si leur amplitude ne varie pas en fonction de l'objet placé en abscisses (homoscédasticité).

Ecriture matricielle
Pour la simplification de l'exposé, on notera encore X la matrice des régresseurs centrés, et a le vecteur $(a^1,\ldots,a^p)'$.
La technique des moindres carrés ordinaire consiste à minimiser la variance des résidus $\sum u_i^2$, fonction de a. Le système des équations normales s'écrit :

$$X'X\, a = X'Y$$

Sa résolution suppose que $X'X$ soit inversible, donc que X soit injective, c'est-à-dire :

- $n > p$

- *les régresseurs ne doivent pas être multicolinéaires (linéairement dépendants).*

Cette dernière condition est vérifiée dès que les régresseurs sont deux à deux statistiquement indépendants.
La solution, appelée estimateur de Gauss-Markov, est alors donnée par :

$$\hat{a} = (X'X)^{-1} X'Y$$

La constante a_0 est estimée par : $\hat{a}_0 = \bar{y} - \sum_{j \leq p} \hat{a}_j \bar{x^j}$
$\hat{a}$ présente les avantages d'être sans biais et de variance minimale parmi les estimateurs linéaires, donc de précision optimale (Tomassone, 1992).
Sa variance est estimée par $V(\hat{a}) = (X'X)^{-1} s^2$, où s^2 est l'estimateur sans biais de σ^2.

$$s^2 = \frac{(Y - \hat{a}_0 1 - X\hat{a})'(Y - \hat{a}_0 1 - X\hat{a})}{n - p - 1}$$

Dans le cas où les régresseurs sont statistiquement indépendants les uns des autres, chaque coefficient est interprétable, et de même signe que le coefficient de corrélation de Y à la variable correspondante. Ceci n'est plus du tout vérifié lorsque les régresseurs sont multicolinéaires. Même lorsque la matrice de covariance est inversible, elle est souvent très mal conditionnée (déterminant proche de 0) : la variance de $\hat{a}$ devient trop grande et les coefficients très instables.

En pratique, il est quasiment impossible de respecter strictement toutes les conditions qui viennent d'être présentées. En effet, s'il arrive de disposer d'observations indépendantes, il est bien rare que les variables explicatives soient deux à deux non-corrélées. Cette exigence est pourtant la plus importante. Ainsi, il ne s'agit plus de vérifier le strict bien fondé des présupposés, mais de s'en écarter le moins possible.

L'étape suivante, qui consiste à expliquer un tableau de plusieurs variables par un autre, fait l'objet de la régression multivariée.

Il existe des techniques pour réduire la forte variabilité des coefficients de régression. R. Tomassone (1992) traite cette question. Entre autres solutions, il propose le principe de la régression orthogonalisée (analyse en composantes principales des régresseurs, puis régressions sur les composantes, variables indépendantes). Les composantes de plus forte variance n'expliquent cependant pas forcément au mieux les variations des variables de Y.
Le régression Partial Least Square, notée régression PLS, comble cette lacune par la construction d'axes aussi représentatifs que possible des variables de X et les plus liés possible aux variables de Y.

2- *Principe de la régression Partial Least Square*

Cette méthode vise à établir une relation linéaire entre deux tableaux de variables quantitatives mesurées sur les mêmes individus, même en présence de corrélations internes dans chaque groupe de variables, surtout lorsque le nombre de variables est très important par rapport au nombre d'observations.
Elle est introduite pour la première fois par Wold (1966), qui met au point l'algorithme NIPALS, destiné à établir des modèles de régression en présence de fortes corrélations entre régresseurs, et l'améliore par la suite (Wold, 1975). Frank & Friedman (1993), compare les performances en matière de prévision des modèles PLS, de régression orthogonale, de régression pas à pas et de régression pseudo-orthogonale (régression orthogonale avec contrainte sur les variances des coefficients), et concluent à une nette supériorité de la méthode PLS sur la régression pas-à-pas, alors que les régressions orthogonales et pseudo-orthogonales s'avèrent équivalentes. D'autres chercheurs, comme Palm & Iemma (1995), Garthwaite (1994), ont complété ce travail. Ce dernier a également conclu à la supériorité de la régression PLS sur les autres techniques lorsque l'on doit traiter beaucoup de variables et que les résidus possèdent une forte variance.
Tout comme la régression multiple classique (R. Tomassone, 1992), cette méthode présente l'avantage de pouvoir s'appliquer à des variables qualitatives. Les lecteurs intéressés pourront se référer à l'article de Tenenhaus & *al.*(1995), qui traite en détails cette question.
Nous utiliserons ses notations. A savoir :

X : matrice des variables explicatives.
Y : matrice des variables expliquées.
n : nombre d'observations.
p : nombre de variables explicatives.
q : nombre de variables à expliquer.
E_0 : version centrée-réduite de X.
F_0 : version centrée-réduite de Y.
E_k : matrice des résidus de la décomposition de E_0 en utilisant k composantes.
F_k : matrice des résidus de la décomposition de F_0 en utilisant k composantes.

Algorithme

On travaille à partir des matrices centrées-réduites E_0 et F_0.

Etape 0

On recherche deux combinaisons linéaires t_1 (premier axe) et u_1 des colonnes de E_0 et F_0 respectivement, les plus liées possibles : $t_1 = E_0 w_1$ et $u_1 = F_0 c_1$, telles que $cov(t_1,u_1)$ soit maximale sous les contraintes $||w_1||^2 = ||c_1||^2 = 1$.

On effectue ensuite deux régressions : E_0 sur t_1 et F_0 sur t_1.

$$E_0 = t_1 p_1' + E_1$$
$$F_0 = t_1 r_1' + F_1,$$

avec $p_1 = t_1' E_0 / ||t_1||^2$ et $r_1 = c_1' F_0 / ||c_1||^2$.

Ainsi, t_1 et u_1 sont les plus représentatives possibles des données de E_0 et F_0 (respectivement) et t_1 explique au mieux les variables de F_0.

Etape 1

On recommence en travaillant non plus sur E_0 et F_0, mais sur les résidus obtenus à l'étape précédante : E_1 et F_1. On obtient les décompositions suivantes :

$$E_0 = t_1 p_1' + t_2 p_2' + E_2$$
$$F_0 = t_1 r_1' + t_2 r_2' + F_2,$$

avec:

$t_2 = E_1 w_2$

$u_2 = F_1 c_2$

$p_2 = t_2' E_1 / ||t_2||^2$

$r_2 = t_2' F_1 / ||t_2||^2$

Etape k

Recherche de $t_k = E_{k-1} w_k$ et $u_k = F_{k-1} c_k$, avec $||w_k||^2 = 1$, $||c_k||^2 = 1$ et $cov(t_k, u_k)$ maximale, puis régressions de E_{k-1} et F_{k-1} sur t_k :

$$E_{k-1} = t_k p_k' + E_k$$
$$F_{k-1} = t_k r_k' + F_k.$$

On peut continuer ainsi jusqu'à l'explication complète de E_0, c'est-à-dire jusqu'au rang de X, mais nous aborderons cette question plus loin.

<u>Détaillons maintenant la recherche des composantes t_1 et u_1</u>

$t_1 = E_0 w_1$ et $u_1 = F_0 c_1$.

On veut maximiser $\theta_1 = cov(t_1, u_1)$ sous les contraintes $||w_1||^2 = 1$ et $||c_1||^2 = 1$.

Le Lagrangien associé à ce problème s'écrit :

$$L(\alpha, \beta, w_1, c_1) = t_1' u_1 - \alpha (w_1' w_1 - 1) - \beta (c_1' c_1 - 1)$$
$$= w_1' E_0' F_0 c_1 - \alpha (w_1' w_1 - 1) - \beta (c_1' c_1 - 1)$$

On obtient le système des équations normales :

$$E_0' F_0 c_1 - 2 \alpha w_1 = 0 \qquad (1)$$
$$F_0' E_0 w_1 - 2 \beta c_1 = 0 \qquad (2)$$

$$w_1' w_1 - 1 = 0 \qquad (3)$$
$$c_1' c_1 - 1 = 0 \qquad (4)$$

En prémultipliant (1) par w_1' et (2) par c_1', et en tenant compte des deux contraintes (3) et (4), il vient :

$$2\alpha = 2\beta = \theta_1$$
$$E_0' F_0 c_1 = \theta_1 w_1$$
$$F_0' E_0 w_1 = \theta_1 c_1$$

et :

$$E_0' F_0 F_0' E_0 w_1 = \theta_1^2 w_1$$
$$F_0' E_0 E_0' F_0 c_1 = \theta_1^2 c_1$$

w_1 est donc le vecteur propre normé de $E_0' F_0 F_0' E_0$ associé à la plus grande valeur propre, et c_1 est l'équivalent pour le produit des matrices de covariance $F_0' E_0 E_0' F_0$.

Quelques propriétés des composantes PLS

* Relations cycliques:

les composantes et facteurs obéissent aux égalités suivantes :

$$t_h = E_{h-1} w_h$$
$$c_h = F'_{h-1} t_h / \theta_h$$
$$u_h = F_{h-1} c_h$$
$$w_h = E'_{h-1} u_h / \theta_h$$

En programmation, on utilise ces résultats plutôt que la diagonalisation des produits de matrices de covariance, dont les dimensions éventuellement très grandes peuvent allonger inutilement les temps de calcul. Les axes sont donc issus d'un algorithme itératif, qui d'ordinaire converge très vite (de l'ordre de 3 ou 4 itérations) :

Initialisation

$u \leftarrow F^j$

Répéter

$w \leftarrow E'u / || E'u||$

$t \leftarrow E w$

$c \leftarrow F't / ||F't||$

$u \leftarrow F c$

Fin.

Cette méthode converge quelque soit le choix initial de u.

Le critère d'arrêt des itérations peut être son nombre, ou la stabilisation de la covariance entre le t et le u (on se fixe un ε et on s'arrête lorsque $|1\text{-cov}^{\text{étape-1}} / \text{cov}^{\text{étape}}| \leq \varepsilon$)

* Relations d'orthogonalité:

La famille des axes $(t_h)_h$ est orthogonale.

La famille des facteurs associés à X, $(w_h)_h$, est orthonormée.
$\forall l,\ \forall h \leq l$, on a: $t_h' E_l = 0$.

Qualité de l'ajustement

Elle se mesure au pourcentage de variance expliquée de Y, ou coefficient de corrélation multiple.
Si on choisit comme norme matricielle la norme euclidienne, c'est-à-dire $||A||^2 = tr(AA')$, on a, en utilisant k axes, les égalités suivantes :

$$||E_0||^2 = \sum_{j \leq k} ||t_j||^2 ||p_j||^2 + ||E_k||^2$$
$$||F_0||^2 = \sum_{j \leq k} ||t_j||^2 ||r_j||^2 + ||F_k||^2$$

On définit le pouvoir explicatif d'un axe h par le coefficient de détermination de F_0 à cet axe, c'est-à-dire :

$$R^2(F_0,t_h) = ||t_h||^2 ||r_h||^2 / ||F_0||^2$$

Le pourcentage de variance de Y expliqué à l'ordre k est:

$$R^2(F_0,t_1,t_2,...,t_k) = \sum_{j \leq k} R^2(F_0,t_j).$$

Choix du nombre d'axes

Il existe plusieurs critères :

- l'évolution des écarts quadratiques moyens sur les échantillons de test et de calage en fonction du nombre d'axes .
- La validation croisée (méthode de validation interne).

Cette dernière, introduite par M. Stone (1974), et S. Geisser (1974), fut vite adaptée à l'algorithme PLS par Wold (1982), puis par Wold *& al.* (1983). Cette technique rencontre un large succès en modélisation dans de nombreux domaines d'application.
Elle consiste à calculer un critère, le PRESS (PRedictive Sum of Square), en fonction du nombre d'axes pris en compte. La dimension du modèle optimal sera celle qui minimisera cette quantité.

Le PRESS correspondant à la sélection de A axes est défini comme suit :

$$PRESS(A) = \sum_{k \leq A} \sum_{j \leq q} \sum_{i \leq n} (Y_i^j - Y_{i\,est}^j)^2$$

où $Y^j_{i\,est}$ est l'estimation de Y^j_i par le modèle PLS à k composantes calculé sur les matrices de départ auxquelles on a retiré les mesures sur la $i^{ème}$ observation.

Cette quantité perd toutefois en fiabilité lorsque le nombre d'observations est trop important par rapport au nombre de variables. Le PRESS devient trop optimiste. Ceci peut s'expliquer par le fait qu'en présence de trop de lignes, en retirer une particulière a une incidence non significative sur les résultats de la régression. L'application de cette technique requiert, de plus, énormément de mémoire machine pour des problèmes de grandes dimensions, et nécessite de longs temps en exécution. Il faut donc l'adapter. Cramer (1988) présente la validation croisée dans le cadre des études des QSAR (Quantitative Structure-Activity Relationships) et enlève, à chaque étape, non pas une observation, mais α %, choisies aléatoirement. L'incidence du choix des lignes sélectionnées sur le résultat semble être une question à approfondir.
Cette méthode ne semble donc pas pertinente dans notre cas, et nous tracerons donc les Eqm en test et en calage en fonction du nombre d'axes.

Reconstitution de la relation variables explicatives-variables expliquées

A la $k^{ème}$ étape, on a $F_0 = t_1 r_1' + t_2 r_2' + ... + t_k r_k' + F_k$.
Les composantes étant elles-mêmes des combinaisons linéaires des variables de départ, on peut retrouver la relation directe.
En effet, on peut montrer par une récurrence que :

$$\forall h \leq k \quad E_{h-1} = E_0(I-w_1p'_1)(I-w_2p'_2)...(I-w_{h-1}p'_{h-1}).$$

Il vient alors:

$$t_h = E_{h-1} w_h = E_0 w_h^*$$

où $w_h^* = (I-w_1p'_1)(I-w_2p'_2)...(I-w_{h-1}p'_{h-1})\, w_h$.

On obtient finalement

$$F_0 = E_0 \sum_{h \leq k} w_h^* r'_h + F_k$$

Il suffit ensuite de revenir aux variables non centrées réduites.

3- Un algorithme de sélection des variables explicatives

Cette question fait l'objet de nombreuses recherches. On peut citer notamment les procédure de sélection pas à pas, ou stepwise, appliquées dans le contexte de la régression multiple (Tomassone, 1992).

Baroni & *al.* (1992), se focalisent plus particulièrement sur la sélection des régresseurs dans les modèles PLS : ils élaborent la technique GOLPE (Generating Optimal Linear PLS Estimations). Celle-ci permet, outre la sélection de variables, d'améliorer la qualité prédictive du modèle (en utilisant le PRESS comme critère), mais s'avère converger lentement. Martens & Naes (1989), eux, mettent au point une méthode basée sur la mise à 0, dans le vecteur des poids des variables, w, des coordonnées jugées incertaines ou trop peu significatives, au regard d'un certain critère. Lindgren & *al.* (1994), reprend cette idée et la développe. Il met sur pied la technique IVS for PLS (Interactive Variable Selection for PLS), qui permet de sélectionner une variable par axe en tronquant chaque w jusqu'à n'obtenir qu'un réel, ceci à chaque étape de l'algorithme de la régression PLS. L'étape suivante s'en trouve donc modifiée.
Ces techniques, quoiqu'efficaces, utilisent massivement la validation croisée, et ne sont donc pas exploitables dans le cas présent.

Nous proposons une démarche en quatre temps :

- L'exécution de PLS avec toutes les variables,
- La sélection d'un nombre d'axes suffisants,
- La sélection des variables représentatives de ces composantes,
- Une régression multiple sur ce dernier ensemble de régresseurs. Après analyse des résultats (calcul des Eqm en test et en calage), deuxième sélection suivant le même processus ou sélection pas à pas (si l'Eqm n'a pas été amelioré)

La première étape consiste à *appliquer la méthode PLS à toutes les données*, c'est-à-dire en faisant participer toutes les variables, avec un nombre d'axes conséquent (par exemple les 2/3 du nombre de variables)

Il nous faut ensuite *réduire le nombre d'axes*.
Nous tracerons à cet effet le graphe des écarts quadratiques moyens, sur les échantillons de test et de calage, en fonction de la dimension du modèle. Ces écarts se stabiliseront au bout d'un certain nombre de composantes et la dimension sélectionnée sera celle au-delà de laquelle les variations de ces quantités ne seront plus significatives.

Vient enfin la *sélection des variables*.
Le lot final doit représenter «correctement» l'ensemble des axes retenus, compte tenu du pouvoir explicatif de chacun d'eux. Ainsi, plus un axe expliquera le tableau Y (bon coefficient de corrélation multiple), plus on choisira de variables pour le représenter. Ces variables seront celles les plus corrélées, en valeur absolue, à l'axe en question.
Cette démarche *a priori* a dû être adaptée toutefois à notre cas précis, caractérisé par un nombre important de variables (90) mais un nombre faible d'axes au pouvoir explicatif significatif (4 ou 5). En effet, la sélection des variables selon le principe proposé en ne

tenant compte que de ces seuls premiers axes conduit à des Eqm élevés. L'information apportée par les suivants ne peut donc pas être négligée.
Aussi opérerons nous en les trois étapes suivantes :

* Détermination, d'après l'évolution des pouvoirs explicatifs, des axes à représenter individuellement (les plus importants).

* Pour chacun d'eux, sélection des variables les plus représentatives, au sens du coefficient de détermination maximal.
Le nombre de variables à retenir par axe doit aller en diminuant (comme le pouvoir explicatif de l'axe) et respecter au mieux les groupes observés sur les cercles de corrélation.

* Pour les suivants, moins importants dans l'explication de Y individuellement, mais non négligeables dans leur ensemble, nous établirons un tableau recensant pour chaque axe les variables les plus représentatives, et nous sélectionnerons les régresseurs bien représentés dans la majorité des systèmes d'axes (t_i,t_j).

Nous effectuerons ensuite une *régression multiple avec ce lot de variables*. Si les résultats sont satisfaisants (amélioration de l'Eqm sur l'échantillon test et résultat satisfaisant sur l'échantillon de calage), nous pourrons procéder à une *deuxième sélection* en suivant la même démarche, une troisième, etc...jusqu'à ce que les Eqm en test et en calage se stabilisent où commencent à augmenter.

La dernière étape consistera en un *stepwise*.

Cette technique, quoiqu'efficace sur notre exemple de prévision des crues, présente l'inconvénient du caractère arbitraire du choix des variables représentant les axes, en particulier lorsque les inter-corrélations entre variables indépendantes sont trop élevées. Pour les premiers, il faut se fixer une limite inférieure du coefficient de détermination avec l'axe en question. Pour les suivants, il s'agit d'une sélection sur critère qualitatif, l'ordre par qualité de représentation par axe jouant un rôle moindre. Cette technique demande donc à être améliorée.

4- Analyse des résidus

En régression PLS, aucune hypothèse n'est posée sur les résidus. Ils permettent toutefois de déceler, à travers une structure persistante, si tout ce qui peut être expliqué l'a été, ou s'il existe des variables «cachées».

On tracera en particulier pour chaque modèle :

- les résidus réduits en fonction du numéro de l'observation,
- les résidus réduits en fonction des estimations des variables de Y,

- les résidus réduits en fonction des variables explicatives.

Etablir tous ces graphes semble peu raisonnable selon le nombre de variables dont nous disposons. Il faut donc sélectionner. Dans le cas de la prévision des crues à Manaus, nous nous limiterons à celui en fonction du temps, celui en fonction de l'estimation Y_{est} de Y et ceux en fonction de deux ou trois variables explicatives (les plus corrélées à Y_{est}). Ils nous permettront notamment de déceler les auto-corrélations éventuelles des résidus réduits (évolution en fonction du numéro de l'observation, c'est-à-dire du temps), et une dépendance éventuelle par rapport aux variables (autres graphes), auquel cas le modèle serait à améliorer.

B/ PRISE EN COMPTE DES DONNEES MANQUANTES

Dans de nombreuses situations expérimentales, les données à traiter sont incomplètes pour de multiples raisons. Lorsque l'on souhaite effectuer une analyse factorielle sur un groupe de variables quantitatives, et que la matrice des données comporte des valeurs manquantes, les méthodes usuelles ne sont plus applicables. Ignorer les lignes incomplètes suppose de gaspiller de l'information. Il importe donc d'adapter ces techniques de façon à exploiter au mieux toute l'information disponible.

Nous essaierons dans un premier temps de dresser une liste des grandes classes de traitements tirés de la littérature. Nous exposerons ensuite les démarches qui nous semblent les plus adaptées à notre problématique et desquelles nous tirerons les différents modèles.

1-Approche bibliographique

D'après Roderick J.A.L. (1992), et le dictionnaire de statistiques par Kotz *& al.*(1981), qui passent en revue les diverses techniques permettant d'intégrer la présence de données manquantes dans l'établissement des modèles, on peut dégager six grands types :

- *Complete-case analysis (CCA)*

C'est le traitement standard. Il s'agit ici tout simplement d'écarter les valeurs manquantes et d'établir les modèles sur les données disponibles. Etudiée par Glynn R.J. *& al.* (1986) dans le cas où les valeurs manquantes sont concentrées sur la matrice des régresseurs, cette pratique présente l'avantage de fournir des modèles de bonne qualité lorsque le pourcentage de données manquantes reste raisonnable, c'est-à-dire ne conduit pas à une trop forte réduction de l'échantillon. Elle entraîne cependant inévitablement un gaspillage d'information.

- *Available-case analysis (ACA)*

On se sert ici du maximum de cases pour estimer les paramètres. Glasser (1964) calcule les estimations des premiers moments en utilisant, pour chaque variable, les données présentes : la moyenne est approchée par la moyenne empirique sur les cases complètes, de même pour la variance, et la covariance entre deux variables par la covariance empirique sur les observations simultanément disponibles.
L'inconvénient de cette procédure est que la matrice de covariance d'un ensemble de régresseurs perd son caractère défini-positif, en particulier en présence de fortes inter-corrélations et n'est alors plus inversible.
Des simulations effectuées sur données plus ou moins inter-corrélées par Haïtovsky (1968) et Kim & Curry (1968) ont permis de conclure à la supériorité de l'ACA sur la CCA lorsque les inter-corrélations sont faibles. Dans le cas contraire, c'est la CCA qui l'emporte.

- *Imputation-based procedures*

On remplace les valeurs manquantes par des estimations, et on applique les méthodes classiques.

Il existe donc autant de possibilités que de procédés d'estimation :

- mean imputation :

Estimation d'une donnée manquante par la moyenne de valeurs pertinemment choisies et disponibles sur la même variable.

- regression imputation :

On effectue une régression de la variable «pathologique» (présentant une donnée manquante pour une observation particulière) sur les variables pour lesquelles cette observation est présente. On prédit ensuite la valeur manquante grâce au modèle obtenu. Mais, si pour cette observation, les valeurs des régresseurs sont loin de leur moyenne respective, l'estimation sera de mauvaise qualité.

Cette technique doit être utilisée avec précaution, tous les contextes ne se prêtant pas à l'application d'une régression. Elle suscite cependant l'idée plus générale d'établir un modèle, plus forcément linéaire, permettant d'expliquer la variable «pathologique» par les autres, et de remplacer la valeur manquante par sa prédiction.

- interpolation imputation :

On trace la courbe de la variable «pathologique» en fonction du rang (numéro) de l'observation pour les variables présentes et on l'interpole à un ordre choisi d'après la régularité du graphe (si l'allure semble linéaire, une interpolation d'ordre 1 suffira). On estime ensuite l' élément manquant par la valeur interpolée.

Attention : ceci suppose que la série soit indexée sur un ensemble strictement ordonné, le temps par exemple.

Toutes ces techniques ne tiennent pas compte des erreurs sur les estimations. Rubin (1978, 1987) propose une solution : la *multiple imputation*. Les lecteurs intéressés pourront se reporter à ses articles.

- *Weighting procedures* :

Il s'agit ici de minimiser l'incidence de la présence de données manquantes en pondérant chaque ligne par un poids, fonction croissante de son nombre de données présentes. On applique ensuite une technique compatible avec la présence de valeurs manquantes.

Il est bien sûr possible de combiner les techniques de substitution et de pondération. On considère alors les cases complétées comme moins fiables que les réelles, et on leur assigne un poids moindre, en suivant la même logique que celle des weighting procedures. En d'autres termes, on calcule d'abord la matrice des poids, on estime les valeurs manquantes, puis on applique la méthode statistique sur les données complétées pondérées.

- *Patterns of missing data* :

Ces techniques utilisent des modèles intégrant directement les données manquantes et basent les inférences sur la maximisation de la vraisemblance. Les algorithmes sont itératifs. On peut citer l'algorithme EM (Expectation-maximisation), qui s'applique à n'importe quel modèle de données manquantes. Elaboré par Dempster *& al.* (1977), suite aux travaux de Orchard & Woodbury (1972), il se décompose en deux étapes :

◊ le E-step : Calcul de la log-vraisemblance du modèle sachant les données observées et les estimations des paramètres de base (moyennes, variances, ...).

◊ Le M-step : Estimation finale des paramètres d'intérêt par la maximisation de la fonction obtenue.

Muthen *& al.* (1987), Little (1988a), et Azen *& al.* (1989) comparent les performances en qualité d'estimation ponctuelle de l'algorithme EM et de l'ACA. Les simulations ont mis en évidence la supériorité de la méthode EM.

- *Bayesian methods :*

Ce type de traitement s'applique surtout aux petits échantillons, sur lesquels les autres méthodes fournissent des résultats médiocres. Une idée est de considérer une vraisemblance *a priori* et de baser les inférences sur la distribution *a posteriori*. La question est épineuse et les travaux sur ce type d'approche appliquée plus particulièrement aux modèles de régression lorsque les données manquantes figurent dans les régresseurs paraissent limités. Nous ne développerons pas plus. Les articles de Chen (1986), et Guttman *& al.* (1983), qui traitent des cas de régression multivariée dans le contexte bayésien, peuvent être consultés pour d'avantage de précisions.

2- Les essais :

L'analyse des cases disponibles fait l'objet du paragraphe A/.

En ce qui concerne la prise en compte des données manquantes, nous construirons les modèles selon plusieurs démarches, puis nous comparerons les résultats.

1.*Combinaison substitution-pondération* :

Nous calculerons d'abord les poids à associer aux lignes, selon le principe que nous allons développer. Nous remplacerons ensuite une donnée manquante sur une variable v, un jour j, une année a, par la moyenne des valeurs disponibles de v le jour j, les autres années. Nous appliquerons enfin la technique de sélection des variables développée au paragraphe A/.

Calcul de la matrice des poids :

Notons :

X	la matrice des variables explicatives, avec données manquantes
n	Nombre de lignes de X
p	Nombre de variables explicatives.

L_c l'ensemble des indices des lignes complètes
n_c le nombre de lignes entièrement connues
L_{na} l'ensemble des indices des lignes incomplètes
na_i le nombre de valeurs manquantes sur la ligne i
Na le nombre total de valeurs manquantes.
p_i le poids de l'observation i .

Les p_i doivent vérifier :

$\forall i \in L_c$, p_i est constant et proportionnel à $1/n$.

$\forall i \in L_{na}$, p_i est proportionnel à $(1 - na_i /p)$

$\sum_{i \le n} p_i = 1$ (3)

Le poids moyen des lignes de L_{na} doit être inférieur à celui des lignes de L_c. (4)

Posons

$$\forall i \in L_c \qquad p_i = \frac{\alpha}{n} \text{ , et}$$

$$\forall i \in L_{na} \qquad p_i = \frac{\beta\left(1 - \frac{na_i}{p}\right)}{n\alpha},$$

avec α et β des paramètres à déterminer.
Les conditions (3) et (4) s'écrivent :

$$\begin{cases} \sum_{i \le n} p_i = 1 \\ \frac{\sum_{i \in L_{na}} p_i}{n - n_c} \le \frac{\sum_{i \in L_c} p_i}{n_c} \end{cases}$$

La résolution de ce système d'inéquations conduit à :

$$\begin{cases} 1 \le \alpha \le \frac{n}{n_c} \\ \forall i \in L_{na} \qquad p_i = \frac{(n - \alpha n_c)(p - na_i)}{n(p(n - n_c) - Na)} \end{cases}$$

Si $\alpha = 1$, la moyenne des poids des lignes complètes est égale à celle des poids des lignes incomplètes. Comme le poids vaut 1/n sur une observation connue, on aura obligatoirement des lignes incomplètes plus pesantes que les lignes complètes. Ce choix est donc à écarter.
Si $\alpha = n/n_c$, les poids des lignes incomplètes sont nuls, et l'étude se ramène à une analyse sur les cases complètes.

Nous choisirons donc un compromis entre ces deux valeurs, c'est-à-dire :

$$\alpha = \frac{1 + n\ n_c}{2}$$

On obtient ainsi les poids suivants :

$$p_i = \frac{(n - n_c)(p - na_i)}{2n(p(n - n_c) - Na)}$$

Estimation des cases manquantes :
Dans le cas où, pour un jour j, les valeurs font défaut pour ce même jour sur toutes les années, aucune estimation ne peut être menée. Les lignes correspondant à ce jour seront alors supprimées, ce qui constitue une perte d'information.

2.*Modèles de valeurs manquantes* :
On applique ici la procédure ACA.
On choisit les conventions suivantes :

*La moyenne et la variance d'une variable avec données manquantes sont approchées par la moyenne et la variance empiriques calculées à partir des données présentes.

*La covariance entre deux variables incomplètes est estimée par la covariance empirique sur les observations simultanément présentes.

* Le produit scalaire entre un vecteur complet et un deuxième incomplet de même dimension est assimilé au produit scalaire des vecteurs réduits aux observations disponibles.

Cette dernière approximation, la plus biaisée puisque n'étant pas divisée par un nombre d'observations, peut entraîner de lourdes erreurs au niveau du calcul des axes. En effet, chaque composante *t* est le résultat d'une série d'itérations et à chaque étape, les résidus issus de *X*, contenant autant de données manquantes que la matrice de départ, sont multipliés par le facteur *w*. Ainsi, étape après étape, de fortes erreurs s'accumulent et le résultat final n'est plus du tout fiable. Pour éviter ce phénomène, quitte à allonger les temps d'exécution, on préfère calculer les facteurs *w* par diagonalisation du produit des matrices de covariance.

Nous établirons ce modèle, avec les pondérations calculées précédemment.

2ème partie :

Application à la prévision des crues de l'Amazone à Manaus, établissement de modèles de prévision à 30 jours

A/GENERALITES

La variable "niveau d'eau à Manaus"

Il s'agit d'une *série chronologique* à temps discret, à pas de temps régulier (le jour julien), non déterministe.

Le limnigramme des années 1978 à 1992 (annexe 2) porte à penser que cette série est *stationnaire* (apparemment sans tendance générale). On le vérifie en traçant les autocorrélogrammes à partir de différentes dates (annexe 2).

Choix du modèle

Etablir un *modèle de type linéaire* constitue une première approche. En l'absence d'éléments, on pense au plus simple. Le travail de H. Thepaut (1994) a en outre montré que ce choix n'est pas aberrant, étant donné la qualité des résultats obtenus. Nous verrons que la technique PLS, permet, entre autres, de "confirmer" la pertinence de cette option.

La stationnarité implique déjà que les coefficients du modèle seront indépendants du temps.

Les données

Les mesures de niveau d'eau que nous utiliserons sont journalières, s'étalent sur 15 ans (de 1978 à 1992), sur 39 stations réparties sur les trois affluents.

La série à Manaus est complète, ainsi que les données sur 10 autres stations.

Les 29 stations restantes sont entachées de *valeurs manquantes*. 18 d'entre elles ont moins de 5% de valeurs manquantes, et 24, moins de 10% de lacunes. Ces chiffres restent donc raisonnables.

La constitution des *échantillons de calage et de test* est faite de manière à ce que chacun d'eux soit le plus représentatif possible de l'échantillon global des données à Manaus (moyennes, écart-types, maxima et minima du même ordre). Cette distinction a été mise au point lors de l'étude menée par les hydrologues (1994).

Les échantillons retenus ont les caractéristiques statistiques suivantes (en cm) :

	moyenne	écart-type	minimum	maximum
calage et test	2749.6	112.98	2542	2942
calage	2740.7	119.97	2542	2897
test	2781	116.22	2652	2942

Présélection de variables explicatives

Appliquer PLS, oui. Mais à quelles variables de départ?

Calculer le temps de propagation d'une onde de crue de chaque station jusqu'à Manaus semble impossible étant donné les considérables apports en eau au bief en question de centaines de ruisseaux et rivières non cartographiés. A cet égard, apprécier cette valeur

par le délai écoulé entre les dates des deux hauteurs maximales sur les deux stations n'est pas raisonnable.
L'une des idées intéressantes de la première investigation sur le sujet consiste à évaluer un "temps de propagation statistique" (le terme est abusif): pour chaque station S, il s'agit de calculer le décalage temporel Δt_s pour lequel la série décalée de Δt_s est la plus corrélée à la station de Manaus. Ces décalages seront appelés les délais de corrélation maximale, et notés DCM.
Les résultats suivants ont été établis :

Affluent	station	DCM	coefficient de corrélation
Rio Madeira	Abuna	-75	0.66
	Porto Velho	-75	0.58
	Fazenda	-60	0.75
	Borda	-45	0.80
Rio Solimoes	Sao Paulo	-46	0.82
	Santo Antonio	-38	0.83
	Itapeua	-10	0.95
	Manacapuru	-3	0.99
Rio Negro	Sao Felipe	0	0.68
	Curicuriari	0	0.67
	Serrinha	0	0.46
	Moura	0	0.89
Rio Branco	Boa Vista	0	0.44

Ces chiffres paraissent cohérents. En effet, il suffit de visualiser la carte des stations utilisées (annexe 1) pour remarquer que le DCM sur le Rio Solimoes ou sur le Rio Madeira augmente avec la distance entre la station et Manaus.
Les DCM nuls sur le Rio Negro ainsi que sur son affluent (le Rio Branco) s'expliquent par les effets de remous présents dans le Rio Negro, qui entraînent une propagation quasi-instantanée des eaux du fleuve, d'aval en amont.

La présélection des variables explicatives s'effectuera selon la règle suivante :

Pour un délai de prévision d, nous pouvons classer l'ensemble des stations en trois groupes :

* celles dont le DCM est inférieur à d
* celles dont le DCM est du même ordre de grandeur que d,
* celles dont le DCM est très supérieur à d.

A la date d'émission de la prévision, les mesures les plus intéressantes sur les stations du premier groupe n'ont pas encore pu être effectuées. On se contentera alors d'utiliser ces stations décalées de d jours, et d'un certain nombre de décalages supplémentaires du même ordre.
Parmi celles du deuxième groupe, les données les plus corrélées avec celles de Manaus sont disponibles et on choisira, pour quelques stations, un bon nombre de décalages autour de d.
Les variables du troisième groupe ne seront pas sélectionnées.

Pour des commodités de rédaction, nous coderons les stations de la façon suivante :

station	abréviation
Manaus	M
Manacapuru	Mc
Itapeua	It
Sao Paulo	SP
Santo Antonio	SA
Curicuriari	Cu
Sao Felipe	SF

Les variables retenues selon ce principe pour la prévision à 30 jours sont :

Affluent	station, décalages	numéro
Rio Negro	M-30	1
	M-31	2
	M-32	3
	M-35	4
	M-37	5
	M-40	6
	M-42	7
	M-45	8
	M-47	9

Affluent	station, décalages	numéro
Rio Negro	M-50	10
	M-55	11
	M-57	12
	M-60	13
	M-62	14
	M-65	15
	M-70	16
	Cu-30	70
	Cu-32	71
	Cu-35	72
	Cu-37	73
	Cu-40	74
	Cu-42	75
	Cu-45	76
	Cu-47	77
	Cu-50	78
	Cu-55	79
	Cu-60	80
	Cu-65	81
	Cu-70	82
	SF- 30	83
	SF-35	84
	SF-40	85
	SF-45	86
	SF-50	87
	SF-55	88
	SF-60	89
	SF-65	90
Rio Solimoes	Mc-30	18
	Mc-31	19
	Mc-32	20
	Mc-35	21
	Mc-37	22
	Mc-40	23
	Mc-42	24
	Mc-45	25
	Mc-47	26
	Mc-50	27
	Mc-52	28
	Mc-55	29
	Mc-57	30

Affluent	station, décalages	numéro
Rio Solimoes	Mc-60	31
	Mc-62	32
	Mc-65	33
	Mc-70	34
	It-30	35
	It-35	36
	It-40	37
	It-45	38
	It-50	39
	It-55	40
	It-60	41
	It-65	42
	It-70	43
	SP-30	44
	SP-32	45
	SP-35	46
	SP-37	47
	SP-40	48
	SP-42	49
	SP-45	50
	SP-47	51
	SP-50	52
	SP-55	53
	SP-60	54
	SP-65	55
	SP-70	56
	SA-30	57
	SA-32	58
	SA-35	59
	SA-37	60
	SA-40	61
	SA-42	62
	SA-47	63
	SA-50	64
	SA-52	65
	SA-55	66
	SA-57	67
	SA-60	68

B/ ANALYSE DES CASES COMPLETES

Nous analyserons, après chaque sélection, les sorties graphiques des exécutions de l'algorithme PLS. Nous étudierons enfin les résidus du modèle final obtenu.

1-Première sélection

Les sorties graphiques que nous allons exploiter pour opérer cette sélection figurent en annexe 3 .

• *La première étape consiste à appliquer PLS à toutes les variables, avec un bon nombre d'axes.*

Les premières représentations à considérer sont les tracés, en fonction du nombre d'axes, des Eqm en test et en calage, et de la variance de M expliquée par axe en fonction du numéro de l'axe.

Graphe des Eqm (annexe 3a) :
Les Eqm chutent d'abord, de 97.93 et 71.9 cm en test et en calage (respectivement) au niveau de l'axe 1, aux valeurs 29.268 et 27 cm pour 13 axes. Ils se stabilisent ensuite jusqu'à atteindre les chiffres de 29.01 et 24.72 (toujours respectivement) pour 60 axes.

Graphe des variances expliquées (annexe 3a) :
L'évolution des inerties expliquées se montre plus brutale. Elle passe de 0.75 pour l'axe 1, à seulement 0.0389 dès le troisième axe. La dégradation se poursuit ensuite plus lentement et se stabilise à un niveau particulièrement faible (de l'ordre de 0.007).

Nous représenterons donc 13 axes : individuellement, les 3 premiers axes, tandis que les 10 suivants, du $4^{ème}$ au $13^{ème}$ feront l'objet d'une représentation groupée. Cette opération s'effectuera d'après la consultation des cercles de corrélation.

• *Analyse des cercles de corrélation*

On peut les visualiser en annexe 3b.

La double auto-corrélation (spatiale et temporelle) apparaît clairement, en particulier dans le plan (t_1,t_2) : on peut discerner nettement 7 groupes de variables, disposés chacun en arc de cercle (on retrouve les 7 stations utilisées).

Pratiquement toutes les variables sont bien représentées sur le premier axe. La moins corrélée est la variable 82 (Cu -70), avec tout de même un coefficient de corrélation de 0.58. 40 des 90 variables utilisées sont corrélées avec un coefficient de corrélation supérieur à 0.90 avec cet axe. Ces quarante régresseurs sont essentiellement répartis sur trois stations : M, Mc, et It. Ils correspondent également, pour chacun de ces

postes, aux plus petits décalages temporels (au-dela de -30). On peut donc *timidement* avancer que *l'axe 1 représente les stations les plus proches de Manaus, à la fois dans le temps et dans l'espace, sur le Rio Solimoes.*

Ces coefficients de corrélation se dégradent dès le *2ème axe*. La variable la mieux représentée est la n° 44 (SP -30), avec un coefficient de 0.62, alors que la 8ème variable la plus corrélée (en valeur absolue) est la 48 (SP -40), avec une valeur du coefficient de 0.5. Ces huit régresseurs concernent les stations *SP et SA, et des décalages voisins de -30.*

Ce n'est qu'à partir du *3ème axe* que commencent à apparaître des stations du Rio Negro. En effet, Cu-60 (variable 80) est la mieux représentée avec un coefficient de 0.47. M-70 (n°17) arrive en troisième position (corrélation de 0.4312) et Mc, poste du Rio Solimoes, est encore bien représentée, mais avec un décalage temporel de -70. En fait, on peut remarquer que plus on avance dans les axes, plus les stations représentées sont loin de M, spatialement ou temporellement. Trois des cinq stations les mieux représentées (coefficient de corrélation supérieur à 0.4) par l'axe 3 concernent Cu, avec de grands décalages temporels (de -55 à -65). Ainsi, on peut schématiser en considérant que *l'axe 3 représente Curicuriari.*

La consultation du tableau de sélection des variables (annexe 3c) nous amène à remarquer que *plus le numéro de l'axe augmente, plus les variables les mieux représentées sont loin de Manaus, qu'il s'agisse de distance spatiale, ou « temporelle ».* Ainsi, les variables qui expliquent le mieux l'évolution de la crue à Manaus pour une prévision à 30 jours sont d'abord celles situées sur le Rio Solimoes, puis celles riveraines du Rio Negro. Rien d'étonnant à cela : Le Rio Solimoes constitue le lit principal de l'Amazone et contribue à lui seul à 36% de l'apport en eaux au niveau de la confluence avec le Madeira. Le Rio Negro n'y apporte qu'une participation de 11% (H. Thepaut, 1994).

Ces idées, ainsi que la représentation des autres axes sont résumées par le tableau placé en annexe 3c.

La sélection :
Les variables finalement sélectionnées figurent en dernière colonne de ce tableau.

- axe 1 :

La sélection des variables résumant l'axe 1 est particulièrement délicate. Ainsi, on essaiera de sélectionner, dans chacun des groupes des 3 stations M, Mc, et It, les variables les mieux corrélées à l'axe. Le but est de perdre le moins d'information possible en représentant ces trois stations (les plus intéressantes).

• axe 2 :
L'axe 2 n'expliquant plus que 0.08 de la variance de M, on le représentera par moins de variables : les 2 les mieux corrélées, à savoir SP-30 et SP-32.

• axe 3 :
L'axe 3, essentiellement représenté par la station Cu, nous amènera à sélectionner les variables 79 et 80 (Cu-55 et -60 respectivement).

La représentation groupée des 10 autres axes ne fait plus intervenir strictement l'ordre par qualité de représentation (coefficient de corrélation « élevé » en valeur absolue) des variables par axes. Représentant les stations les plus lointaines (aux sens spatial et temporel), leur importance quant à l'explication du signal à Manaus est moindre.

La sélection aboutit finalement à un lot de 25 variables, ce qui constitue une réduction du plus du tiers du nombre de régresseurs d'origine. En voici la liste :

affluent	variable	numéros (même ordre)
Rio Solimoes	Mc-30, -31, -32, -35, -37, -40	18 à 23
	It-35, -40, -45	36 à 38
	SP-30, -32	44, 45
Rio Negro	M-30, -31,-32, -35, -37	1 à 5
	Cu-40, -42, -55, -60, -70	74, 75, 79, 80, 82
	SF-40, -55, -60	85, 88, 89

2- 2ème sélection

Les graphes relatifs à cette deuxième section sont répertoriés en annexe 4.

La *régression multiple* appliquée à cette dernière collection de variables mène à des Eqm en test et en calage respectifs de 28.41 (cm) et 29.44 (cm). Ce résultat s'avérant encore satisfaisant (amélioration sur l'échantillon test, petite dégradation sur celui de calage), on peut tenter une deuxième sélection suivant la même démarche.

On effectue donc une deuxième régression PLS, avec les 25 variables sélectionnées.

• *Graphes des Eqm et de la variance expliquée* (annexe 4a) :
Le graphe des Eqm nous permet de sélectionner le nombre d'axes qu'il faudra représenter, et celui de la variance expliquée de M par axe nous signifiera les axes à représenter individuellement ou collectivement.
Ici encore, on choisira 13 axes (Eqm en test et en calage de 30.91 et 31.76 respectivement), dont seulement les 3 premiers seront à représenter individuellement.

- *Analyse des cercles de corrélation* (annexe 4b):

Les groupes de variables sont ici nettement mieux séparés.
Les corrélations maximales se dégradent également nettement plus vite. En effet, dès le 5eme axe, ces chiffres ne dépassent que rarement 0.3. Ceci semble indiquer que la première sélection a été « correcte » : toute l'information est concentrée sur les axes 1, 2, et 3.

L' *axe 1* est très corrélé avec les variables issues de *M, Mc et It*, comme auparavant.
Les variables sélectionnées pour cet axe se feront, comme au paragraphe précédant, de manière à représenter ces trois stations.

Sur le *deuxième axe*, on repère surtout les stations *Cu* (variables 79, 80, 82) et *SP* (n° 44 et 45). L'inertie expliquée étant moindre (0.110 contre 0.770 pour le premier axe), on retiendra moins de variables.

L'axe 3 paraît quant à lui assimilable à la station *SF* : SF-30 et SF-40 sont les deux variables les mieux représentées, avec des coefficients de corrélation de 0.3632 et 0.4156 respectivement.

Les variables les mieux représentées sur les axes suivants sont indiquées sur le tableau de sélection de l'annexe 4 (4c).

On retiendra la liste suivante, réduite à 14 variables :

affluent	variables	numéro
Rio Solimoes	Mc-30, -31, -32	18 à 20
	It-30	36
	SP-30	44
Rio Negro	M-30, -31	1, 2
	Cu-55, -60, -70	79, 80, 82
	SF-30, -40, -55, -60	83, 85, 88, 89

Une régression multiple menée avec cet ensemble de régresseurs a abouti à des Eqm de 37.08 en test, et de 34.59 en calage. *Il y a donc dégradation des résultats.*
Cela peut s'expliquer par le fait que le Rio Negro soit plus représenté que le Solimoes (9 variables contre 5) alors que ce dernier apporte davantage dans le signal de crue à Manaus.

Nous préférons alors à ce niveau effectuer une régression pas à pas sur les 25 variables sélectionnées auparavant.

Le graphe des Eqm en fonction du pas du stepwise (annexe 5) nous montre une stabilisation de ces critères aux valeurs de 29.46 (calage) et de 27.37 (test), au niveau du 17^{eme} pas.
Le modèle final sera donc le suivant :

affluent	variable	coefficient
Rio Negro	M-30	4.52507
	M-31	-2.05227
	M-35	-0.67935
	Cu-40	-0.24872
	Cu-42	-0.02194
	Cu-55	-0.08274
	Cu-70	-0.02886
	SF-30	0.09911
	SF-60	0.05215
	SF-40	0.03940
Rio Solimoes	Mc-30	2.41251
	Mc-31	-2.37289
	It-45	0.26504
	It-35	-0.14256
	SP-30	-0.58061
	SP-32	-0.58061
Constante		-594.95877

Eqm sur l'échantillon test : 27.37 cm
Eqm sur l'échantillon de calage : 29.46 cm

3- Analyse des résidus

Les graphes des résidus, consultables en annexe 5, nous permettent de tirer les conclusions suivantes :

- *Les résidus sont centrés*

Leur répartition de part et d'autre de l'axe des abscisses semble équilibrée.

- *Les résidus sont auto-corrélés*

On peut le remarquer sur tous les graphes présentés. En effet, en suivant chronoliquement les points, on peut encore déceler une certaine structure, car le résidu calculé pour un jour donné ne s'éloigne pas de façon significative de celui calculé la veille ou la semaine d'avant.
Il semble de toutes façons très difficile, étant donné l'intensité des auto-corrélations des variables de départ, d'aboutir à des résidus indépendants sans faire appel à des modification du modèle par des procédés auto-régressifs sur les résidus(principe exposé dans la première partie, A/, § 1)

- *lien avec l'estimation de Manaus et certaines variables explicatives*

Le nuage des résidus se concentre de plus en plus autour de l'axe des abscisses lorsque les valeurs des côtes estimées augmentent. Ainsi, ilsemble y avoir une certaine corrélation. En fait, ceci révèle une meilleure qualité d'approximation pour les cotes importantes. Ceci peut s'expliquer par les plus faibles variances du signal de crue à Manaus à l'étale de la crue.
La même évolution est constatée relativement aux variables explicatives M-30, M-35, Cu-55, SF-40. Ceci peut s'expliquer par la forte corrélation existant entre ces variables et l'estimation de Manaus.

- *étude d'une relation éventuelle avec l'amplitude de la crue à Manaus*

L'évolution des résidus par rapport au temps présente une faible tendance générale de type sinusoïdal. La question a alors été posée d'une corrélation éventuelle avec l'amplitude du signal à Manaus. Les deux derniers graphes de l'annexe 5 concernant ce lien montre que non.

C/ MODELES DE DONNEES MANQUANTES

Nous avons effectué le remplacement de valeurs manquantes dans la matrice des 90 variables originelles spécifiées au §B/, par des estimations par moyenne, comme expliqué en première partie, § B/1.

Les résultats graphiques, cercles de corrélations, graphes des Eqm, des variances expliquées figurent en annexe 6, pour le modèle avec pondération, et en annexe 7 pour le modèle obtenu sans pondérer les observations.

Les similitudes avec ceux résultant de l'étude sur cases complètes sont frappantes.
Nous nous dispenserons donc d'effectuer les sélections relatives à ces données, considérant que l'on ne peut aboutir qu'aux mêmes résultats.

Ces ressemblances peuvent s'expliquer facilement par le faible pourcentage de données manquantes. En effet, parmi les 7 stations utilisées, seules 3 comportent des données manquantes, et en faible proportion (Sao Paulo, avec 4%, Santo-Antonio, 1%, et Itapeua, 2%). Il en résulte que les deux modèles ACA (avec et sans pondération), après éliminations de lignes handicapant les estimations, ont été calés sur presque 1200 lignes, ce qui est proche des 1137 observations caractérisant le modèle CCA. L'effet de la technique de pondération n'a pas pu être véritablement apprécié sur cette application.

Conclusion

La régression Partial Least Squares a permis d'établir un modèle de prévision des hauteurs d'eau de l'Amazone à Manaus avec un délai de 30 jours.
Les performances de ce dernier évaluées par l'écart quadratique moyen en calage et en validation sont respectivement de 23.46 cm et 27.37 cm, ce qui est tout à fait acceptable pour le délai de prévision considéré.

Toutefois il faut souligner les limites de la méthode PLS rencontrées dans ce type d'étude. Le choix des régresseurs est rendu délicat par la présence de corrélations spatiales et temporelles particulièrement fortes. Le premier axe identifié par la méthode porte presque toute l'information, et l'interprétation de cette composante s'en trouve compliquée.
La technique de sélection de variables proposée fournit des résultats convenables mais mérite d'être améliorée dans ce contexte. Elle permet notamment de choisir des variables explicatives même en grand nombre à l'origine, et même en présence de données manquantes.

Cette étude a souligné de même les limites de la méthode de validation croisée dans le cas de nombreuses observations. Une validation croisée "par groupe" demande à être mise au point, par la détermination d'un critère de sélection "optimal" des lignes à retirer d'une étape à l'autre.

Enfin, la recherche d'intervalles de confiance n'a pas pu être abordée dans le cadre de ce stage. On pourrait pallier à ce problème en utilisant la technique du bootstrap, qui permet de construire de tels intervalles de façon empirique.

Bibliographie

BARONI M. & al . (1992), Chemometrics, 6, 347-356.

CRAMER R. D. (1988), *Crossvalidation, bootstrapping, and Partial Least Squares compared with multiple regression in conventionnel QSAR studies.* Quantitative Structure-Activity Relationships, 7, 18-25.

FRANK I.E. & FRIEDMAN J.H.(1993), *A statistical view of some chemometrics regression Tools*, Technometrics, 35, 2, 109-135.

GARTHWAITH P.H. (1994), *An interpretation of PLS,* Journal of the American Statistical Association, 89, 425, 122-127.

GEISSER S. (1974), Biometrika, 61, 101-107.

KOTZ S., JOHNSON N.L.(1981), dictionnaire de statistiques, Ed. Campbell BRED.
LINDGREN F., GELADI, RANNA, WOLD (1994), *Interactive variable sélection (IVS) for PLS. Part I: Theory and algorithms,* Journal of Chemometrics, Vol. 8, 349-363.

LITTLE R.J.A. (1988), *Robust estimation of the mean and covariance matrix from data with missing values*, Applied statistics, 37, 23-29.

MARTENS H. & NEAS T. (1989), *Multivariate calibration,* Wiley, Chichester.

MUTHEN B., KAPLAN D. & HOLLIS M. (1987), *On structural equation modeling with missing data that are not missing completely at random*, Psychometrica, 52, 431-462.

ORCHARD T.& WOODBURY M.A.(1972), *A missing information principle : theory and applications*, proceeding of the sixth Berkeley symposium on mathematical statistics and probability, 1, 697-715.

PALM R.& IEMMA E.F. (1995), *Quelques alternatives à la régression classique dans le cas de la colinéarité,* Revue de Statistique Appliquée.

RICHARD D. CRAMER & al.(1988), *Crossvalidation, Bootstrapping, and Partial Least Squares Compared with Multiple Regression in Conventionnal QSAR Studies,* Tripos Associates, 6548 Clayton Road, St. Louis, MO 63117.

RODERICK J.A.L. (1992), *Regression with missing X's : A review,* Journal of the american Statistical Association, N°420, vol.87, 1227-1237.

SAPORTA G.(1990), Probabilités, analyse des données et statistique, éd. Technip.

STONE M. (1974), J. R. Stat. Soc. B, 36, 111-133.

TENENHAUS M.,GAUCHI J.P., MENARDO C. (1995), *Régression PLS et applications,* Revue de Statistique appliquée, XLIII (1), 7-63.

TENENHAUS M.(1995), *Nouveaux regards sur la régression PLS, théorie et pratique, partie 1 : Nouvelles méthodes de régression PLS.*

TOMASSONE R., AUDRAIN S., LESQUOY-de TURCKHEIM E., MILLIER C.(1992), La régression, nouveaux regards sur une ancienne méthode statistique, éd. Masson, coll. Inra.

WOLD H. (1966), *Estimation of principal components and related models by iterative Least Squares.* In Multivariate Analysis, ed. P.R. Krishnaiah, New York : Academic Press.

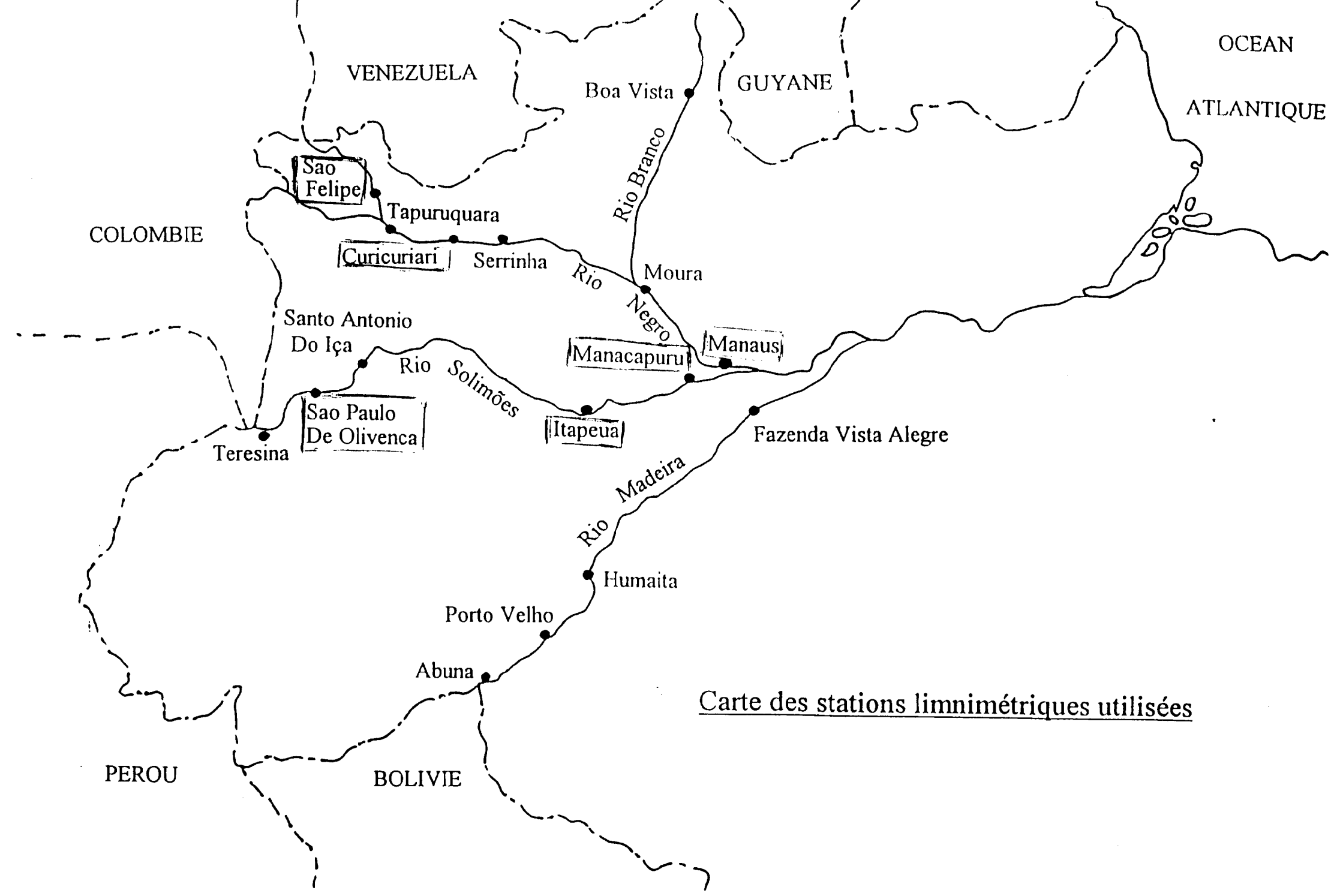
WOLD H. (1975), *Soft modelling by latent variables : the non-linear iterative Partial Least Squares approach*, In Perspectives in probability and statistics. Papers in Honour of M.S. Bartlett, ed. J. Gani, London : Academic Press.

WOLD H. & JORESKOG K. (1982), *System under indirect observation : causality, structure, prediction,* North-Holland, Amsterdam.

WOLD S. & al. (1983), Food research and data analysis, ed. by Martens & Russwurm, 147-188, Applied science, London.

Annexe 1

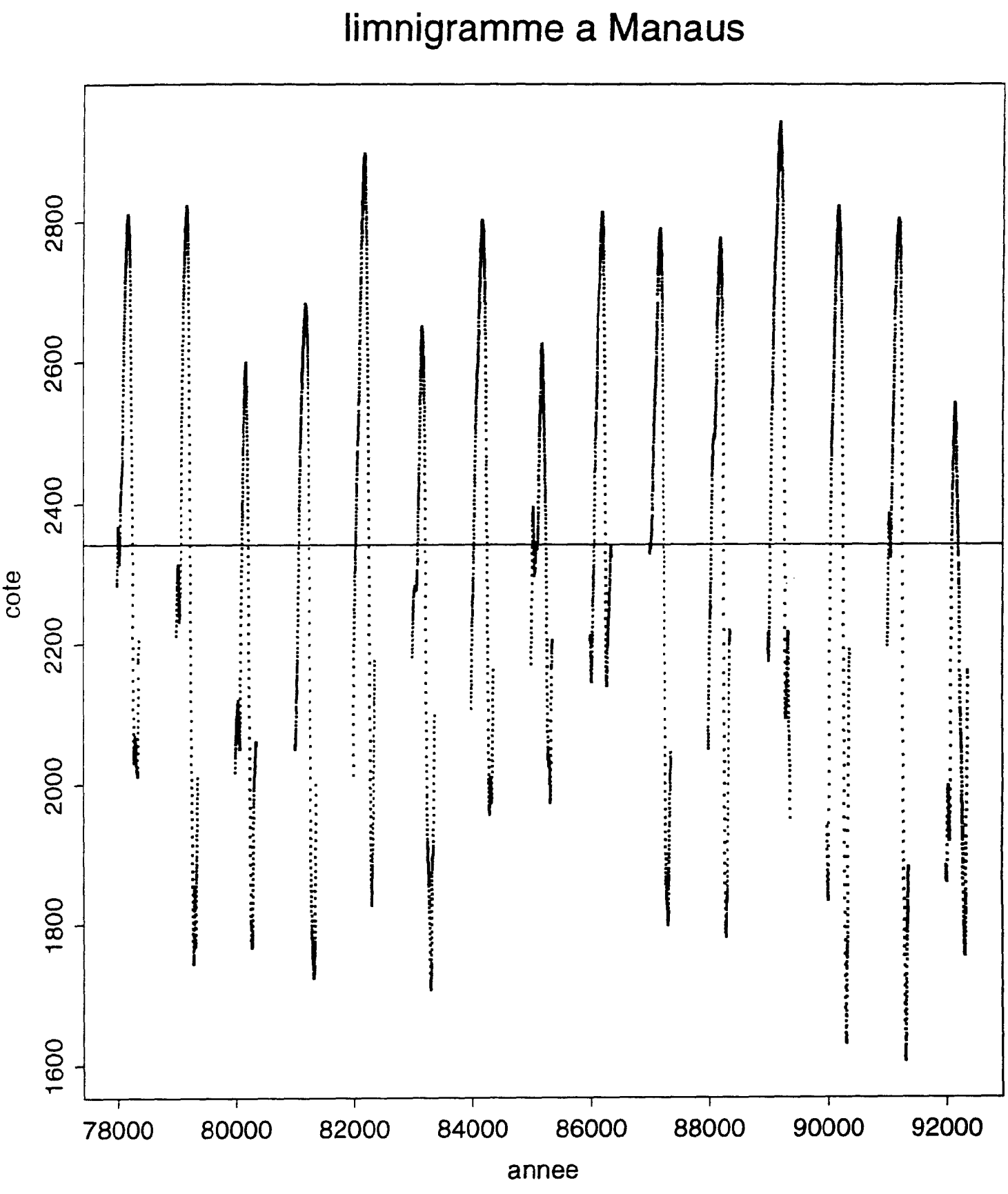
Carte des stations limnimétriques

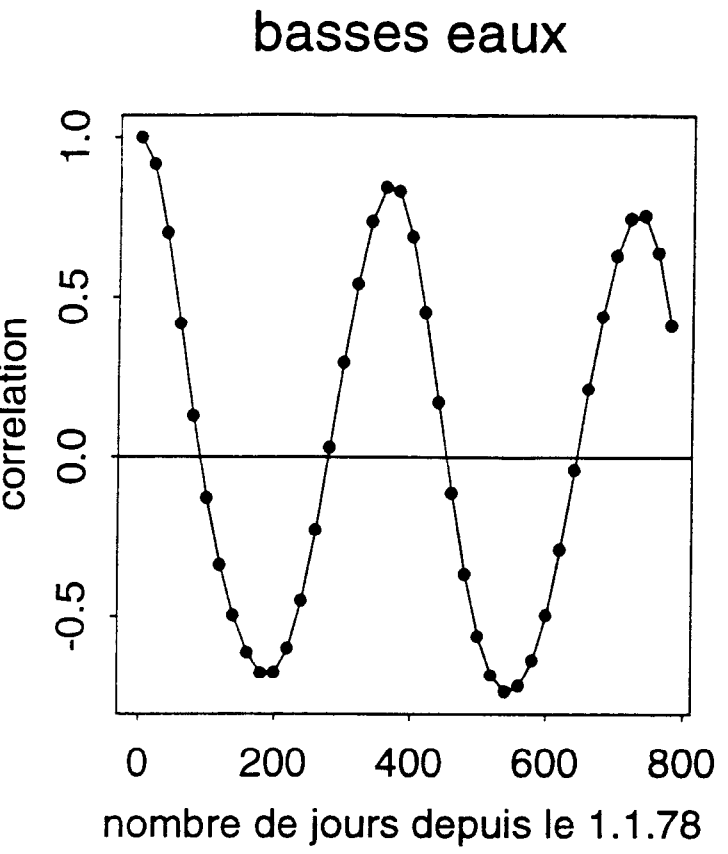


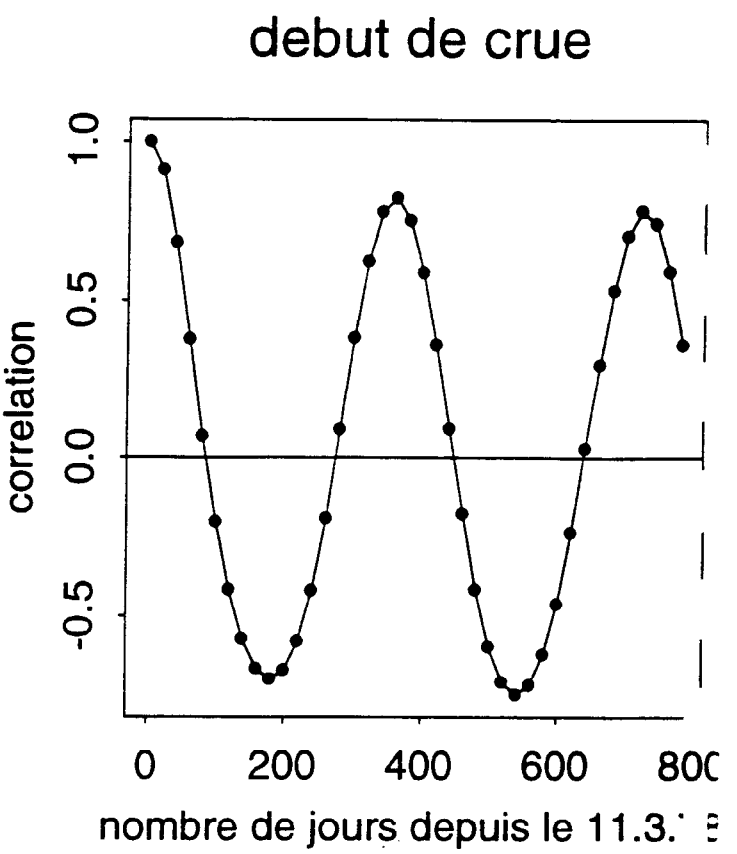
Carte des stations limnimétriques utilisées

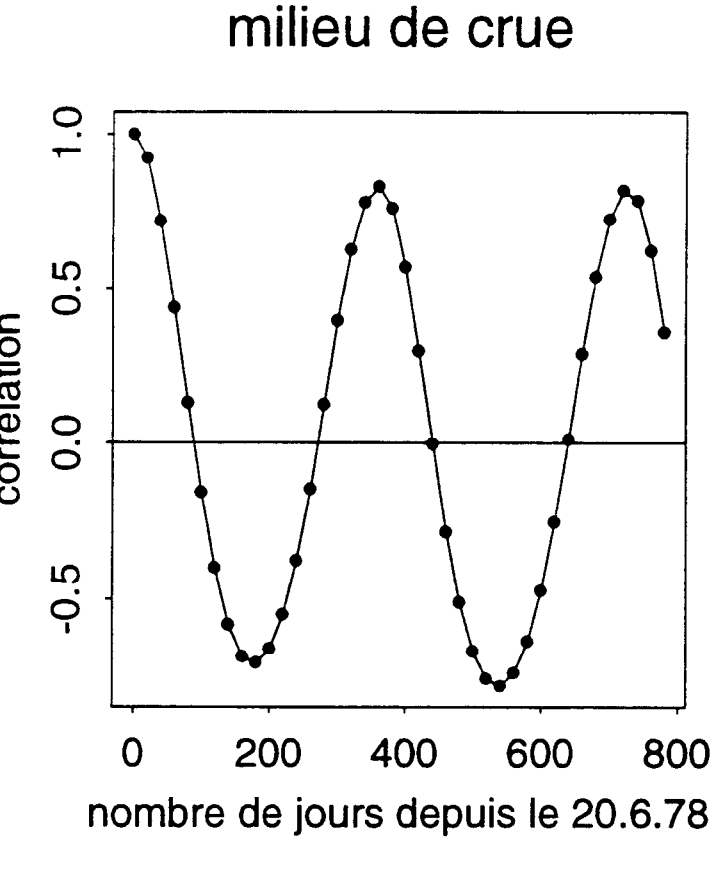
Annexe 2

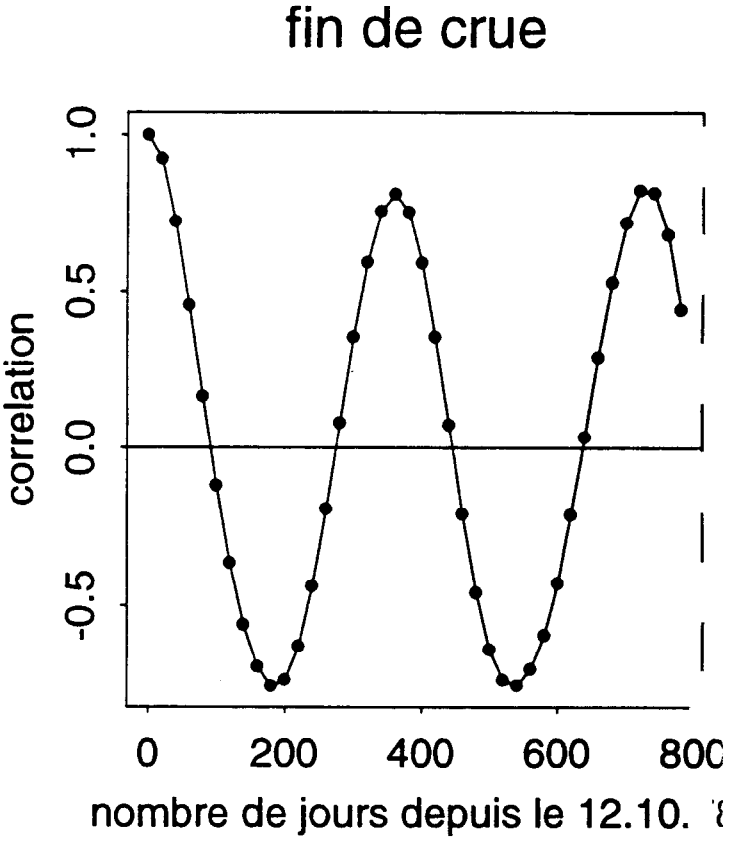
La série à Manaus


limnigramme a Manaus
cote
1600
1800
2000
2200
2400
2600
2800
78000
80000
82000
84000
86000
88000
90000
92000
annee


basses eaux
correlation
1.0
0.5
0.0
-0.5
0
200
400
600
800
nombre de jours depuis le 1.1.78

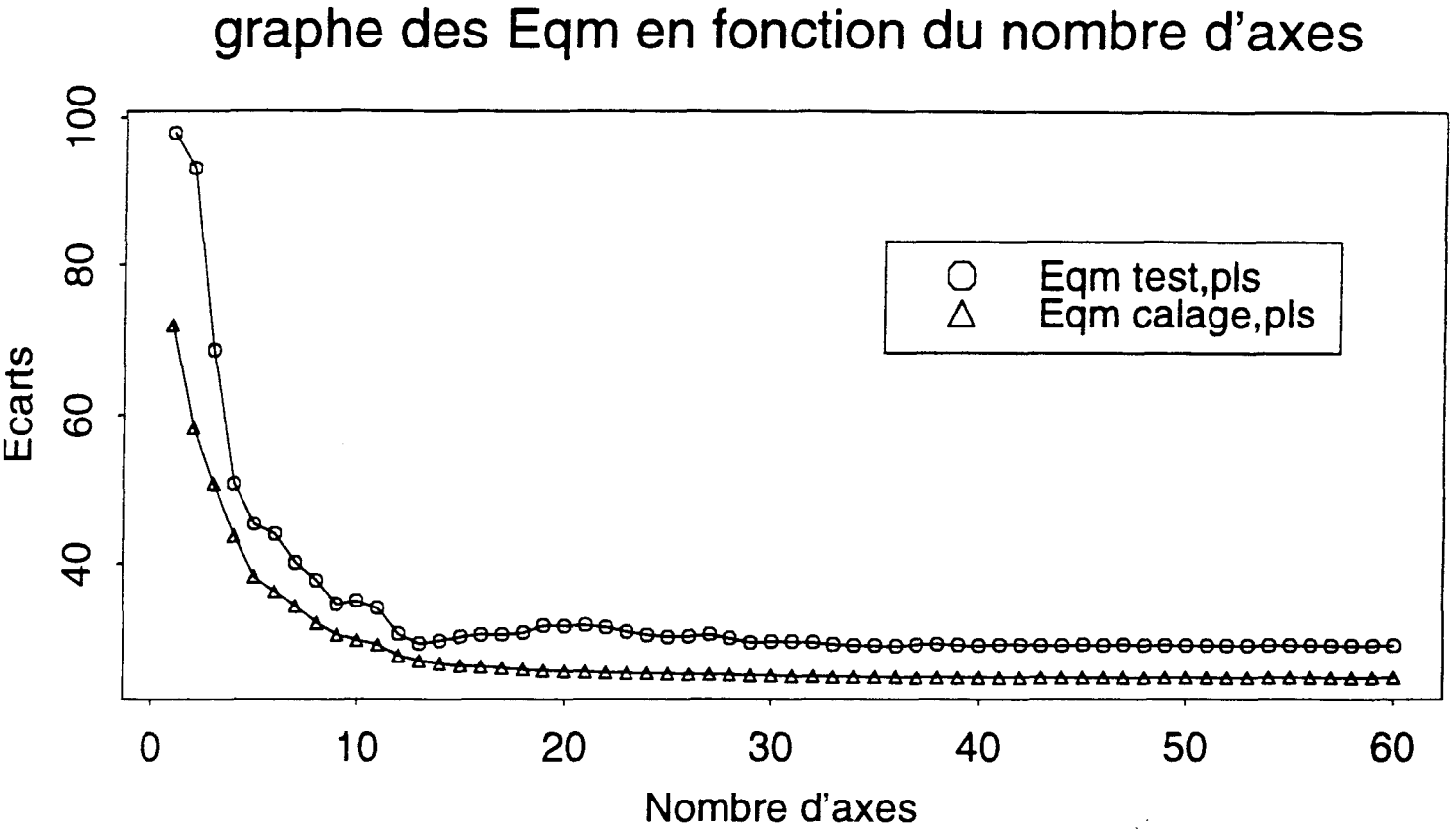

debut de crue
correlation
1.0
0.5
0.0
-0.5
0
200
400
600
nombre de jours depuis le 11.3.

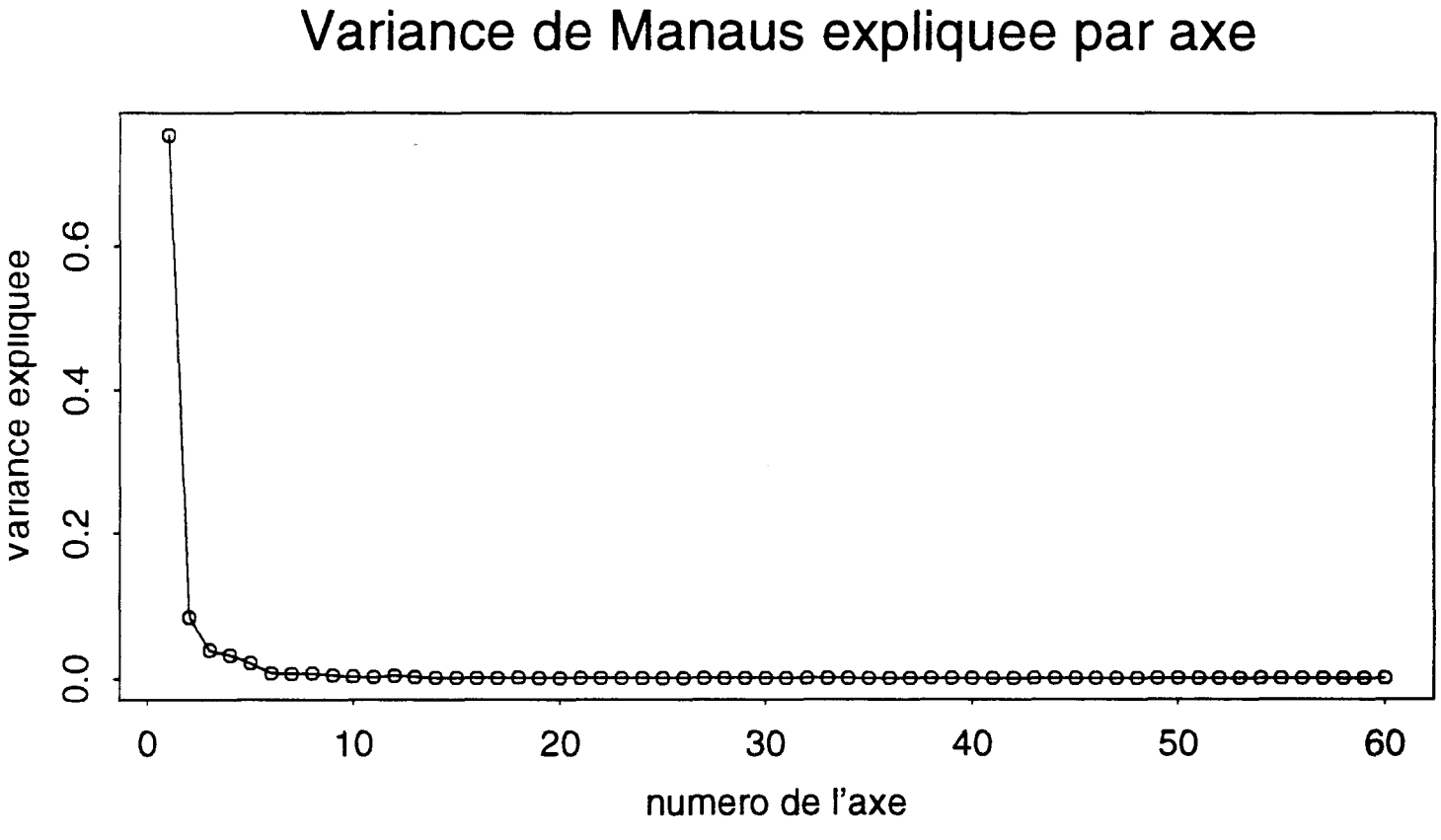

milieu de crue
correlation
1.0
0.5
0.0
-0.5
0
200
400
600
800
nombre de jours depuis le 20.6.78

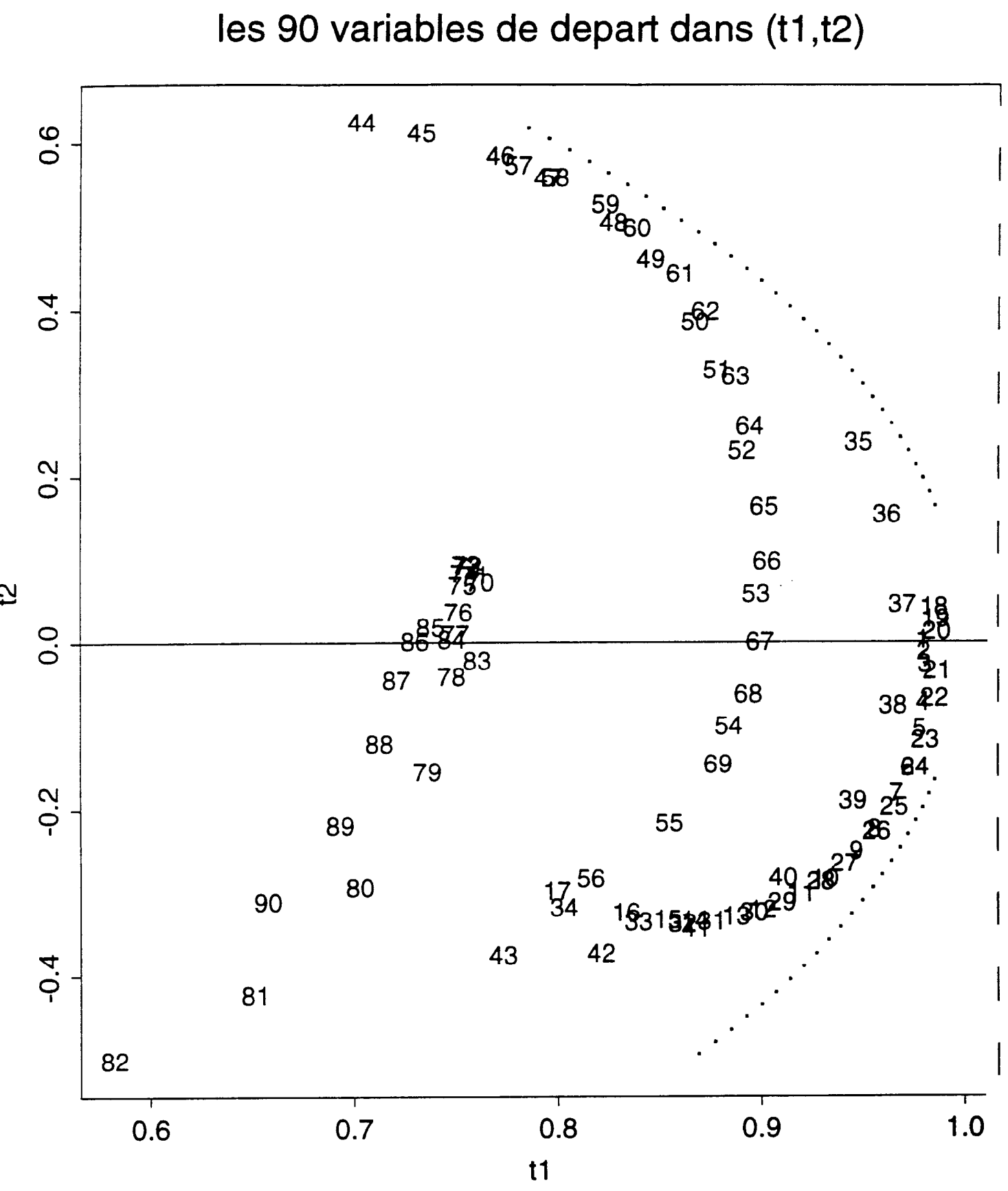

fin de crue
correlation
1.0
0.5
0.0
-0.5
0
200
400
600
nombre de jours depuis le 12.10.

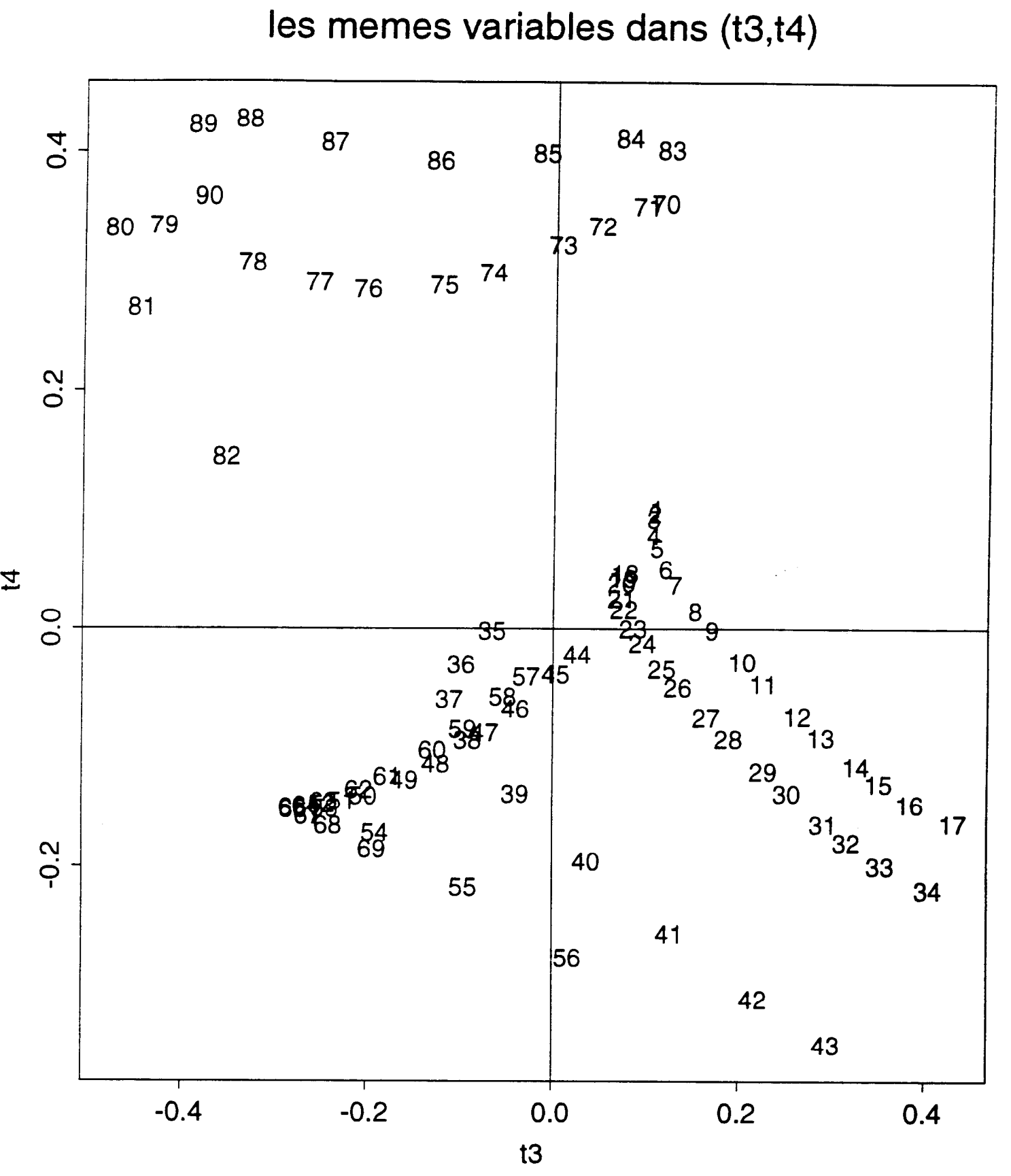
Annexe 3

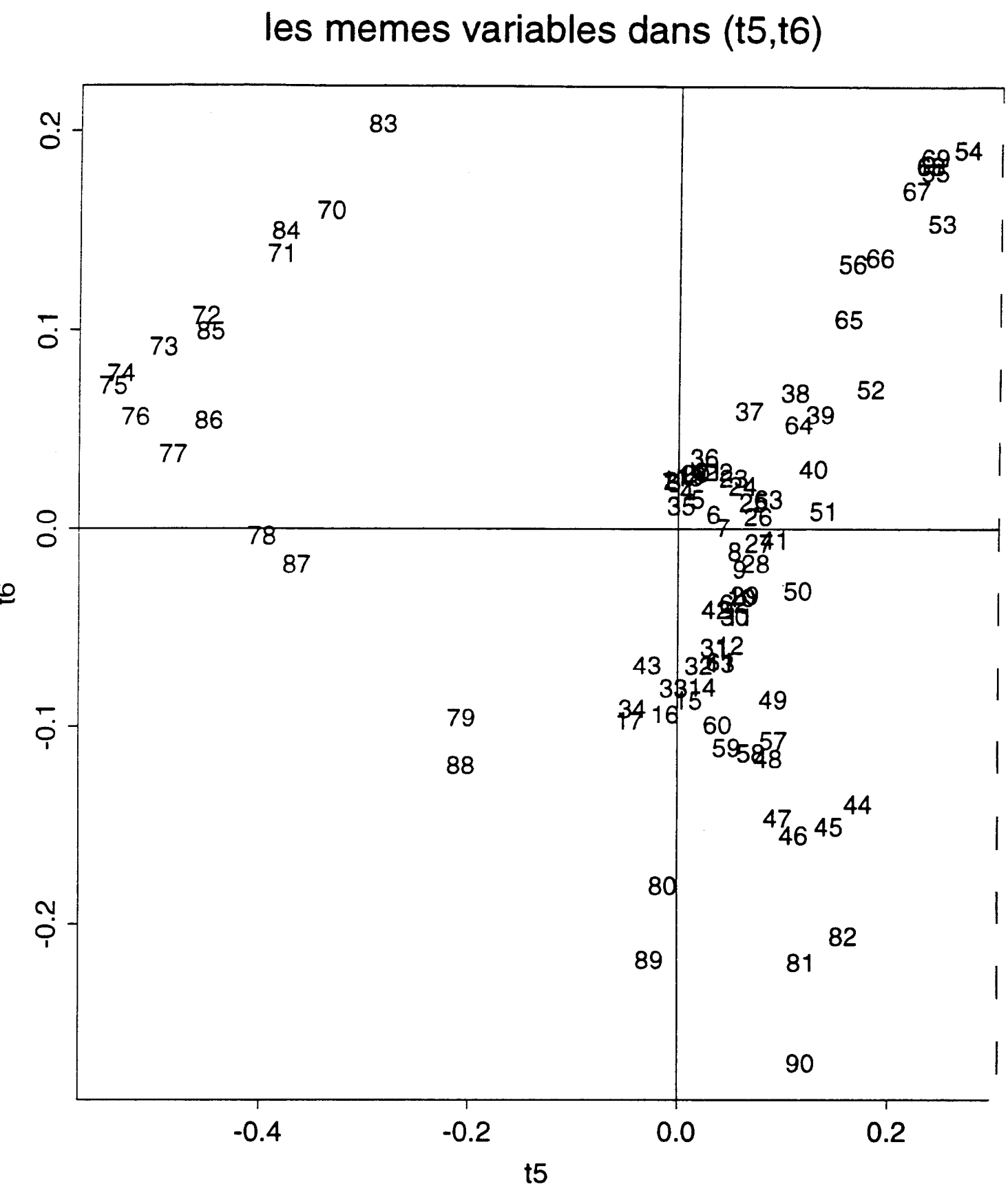
Analyse des cases complètes première sélection

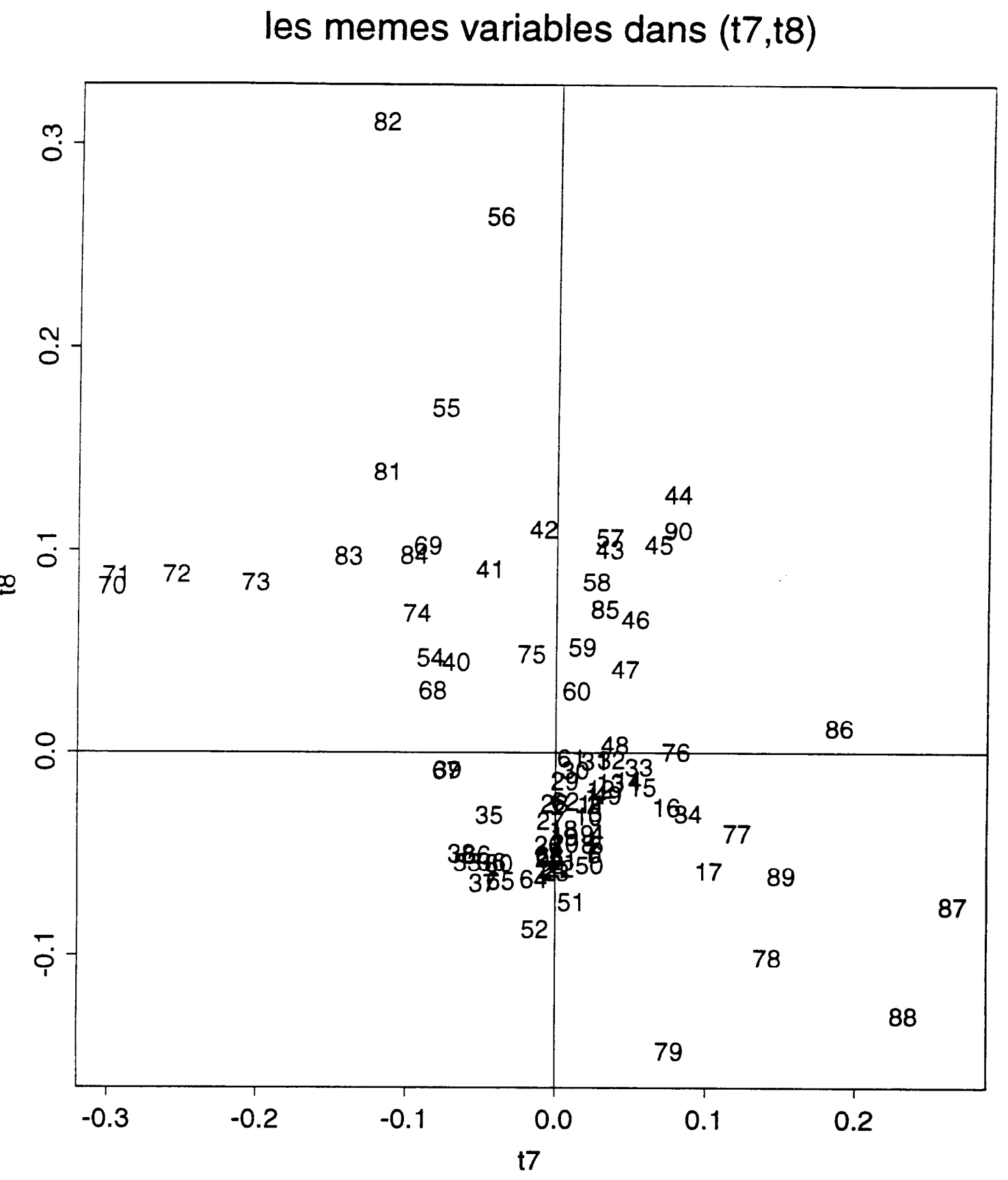

graphe des Eqm en fonction du nombre d'axes
Ecarts
40
60
80
100
Eqm test,pls
Eqm calage,pls
0
10
20
30
40
50
60
Nombre d'axes

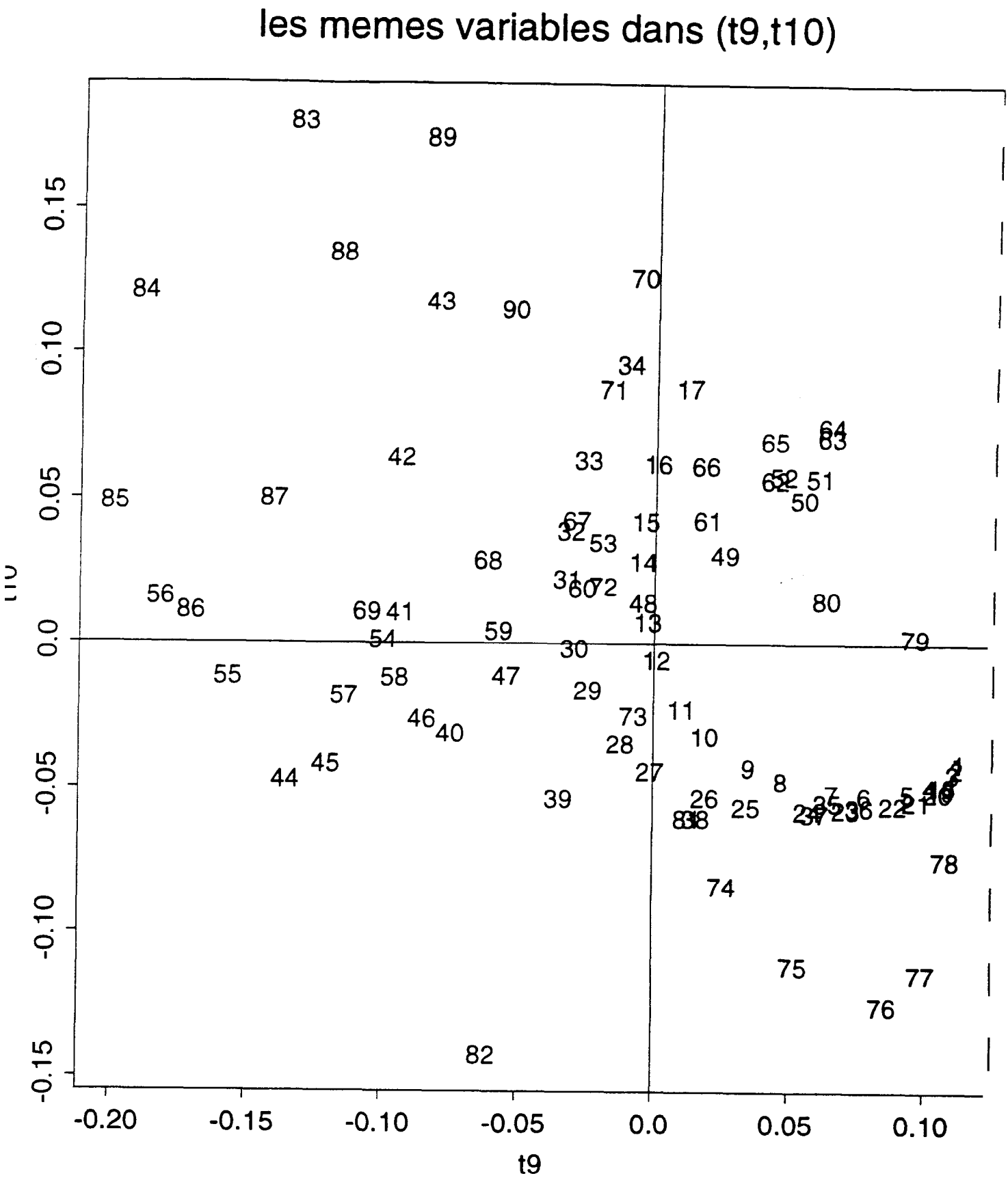

Variance de Manaus expliquee par axe
variance expliquee
0.0
0.2
0.4
0.6
0
10
20
30
40
50
60
numero de l'axe


les 90 variables de depart dans (t1,t2)
t2
t1
0.6
0.4
0.2
0.0
-0.2
-0.4
0.6
0.7
0.8
0.9
1.0

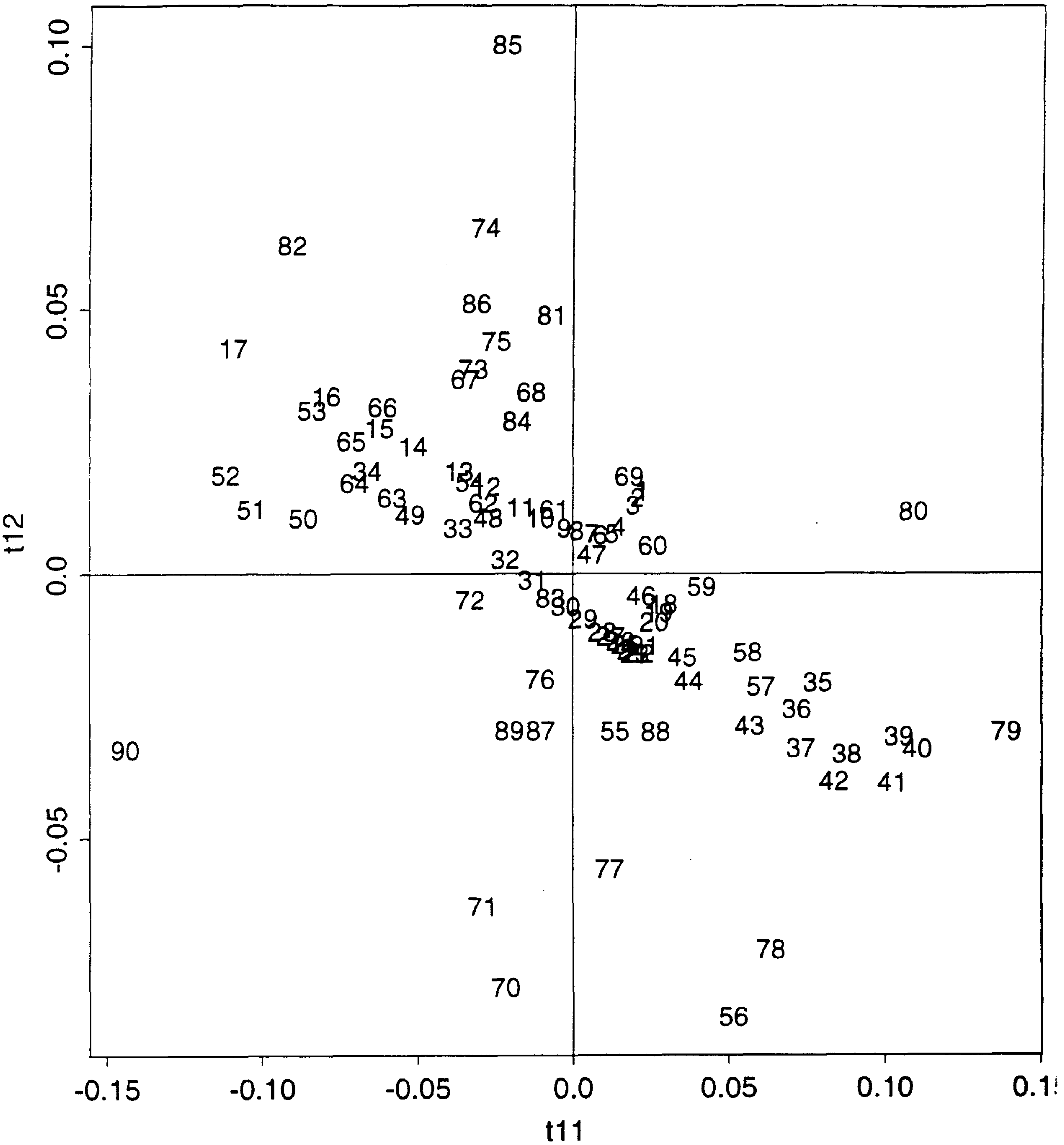

les memes variables dans (t3,t4)
t4
t3
0.4
0.2
0.0
-0.2
-0.4
-0.2
0.0
0.2
0.4

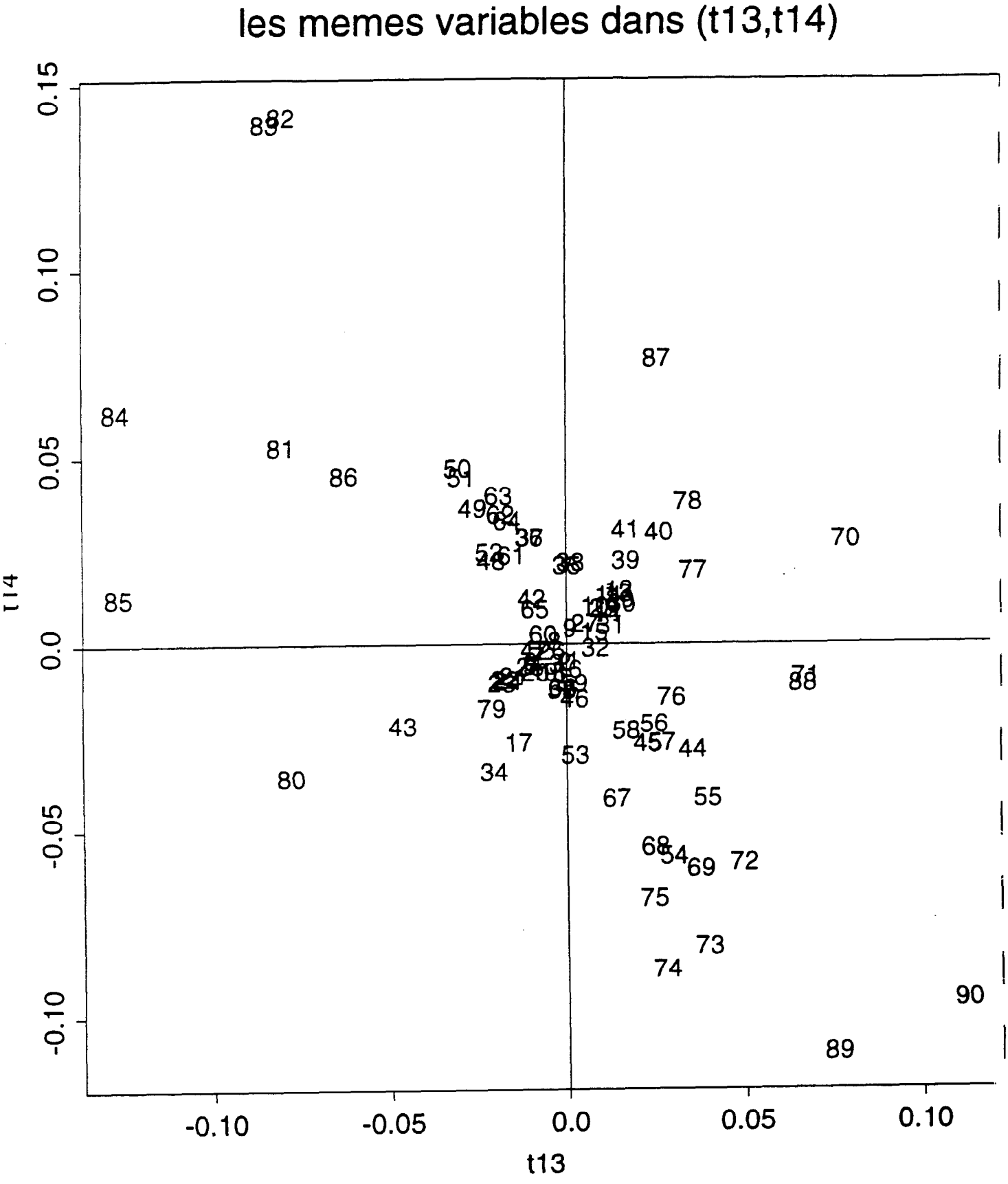

les memes variables dans (t5,t6)
t6
t5
0.2
0.1
0.0
-0.1
-0.2
-0.4
-0.2
0.0
0.2
83
70
84
71
72
85
73
74
75
76
86
77
78
87
79
88
80
89
90
82
81
44
47
46
45
49
60
59
57
58
48
43
54
67
53
56
66
65
38
52
37
39
64
40
51
50

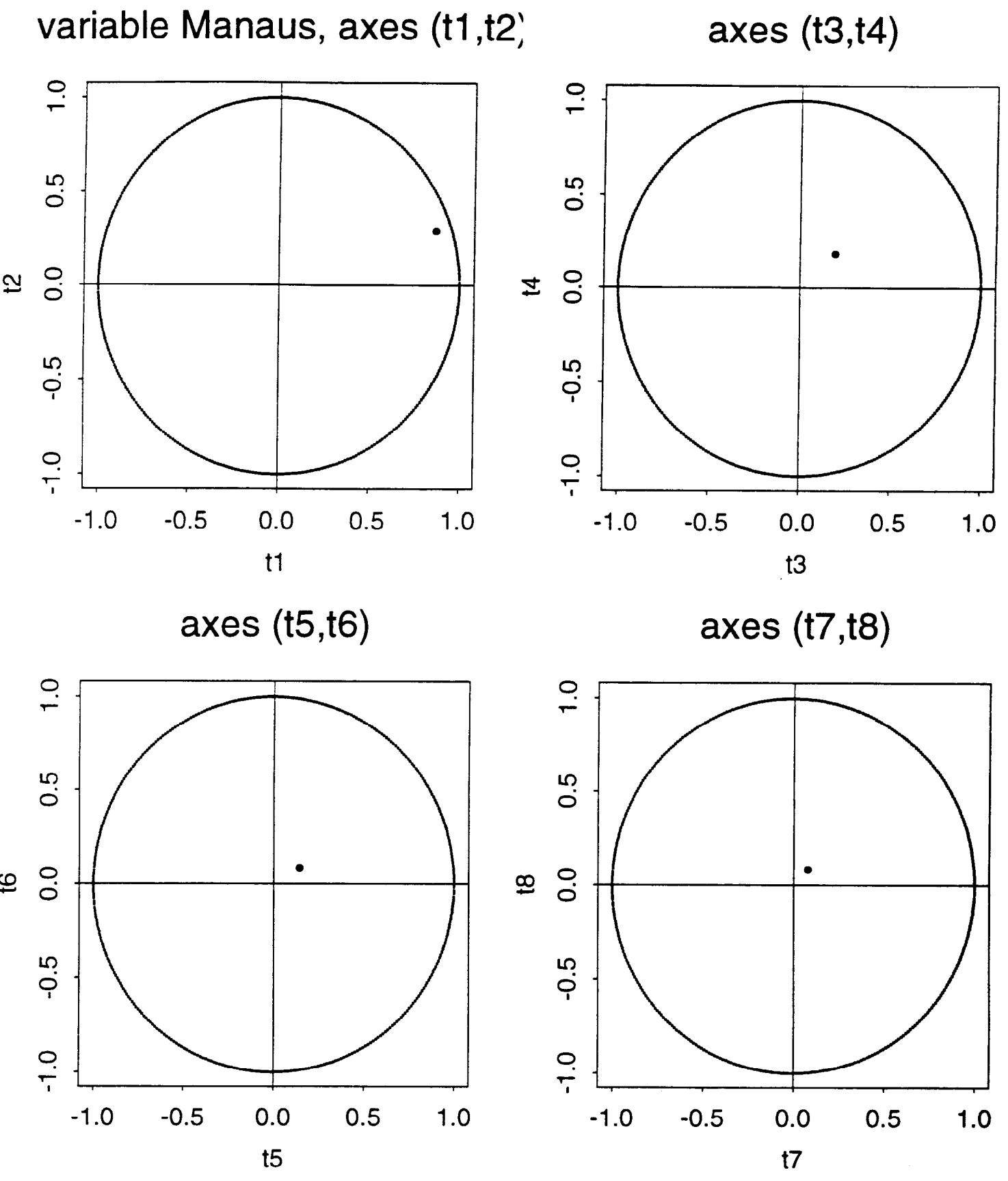

les memes variables dans (t7,t8)
t8
t7
0.3
0.2
0.1
0.0
-0.1
-0.3
-0.2
-0.1
0.0
0.1
0.2
82
56
55
81
44
42
90
69
83
84
41
45
71
72
73
70
58
74
85
46
54
40
75
59
47
68
60
86
48
76
35
16
34
77
17
89
51
87
52
78
88
79


les memes variables dans (t9,t10)
t10
t9
0.15
0.10
0.05
0.0
-0.05
-0.10
-0.15
-0.20
-0.15
-0.10
-0.05
0.0
0.05
0.10
83
89
88
84
43
90
70
34
71
17
42
33
16
66
65
64
63
85
87
51
50
67
15
61
32
53
68
14
49
56
86
31
60
72
48
80
69
41
13
54
59
30
79
12
55
58
47
29
57
73
11
46
40
28
10
45
44
27
9
8
39
26
25
74
78
75
77
76
82

les memes variables dans (t11,t12)




les memes variables dans (t13,t14)
t14
t13
0.15
0.10
0.05
0.0
-0.05
-0.10
-0.10
-0.05
0.0
0.05
0.10
82
87
84
81
86
78
41
40
70
39
77
85
42
32
71
88
76
79
43
58
56
57
44
17
53
34
80
67
55
68
54
69
72
75
73
74
90
89

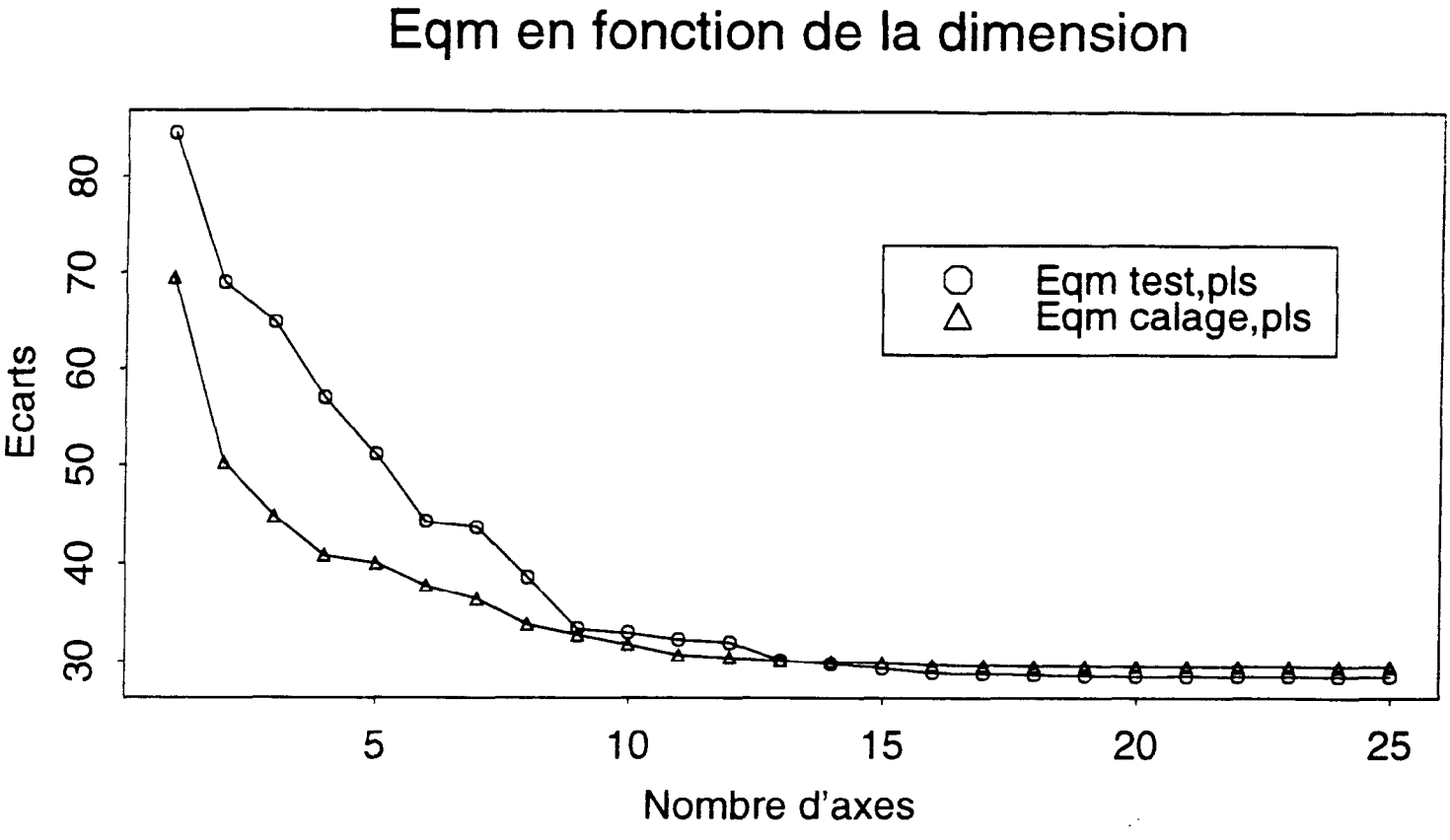

variable Manaus, axes (t1,t2)
axes (t3,t4)
axes (t5,t6)
axes (t7,t8)
t1
t2
t3
t4
t5
t6
t7
t8
-1.0
-0.5
0.0
0.5
1.0

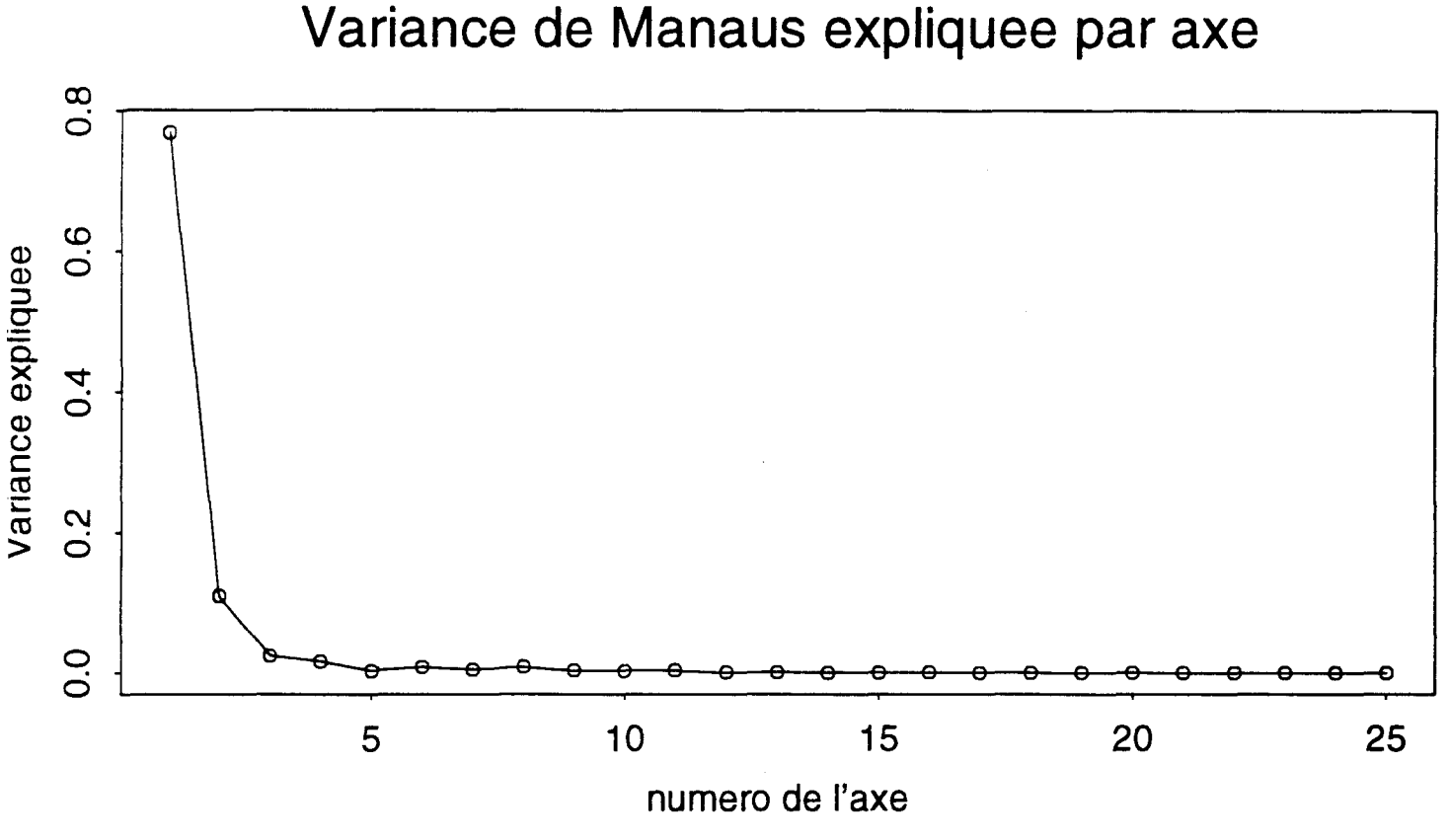
Variables	Axes ->1	2	3	4	5	6	7	8	9	10	11	12	13	SELECTION
1	10													X
2	8													X
3	7													X
4	9													X
5	11													X
6	13													
7														
8	19													
9	21													
10	34													
11	36													
12	40													
13														
14														
15			3											
16														
17														
18	5													X
19	3													X
20	2													X
21	1													X
22	4													X
23	6													X
24	12													
25	15													
26	18													
27	33													
28	35													
29	38													
30														
31														
32			5											
33														
34														
35	20													
36	17													X
37	14													X
38	16													X
39	22													
40	37													
41														
42														
43														
44		1												X
45		2												X
46		3												
47		6												
48		8												
49														
50														
51														

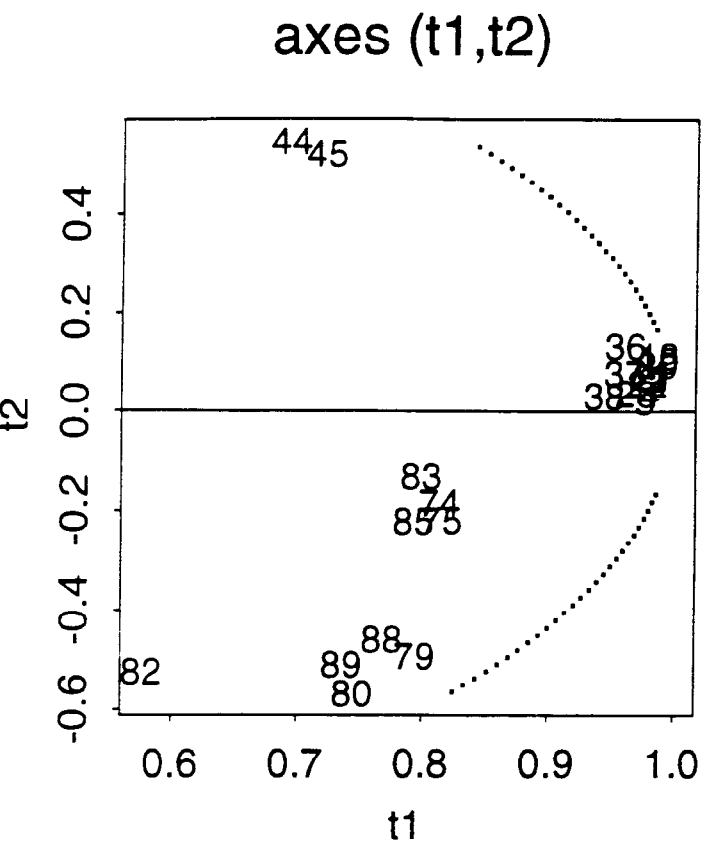
52														
53														
54					X						X			
55					X									
56							X							
57		4						X						
58		5												
59		7												
60														
61														
62														
63														
64														
65	39													
66	38													
67					X									
68					X						X			
69					X									
70						X								
71						X								
72						X								
73				X		X								
74				X										X
75				X										X
76				X										
77				X						X				
78										X				
79			4											X
80			1		X				X				X	X
81			2		X				X			X	X	
82					X		X					X		X
83				X	X			X			X	X	X	
84				X				X						
85				X		X		X						X
86				X		X		X						
87				X						X		X	X	
88				X					X					X
89				X	X				X					X
90					X									

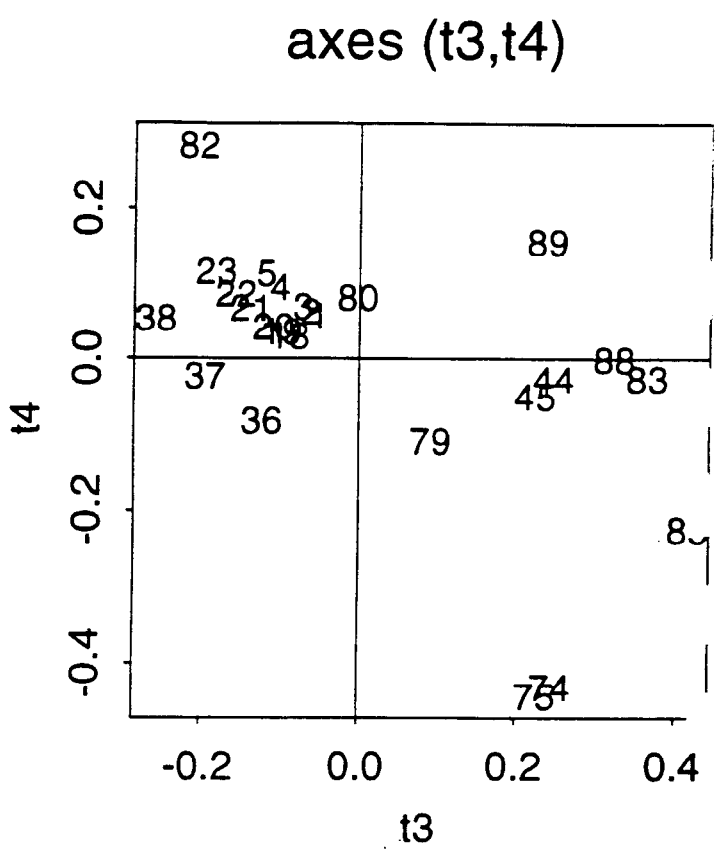
Annexe 4

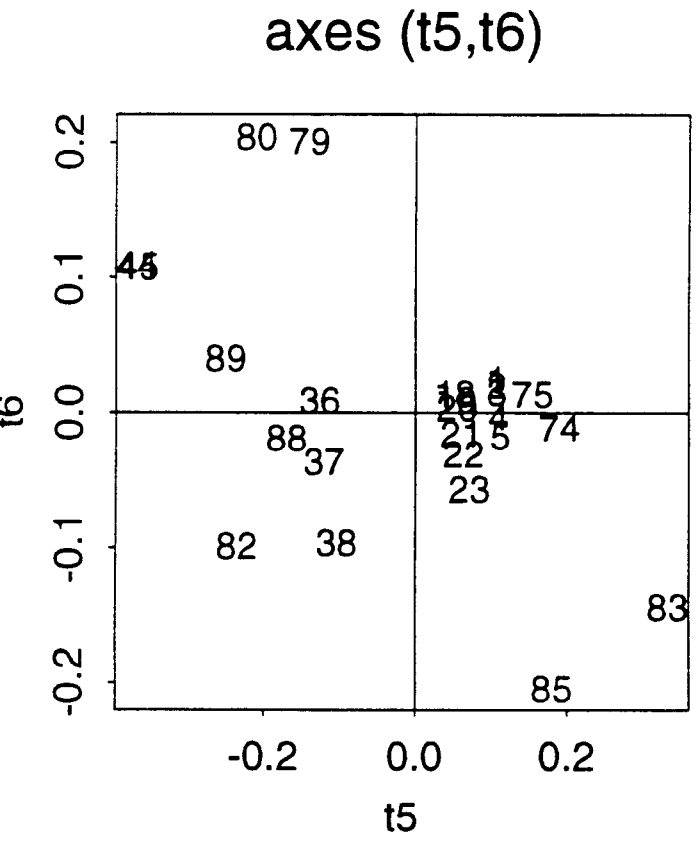
Analyse des cases complètes deuxième sélection

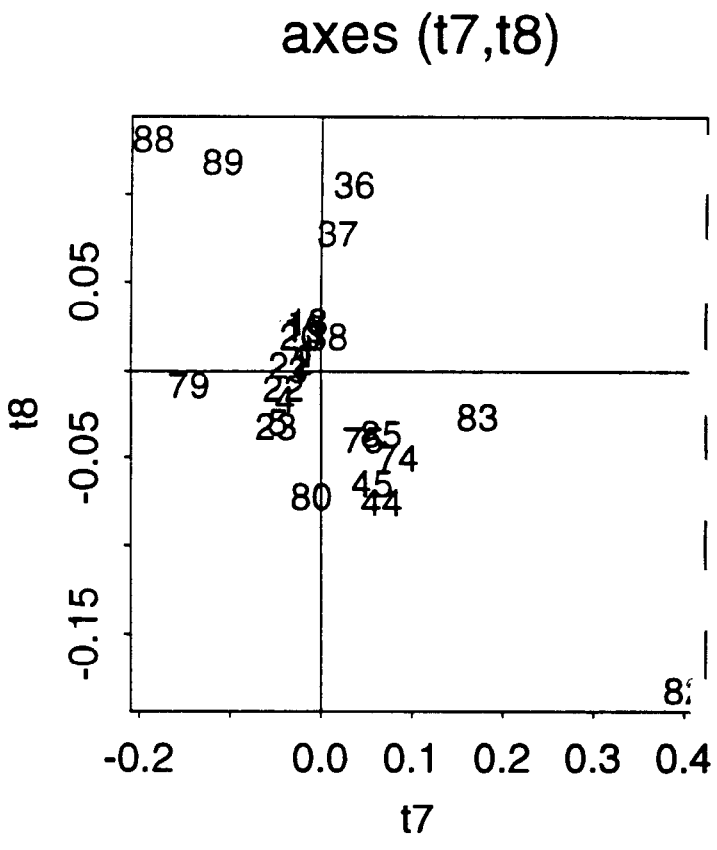

Eqm en fonction de la dimension
Ecarts
30
40
50
60
70
80
Eqm test,pls
Eqm calage,pls
5
10
15
20
25
Nombre d'axes


Variance de Manaus expliquee par axe
Variance expliquee
0.0
0.2
0.4
0.6
0.8
5
10
15
20
25
numero de l'axe

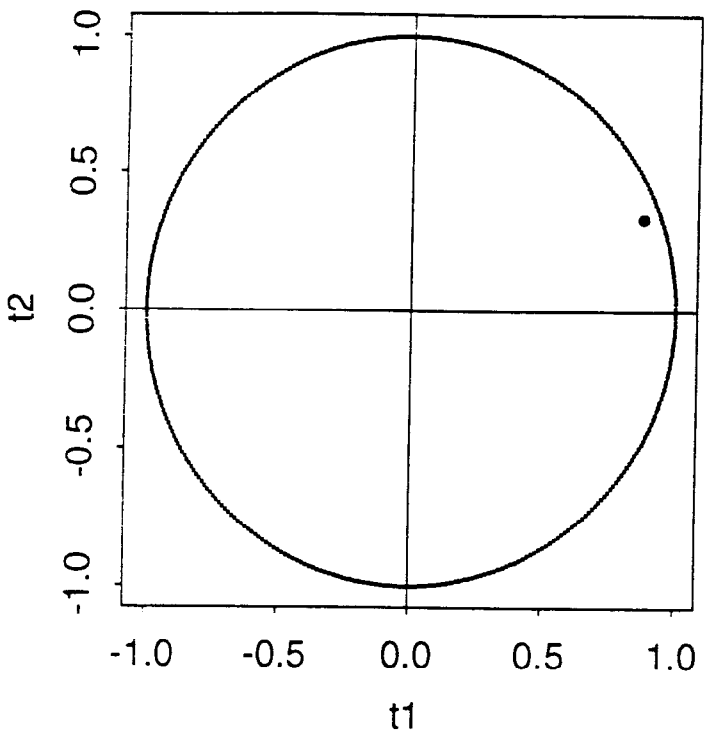

axes (t1,t2)
t1
t2
0.6
0.7
0.8
0.9
1.0
0.4
0.2
0.0
-0.2
-0.4
-0.6


axes (t3,t4)
t3
t4
-0.2
0.0
0.2
0.4
0.2
0.0
-0.2
-0.4

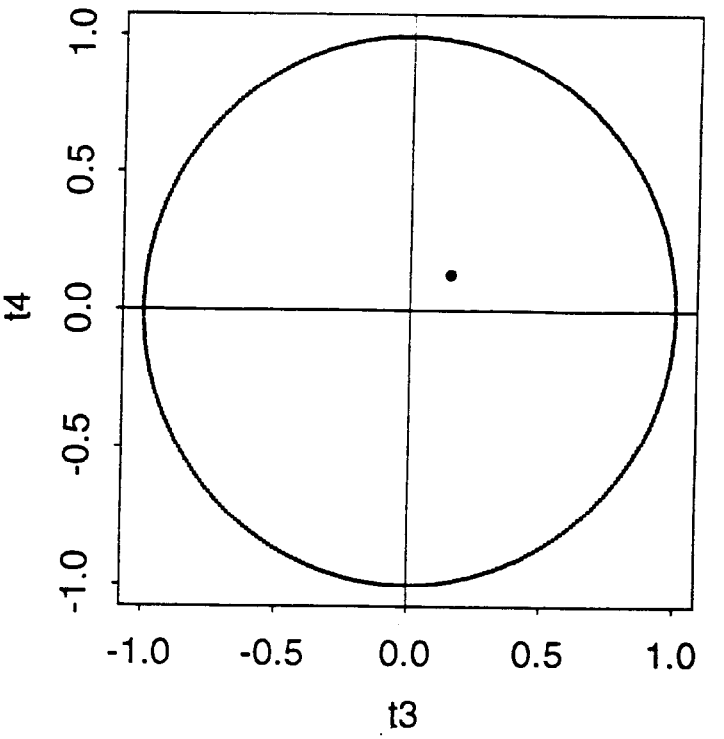

axes (t5,t6)
t5
t6
-0.2
0.0
0.2
0.2
0.1
0.0
-0.1
-0.2


axes (t7,t8)
t7
t8
-0.2
0.0
0.1
0.2
0.3
0.4
0.05
-0.05
-0.15

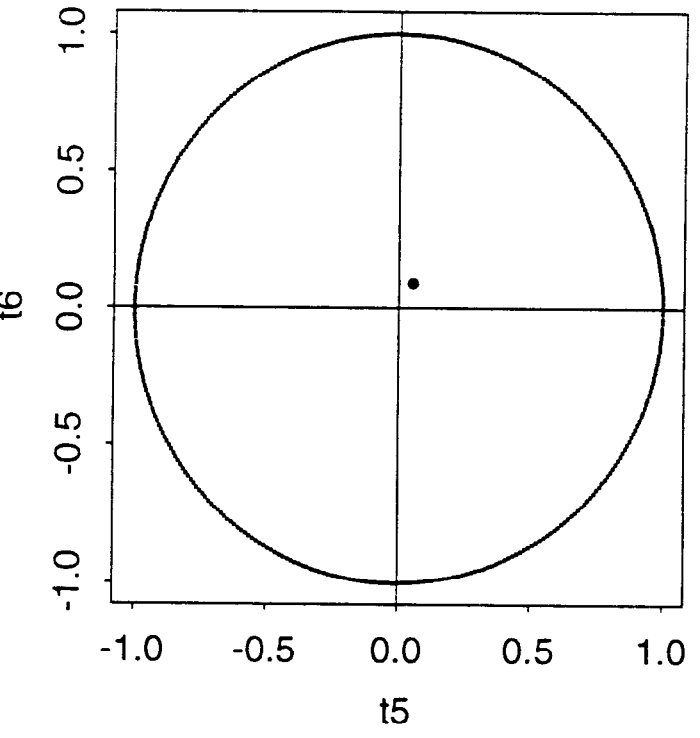
variable Manaus, axes (t1,t2)



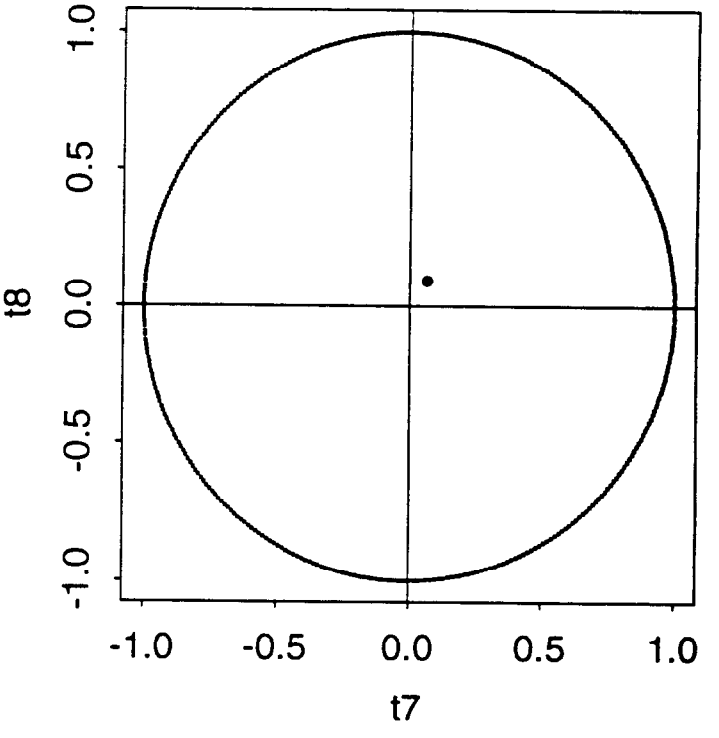
axes (t3,t4)



axes (t5,t6)



axes (t7,t8)



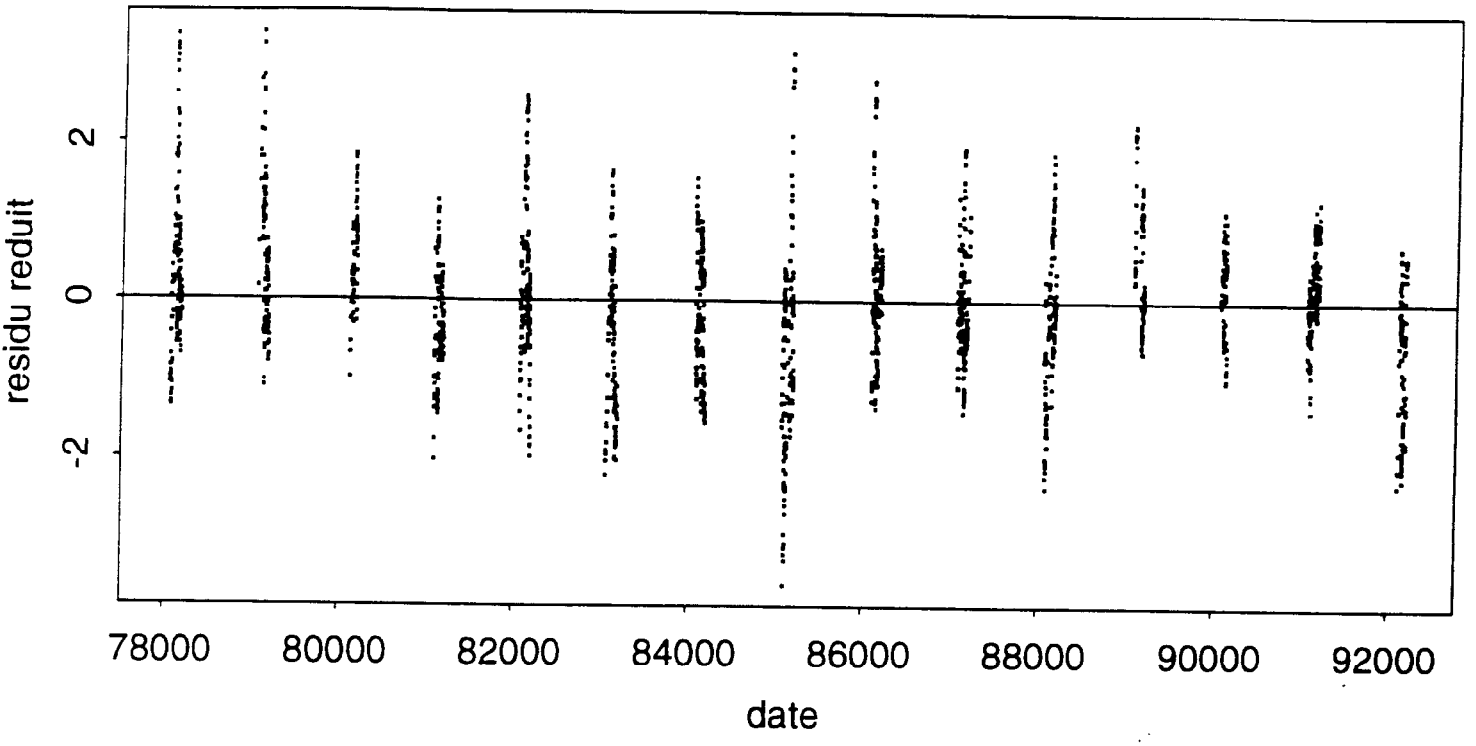
Semables

N° originel	1	2	3	4	5	6	7	8	9	10	11	12	13	SELECTION
1	1								X					X
2	4								X					X
3	6													
4	8			X									X	
5	10			X									X	
18	2												X	X
19	3												X	X
20	5												X	X
21	7													
22	9			X										
23	11			X		X	X							
36	12							X	X		X	X		X
37	13							X	X	X			X	
38	14					X			X	X	X	X		
44		2			X	X	X	X			X	X		X
45		4			X	X	X	X						
74				X	X		X	X		X	X			
75				X	X			X		X	X			
79		6		X		X	X			X	X	X	X	X
80		1			X	X		X	X	X		X	X	X
82		3		X	X	X	X	X	X	X	X	X		X
83			2		X	X	X		X	X	X	X		X
85			1	X	X	X	X		X	X	X	X		X
88		7	3		X		X	X	X		X	X	X	X
89		5		X	X	X	X	X		X	X	X	X	X

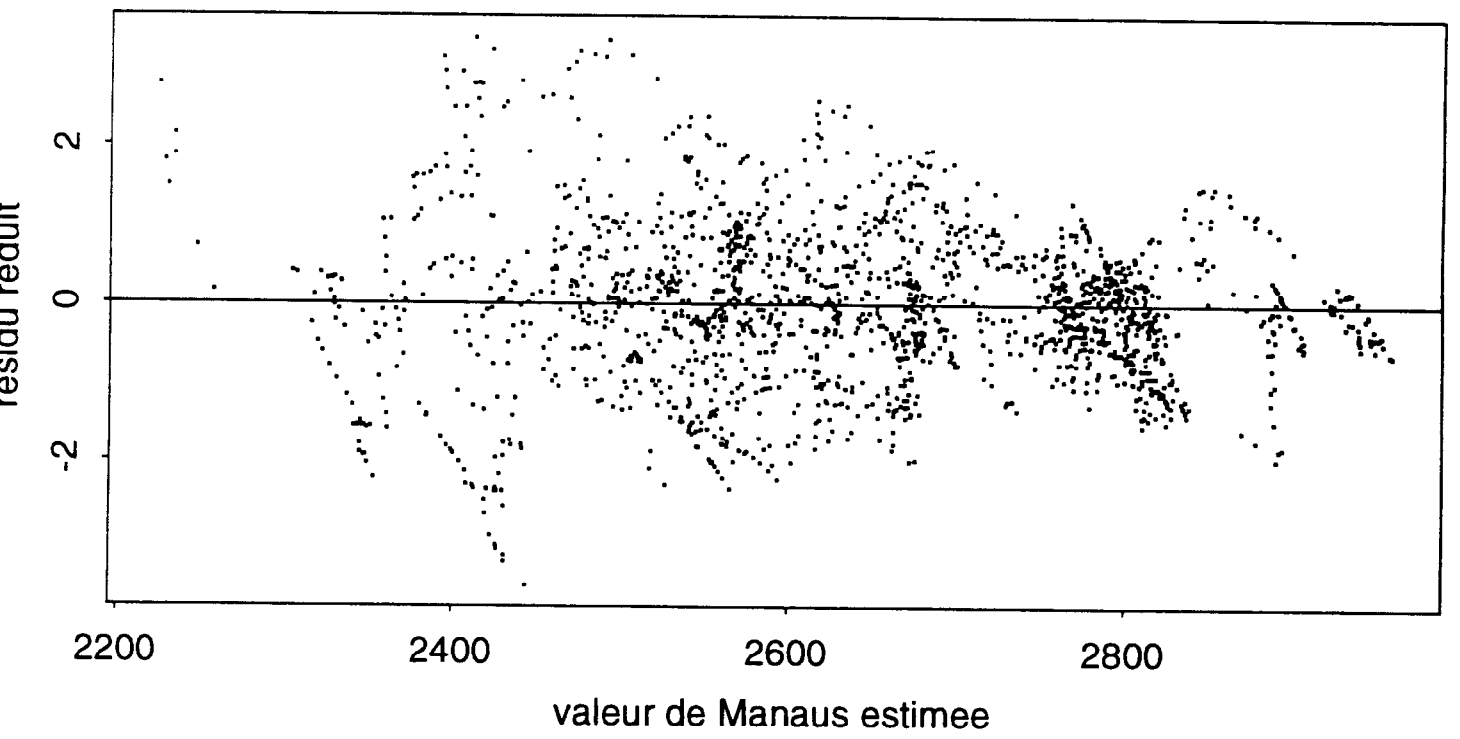
Annexe 5

Analyse des cases complètes graphes des résidus

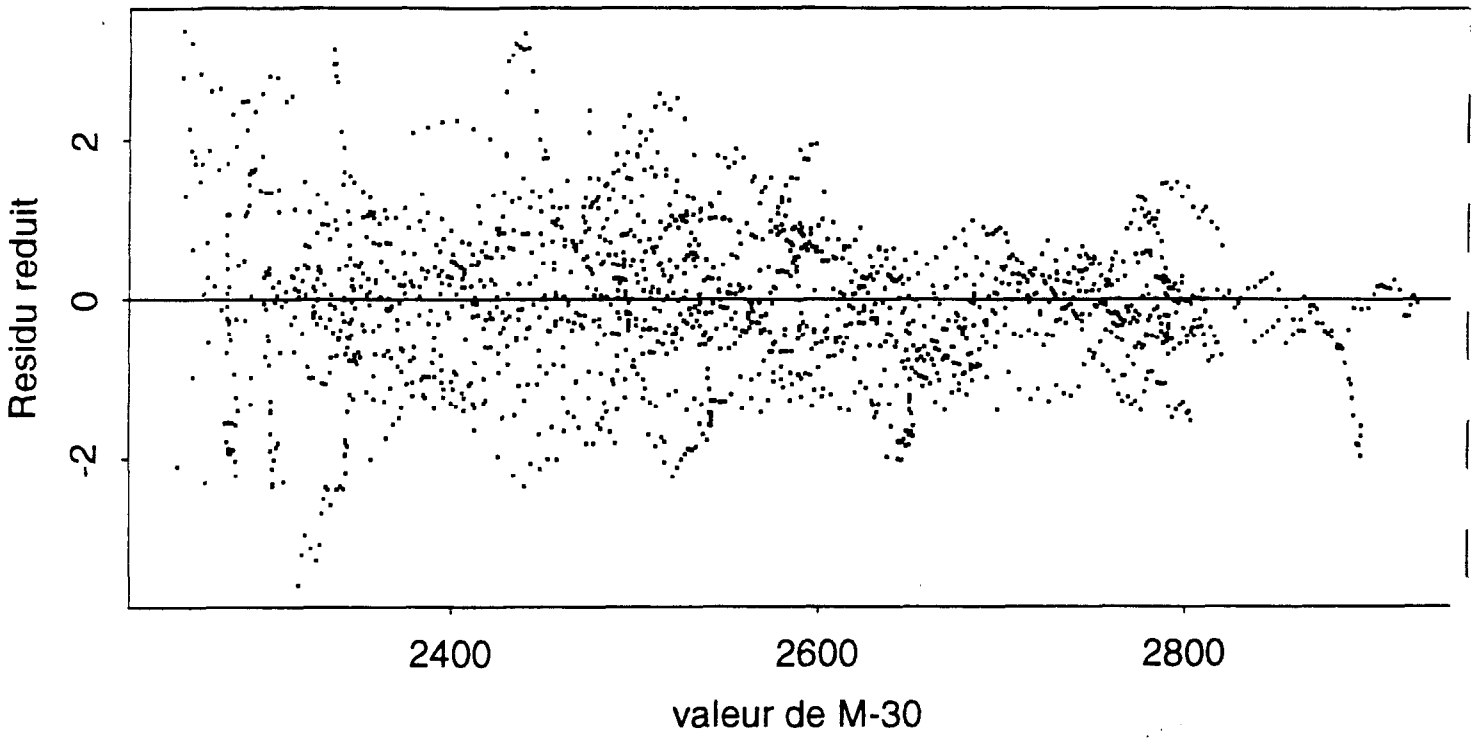
Residus reduits en fonction du temps



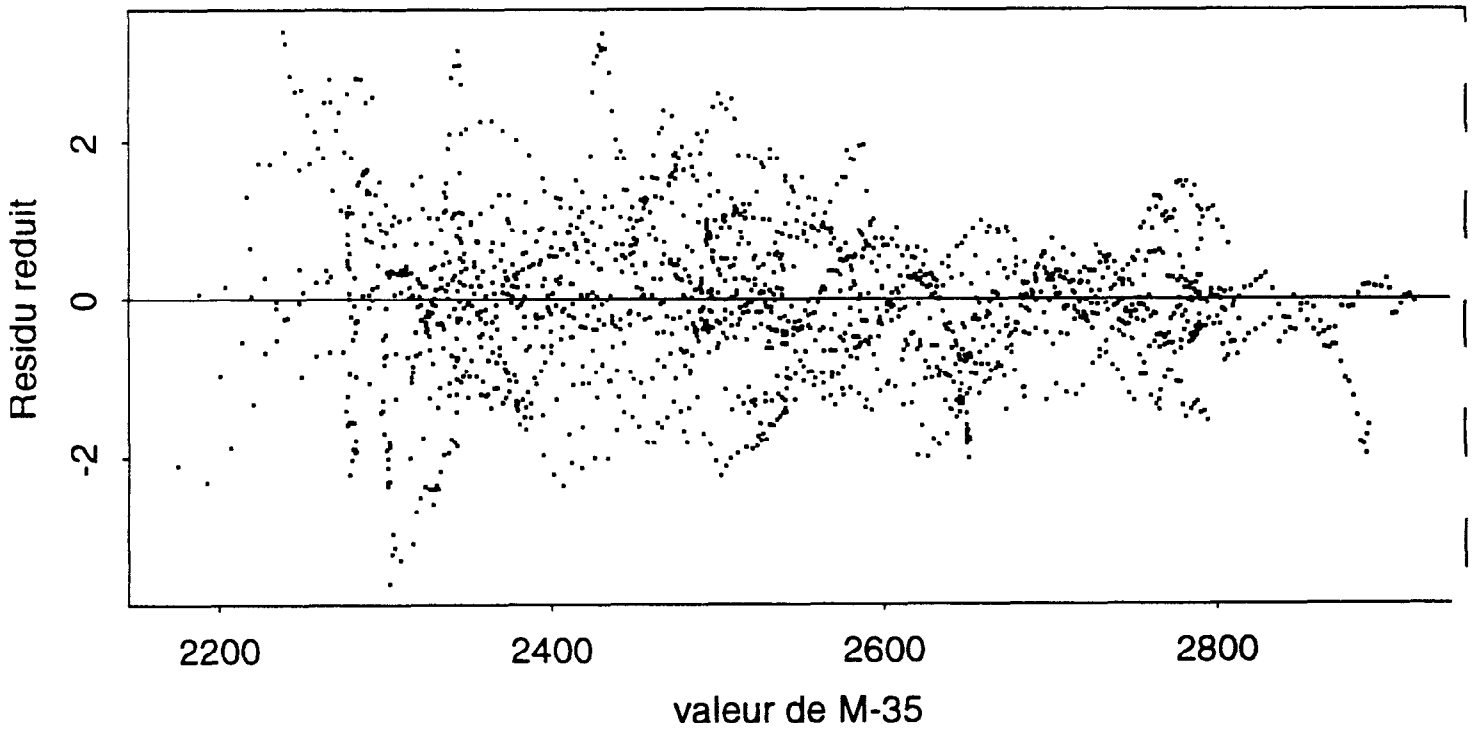
Residus reduits en fonction de l'estimation de Manaus



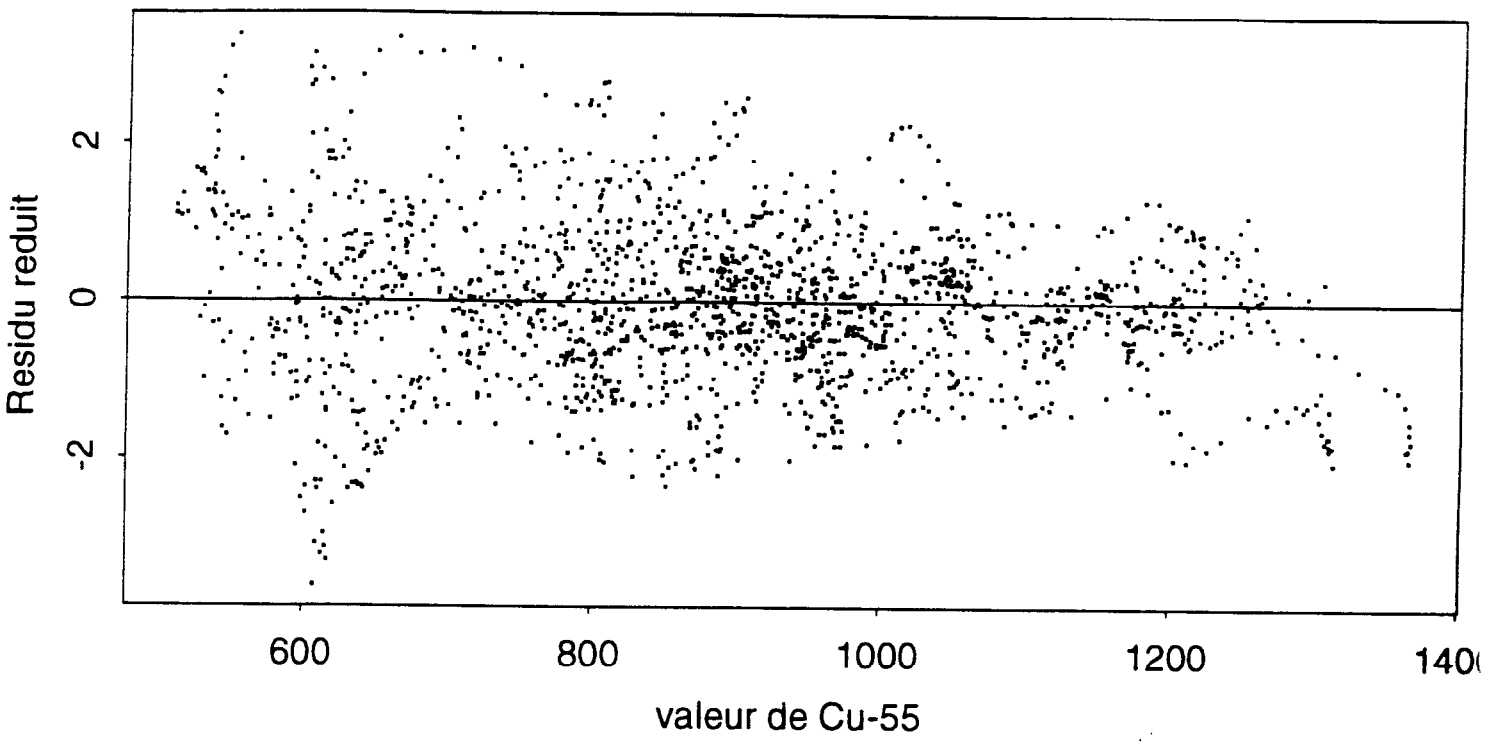
Residus reduits en fonction de M-30



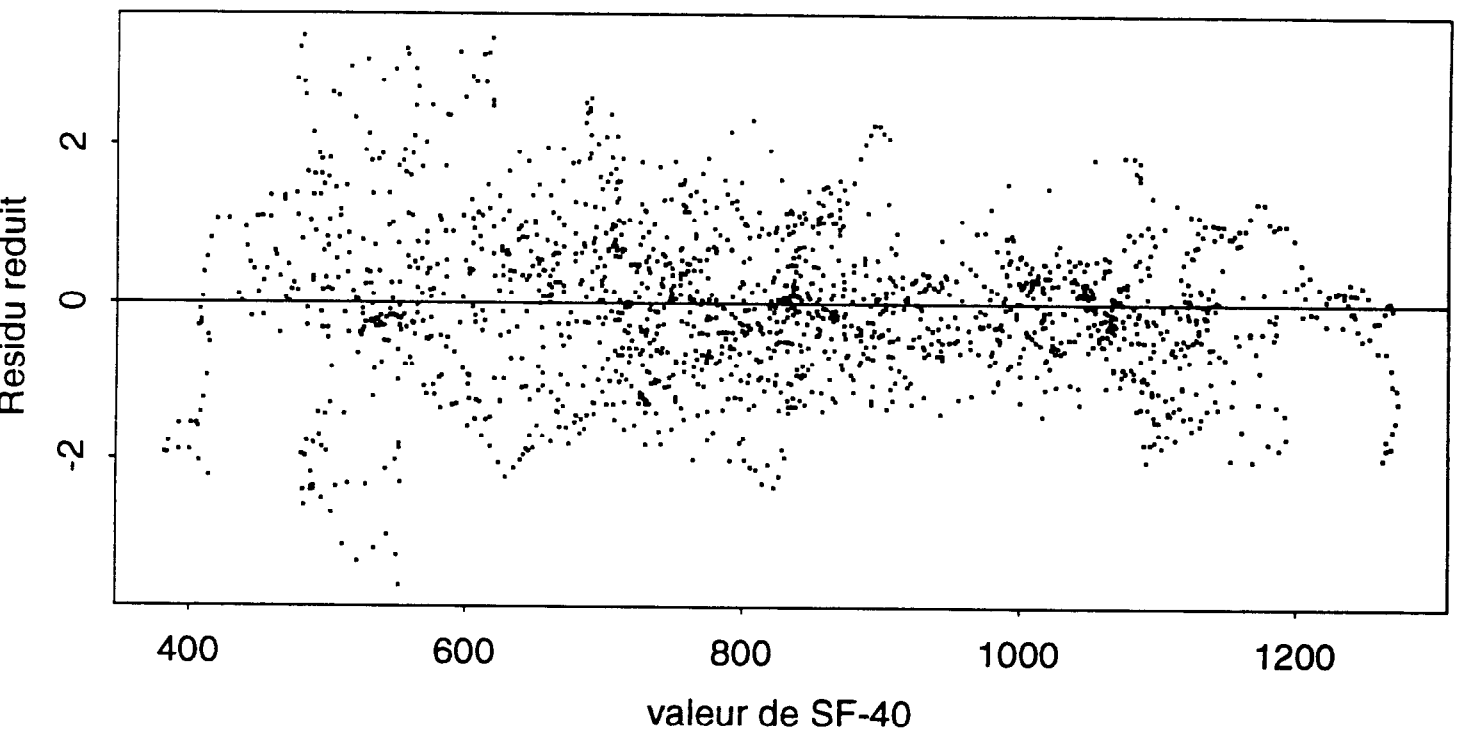
Residus reduits en fonction de M-35

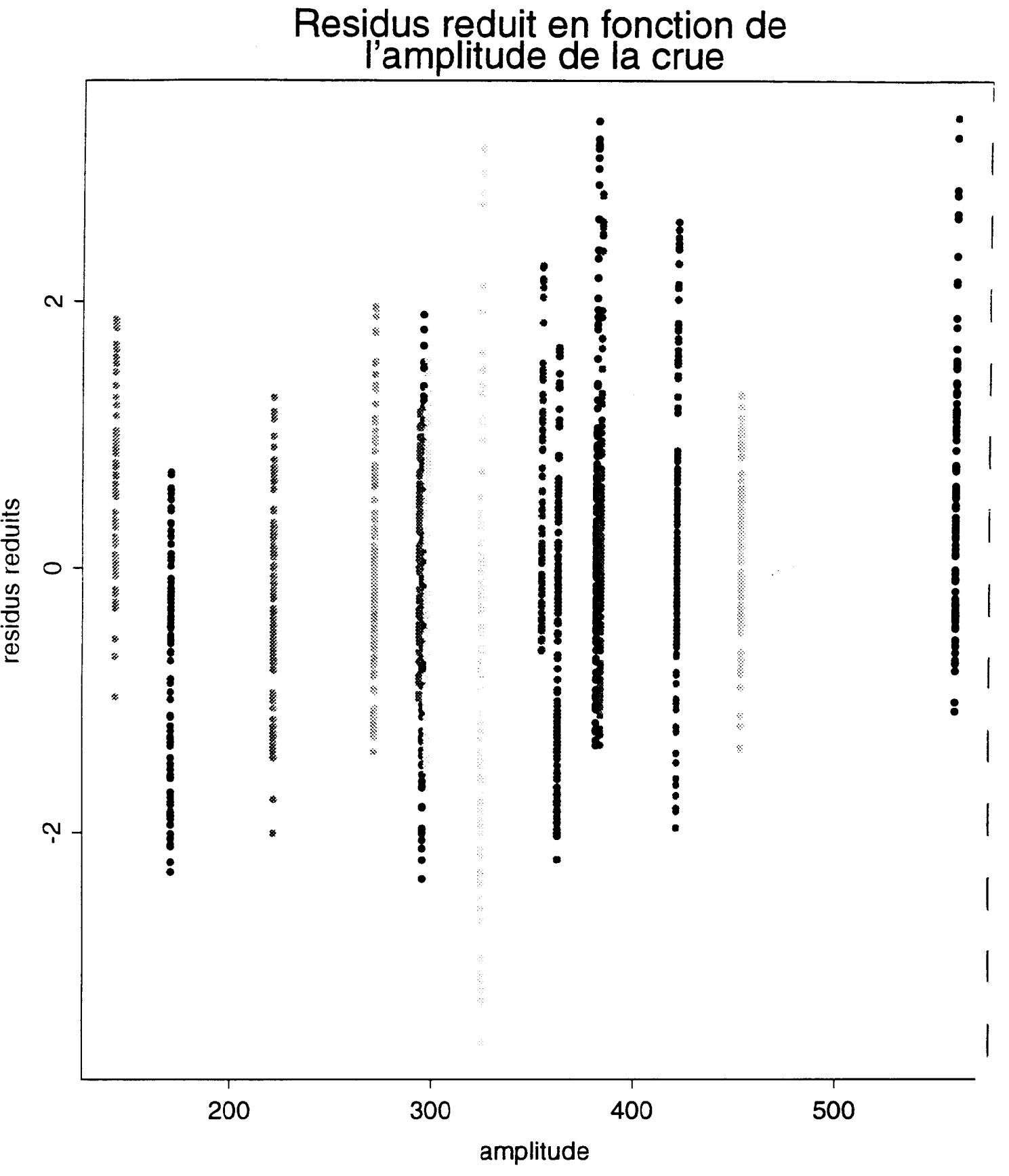


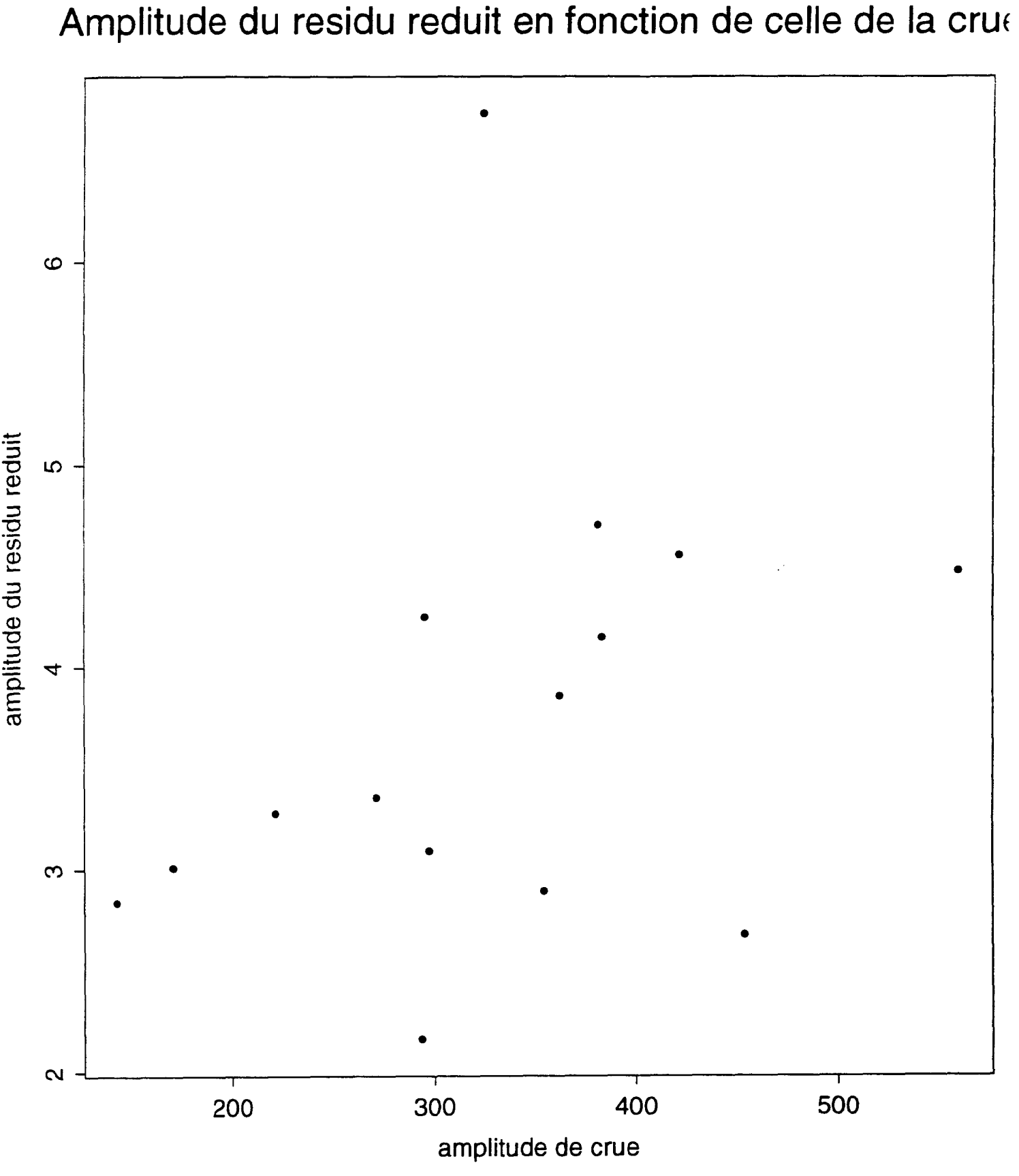
Residus reduits en fonction de Cu-55



Residus reduits en fonction de SF-40

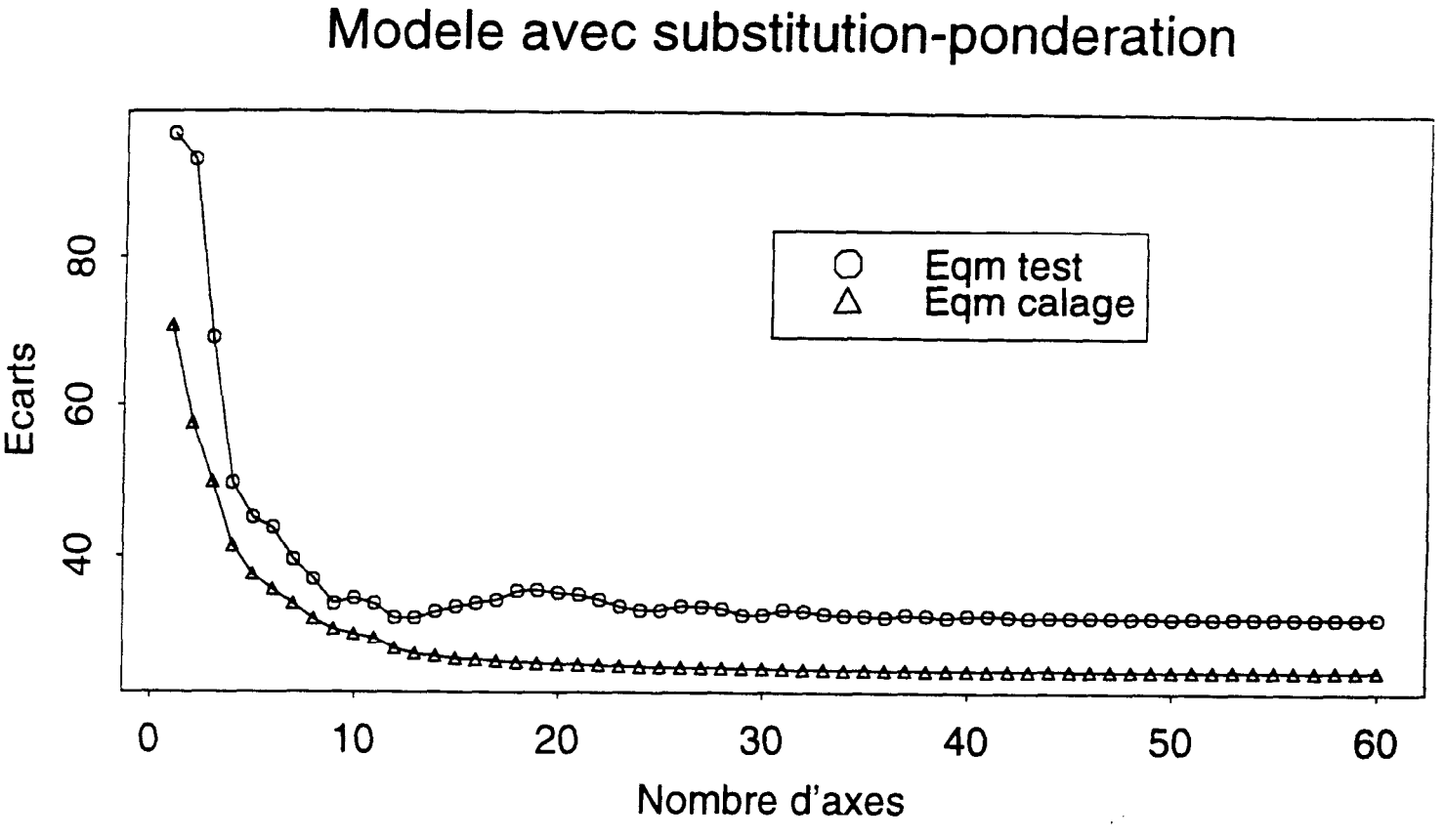


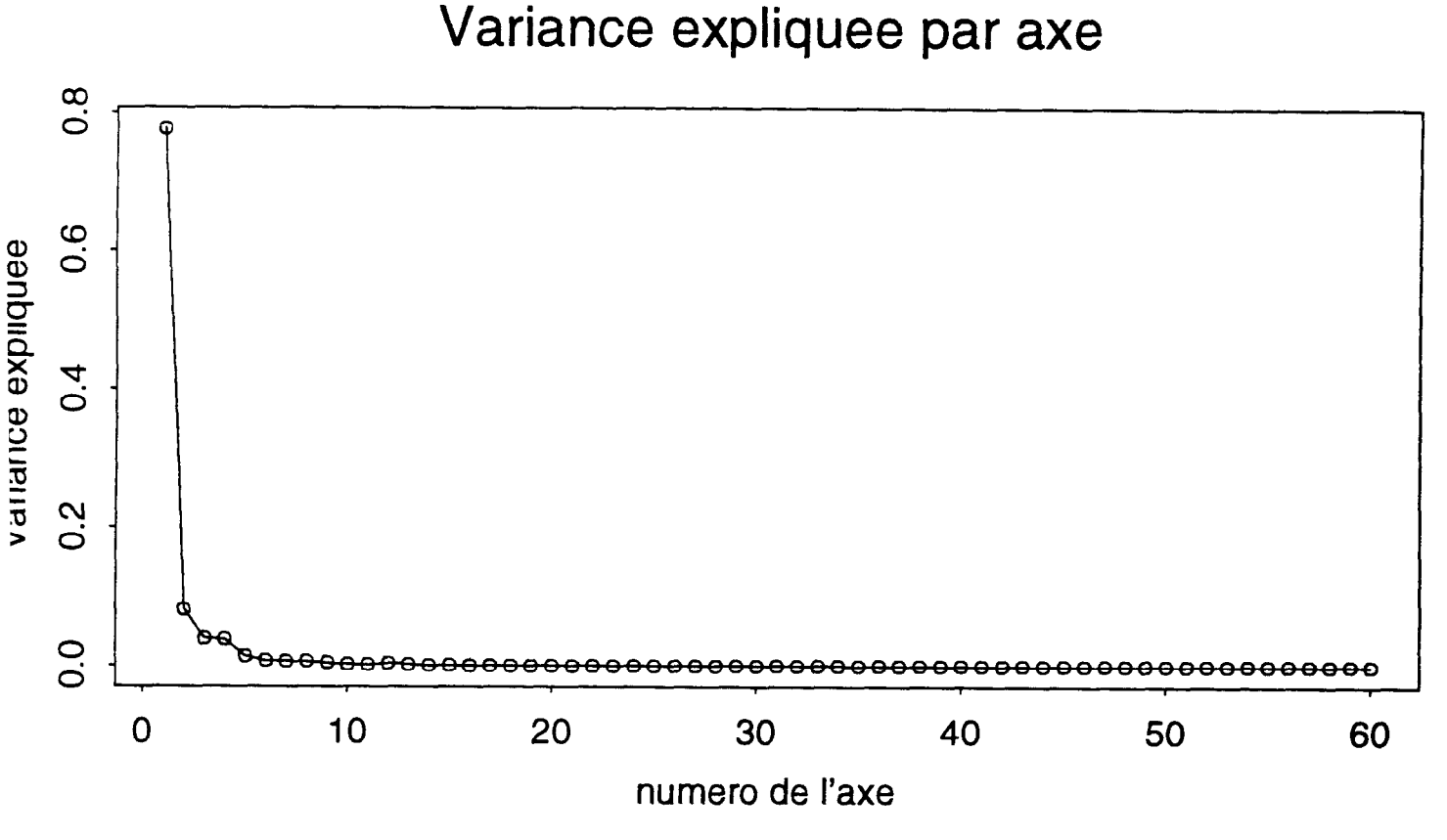

Residus reduit en fonction de
l'amplitude de la crue
residus reduits
2
0
-2
200
300
400
500
amplitude

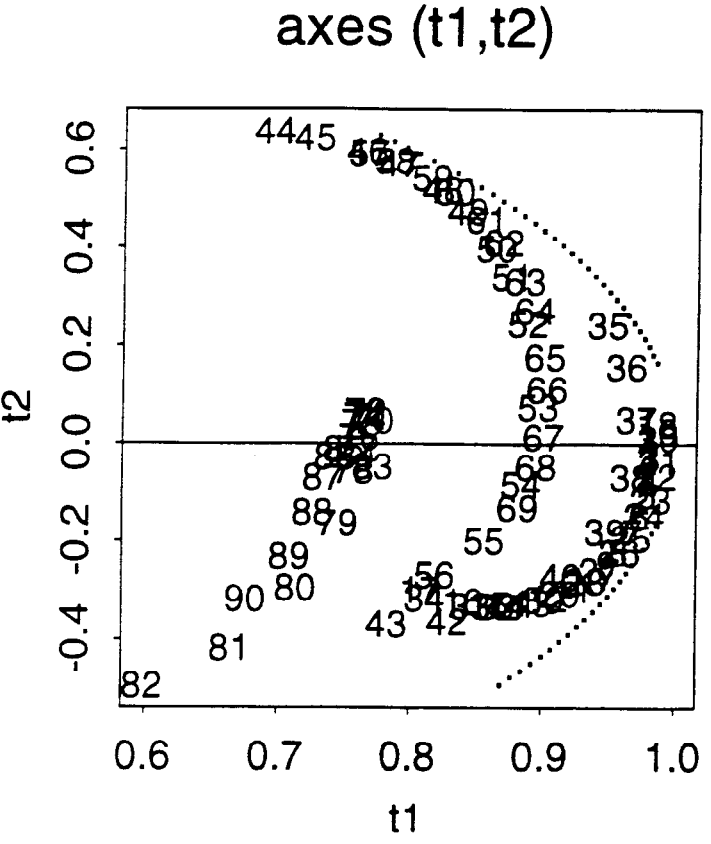

Amplitude du residu reduit en fonction de celle de la crue
amplitude du residu reduit
2
3
4
5
6
200
300
400
500
amplitude de crue

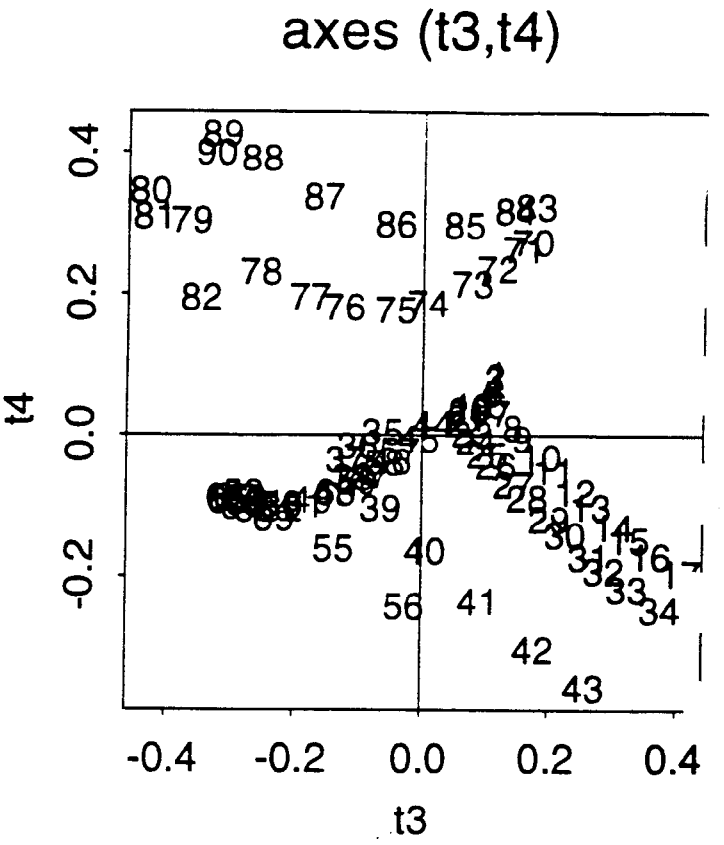
Annexe 6

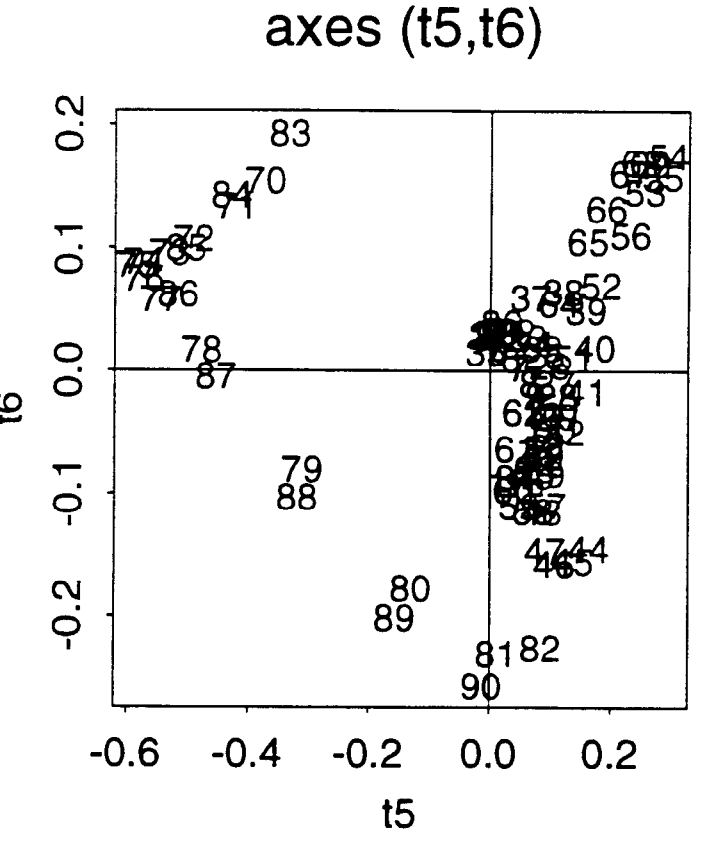
Modèle de substitution avec pondération

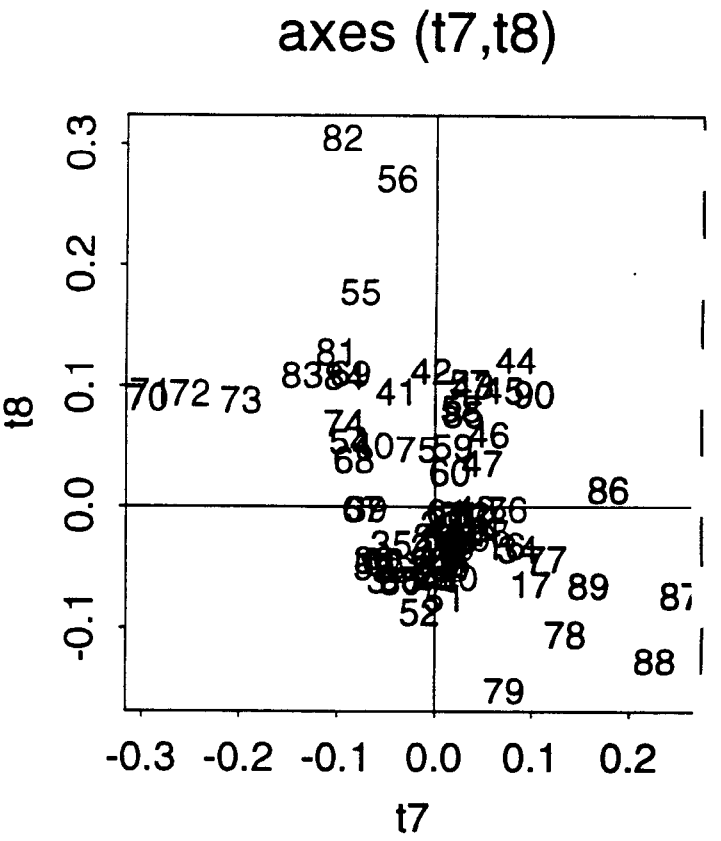

Modele avec substitution-ponderation
Ecarts
40
60
80
Eqm test
Eqm calage
0
10
20
30
40
50
60
Nombre d'axes


Variance expliquee par axe
variance expliquee
0.0
0.2
0.4
0.6
0.8
0
10
20
30
40
50
60
numero de l'axe


axes (t1,t2)
t1
t2


axes (t3,t4)
t3
t4

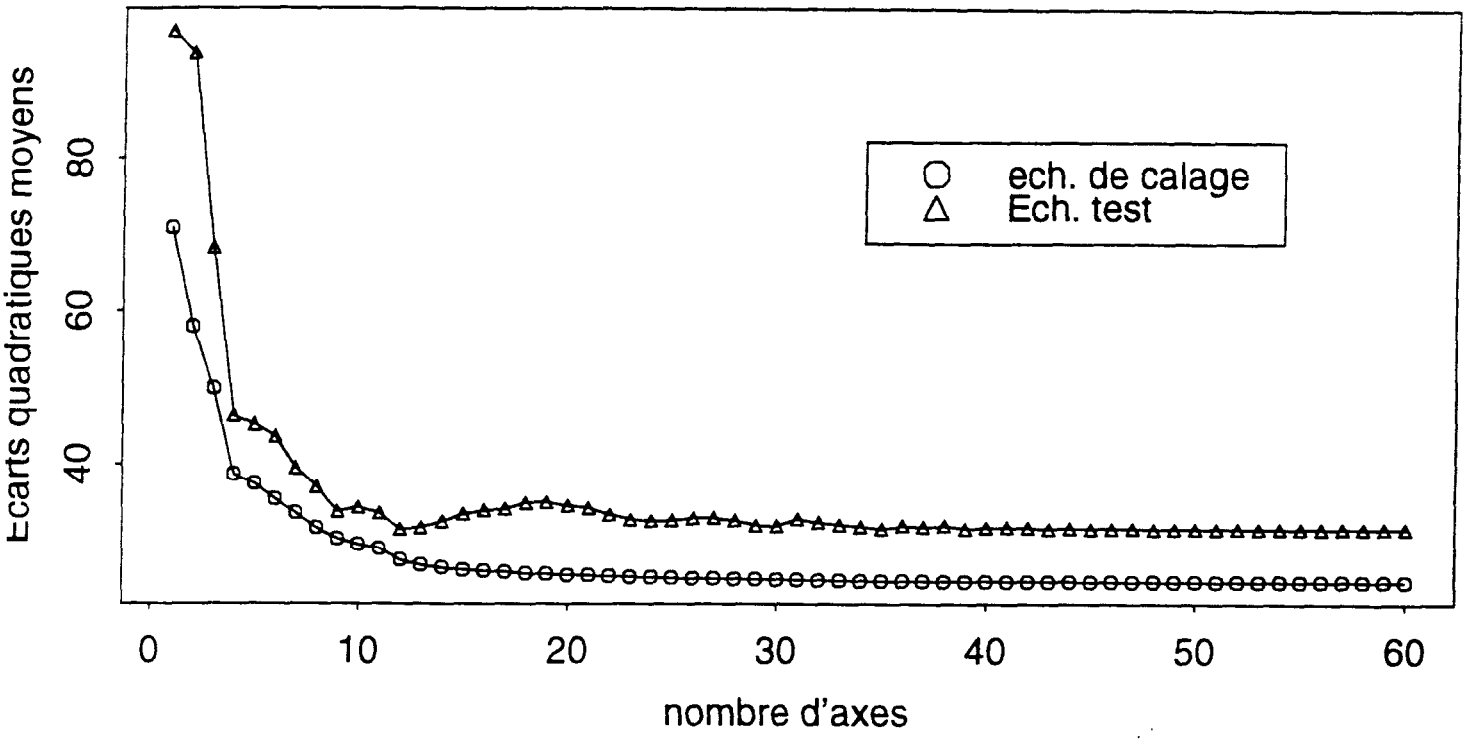

axes (t5,t6)
t5
t6


axes (t7,t8)
t7
t8

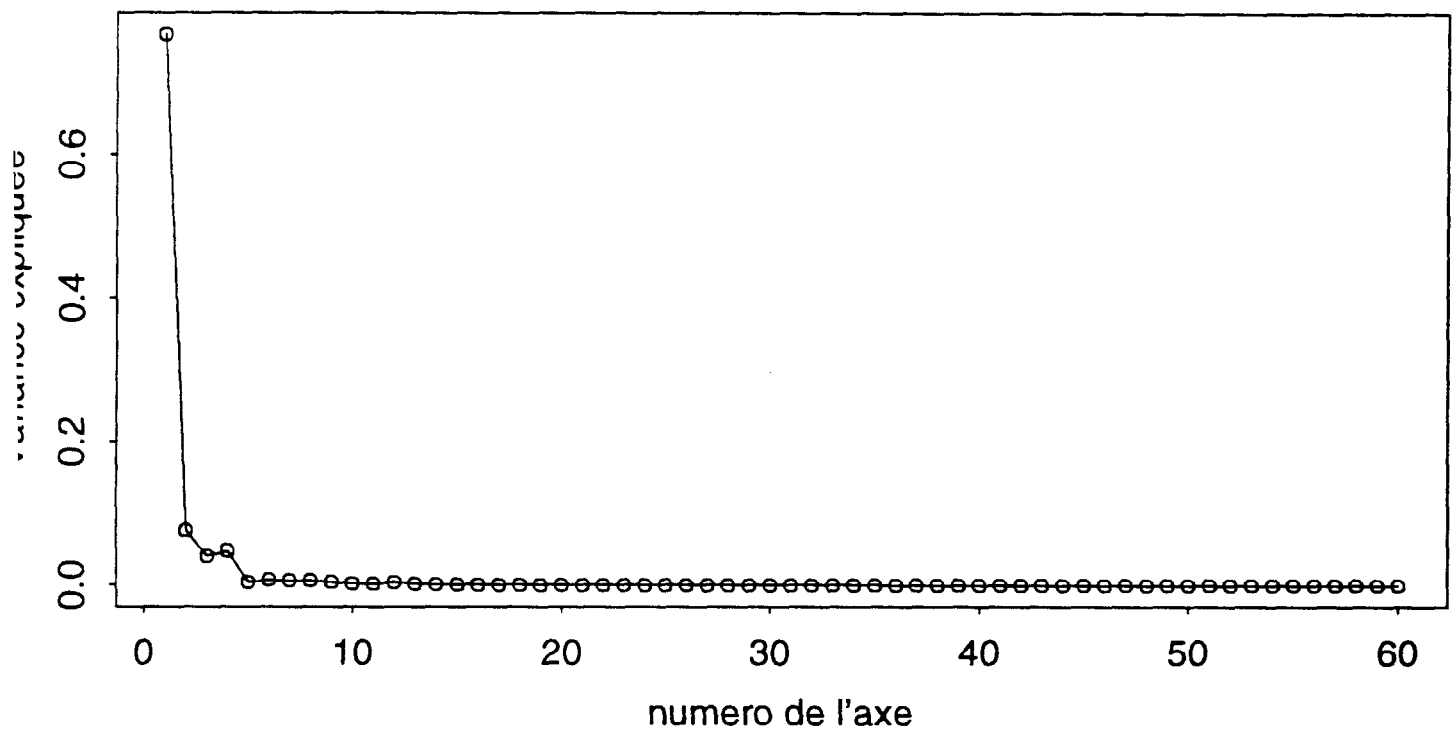
Annexe 7

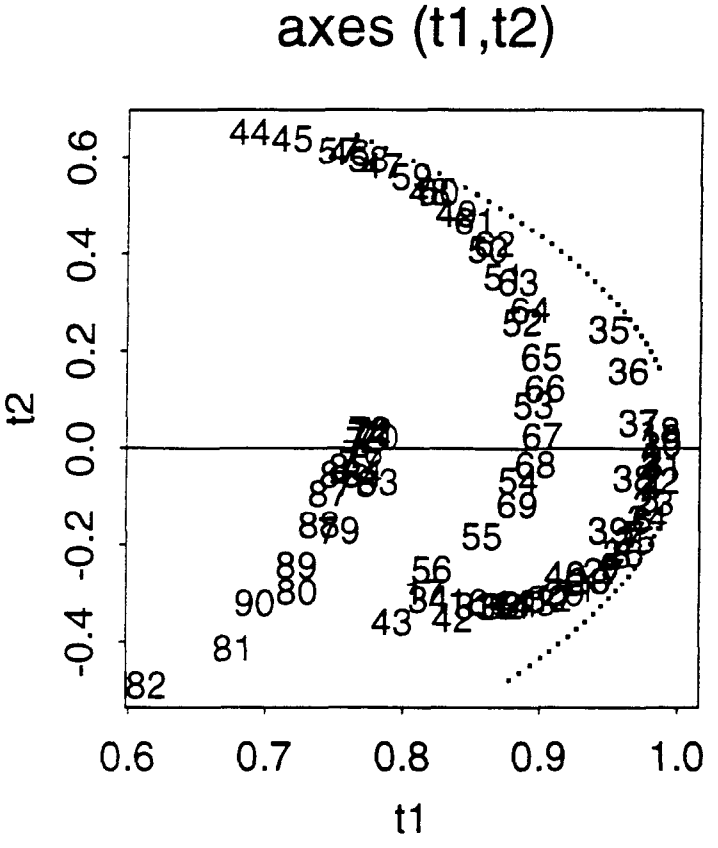
Modèle de substitution sans pondération

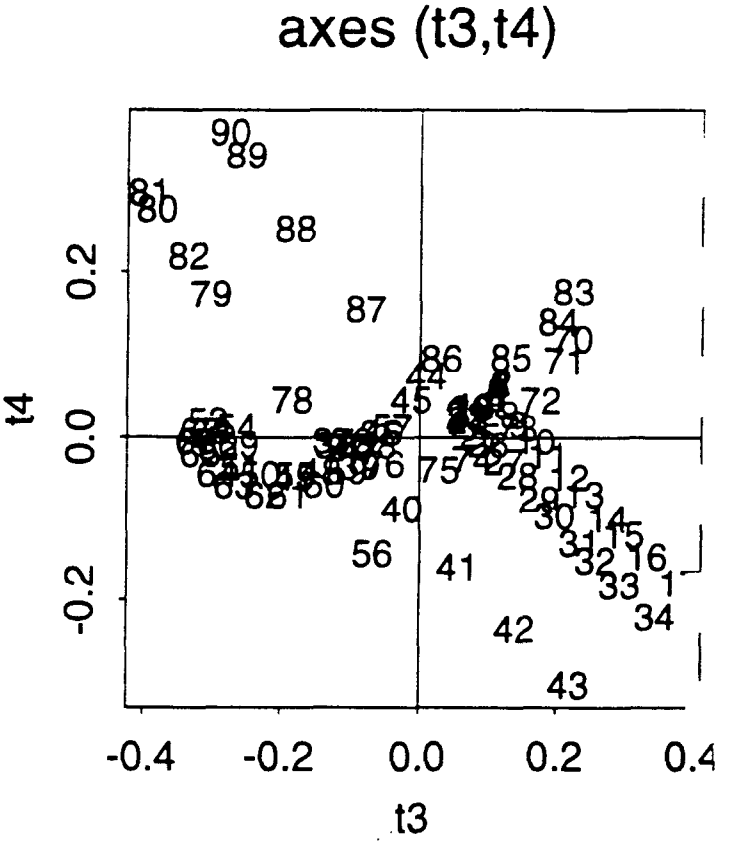
Modele avec substitution, sans ponderation

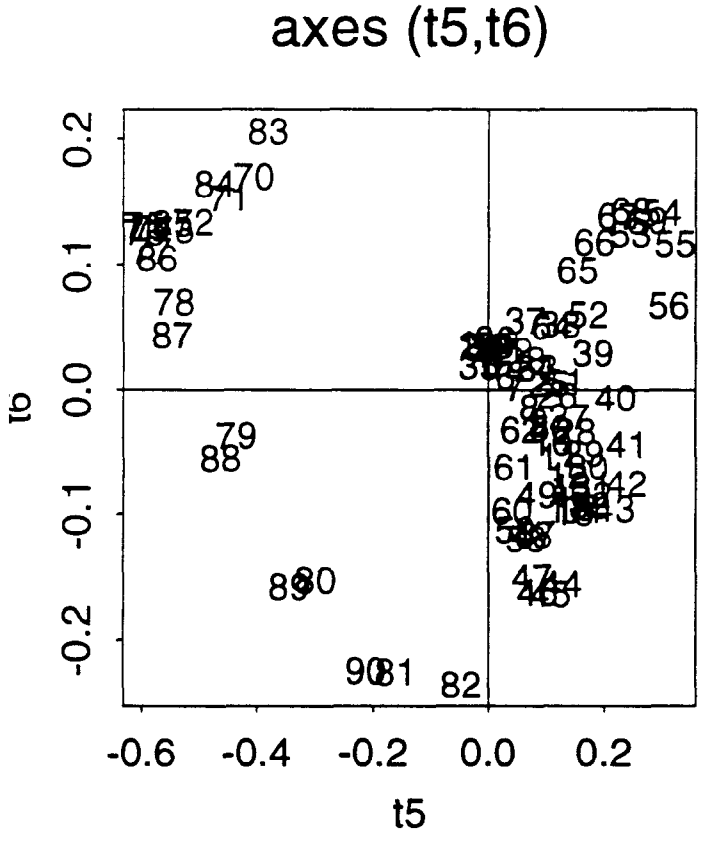


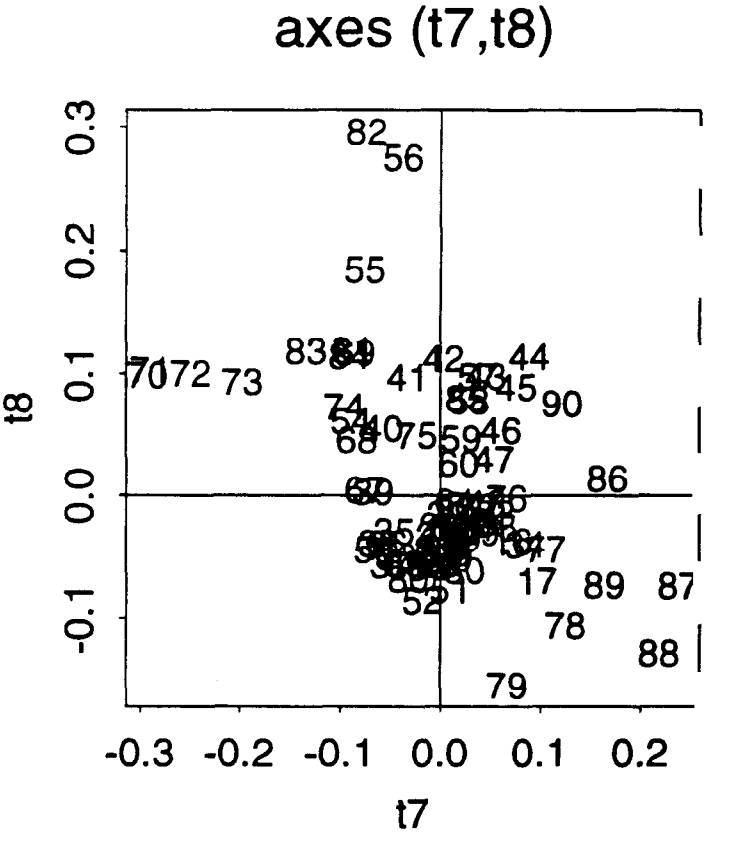
Variance expliquee par axe




axes (t1,t2)
t1
t2
0.6
0.7
0.8
0.9
1.0
-0.4
-0.2
0.0
0.2
0.4
0.6


axes (t3,t4)
t3
t4
-0.4
-0.2
0.0
0.2
0.4


axes (t5,t6)
t5
t6
-0.6
-0.4
-0.2
0.0
0.2
-0.2
-0.1
0.0
0.1
0.2


axes (t7,t8)
t7
t8
-0.3
-0.2
-0.1
0.0
0.1
0.2
-0.1
0.0
0.1
0.2
0.3

RESUME

Manaus, capitale de l'état d'Amazonie, au Brésil, doit faire face annuellement aux lourdes conséquences économiques autant qu'humaines des crues de l'Amazone. La prévision des hauteurs d'eau du fleuve à Manaus est une nécessité pour mettre en oeuvre les mesures de protection des personnes et des biens qui s'imposent. Cette prévision est fonction des cotes enregistrées quotidiennement aux stations de mesure situéees sur l'Amazone et ses affluents, en amont de Manaus. Un modèle de nature statistique utilisant cette information hydrométrique a été recherché. La régression Partial Least Squares, *a priori* adaptée au cas de variables fortement corrélées, a été utilisée. Une adaptation de cette méthode est proposée pour tenir compte de données manquantes, et pour sélectionner un nombre raisonnable de prédicteurs à des fins opérationnelles.

MOTS CLE

- Régression PLS, données manquantes, auto-corrélations, sélection des variables.

- PLS Regression, missing values, auto-correlations, variable selection.