REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

ECOLE NATIONALE POLYTECHNIQUE

Département HYDRAULIQUE

PROJET DE FIN D'ETUDES

THEME

CONTRIBUTION A L'ETUDE ET A LA SIMULATION DES PARAMETRES HYDROMETEROLOGIQUES PAR L'ANALYSE EN COMPOSANTES PRINCIPALES (ACP)

Proposé par :
M.A.BERMAD
M.N DECHEMI

Dirigé par :
M.A.BERMAD
M.N DECHEMI

Etudié par :
Melle.A.HAMRICHE
Melle.K.TACHET

Promotion Juillet 1993

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

ECOLE NATIONALE POLYTECHNIQUE

Département HYDRAULIQUE

PROJET DE FIN D'ETUDES

THEME

CONTRIBUTION A L'ETUDE ET A LA SIMULATION DES PARAMETRES HYDROMETEROLOGIQUES PAR L'ANALYSE EN COMPOSANTES PRINCIPALES (ACP)

Proposé par :
M.N DECHEMI
M.A.BERMAD

Dirigé par :
M.N DECHEMI
M.A.BERMAD

Etudié par :
Melle.A.HAMRICHE
Melle.K.TACHET

Promotion Juillet 1993

ERRATUM

Page 21 :

$$[R]^m X = \lambda_1^m \Sigma c_j [R]^m V_j = \sum_{j=1}^{P} c_j \lambda_j^m V_j \qquad \text{(II.3.2.a)}$$

$$[R]^m X = \lambda_1^m \sum_{j=1}^{P} c_j \left(\frac{\lambda_j}{\lambda_1}\right)^m V_j \qquad \text{(II.3.2.a)}$$

$$[R]^m X = \lambda_1^m c_1 V_1 + \lambda_1^m \sum_{j=1}^{P} c_j \left(\frac{\lambda_j}{\lambda_1}\right)^m V_j \qquad \text{(II.3.2.b)}$$

Page 22 :

$$X_1 = \frac{Y_1}{\|Y_1\|}$$

Page 67 :

$$\sigma_{Xj} = COR\,(C'_1, V_j) = \frac{\sigma_{Xi}}{\sigma_{C'1}}\ COR\,(C'_1, X_j)$$

Page 70 :

$$Z_1 = (-2 \ln U_1)^{1/2} \sin 2\pi U_2$$

$$Z_2 = (-2 \ln U_2)^{1/2} \cos 2\pi U_1$$

Page 72 :

Lire " dans N(0,1) " *au lieu de* " par les Fct.Rpt ".

A ma mère et a mon père
dont les sacrifices à mon égard
n'ont de compensation que mon admiration et ma profonde affection.

A ma grand-mère

A mes frères.

A mes soeurs et belles soeurs.

A mes nièces et mon neveu WAlid

A mes amis

A tous ceux qui me sont chers

KARIMA

A ma mère et mon père, êtres chers et dévoués.

A mes frères et Soeurs

A mes amis

AMEL

AVANT-PROPOS

Nous tenons à remercier *MM. N.DECHEMI* et *A.BERMAD,* pour nous avoir proposé ce sujet et accepté d'être nos encadreurs tout le long de ce laborieux travail. Leurs conseils "guides" et leurs éclaircissements, nous ont permis d'aboutir à ces résultats qui ne seraient ce qu'ils sont sans leurs apports scientifiques et l'abnégation qu'ils ont montré au cours de l'encadrement de ces travaux. Qu'ils trouvent en ces quelques lignes l'expression de notre profonde gratitude et entière reconnaissance.

Nous tenons également à remercier Messieurs les membres du jury qui nous honorent de leur présence, en l'occurrence :

M. OUABDESSALAM
Professeur et Maître de Conférences à l'Ecole Nationale Polytechnique
Ancien Directeur de l'Ecole Nationale Polytechnique.

M. A.SOUAMES
Docteur
Directeur des Etudes à l'Ecole Nationale Supérieure d'Administration et de Gestion.

M. A.KHAMARI
Chargé de cours à l'Institut National d'Agronomie.

M. M.BERKANI
Docteur chargé de cours à l'Ecole Nationale Polytechnique
Chef du Département d'Hydraulique.

M. N.DECHEMI
M. A.BERMAD

Nous n'ometterons pas de remercier *M.F.SADAT* de l'Agence Nationale des Barrages, qui a mis à notre disposition les moyens nécessaires pour atteindre nos objectifs.

Enfin nos vifs remerciements s'adressent à *Melle D.SOUAG* et *MM. M.HELLAL* et *R.GHEZAL* Ingénieurs Polytechniciens pour leurs aides précieuses, leurs présences constantes et leurs soutien moral.

ملخص

الهدف من هذا العمل هو دراسة وتمثيل العوامل «الهيدروجوية» باستعمال التحليل بالمركبات الأساسية.
يخصص الجزء الأول إلى نشاء وتقديم برنامج طريقة التحليل بالمركبات الأساسية، بينما يعنى الجزء الثاني بالتطبيقات في ميدان الهيدرولوجيا. نماذج تمثيل عديدة مقترحة لدراسة مختلف العوامل: عشوائية كانت أم دورية. تتعرض بعدها لدراسة تأثير عدد المركبات الأساسية المستعملة وحجم العينة على نتائج التمثيل.

RESUME

L'étude et la simulation dés paramètres hydrométerologique par l'Analyse en Composantes Principales (ACP) a constitué l'objectif de ce travail.
La première partie est consacrée à la présentation et l'élaboration du programme de l'ACP alors que la seconde est réservée essentiellement aux applications.
Différents modèles de simulation traitant un phénomène aléatoire ou cyclique sont proposés.
L'étude de l'influence du nombre de Composantes Principales utilisé et la taille de l'échantillon sur les résultats est abordée.

SYNOPSIS

The study and simulation of the hydroclimatological parameters by the main Compoments analysis (PCA) is the aim of this work.
The first part is concerned with the presentation and elaboration of the (PCA) program, whereas the second part deal mainly with the applications.
Different simulating patterns dealing with an aleatory or cyclique phenomena shall be propound.
The study of the effect of the number of the main components used, and the size of the sample on the resultas is dealt with.

SOMMAIRE

DEUXIEME PARTIE

INTRODUCTION GENERALE

Aux aléas de la sécheresse qui sévit actuellement en Algérie, s'ajoute une situation de surexploitation de la majorité des nappes et barrages existants et ce, à cause d'une demande de plus en plus croissante pour les besoin du développement.

L'optimisation de la gestion des ressources en eau, quelles que soient les performances des algorithmes de calcul proposés, n'a de sens que par référence à un état donné de l'information hydrologique : il faut donc extraire le maximum de l'information contenue dans les échantillons de mesures disponibles effectuées sur le terrain. s'appuyant sur des bases simples, l'Analyse en composantes Principales (ACP) des données observées se révèle alors utile et pratique.

L'objectif de notre travail est donc, d'étudier les variables hydrologiques et proposer des modèles de simulation pour les différents paramètres hydrométérologiques par l'ACP. La génération de longues séries synthétique de ces différents paramètres constitue, pour les gestionnaire de la ressources en eau d'aide à la décision

PREMIERE PARTIE

Cette première partie est consacrée à l'exposé de la théorie de l'Analyse en Composantes Principales, on élaborera un programme permettant l'utilisation de cette technique dans le domaine de l'Hydrologie, auquel on joindra un exemple d'éxécution.

CHAPITRE I
THEORIE DE L'ANALYSE EN COMPOSANTES PRINCIPALES

I.1. INTRODUCTION A L'ANALYSE DES DONNEES

Les méthodes d'analyse de données ont largement démontré leur efficacité dans l'étude de grandes masses d'informations.Ce sont des méthodes multidimensionnelles, en opposition aux méthodes de statistique descriptives simples ; qui ne traitent qu'une ou deux variables à la fois. Elles permettent la confrontation d'un ensemble d'informations, ce qui est infiniment plus riche que leur examen séparé.Les représentations simplifiées de grands tableaux de données que ces méthodes permettent d'obtenir, s'avèrent un outil de synthèse remarquable.

L'analyse de données (ADD) a été définie par J.P. BENZECRI (voir Réf N°7) comme étant : ***"un outil pour dégager de la langue des données le pur diamant de la véridique nature"***. Parmi les méthodes utilisées, on étudiera en particulier, celles de l'analyse factorielle à partir de laquelle plusieurs variantes se sont développées :

■ L'analyse factorielle des correspondances (AFC) : méthode de description des données qualitatives, proposée pour l'étude des tableaux de contingence (croisement de deux caractères nominaux).

L'analyse discriminante (AD) : possède des motivations descriptives et décisionnelles ; elle cherche à déterminer les axes donnant une meilleur discrimination du nuage de points.

- **La variable à expliquer est nominale (qualitative).**
- **Les variables explicatives sont métriques (mesurables).**

■ L'analyse de variance (AV) : applicable dans le cas inverse de l'analyse discriminante c'est-à-dire :

- **Variable à expliquer mesurable.**
- **Variable explicative nominale.**

■ L'analyse des proximités : se base sur les tableaux de distance (tableau carré symétrique). Cette méthode est dite aussi procédé non métrique.

L'analyse en composantes principales (ACP) : permet de visualiser l'information contenue dans un tableau de données quantitatives.

I.2 ANALYSE EN COMPOSANTES PRINCIPALES (ACP)

I.2.1 HISTORIQUE

Les premières analyses à plusieurs variables remontent au tout début du XXieme siècle (PEARSON 1901 , SPEARMAN 1904,BURT 1909).

Les bases théoriques de presque toutes celles utilisées actuellement ont été établies, de façon quasi définitive vers 1930 dans un trés court laps de temps (MAHALANOBIS 1927 ; FISHER 1928 ,1936 HOTTELING 1931, 1933, 1936 ; BARCETTE, 1933) avec des compléments importants aprés 1950 (RAO 1925, ANDERSON 1958).

Faute de moyens de calcul suffisants, ces méthodes n'ont pu être utilisées que de façon extrêmement limitée.

Jusqu'aux années 60, ces méthodes étaient perfectionnées et s'enrichissaient de variantes mais toutes, restaient inabordables par les praticiens car elles nécessitaient une masse considérable de calcul.

C'est l'apparition, puis l'extraordinaire développement des ordinateurs qui permirent la vulgarisation des techniques statistiques d'analyses des données.

I.2.2 DEFINITIONS

Dans le domaine d'hydrologie l'étude des phénomènes hydrométérologiques et plus particulièrement le traitement de données constitue une tâche trés complexe. Devant la quantité importante de données, l'analyse en composante principale (ACP) s'impose en tant que technique d'analyse des données.

Dans ce contexte G. KENDALL écrit : ***"dans l'analyse en composantes, avec l'espoir de réduire le nombre des dimensions suivant lesquelles varient les données et aussi parfois d'interpréter les composantes".*** (voir Réf N° 2). Autrement dit l'ACP permet de réduire l'information en ne conservant que ses variations essentielles.

I.2.2.1 Notions d'individu et de caractère :

On distingue généralement deux ensembles : les individus et les caractères relatifs à ces individus. Le terme "individu" peut désigner selon les cas : une année d'observations ou une autre unité de temps. L'ensemble des individus peut provenir d'un échantillonnage dans une population ou, il peut s'agir de la population toute entière (cas rare surtout en hydrologie).

L'individu "i" est décrit par le vecteur appartenant à R^P

$$X_i = \{ X_{ij} / j=1 \text{ à } P \}$$

De plus chaque individu "i" est muni d'un poids P_i tel que :

$$\forall i \quad i=1 \text{ à } N \qquad \begin{cases} \text{on a } P_i > 0 \text{ (N : nombre d'individus)} \\ \text{et} \\ \sum_{i=1}^{N} P_i = 1 \end{cases}$$

(en général $P_i = 1/N$ pour i =1à N, le cas $P_i \neq 1/N$ peut se présenter si les individus n'appartiennent pas à la même population)

Sur un individu, on relève un certain nombre de caractères, dits aussi variables, désignant en général un paramètre intervenant dans un phénomène complexe à étudier, exemple : répartition spatiale ou temporelle des précipitations dans un bassin versant.

Le caractère (ou variable) "j" est décrit par le vecteur de $\mathbb{R}^N$

$$X_j = \{ X_{ij} / \ i=1\text{à } N\}$$

Ainsi si l'ensemble des individus doit être homogène, l'ensemble des variables peut être hétérogène.

I.2.2.2. Présentation des données

On dispose d'un ensemble de données qui peuvent être rangées dans un tableau à double entrée de N lignes et P colonnes.

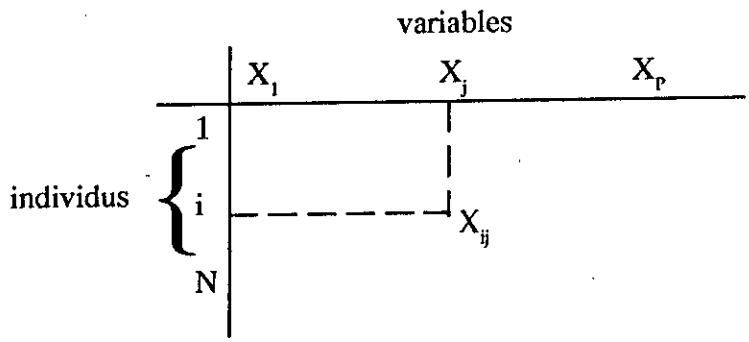


Tableau "individu x caractère"

Au croisement de la ligne "i" et de la colonne"j" se trouve la valeur prise par la variable j sur l'individu i notée X_{ij}.
Ceci étant, il est utile de présenter la technique de l'ACP sous ses deux aspects : algébrique et géométrique.

I.2.2.3. Définition algébrique

L'ACP considère P variables pour lesquelles on dispose de N individus, donc la matrice [X], résultat du croisement "NxP" est la matrice de données

$$[X] = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2P} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{iP} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{N1} & x_{N2} & \cdots & x_{Nj} & \cdots & x_{NP} \end{bmatrix}$$

Définissons les paramètres statistiques des variables :

Moyenne

$$\overline{X_j} = \sum_{i=1}^{N} P_i X_{ij} = (1/N) \sum_{i=1}^{N} X_{ij}$$

$\overline{X_j}$: moyenne de la j^ieme variable

- **Variance :**

$$\sigma_{Xj}^2 = Var_j = \sum_{i=1}^{N} P_i (X_{ij} - \overline{X_j})^2 = (1/N) \sum_{i=1}^{N} (X_{ij} - \overline{X_j})^2$$

Var j : Variance de la j^ieme variable

σ_{Xj} : écart type de la j^ieme variable

- **Covariance :**

$$COV(X_j, X_{j'}) = \sum_{i=1}^{N} P_i (X_{ij} - \overline{X_j}) (X_{ij'} - \overline{X_{j'}})$$
$$= (1/N) \sum_{i=1}^{N} (X_{ij} - \overline{X_j}) (X_{ij'} - \overline{X_{j'}})$$

COV $(X_j, X_{j'})$: covariance de la variable j avec la variable j'.

- **Corrélation :**

$$COR (X_j, X_{j'}) = COR (X_j, X_{j'}) / [Varj \times Varj']^{1/2}$$

COR $(X_j, X_{j'})$: corrélation entre la variable j et la variable j'.

Le tableau [X] de départ est remplacé par un tableau [Y] (individus x nouvelles variables) en réduisant le nombre de variables nécessaire pour décrire les individus, avec une perte minimale d'informations. Ces nouvelles variables sont appelées composantes principales (ou CP)
Calculer les composantes principales notées Cj revient à déterminer P relations linéaires entre les variables X_j :

$$\begin{cases} C_1 = a_{10} + a_{11} X_1 + \ldots\ldots\ldots\ldots + a_{1P} X_P \\ C_2 = a_{20} + a_{21} X_1 + \ldots\ldots\ldots\ldots + a_{2P} X_P \\ \\ C_j = a_{j0} + a_{j1} X_1 + \ldots\ldots\ldots\ldots + a_{jP} X_P \end{cases}$$

C_j : j^ieme composante principale
X_j : Vecteur variable
a_{jk} : coefficient du système

On note au passage que les termes a_{j0} désignent le vecteur permettant la translation de l'origine de l'ancien repère vers le centre de gravité du nuage de points. Un centrage des données initiales annule les coefficients a_{j0} $\forall j$.

I.2.2.4. Définition géométrique :

L'Analyse en Composantes Principales est puissante par son support géométrique ; la méthode consiste à rechercher un premier axe qui soit le plus prés possible de tous les points au sens des moindres carrés : tel que la **somme des carrés** des distances des N points à cet axe soit **minimale** ; ou encore la projection de ces derniers sur cet axe aie une dispersion maximale. Cet axe est appelé **"axe factoriel"**.

Un second axe est obtenu après projection des N points sur un hyperplan orthogonal au premier axe, tel que la dispersion des projections des N points sur celui-ci soit toujours maximale, et le processus se réitère P fois.

On obtient ainsi un nouveau système d'axes défini par les nouvelles variables dites Composantes Principales.

I.2.3 FORMULATION MATHEMATIQUE DU PROBLEME :

La recherche des composantes principales est faite sous deux contraintes :

Elles doivent être indépendantes , c'est-à-dire, prise deux à deux, elles présentent obligatoirement des corrélations nulles.

Les axes factoriels doivent être déterminés par ordre d'importance décroissante. Le premier axe expliquera le maximum de la variance totale tandis que le second expliquera le maximum de la variance résiduelle non expliquée par le premier ; jusqu'au dernier axe. Mais l'expérience a montré qu'un nombre Q d'axes nettement inférieur à P suffit pour donner le maximum d'informations.

Optimisation du problème

Déterminer un sous espace W, revient à choisir canoniquement dans R^P un référentiel défini par un vecteur constant $\vec{A}$ de R^P et un système orthonormé de vecteurs $\vec{V}_1,\ldots,\vec{V}_Q$ formant une base du sous espace vectoriel W.

Le nouveau référentiel est obtenu à la suite d'une translation plus une rotation .On définit donc W par :

$$W = \vec{A} + [V_1, \ldots, V_Q]$$

Tel que $\vec{A}$ définit le vecteur de translation de l'origine du repère initial vers celui du nouveau repère, et la séquence $\{V_1, \ldots, V_Q\}$ les vecteurs permettant la rotation autour de l'origine translaté.

W sera choisi de telle sorte que la déformation en projection soit minimum, et les Q nouvelles variables seront alors les composantes des projections des points du nuage dans ce référentiel.

On écrira alors pour tout individu i

$$\vec{C}_i = \vec{A} + \sum_{k=1}^{Q} C_{ik} \cdot \vec{V}_k$$

avec :

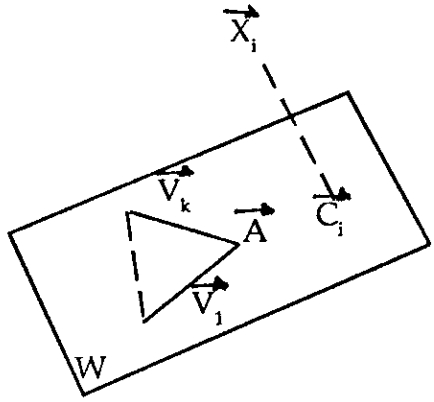
$\vec{C}_i$: projection orthogonale de X_i sur W

$\vec{V}_k$: vecteur propre de rang k

C_{ik} : coefficient des composantes principales.

La déformation globale en projection sur W est mesurée par l'inertie autour de W, notée I_w et définie par :

$$I_w = \sum_{i=1}^{N} P_i \, d^2 (\vec{X_i}, \vec{C_i})$$



$d(\vec{X_i}, \vec{C_i})$: distance entre le point défini par le vecteur X_i et celui défini par sa projection sur W.

La recherche du référentiel W se fait sous l'hypothèse de minimisation de la déformation globale, ce qui revient à minimiser l'inertie I_w.

Toute solution W du problème contient l'extrémité du vecteur G définissant le centre de gravité du nuage des individus.

Soit A le point défini par l'extrémité du vecteur A appartenant à R^p on notera W_A le sous espace passant par A.

Utilisons le théorème de HUYGENS pour montrer que :

pour tout vecteur $\vec{A}$ de R^p

$$I_{WA} \geq I_{WG}$$

Ce théorème s'énonce ainsi :

Soit W_A un sous espace passant par A, W_G le sous espace parallèle à W_A et contenant le centre de gravité G du nuage des individus.
on a :

$$I_{WA} = I_{WG} + d^2(G., W_A)$$

Le théorème ci-dessus montre donc que :

L'inertie du nuage des individus par rapport à l'ensemble des sous espaces parallèles à W_0 est minimum pour le sous espace W_G contenant le centre de gravité du nuage.

Le **sous espace optimal** W cherché contient donc le centre de gravité du nuage

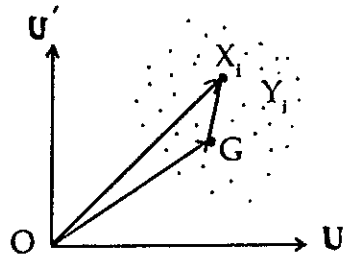
$$W = W_G$$

On peut supposer les données centrées. En effet le point individu du tableau centré s'écrit :

$$\begin{aligned}\overrightarrow{OX_i} - \overrightarrow{OG} &= \overrightarrow{OX_i} + \overrightarrow{GO} \\ &= \overrightarrow{GO} + \overrightarrow{OX_i} \\ &= \overrightarrow{GX_i} = \overrightarrow{Y_i}\end{aligned}$$

$$X_i - G = Y_i$$

(voir figure ci-contre)

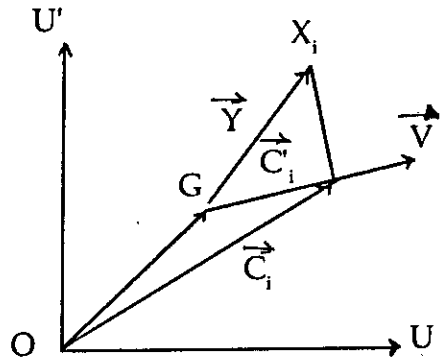


Si C_i est la projection de X_i sur W_G et C'_i la projection de Y_i sur W_0 alors

$$C'_i = C_i - G$$

$$\overrightarrow{C'_i} = \overrightarrow{C_i} - \overrightarrow{OG}$$

(voir figure I.2)



et il est équivalent de minimiser la quantité

$$\sum_{i=1}^{N} P_i \, d^2(X_i, C_i)$$

ou la quantité

$$\sum_{i=1}^{N} P_i \, d^2(Y_i, C'_i)$$

Car tout simplement :

$X_i - C_i = Y_i - C'_i$ (voir Figure ci-contre)

Dans la suite nous supposerons toujours que les données sont centrées c'est-à-dire que :
pour tout i , $Y_i = X_i$, G=0 et $W=W_g = W_0$

Formulation matricielle :

L'objectif de l'ACP étant de maximiser la variance la formulation matricielle du problème est la suivante :

Soit les matrices colonnes V_1 , V_2 , ...V_Q de dimension (N x 1) représentant l'hyperplan formé par les axes principaux vérifiant les conditions de normalité et d'orthogonalité :

$$\forall i,j \qquad \begin{matrix} i=1 \text{ à } Q \\ j=1 \text{ à } Q \end{matrix} \qquad \begin{matrix} V_i^t \, . \, V_i = 1 \\ V_i^t \, . \, V_j = 0 \end{matrix}$$

On veut maximiser la quantité : $\sum_{j=1}^{Q} \text{Var}\,(C_j)$

Sachant que $\text{Var}(C_j) = V_j^t \, . \, [M] \, . \, [R] \, . \, [M] \, . \, V_j$

C_j : CP
R^j : matrice des covariances des variables (X_1 , X_2 ,...X_P)
M : métrique définissant le produit scalaire sur l'espace R^P.

Choix de la métrique

La métrique M possède deux options classiques :

- M=I : matrice identité

La covariance sera utilisée afin de quantifier les relations inter-variables, on parlera alors d'une ACP CANONIQUE.

- $M=D_{1/\sigma^2}$

$$[D_{1/\sigma^2}] = \begin{bmatrix} 1/\sigma_1^2 & & 0 \\ & 1/\sigma_j^2 & \\ 0 & & 1/\sigma_P^2 \end{bmatrix}$$

On utilise généralement cette métrique pour palier le problème d'hétérogénéités des caractères (variables) et éviter l'influence du choix d'unité des variables. Dans ce cas on parlera d'ACP NORMEE ; elle est équivalente à une ACP CANONIQUE effectuée sur des variables centrées réduites.
Les données ainsi transformées se présentent sous forme d'une matrice dont toutes les variables sont de moyenne nulle et d'écart type unité.

I.2.4 PROCEDE D'APPLICATION DE L'ACP

I.2.4.1 Calcul de la matrice de covariance

La matrice de covariance (notée [R]) étant la matière première de l'ACP, elle est obtenue en appliquant la relation suivante :

$$[R] = (1/n)\ [X]^t . [M] . [X]$$

[R] : matrice de covariance de dimension (PxP)
[X] : matrice de données centrées.
$[X]^t$: matrice transposée de [X]
[M] : métrique

I.2.4.2 Recherche des axes principaux

Le but principal est de construire un nouveau système d'axes avec un minimum de variables assurant un maximum de variance.

I.2.4.2.1. Recherche du premier axe

Tel que mentionné précédemment, la contribution maximale est donnée par le premier axe principal ; on doit donc maximiser la variance relative à celui-ci. La recherche du premier axe principal consiste à résoudre le problème

$$\begin{cases} \text{Max Var } (C_1) \\ V_1^t \,.\, [M] \,.\, V_1 = 1 \end{cases}$$

On peut exprimer la variance de C_1 a l'aide de la matrice des covariances R du vecteur aléatoire $X = (X_1, X_2, \ldots X_j, \ldots X_p)$

$$\text{Var } (C_1) = V_1^t.[M].[R].[M].V_1$$

En utilisant la méthode des multiplicateurs de LAGRANGE on peut écrire que

$$L = V_1^t.[M].[R].[M].V_1 - \lambda_1 (V_1^t.[M].\, V_1 - 1).$$

La dérivée par rapport à V est nécessairement nulle :

$$dL/dV_1 = 2.[M].[R].[M].V_1 - 2\lambda_1 [M]\, V_1 = 0$$

Puisque la matrice [M] est inversible :

$$[R] \,.\, [M].\, V_1 = \lambda_1 V_1.$$

Tout vecteur T non nul transformé par une matrice donnée [A] en T, c'est-à-dire vérifiant

$$[A].T = \lambda_1 T$$

est appelé vecteur propre de la matrice A associé à la valeur propre λ_i (voir Réf N° 11)

Donc obligatoirement V_1 est le vecteur propre de la matrice [R][M]. Il suffit, pour maximiser la variance de C_1 de choisir comme vecteur V_1, le vecteur propre associé à la plus grande valeur propre λ_1 de la matrice [R][M].

I.2.4.2.2.Recherche du second axe

On cherche à déterminer le vecteur unitaire V_2 tel que la composante C_2 soit de variance maximale et soit non correlée à C_1
On a comme précédemment :

$$\begin{cases} \text{Var } (C_2) = V_2^t.[M].[R].[M].V_2 \\ V_2^t.[M].\, V_2 = 1 \\ \text{COV } (C_1, C_2) = 0 \end{cases}$$

Exprimons COV (C_1, C_2) on a :

$$\text{COV } (C_1, C_2) = V_1^t \,.\, (M).[R] \,.[M].V_2$$

Comme la covariance ne tient pas compte de l'ordre on a :

$$COV\ (C_1,C_2) = COV(C_2,C_1) = V_2^t.[M].[R].[M].V_1$$

or on sait que V_1 est un vecteur propre de [R].[M]. associé à la valeur propre λ_1 ; on en déduit :

$$COV\ (C_1,C_2) = \lambda_1\ V_2^t.V_1 = 0$$

Une covariance nulle entre C_1 et C_2 est équivalente à l'orthogonalité des vecteurs V_1 et V_2:

$$COV\ (C_1,C_2) = 0 \iff V_1.V_2 = 0$$

En appliquant la même méthode pour la recherche du deuxième axe ; on aura:

$$L = V_2^t.[M].[R].[M].V_2 - \lambda_2\ (V_2^t.[M].V_2 - 1) - \mu\ (V_2^t.[M].V_1)$$

$$dL/dV_2 = 2\ [M].[R].[M].V_2 - 2\ \lambda_2 [M].V_2 - \mu[M].V_1 = 0$$

En simplifiant par [M] on obtient :

$$2\ [R].[M].V_2 - 2\ \lambda_2 V_2 - \mu V_1 = 0$$

On multiplie à gauche par $V_1^t[M]$ on obtient :

$$2\ V_1^t[M].[R].[M].V_2 - 2\lambda_2\ V_1^t[M].V_2 - \mu\ V_1^t[M].V_1 = 0$$

or $\quad V_1^t.[M]\ V_2 = 0 \qquad$ (par hypothèse)

$$V_1^t.[M].[R].V_2 = V_2^t[M].[R].[M].V_1 = l_1 V_2^t[M].V_1 = 0$$

$V_1^t.\ [M].\ V_1 = 1$ puisque le vecteur V_1 est unitaire.

Le multiplicateur de LAGRANGE μ est donc nul et l'on est ramené au problème précèdent . On peut donc énoncer la définition suivante :

Le second axe est défini par le vecteur V_2: vecteur propre unitaire de la matrice [R] . [M] orthogonal à V_1 et associé à la plus grande valeur propre λ_2 inférieur ou égale à λ_1.

I.2.4.2.3.Recherche des autres axes

En itérant le procédé, on détermine donc les valeurs propres et les vecteurs propres de la matrice [R].[M] pour obtenir la l^{ieme} composante principale C_l.

Le vecteur propre unitaire de la matrice [R].[M] définit le l^{ieme} axe orthogonal à $V_1, V_2, \ldots V_{l-1}$; et associé à la l^{ieme} plus grande valeur propre λ_l. On constate que la mise en équation de ces régles aboutit aux résultats suivants :

• On appelle l^{ieme} vecteur principal: le vecteur propre unitaire V_l de la matrice [R]. [M] associée, qui fournit les coefficients qui pondèrent les variables initiales pour le calcul des composantes principales.

- **On appelle $l^{ième}$ axe principal, la droite engendrée par le $l^{ième}$ vecteur principal**
- **Chaque composante C_k est portée par le $k^{ième}$ axe principal.**
- **La dispersion des projections des variables sur la composante C_k est mesurée par la valeur propre λ_2**
- **Les valeurs sont rangées par ordre décroissant : $\lambda_1 > \lambda_2 > \lambda_3 > \ldots\ldots\ldots > \lambda_Q$**
- **Les CP sont rangées de 1 à Q dans l'ordre des valeur propres.**
- **La moyenne de chaque CP est nulle.**

Si on veut normer les CP c'est-à-dire, imposer à chacune d'entre elles, d'avoir un écart type unité, il suffira de diviser par la variance expliquée $(\lambda_k)^{1/2}$ correspondante.

I.2.4.3 Calcul des composantes principales

On désigne par CP la projection du nuage de points initial sur le nouveau système d'axes, fourni par les vecteurs propres. Notons [C] la matrice des CP.

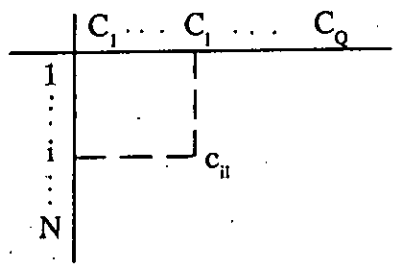
$$[C] = [A]^t.[X]$$

[X] : matrice des données initiales
$[A]^t$: matrice transposée de la matrice composée par les vecteurs propres.

I.2.5 PRINCIPAUX RESULTATS :

En général une Analyse en Composantes Principales fournit trois sources de renseignements, toutes nécessaires à l'interprétation :

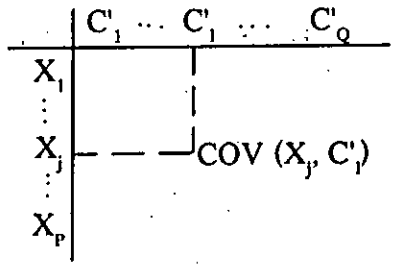
- Un tableau de vecteurs propres et valeurs propres
- Un tableau des corrélations des individus avec les axes factoriels qu'on notera [AI].



[AI] : Tableau "Individu x CP"

c_{il} : projection de l'individu i sur l'axe principal de rang l.

● Un tableau des Corrélations des variables aux axes principaux noté [AV].



[AV] : Tableau "Variable x CP".

$c'_l = c_l / (\lambda_l)^{1/2}$
c'_l : composante normée de c_l
λ_l : l^{ieme} valeur propre.

I.2.6 REPRESENTATIONS GRAPHIQUES

Une fois les résultats numériques obtenus on passe à la représentation graphique.

I.2.6.1 graphique des variables

L'examen de ce graphique permet d'observer l'organisation des variables sur le plan principal choisi.

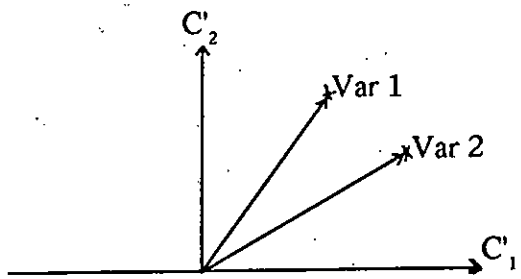


Figure I.1 Représentation des variables

On peut représenter chaque variable par un point dont les coordonnées sont les covariances avec les CP réduites.(car dans les représentations graphiques les axes sont orthonormés).

$$COV\ (X_j\ ,\ C'_k) = (\lambda_k)^{1/2} V_k\ (j)$$

Dans l'hypothèse où les variables sont bien représentées dans un plan formé par deux CP :

● La distance d'une variable à l'origine des axes est égale à sa norme (écart type).

$$\|Xj\|^2 = \sum_{i=1}^{Q} P_i X^2_{ij}$$

● Le cosinus de l'angle formé par deux variables est égale à leurs coefficients de corrélation

$$\langle X_j, X_{j'} \rangle = \|X_j\| \, \|X_{j'}\| \cos\theta$$

$$COV \langle X_j, X_{j'} \rangle = \sigma_{xj} \, \sigma_{xj'} \cos\theta$$

$$\cos\theta = [cov(X_j, X_{j'}) / \sigma_{xj} \, \sigma_{xj'}] = COV(X_j, X_{j'}).$$

- **Deux variables fortement corrélées et éloignées l'une de l'autre possèdent une covariance élevée négative.**

Dans la pratique l'interprétation est moins précise, il est indispensable d'examiner la qualité de la représentation de chaque variable avant d'en tirer des conclusions fiables.

Cas particulier : Cercle de corrélation
Afin d'éviter un examen de la qualité de représentation, considérons la variable centrée réduite Y_j au lieu de la variable initiale X_j.

$$\text{De ce fait : } Var(Y_j) = \|Y_j\|^2 = \sum_{i=1}^{N} P_i \, Y_j^2(i) = 1$$

Cette équation signifie que le point représentatif de la variable Y_j se trouve sur une sphère de rayon unité . Les coordonnées de la variable Y_j dans le plan factoriel sont calculés par :

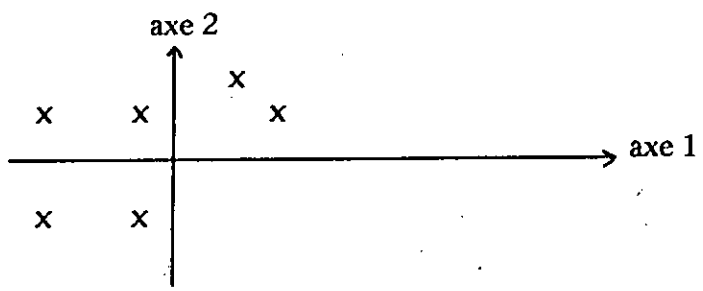
$$COV(Y_j, C'_l) = COV\left(\frac{X_j}{(Varj)^{1/2}}, C'_l\right) = COR(X_j, C'_l)$$

C'_l = composante principale normée
Varj : variance de la j^{ieme} variable.

C'est pourquoi le cercle unité tracé sur chacun des plans étudié s'appelle cercle de corrélation. On peut alors évaluer directement la qualité de la représentation de chaque variable en traçant un cercle de rayon unité sur le plan examiné. La variable sera d'autant mieux expliquée, que son point représentatif se rapproche du cercle et inversement.

I.2.6.2 Graphique des individus

Ce graphique permet d'observer la répartition des individus selon les axes principaux choisis, cette représentation est le résultat de la projection du nuage des individus sur les axes factoriels.



Figurel. 2 : Représentation des individus

Deux individus seront proches dans l'espace R^p, s'ils sont proches dans le plan factoriel et bien représentés dans ce dernier.Dans le graphique des individus on s'intéresse aux distances inter-individuelles qu'on peut interpréter comme étant des ressemblances.

I.2.6.3 Aides d'interprétation

En vue d'une bonne interprétation et donc, une meilleure compréhension, il serait utile de définir quelques paramètres, qui quantifieront la qualité de la représentation.

- **INR = inertie**

L'inertie est une mesure de dispersion du nuage, elle permet d'avoir la qualité globale de la représentation. Elle est mesurée par la variance expliquée par le plan principal choisi.

$$INR = (\lambda_1 + \lambda_2) / \sum_{i=1}^{Q} \lambda_i$$

$\lambda_{1,2}$: Valeurs propres correspondant respectivement au premier et au second axe factoriel.

$\sum_{i=1}^{Q} \lambda_i$: variance totale apportée par les Q CP retenues

- **CO2= cosinus² de l'angle**

Pour un axe donné, la valeur du cosinus carré de l'angle formé par un point et l'axe, mesure la qualité relative de la représentation sur cet axe.

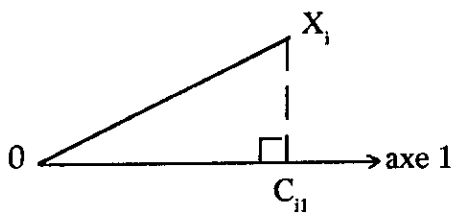


Figure I.3 : Cas d'une représentation d'individus

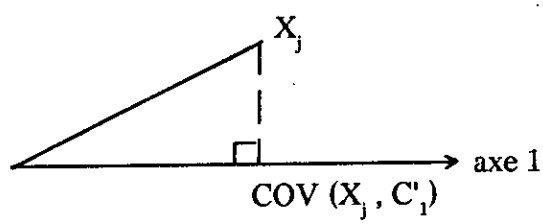


Figure I.4 : Cas d'une représentation de variable

Une bonne représentation du point sur l'axe correspond à une valeur de CO2 voisine de 1.

- **CTR = taux de variance dû à l'individu i relativement à la variance de l'axe considéré**

$$CTR = P_i C^2_k(i) / \lambda_k$$

P_i : poids du i^{ieme} individu
$C_k(i)$:projection du i^{ieme} individu sur la k^{ieme} composante
λ_k : variance de la k^{ieme} composante

Plus CTR est voisin de 1 plus l'individu i contribue a la détérmination du k^{ieme} axe, donc CTR permet d'apprécier l'importance de l'individu i dans la détermination de l'axe principal K.
Noter que

$$\sum_{i=1}^{N} P_i c^2_k(i) / \lambda_k = 1$$

I.3 CONCLUSION

L'Analyse en Composante Principale permet d'établir deux sortes de bilan :

- Un bilan de liaisons entre variables faisant l'objet des questions suivantes

  - **Quelles sont les variables qui sont liées positivement et celles qui s'opposent (liées négativement) ?**
  - **Existe t-il des groupes de variables corrélées entre elles?**
  - **Peut-on mettre en évidence une typologie de variables ?**

- Un bilan de ressemblance entre individus répondant aux questions suivantes.

  - **Quels sont les individus qui se ressemblent et ceux qui diffèrent ?**
  - **Existe t-il des groupes homogènes d'individus ?**
  - **Peut-on mettre en évidence une typologie des individus ?**

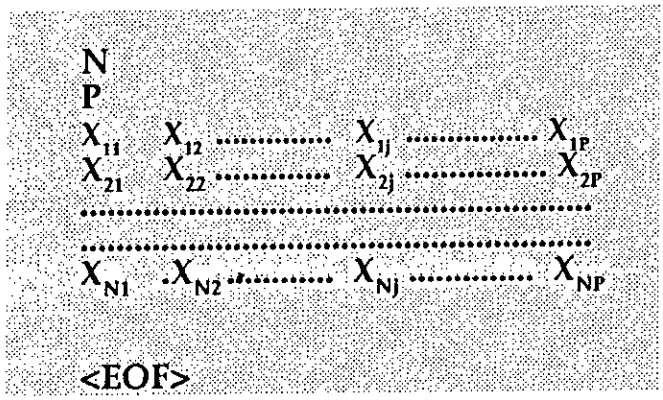
En dehors de l'analyse descriptive citée plus haut, l'ACP ouvre les portes sur un domaine beaucoup plus intéressant qui est l'Analyse Opérationnelle : cette dernière traite les problèmes de prévision d'optimisation et de simulation.

CHAPITRE II
ELABORATION DU MODELE
(ACP programmation et exemple d'exécution)

Dans ce chapitre on traitera de la programmation de l'Analyse en Composantes Principales. Le logiciel proposé est constitué d'une suite de sous programmes traitant chacune une partie bien spécifique du problème. Le modèle élaboré est très simple à utiliser et affranchit l'utilisateur du format des fichiers INPUT, vu que la lecture de ceux-ci se fait sous format libre.
On explicitera dans ce chapitre les différentes phases du programme par un organigramme clair et complet. Un exemple d'exécution largement commenté terminera ce chapitre.

II.1 FICHIER DE DONNEES (INPUT)

Le fichier de données doit être construit comme suit :



N,P : sont des nombres entiers désignant respectivement le nombre d'individus et le nombre de variables.
X_{ij} : données à introduire.

L'introduction des données se fait ligne par ligne et chacune d'elle représente un individu, cela se fait sous contrainte, qu'il y'ait plus de lignes que de colonnes, pour que le nuage de points se comporte comme un corps solide et qu'on puisse parler d'inertie et de centre de gravité.

II.2 STRUCTURE DE LA CHAINE ACP

Avant de présenter l'organigramme de la chaîne ACP, voici la liste des sous programmes du modèle avec une brève explication.

Sous programmes	Tâches effectuées
SAISIE - DES - DONNEES	Introduction directe des données ou bien lecture dans un fichier déjà existant.
MOY -VAR	Calcul des moyennes et variances des différentes variables.
OPTION	Choix de la métrique : •ACP normée •ACP canonique.
CALCUL	Détermination de la matrice de covariance ainsi que la matrice de corrélation.
DIAGONALISATION	Calcul des valeurs et vecteurs propres
EDIT/RESULTAS	Visualisation et stockage des résultats obtenus.
ANAL / VARIABLES	Etude détaillée des variables.
EDIT/AV	Edition et sauvegarde des coordonnés des variables dans les plans principaux.
CERCLE/CORRELATION	Représentation graphique des variables.
ANAL/INDIVIDUS	Etude détaillée des individus.
EDIT/AI	Edition et sauvegarde des coordonnées des individus dans les plans principaux
CONTRIB/IND	Contribution de chaque individu à la formation des CP
QUAL/REP/IND	Evaluation de la qualité de représentation des individus.
TRACÉ/PLAN-FACT	Projection du nuage d'individus dans le plan principal.

II.3 METHODE DE DIAGONALISATION

Le calcul numérique des valeurs propres est un sujet fort complexe, et il existe de très nombreux développements numériques à ce problème. Les procédés couramment utilisés se basent sur la méthode de diagonalisation de JACOBI.
Cette méthode consiste à faire subir à la matrice initiale [R] une suite de variations planes qui la transforme en une matrice diagonale.

Formulons le processus de JACOBI : soit la matrice [R] d'ordre P symétrique c'est-à-dire :

$$[R]^t = [R]$$

Admettons qu'il est possible de trouver une matrice [S] d'ordre P orthogonale dont les élément s_{ik} ont la propriété de transformer la matrice [R] en une matrice diagonale, ceci peut s'écrire mathématiquement :

$$[S]^t.[R].[S] = [D]$$

Il a été démontré, (R.L. BURDEN 1986) que si cette matrice [S] orthogonale qui transforme une matrice réelle symétrique en une matrice diagonale existe, les éléments de la diagonale de [D] peuvent être adoptés comme les valeurs propres de [R], et les colonnes de [S] comme les vecteurs propres associés aux valeurs propres.

La difficulté de cette méthode est justement de trouver la matrice [S] ; étant donné que chaque transformation orthogonale affecte non seulement les éléments pivots, on ne peut espérer en général qu'une seule série de transformations sur ces derniers, suffise pour obtenir la matrice [D]. Pour cette raison, la méthode de JACOBI est une méthode itérative qui est appliquée jusqu'à ce que la précision désirée soit obtenue.

Le calcul explicite des coefficients du polynôme caractéristique devient rapidement très laborieux lorsque la matrice [R] est d'ordre élevé, et il peut même introduire des erreurs d'arrondis importantes faussant ainsi les racines trouvées du polynôme.

Cette méthode directe comme celle de JACOBI est rarement utilisée et pas avantageuse. Ces procédés calculent tous les valeurs propres.

Quand en 1931 THURSTONE entreprit d'appliquer la méthode des moindres carrés à l'Analyse Factorielle, les moyens de calcul étaient rudimentaires ; et les algorithmes de diagonalisation eux même peu étudiés.
C'est en 1933 que HOTELING, ayant en vue l'Analyse Factorielle, proposa de rechercher les vecteurs propres par itération et orthogonalisation.
Ainsi une nouvelle méthode dite **méthode des puissances iterées** a été mise au point dans ce but. Cette méthode permet de calculer le module de la plus grande valeur propre et le vecteur propre associé, sans utiliser le développement de son déterminant caractéristique.

Avant de formuler le processus des **puissances iterées**, rappelons le théorème suivant :

Si [A] est une matrice symétrique, ses valeurs propres sont réelles et ses vecteurs propres correspondant à des valeurs propres distinctes sont orthogonaux.

Soit la même matrice [R] dont les valeurs propres $\lambda_1\lambda_2,..\lambda_P$ sont distinctes. Puisque leurs vecteurs propres $V_1, V_2, ...V_P$ associés sont orthogonaux (théorème), chaque vecteur X à P composantes $(X_1, X_2,...X_P)$ peut être présenté de la manière suivante :

$$X = c_1V_1 + c_2V_2 + ...+ c_PV_P$$

où c_j : nombre réel

on peut écrire aussi

$$X = \sum_{j=1}^{P} c_jV_j \qquad \text{(II.3.1)}$$

Du point de vue matricielle l'expression (II.3.1) est équivalente à

$$X = [V].[C]^t$$

Multiplions les deux membres de l'équation (II.3.1) par [R], $[R]^2$,...$[R]^m$

$$[R].X = \sum_{j=1}^{P} c_j [R] V_j = \sum_{j=1}^{P} c_j \lambda_j V_j$$

$$[R]^2.X = \sum_{j=1}^{P} c_j [R]^2 V_j = \sum_{j=1}^{P} c_j \lambda_j^2 V_j$$

$$\vdots$$

$$[R]^m.X = \sum_{j=1}^{P} c_j [R]^m V_j = \sum_{j=1}^{P} c_j \lambda_j^m V_j \qquad \text{(II.3.2)}$$

Mettons en facteur le terme λ_1^m dans l'équation (II.3.2)

$$[R]^m.X = \lambda_1^m \sum_{j=1}^{P} c_j \left(\frac{\lambda_1}{\lambda_1}\right)^m V_j \qquad \text{(II.3.2.a)}$$

donc

$$[R]^m.X = \lambda_1^m c_1V_1 + \lambda_1^m \sum_{j=1}^{P} c_j \left(\frac{\lambda_1}{\lambda_1}\right)^m V_j \qquad \text{(II.3.2.b)}$$

Le fait que $\lambda_1 > \lambda_j \ \forall j = 2$ à P (chap I § I.2.4.2.2) implique que :

$$\lim_{m \to +\infty} \left(\frac{\lambda_j}{\lambda_1}\right)^m = 0$$

$$\lim_{m \to +\infty} [R]^m . X = \lim_{m \to +\infty} \lambda_1 c_1V_1 \qquad \text{(II.3.3)}$$

Choisissons un vecteur initial X_0 normé, soit le vecteur Y_1 tel que

$$Y_1 = [R].\ X_0 \qquad \text{(II.3.4)}$$

Définissons le scalaire μ_1, comme étant le rapport des deux vecteurs Y_0 et X_0 :

$$\mu_1 = \frac{Y_1}{X_0}$$

Sachant que l'expression (II.3.4) est équivalente à l'expression (II.3.2.b) pour m=1, remplaçons Y_1 par cette dernière et X_0 par la formule (II.3.1)

On obtient :

$$\mu_1 = \frac{Y_1}{X_0} = \frac{c_1\lambda_1 V_1 + \sum_{j=2}^{P} c_j\lambda_j V_j}{c_1 V_1 + \sum_{j=2}^{P} c_j V_j}$$

$$= \lambda_1 \left[\frac{c_1 V_1 + \sum_{j=2}^{P} c_j \left(\frac{\lambda_j}{\lambda_1}\right) V_j}{c_1 V_1 + \sum_{j=2}^{P} c_j V_j} \right]$$

Soit X_1 le vecteur Y_1 normé

$$X_0 = \frac{Y_1}{\|Y_1\|}$$

De même on peut définir un vecteur Y_2 tel que :

$$Y_2 = [R].\ X_1$$

et par la suite

$$\mu_2 = \frac{Y_2}{X_1}$$

De la même manière on peut définir des séquences de vecteurs $\{X_m\}_{m=0}^{\infty}$ et $\{Y_m\}_{m=1}^{\infty}$ et une séquence de scalaires $\{\mu_m\}_{m=1}^{\infty}$ tel que :

(II.3.5)

$$Y_m = [R] . X_{m-1} \qquad \text{(II.3.5 a)}$$

$$\mu_m = \frac{Y_m}{X_{m-1}} = \lambda_1 \left[\frac{c_1 V_1 + \sum_{j=2}^{P} c_j \left(\frac{\lambda_j}{\lambda_1}\right)^m V_j}{c_1 V_1 + \sum_{j=2}^{P} c_j \left(\frac{\lambda_j}{\lambda_1}\right)^{m-1} V_j} \right] \qquad \text{(II.3.5 b)}$$

$$X_m = \frac{Y_m}{\|Y_m\|} \qquad \text{(II.3.5 c)}$$

En normant les vecteurs X_m à chaque étape on arrive à éviterque les composantes ne deviennent trop grandes.

En examinant l'équation (II.3.5.b) et sachant que

$\frac{\lambda_j}{\lambda_1} < 1 \; (\forall j = 2,..N)$ alors :

$$\lim_{m \longrightarrow +\infty} \mu_m = \lambda_1$$

et la séquence des vecteurs $\{X_m\}_{m=0}^{\infty}$ tend vers un **vecteur propre X_1**, de norme unité, associé a λ_1.

En appliquant de nouveau l'algorithme de la puissance itérée à une matrice [R*] possédant les valeurs propres $0, \lambda_2, ... \lambda_P$, on détermine λ_2 et son vecteur propre associé.

[R*] est définie comme suit :

$$[R^*] = [R] - \lambda_1 . X_1 . X_1^t$$

où X_1 : vecteur propre normé de [R] associé à λ_1.

On vérifie que [R*] possède les valeurs propres λ_k (k = 1 à P) avec $\lambda_1 = 0$. Comme la multiplication des matrices est associative et :

$$X_1^t . X_1 = 1 \text{ (hypothèse de normalité)}$$

On trouve

$$\begin{aligned} [R^*] &= [R] . X_1 - \lambda_1 (X_1 . X_1^t) X_1 \\ &= [R] . X_1 - \lambda_1 X_1 \\ &= \lambda_1 X_1 - \lambda_1 X_1 \\ &= 0 \end{aligned}$$

Ainsi pour k=1, zéro (0) est une valeur propre de [R*].

pour $k \neq 1$ on a :

$$X_1^t.X_k = 0 \text{ {hypothèse d'orthogonalité}}$$

d'où

$$[R^*].X_k = R.X_k - \lambda_1.(X_1.X_1^t).X_k$$

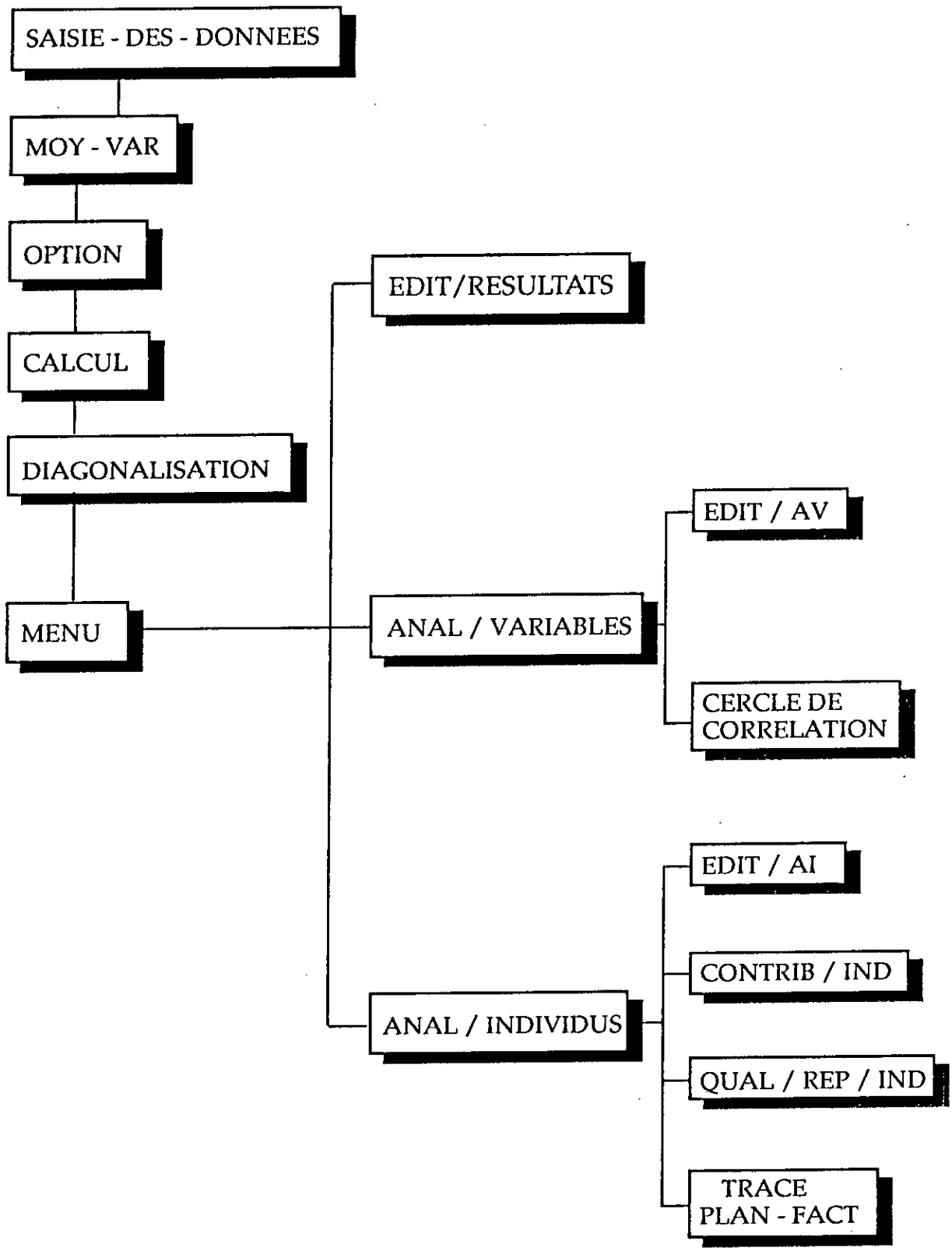
$$= \lambda_k.X_k - \lambda_1.X_1.(X_1^t.X_k)$$

$$= \lambda_k.X_k$$

Ce qui montre que les coefficients λ_k pour $k \neq 1$ sont des valeurs propres de $[R^*]$. On détermine ainsi la seconde valeur propre λ_2 et le vecteur propre associé.

On procède de façon analogue pour calculer les autres valeurs propres et les vecteurs correspondants.

II.4 ORGANIGRAMME GENERAL



II.5 Exemple d'exécution (Fichier output)

On se propose d'étudier un tableau (individus x caractères) défini par dix (10) individus et six (6) variables, où les variables considérées sont des pluies mensuelles mesurées en [mm], relevées en six (6) stations existant au niveau du bassin de SEBAOU : durant une période d'observation de dix(10)années.

Variables / Individus	St1	St2	St3	St4	St5	St6
1	543	716	331	148	438	530
2	520	700	329	143	400	530
3	516	676	323	135	400	494
4	514	630	321	135	400	490
5	502	630	317	126	390	430
6	487	622	314	123	385	425
7	480	617	305	110	376	409
8	475	615	304	74	367	380
9	454	614	395	52	310	376
10	440	609	294	44	305	360

"Tableau de données initiales"

Le fichier de sortie comporte deux grandes parties. La première consacrée aux paramètres statistiques (moyennes, écarts types, matrice de corrélations et matrice de covariances). La seconde à la diagonalisation (détermination des valeurs et vecteurs propres) et à la projection sur les plans factoriels.

Le premier traitement consiste à déterminer la moyenne et la variance pour les différentes variables, on transforme ensuite les données initiales en données centrées réduites.

Après calcul de la matrice de corrélation et la diagonalisation de cette dernière, on aboutit aux valeurs propres et aux vecteurs propres associés formant une nouvelle base orthonormée.

Vu que le but principal de l'ACP est de donner le maximum d'information avec un minimum de composantes, il serait judicieux de ne considérer que les deux premières composantes, expliquant prés de 98% de l'information totale.

```
CARACTERISTIQUES DU FICHIER  : SEBAHOU.DAT

NOMBRE D'OBSERVATIONS: 10

NOMBRE DE VARIABLES: 6


  DEFINITION DES VARIBLES :
A0   :  C021601
A1    : C021701
A2    : C151006
A3    : C150306
A4    : C150111
A5    : C150110
            ########################################
            ##  1-            BRUTE               ##
            ##  2-          CENTREE               ##
            ##  3-       CENTREE REDUITE          ##
            ##  4-          REDUITE               ##
            ##  5-          SORTIE                ##
            ########################################

            TYPE D'ACP: 3

PARAMETRES STATISTIQUES (au 1/10 ème prés)
----------------------
```

variable	moyennes (mm)	ecart type (mm)
A0	493.100	31.866
A1	642.900	39.304
A2	313.300	13.292
A3	109.000	38.317
A4	377.100	41.251
A5	442.400	64.066

MATRICE DE CORRELATION

	A0	A1	A2	A3	A4	A5
A0	1.0000					
A1	0.8392	1.0000				
A2	0.9797	0.8523	1.0000			
A3	0.9433	0.7088	0.9520	1.0000		
A4	0.9614	0.7253	0.9345	0.9517	1.0000	
A5	0.9491	0.8976	0.9652	0.8967	0.8568	1.0000

MATRICE DE COVARIANCES

	A0	A1	A2	A3	A4	A5
A0	1.0000					
A1	0.8392	1.0000				
A2	0.9797	0.8523	1.0000			
A3	0.9433	0.7088	0.9520	1.0000		
A4	0.9614	0.7253	0.9345	0.9517	1.0000	
A5	0.9491	0.8976	0.9652	0.8967	0.8568	1.0000

VECTEUR PROPRE (matrice de passage)

	AXE1	AXE2	AXE3	AXE4	AXE5	AXE6
A0	0.423	-0.086	0.189	-0.546	0.049	-0.691
A1	0.373	0.766	0.337	0.388	0.059	-0.078
A2	0.424	-0.027	-0.186	-0.026	-0.869	0.174
A3	0.407	-0.419	-0.343	0.644	0.207	-0.287
A4	0.405	-0.413	0.613	-0.033	0.181	0.505
A5	0.415	0.242	-0.568	-0.366	0.405	0.385

	VALEURS PROPRES	CONTRIBUTION	CONTRIBUTION CUMULEE
1	5.4805	91.3424	91.3424
2	0.3801	6.3354	97.6779
3	0.0849	1.4149	99.0927
4	0.0312	0.5194	99.6121
5	0.0156	0.2599	99.8721
6	0.0077	0.1279	100.0000

ETUDE DES VARIABLES

##1er colonne projection des variables sur les axes principaux----------#
##2iemcolonne coefficient de determination----------#

	AXE1		AXE2		AXE3		AXE4		AXE5		AXE6	
A0	0.99##	0.981	-0.05##	0.003	0.06##	0.003	-0.10##	0.009	0.01##	0.000	-0.06##	0.004
A1	0.87##	0.763	0.47##	0.223	0.10##	0.010	0.07##	0.005	0.01##	0.000	-0.01##	0.000
A2	0.99##	0.985	-0.02##	0.000	-0.05##	0.003	-0.00##	0.000	-0.11##	0.012	0.02##	0.000
A3	0.95##	0.909	-0.26##	0.067	-0.10##	0.010	0.11##	0.013	0.03##	0.001	-0.03##	0.001
A4	0.95##	0.901	-0.25##	0.065	0.18##	0.032	-0.01##	0.000	0.02##	0.001	0.04##	0.002
A5	0.97##	0.942	0.15##	0.022	-0.17##	0.027	-0.06##	0.004	0.05##	0.003	0.03##	0.001

ETUDE DES INDIVIDUS

##1er colonne coordonnees des individus sur les axes principaux----------#
##2iem colonne cosinus carres(qualite de la representation)----------#

1	3.50###	0.96	0.55###	0.02	0.45###	0.02	-0.06###	0.00	0.06###	0.00	-0.01###	0.00
2	2.55###	0.91	0.74###	0.08	-0.31###	0.01	0.13###	0.00	-0.06###	0.00	0.06###	0.00
3	1.76###	0.98	0.24###	0.02	-0.07###	0.00	0.04###	0.00	0.02###	0.00	-0.04###	0.00
4	1.21###	0.67	-0.66###	0.20	-0.41###	0.08	-0.35###	0.06	0.05###	0.00	0.04###	0.00
5	0.34###	0.20	-0.64###	0.70	0.04###	0.00	0.06###	0.01	-0.18###	0.05	-0.16###	0.04
6	-0.14###	0.04	-0.69###	0.84	-0.08###	0.01	0.23###	0.09	-0.09###	0.01	0.07###	0.01
7	-0.90###	0.64	-0.58###	0.26	0.09###	0.01	0.19###	0.03	0.27###	0.06	0.00###	0.00
8	-1.68###	0.89	-0.23###	0.02	0.50###	0.08	-0.17###	0.01	-0.09###	0.00	0.09###	0.00
9	-3.07###	0.96	0.62###	0.04	-0.12###	0.00	-0.11###	0.00	0.07###	0.00	-0.13###	0.00
10	-3.58###	0.97	0.64###	0.03	-0.09###	0.00	0.05###	0.00	-0.06###	0.00	0.07###	0.00

Pour l'étude des variables, on détermine la covariance de ces dernières avec les axes principaux, ce qui nous donne les coordonnées de chaque variable dans la nouvelle base. De même, pour l'étude des individus, la projection de leur nuage dans la nouvelle base donne de nouvelles variables, dites aussi, Composantes Principales.

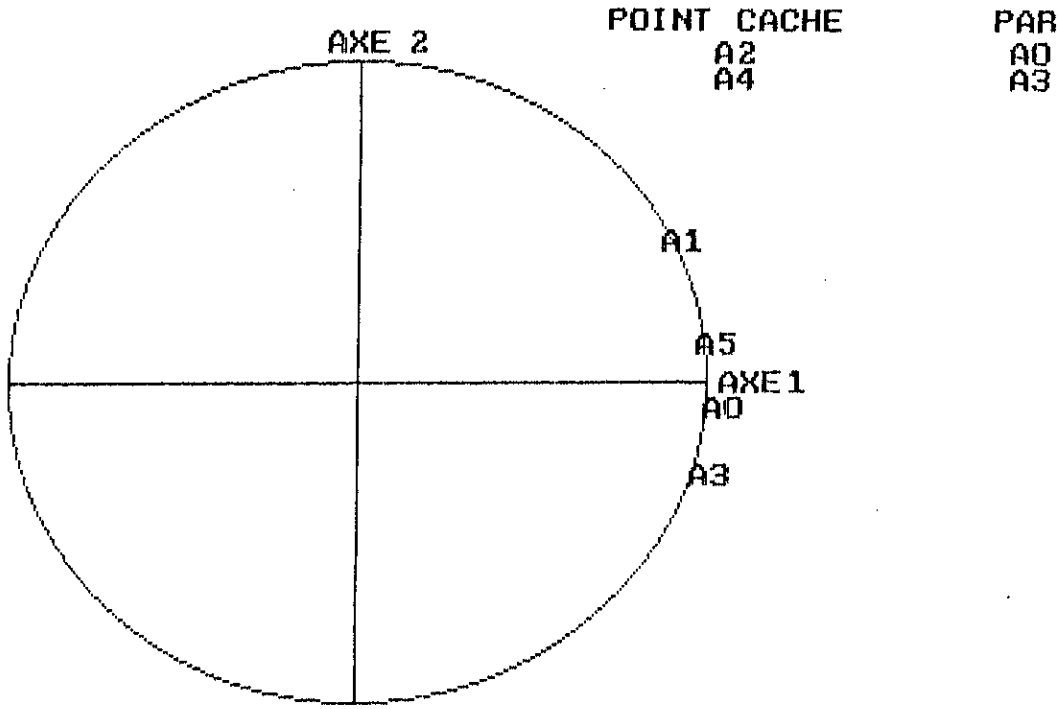
La projection des variables et des individus respectivement, dans le cercle de corrélation (C'_1 x C'_2) et la plan factoriel (1 x 2), sert de base à l'interprétation des résultats obtenus ; les axes 1 et 2 expliquent 98% de la variance totale. La contribution de chaque axe à cette variance définit le pouvoir explicatif de la composante correspondante à laquelle on attribue un sens physique, soit par exemple, la moyenne caractérisant le bassin pour la première composante.

Toutes les variables sont bien représentées sur le cercle de corrélation puisqu'elles sont proches de celui-ci. Il apparait une orientation naturelle vers l'axe 1. Ceci s'explique par la faible variation de la moyenne de chaque station par rapport à la moyenne du bassin.

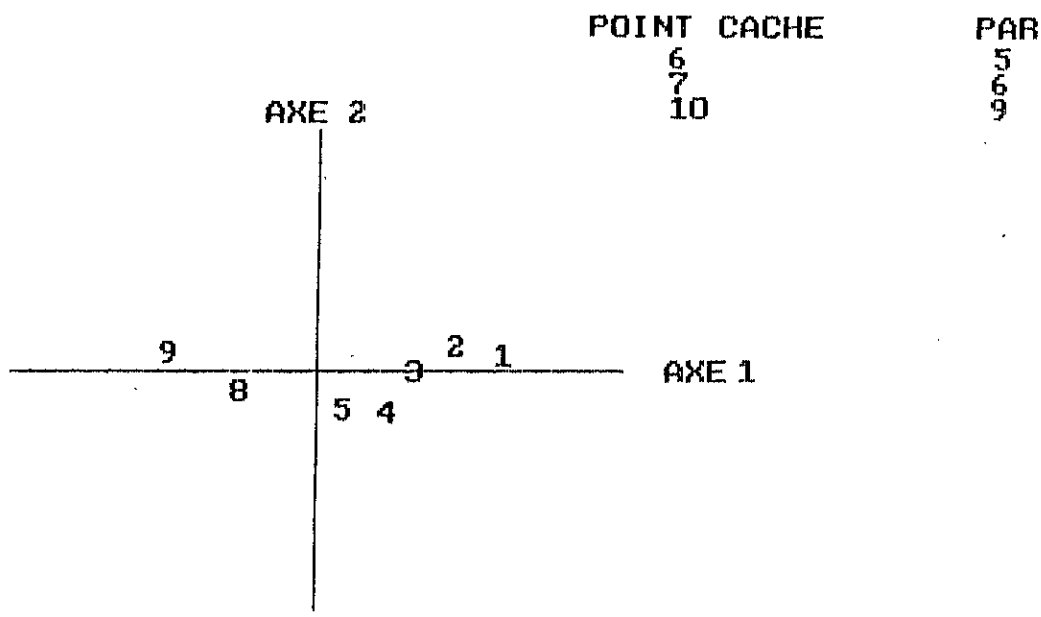
On notera au passage que deux variables proches dans le plan sont proches dans la réalité (bien corrélées). De même, deux individus proches dans plusieurs plans, sont proches dans l'espace.

Exemple
L'individu 9 et 10 ont mêmes coordonnées par rapport a l'axe 1 et 2, effectivement l'historique montre une moyenne interstation très proche.

Analyse des variables



Analyse des individus



DEUXIEME PARTIE

Dans cette partie, on présentera quelques exemples d'application de l'ACP, en Analyse Descriptive et en Analyse Opérationnelle.

Pour l'Analyse Descriptive, deux exemples sont exposés :

- ***L'un traitant des variables homogènes,***
- ***L'autre des variables hétérogènes.***

La simulation des débits mensuels par L'ACP est une application en Analyse Opérationnelle.

INTRODUCTION

En général, le domaine d'utilisation de l'ACP est très large, sans inconvénients, sans dangers et ne nécessite pas au préalable d'hypothèses restrictives.

Si les conditions de normalité ne sont ni impératives ni nécessaires pour pratiquer l'ACP, le type de variables utilisées influe par contre sur les résultats obtenus.

En général la technique de l'ACP permet de traiter des variables caractérisant deux types de phénomènes :

- **Un phénomène spatial** (champ de température de l'air de pressions atmosphérique, précipitations, débits...etc dans un région ou pays).

- **Un phénomène temporel** (chronique des débits, pression ou températures journalières - hebdomadaires - décadaires, etc... en un lieu fixe).

Avant de présenter les exemples d'application, il est utile de distinguer les deux aspects principaux de l'ACP.

- **L'Analyse Descriptive** : s'intéresse à la structure de la matrice d'observations, basée principalement sur la représentation graphique des résultats trouvés par l'ACP et ce, afin de visualiser certains problèmes non décelables à l'état brut, vue l'indépendance des variables (exemple : le phénomène de redondance).On peut aussi déterminer la variable principale qui régit le phénomène étudié grâce à la représentation cartographique qui suggère une interprétation physique en s'appuyant sur la géographie, le climat, le relief etc...

- **L'Analyse Opérationnelle** : L'ACP fournit un outil mathématique en apparence mais très bien adapté aux problèmes rencontrés en hydrologie à savoir :

  - **Optimisation d'un réseau de mesures.**
  - **Critique et reconstitution des données.**
  - **Prévision**
  - **Simulation.**

CHAPITRE III
APPLICATION DE L'ACP DANS LE DOMAINE DESCRIPTIF

III.1 INTRODUCTION

Dans le domaine descriptif on peut étudier deux types de phénomènes

- **Temporel et,**
- **Spatial**

Le premier type met en évidence des relations entre des observations à deux époques distinctes de l'année alors que le second fait ressortir la notion de proximité inter-stations.

L'application portera sur deux types de variables :

- **L'une homogène, caractérisant un phénomène spatial, par exemple : les précipitations relevées sur six (06) stations du Nord de l'Algérie.**
- **L'autre hétérogène : soit l'ensemble des paramètres intervenant dans la détermination de l'évapotranspiration.**

III.2 INFLUENCE DE LA TRANSFORMATION DES VARIABLES SUR LA STRUCTURE DES CP

La matrice de covariance constitue le point de départ de toute Analyse en Composantes Principales, c'est elle en effet qui après diagonalisation permet de calculer les valeurs et vecteurs propres.

Or la valeur du coefficient de covariance entre les variables X_j et X_k dépend essentiellement des couples d'observations (X_{ij} et X_{ik}), et en particulier, de la configuration de ce nuage de points dans le plan (X_j,X_k). Toute transformation appliquée aux P variables X_j modifiera les coefficients de covariance et de ce fait la structure des composantes principales.

Il existe des transformations usuelles pour l'Analyse des données à savoir :

Variable centrée :

posons $$Y=X_j - \overline{X}_j \qquad \text{(II.2.1)}$$

avec X_j : variable initiale
$\overline{X}_j$: moyenne correspondant à X_j
Y_j : variable transformée

Celle-ci permet la translation de l'origine du repère au centre de gravité du nuage de points.

Variable réduite

posons $$Y_j=X_j / \sigma_{Xj} \qquad \text{(III.2.2)}$$

avec σ_{Xj} : écart type de la variable X_j

La présente transformation permet de standardiser les variables (ne considère que leurs variations absolues).

Variable centrée réduite

posons $$Y_j = [X_j - \overline{X}_j] / \sigma_{Xj} \qquad \text{(III.2.3)}$$

Cette dernière réunit les deux premières (**III.2.1 et III.2.2**) et permet donc d'avoir d'une part, l'origine du repère confondue avec le centre de gravité du nuage de points, et d'autre part, des variables standards qui n'interviennent que par leurs variations absolues.

Pour étudier l'influence de ces différentes transformations on a choisi deux exemples d'application : le premier traitant des variables homogènes (répartition spatiale de la pluie), le second, des variables hétérogènes (variation de l'ETP en fonction des paramètres climatiques).

Sur ces deux exemples, quatre types d'ACP ont été effectuées; à savoir :

- **ACP sur variables brutes ;**
- **ACP sur variables centrées ;**
- **ACP sur variables réduites ;**
- **ACP sur variables centrées-réduites.**

Le traitement des variables brutes et des variables centrées donne lieu à la même matrice de covariance et par conséquent, les résultats de la diagonalisation sont identiques.

En effet, considérons la variable brute X et la variable centrée Y **résultant de la transformation (III.2.1)**, et montrons que la covariance de deux variables brutes (X_j, X_k) est équivalente à celle des deux variables centrées respectives (Y_j, Y_k).

$$\left.\begin{array}{l} Y_j = X_j - \overline{X_j} \\ Y_k = X_k - \overline{X_k} \end{array}\right\} \Rightarrow \left\{\begin{array}{l} \overline{Y_j} = 0 \\ \overline{Y_k} = 0 \end{array}\right.$$

$$COV(Y_j, Y_k) = (1/N)\sum_{i=1}^{N} (Y_{ij} - \overline{Y_j})(Y_{ik} - \overline{Y_k})$$

$$= (1/N)\sum_{i=1}^{N} (Y_{ij} \cdot Y_{ik})$$

En remplaçant chaque variable par son expression on obtient :

$$COV(Y_j, Y_k) = (1/N)\sum_{i=1}^{N} (X_{ij} - \overline{X_j})(X_{ik} - \overline{X_k})$$

$$= COV(X_j, X_k)$$

d'où

$$\mathbf{COV(Y_j, Y_k) = COV(X_j, X_k)}$$

Les mêmes constatations sont faites lors du traitement des variables réduites et centrées réduites.

Soit Y_j la variable obtenue par **transformation (III.2.2)** et Z_j celle obtenue par la **transformation (II.2.3).** Montrons que la covariance de deux variables réduites est égale à celle des deux variables centrées réduites, qui n'est autre que la corrélation entre les deux variables brutes correspondantes.

$$\left\{\begin{array}{l} Y_j = X_j / \sigma_{Xj} \\ Y_k = X_j / \sigma_{Xk} \end{array}\right. \qquad \left\{\begin{array}{l} Z_j = (X_j - \overline{X_j}) / \sigma_{Xj} \\ Z_k = (X_k - \overline{X_k}) / \sigma_{Xk} \end{array}\right.$$

Déterminons respectivement la moyenne et l'écart type des variables Y et Z.

- **Moyenne de Y**

$$\overline{Y_j} = (1/N)\sum_{i=1}^{N} Y_{ij}$$

En remplaçant Y_j par l'expression (III.2.2), on a :

$$\overline{Y_j} = (1/N)\sum_{i=1}^{N} X_{ij} / \sigma_{Xj}$$

$$\overline{Y_j} = (1/\sigma_{Xj}) \left[(1/N)\sum_{i=1}^{N} X_{ij} \right]$$

d'où

$$\overline{Y_j} = (1/\sigma_{Xj})\ \overline{X_j} \qquad \text{(III.2.4)}$$

- **Ecart type de Y**

$$\sigma_{Yj}^2 = (1/N)\sum_{i=1}^{N} (Y_{ij} - \overline{Y_j})^2$$

En substituant Y_j et Y_j par les expressions **(III.2.2) et (III.2.4)**, on obtient :

$$\sigma_{Yj}^2 = (1/N)\sum_{i=1}^{N} \left[(X_{ij} / \sigma_{Xj}) - (\overline{X_j} / \sigma_{Xj}) \right]^2$$

$$= (1/N)\sum_{i=1}^{N} \left[(X_{ij} - \overline{X_j}) / \sigma_{Xj} \right]^2$$

$$= (1/\sigma_{Xj})^2 \left[\sum_{i=1}^{N} (X_{ij} - \overline{X_j})^2 \cdot 1/N \right]$$

$$= (1/\sigma_{Xj})^2\ \sigma_{Xj}^2$$

d'où

$$\sigma_{Yj}^2 = 1 \qquad \text{(III.2.5)}$$

- **Moyenne de Z**

$$\overline{Z}_j = \frac{1}{N} \sum_{i=1}^{N} Z_{ij}$$

En remplaçant Z_j par l'expression **(III.2.3)**, on a :

$$\overline{Z}_j = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{X_{ij} - \overline{X}_j}{\sigma_{xj}} \right)$$

$$= \frac{1}{\sigma_{xj}} \left[\frac{1}{N} \sum_{i=1}^{N} (X_{ij} - \overline{X}_j) \right]$$

$$= \frac{1}{\sigma_{xj}} \left[\frac{1}{N} \sum_{i=1}^{N} X_{ij} - \frac{1}{N} \sum_{i=1}^{N} \overline{X}_j \right]$$

$$= \frac{1}{\sigma_{xj}} \left[\overline{X}_j - \frac{N}{N} \overline{X}_j \right]$$

$$= 0$$

d'où

$$\overline{Z}_j = 0 \qquad \text{(III.2.6)}$$

- **Ecart type de Z**

$$\sigma^2_{Zj} = \frac{1}{N} \sum_{i=1}^{N} (Z_{ij} - \overline{Z}_j)^2$$

Suivant l'équation **(III.2.6)** on a :

$$\sigma^2_{Zj} = \frac{1}{N} \sum_{i=1}^{N} Z_{ij}^2$$

En remplaçant Z_j par l'expression **(III.2.3)** on obtient

$$\sigma^2_{Zj} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_{ij} - \overline{X}_j}{\sigma_{Xj}}\right)^2$$

$$= \frac{1}{\sigma^2_{Xj}}\left[\frac{1}{N}\sum_{i=1}^{N}(X_{ij} - \overline{X}_j)^2\right]$$

$$= \frac{1}{\sigma^2_{Xj}} \cdot \sigma^2_{Xj}$$

d'où

$$\sigma_{Zj}^2 = 1 \qquad \text{(III.2.7)}$$

Evaluons maintenant la covariance entre Y_j et Y_k ; Z_j et Z_k

$$COV(Y_j, Y_k) = \frac{1}{N}\sum_{i=1}^{N}(Y_{ij} - \overline{Y}_j)(Y_{ik} - \overline{Y}_k)$$

En remplaçant la moyenne Y par l'expression **(III.2.4)** et la variable Y par celle donnée en **(III.2.2)**, on obtient :

$$COV(Y_j, Y_k) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_{ij}}{\sigma_{Xj}} - \frac{\overline{X}_j}{\sigma_{Xj}}\right)\left(\frac{X_{ik}}{\sigma_{Xk}} - \frac{\overline{X}_k}{\sigma_{Xk}}\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_{ij} - \overline{X}_j}{\sigma_{Xj}}\right)\left(\frac{X_{ik} - \overline{X}_k}{\sigma_{Xk}}\right)$$

$$= \frac{\frac{1}{N}\sum_{i=1}^{N}(X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k)}{\sigma_{Xj} \cdot \sigma_{Xk}}$$

$$= \frac{COV(X_j, X_k)}{\sigma_{Xj}\ \sigma_{Xk}}$$

donc

$$COV(Y_j, Y_k) = COR(X_j, X_k)$$

$$COV(Z_j, Z_k) = \frac{1}{N} \sum_{i=1}^{N} (Z_{ij} - \overline{Z_j})(Z_{ik} - \overline{Z_k})$$

D'après **(III.2.6)** on a :

$$COV(Z_j, Z_k) = \frac{1}{N} \sum_{i=1}^{N} Z_{ij} \cdot Z_{ik}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left(\frac{X_{ij} - \overline{X_j}}{\sigma_{Xj}}\right)\left(\frac{X_{ik} - \overline{X_k}}{\sigma_{Xk}}\right)$$

$$= \frac{\frac{1}{N} \sum_{i=1}^{N} (X_{ij} - \overline{X_j})(X_{ik} - \overline{X_k})}{\sigma_{Xj} \cdot \sigma_{Xk}}$$

$$= \frac{COV(X_j, X_k)}{\sigma_{Xj}\ \sigma_{Xk}}$$

d'où

$$COV(Z_j, Z_k) = COR(X_j, X_k)$$

En conclusion on a :

$$COV(Y_j, Y_k) = COV(Z_j, Z_k) = COR(X_j, X_k)$$

III.3 ETUDE DU PHENOMENE PLUVIOMETRIQUE SUR LE LITTORAL ALGERIEN

III.3.1 PRESENTATION DES VARIABLES ET RESULTATS DE L'A.C.P

On a procédé à une A.C.P *normée* (variables centrées réduites) sur les six (06) stations pluviométriques situées toutes au Nord de l'Algérie donc soumises au climat méditerranéen GHAZAOUET, ORAN, ALGER, EL KALA, SKIKDA et ANNABA, et les 384 mois équivalents à 32 années d'observations.

Les valeurs propres $\lambda(K)$ obtenues sont les suivantes :

Composante K	1	2	3	4	5	6
Valeur propre λ(k)	3,89	1,13	0,37	0,27	0,21	0,13

Ce qui ramené en pourcentage de la variance expliquée ($\lambda(K)/6$) donne :

Composante K	1	2	3	4	5	6
Variance expliquée en %	64,77	18,76	6,25	4,52	3,51	2,19

Enfin les pourcentages cumulés de variance expliquée sont :

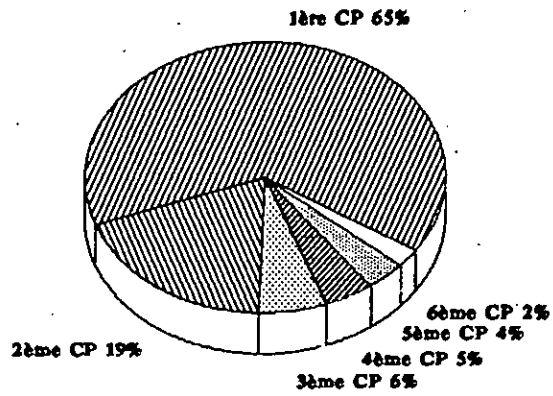
Composante 1 à K	1	2	3	4	5	6
Pourcentage cumulé	64,77	83,53	89,78	94,30	97,81	100,00

On constate donc que la première composante accapare à elle seule plus de la moitié de la variance totale **(voir figure III.3.1)**, et que les deux premières composantes permettent d'expliquer plus de (4/5) de la variance totale. Les composantes suivantes expliquent une part très faible de la variance, environ 16 %, et de ce fait leur rejet n'entraîne qu'une perte minime de l'information.

Les résultats obtenus lors d'une A.C.P. effectuée sur des données brutes, c'est-à-dire n'ayant subi aucune transformation au préalable sont :

Les valeurs propres :

C.P. (K)	1	2	3	4	5	6
Valeurs propres λ(K)	13098,6	2364,1	1466,0	998,1	420,9	333,1



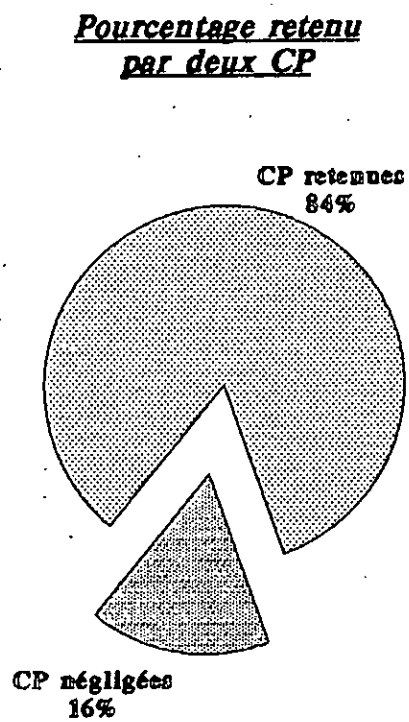


Figure III.3.1 : Choix du nombre de CP

Les pourcentages de la variance expliquée :

C.P. 1 à K	1	2	3	4	5	6
Variance expliquée en %	70,12	12,66	7,84	5,34	2,26	1,78

Les pourcentages cumulés :

C.P. 1 à K	1	2	3	4	5	6
% cumulé	70,12	82,78	90,62	95,96	98,22	100,00

On remarque une augmentation de la contribution du premier axe par rapport au cas précédent (A.C.P. *normée*), la différence est d'environ 6 %, cet écart est récupéré par les autres composantes de façon que la somme des contributions de toutes les C.P. reste égale à 100 %.

Les écarts types des variables initiales :

Stations	ALGER	ANNABA	SKIKDA	GHAZAOUET	EL KALA	ORAN
Ecart type (mm)	62,98	51,21	64,40	41,34	68,96	38,32

Les stations de SKIKDA et EL KALA possèdent des écarts types relativement élevés, tandis que ceux des stations d'ORAN et de GHAZAOUET sont plus faibles. Un simple calcul statistique permet d'évaluer l'erreur de ces écarts types qui est de 12,9 mm. Ceci implique une variabilité notable et une dispersion assez importante des écarts types.

C'est pour cela que, lors d'une ACP *Brute*, ce sont les stations qui possèdent des écarts types élevés qui contribuent fortement à la détermination des axes principaux, et les stations ayant de faibles écarts types se trouvent écartés voire même négligées, malgré leur importance dans le phénomène étudié.

III. 3. 2 . INTERPRETATIONS GRAPHIQUES

Pour l'étude de la répartition spatiale de la pluie, différentes questions peuvent être posées :

- **Quels sont les paramètres qui expliquent la corrélation inter-stations (positive ou négative) ?**

- **Peut-on dégager des groupes de stations tout en donnant les explications et interprétations nécessaires ?**

Au cours de l'étude, chaque variable est affectée d'un code qui facilite sa manipulation

Variables	Codes affectés
ALGER	A0
ANNABA	A1
SKIKDA	A2
GHAZAOUET	A3
EL KALA	A4
ORAN	A5

En passant aux représentations graphiques, considérons en premier lieu le cercle de corrélation sur le plan défini par les deux premières composantes **(figure III - 3.2.)**

On a reporté sur ce cercle les différentes stations en fonction de leur coefficient de corrélation avec les deux axes considérés.

On constate que toutes les stations sont corrélées positivement avec la première composante et que ce coefficient décroît lorsque l'on s'éloigne de l'axe 1.
Le résultat étant tout à fait prévisible puisque la première composante explique le comportement *"MOYEN"* du phénomène pluviométrique sur le nord de l'Algérie.

Des études pluviométriques (Réf.N°14) donnent une moyenne mensuelle d'environ 60 mm sur le nord de l'Algérie.
Il est donc normal que la station la mieux corrélée avec le premier axe soit celle de SKIKDA du fait qu'elle se trouve dans la région la plus pluvieuse d'Algérie, l'EST.
L'examen du tableau des moyennes mensuelles des différentes stations :

Stations	ALGER	ANNABA	SKIKDA	GHAZAOUET	EL KALA	ORAN
Moyenne (mm)	61,4	52,4	64,5	30,1	65,7	32,2

permet de voir la dispersion de celles-ci par rapport à la moyenne générale, 51 mm. Cette dispersion est évaluée à environ 16 mm et est par conséquent assez importante ; on peut l'attribuer à la distance séparant l'EST de l'OUEST(1 200 km).

En terme de corrélation, l'étude du **tableau III - 3.1.** des projections des variables sur les axes principaux montre une assez forte corrélation avec le premier axe : le coefficient de détermination varie de 0,57 à 0,77 Celui-ci décroît très vite pour les autres axes allant même jusqu'à s'annuler.

D'autre part, à partir de la représentation graphique donnée par **la figure III - 3.2.**, il apparaît une distinction évidente entre les stations : ANNABA, SKIKDA, EL KALA, et celles de GHAZAOUET, et ORAN, alors que la station d'ALGER se positionne entre les deux groupes résultants.
La seconde composante a donc tendance à opposer le groupe EST à celui de l'OUEST.

Les résultats présentés dans **le tableau III - 3.2.** mettent en évidence, d'une part le faible lien (coefficient de détermination insignifiant) entre les stations de l'EST et celles de l'OUEST. En effet, les coefficients de corrélation entre la station de SKIKDA (à l'EST) et celle de GHAZAOUET (à l'OUEST) sont de 0,5 (coefficient de détermination est de 0,25), et d'autre part une forte liaison entre les stations appartenant au même groupe. Le coefficient de corrélation entre la station de ANNABA et celle de SKIKDA par exemple, appartenant au groupe de l'EST est de 0,864. Pour ce qui est de la station d'ALGER qui se trouve au centre, à mi-chemin entre les deux groupes, ses coefficients de corrélation avec la station de SKIKDA (EST) et celle d'ORAN (OUEST) sont respectivement de 0,72 et 0,62.

	A0	A1	A2	A3	A4	A5
1	0.86	0.86	0.88	0.64	0.80	0.76
2	0.00	-0.37	-0.31	0.69	-0.33	0.55
3	-0.40	0.00	-0.09	0.19	0.41	-0.04
4	-0.29	0.23	0.23	0.09	-0.27	0.01
5	0.12	0.03	-0.01	0.27	-0.04	-0.35
6	0.02	0.25	-0.26	-0.02	-0.03	0.03

Tableau III.3.1 : Projection des variables sur les axes principaux

Figure III-3.2. : Cercles de corrélation (Pluie)

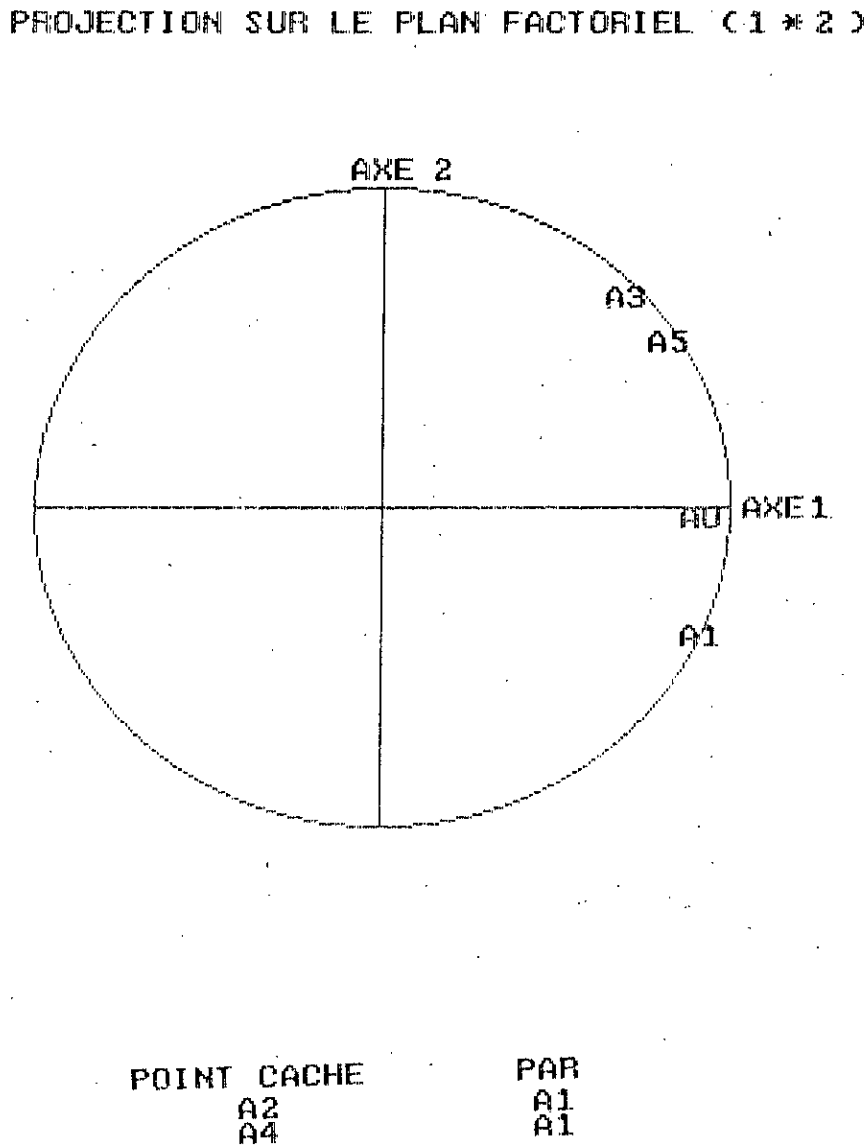
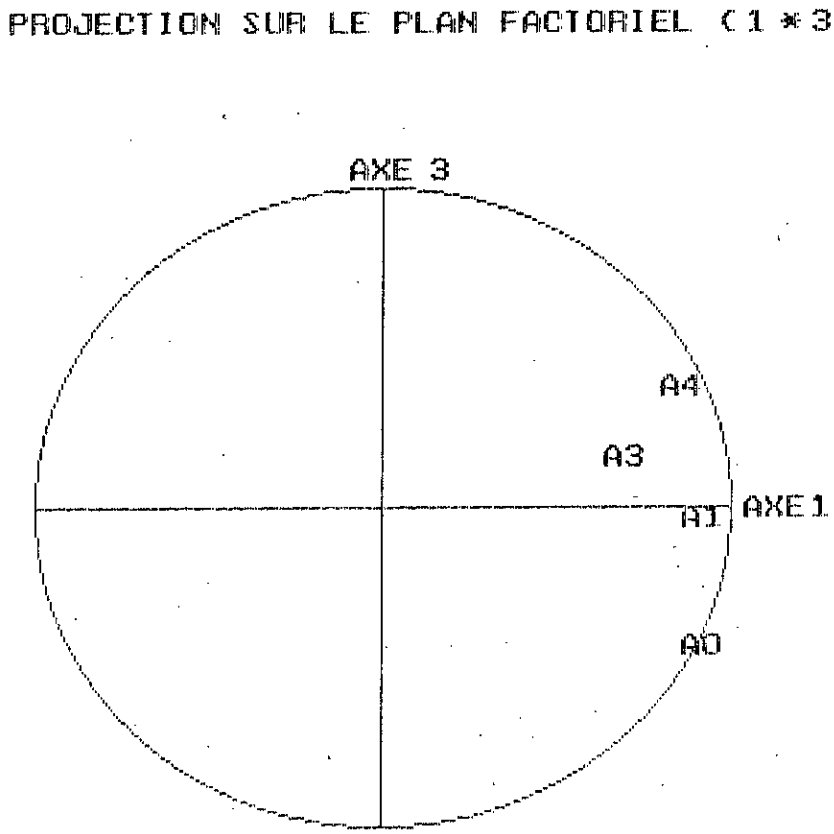


Figure III-3.2. : Cercles de corrélation (Pluie)



POINT CACHE	PAR
A2	A1
A5	A1

Figure III-3.2. : Cercles de corrélation (Pluie)

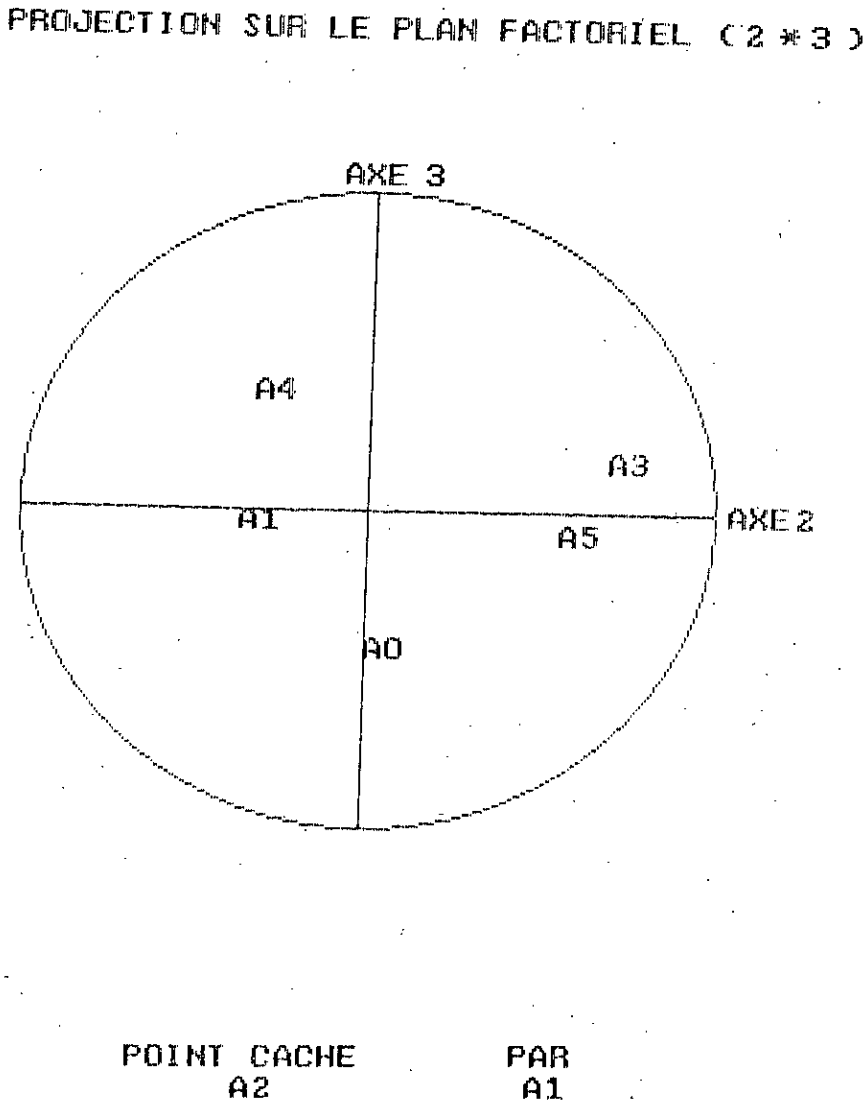
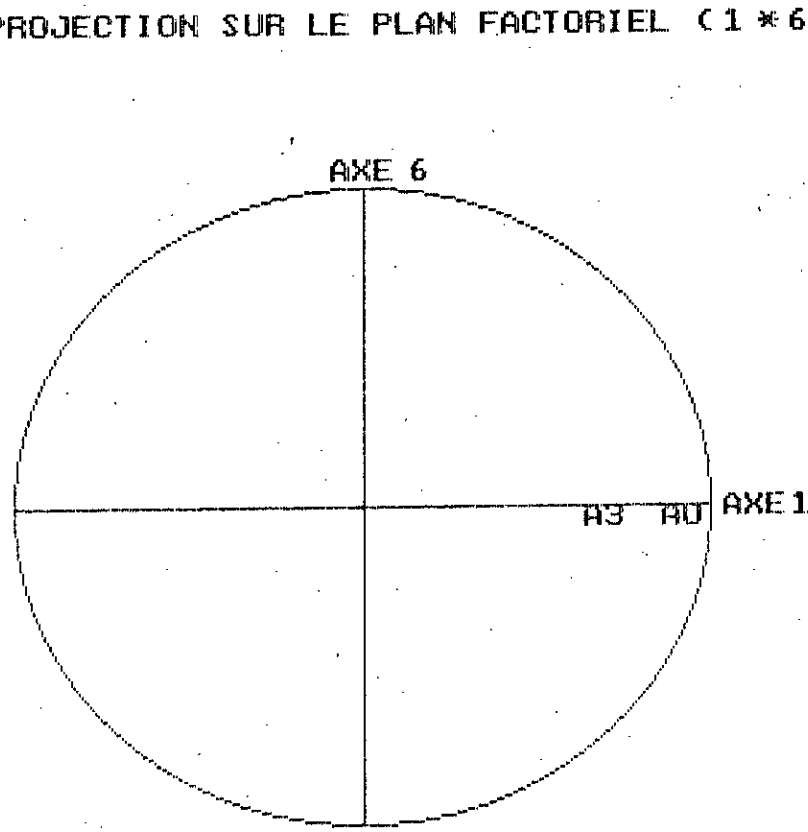


Figure III-3.2. : Cercles de corrélation (Pluie)



POINT CACHE	PAR
A1	A0
A2	A0
A4	A0
	A0

	A0	A1	A2	A3	A4	A5
A0	1					
A1	0.684	1				
A2	0.722	0.864	1			
A3	0.482	0.320	0.352	1		
A4	0.600	0.745	0.718	0.328	1	
A5	0.622	0.446	0.500	0.764	0.420	1

Tableau III.3.2 : Matrice de corrélation

En conclusion, on peut dire que la station d'ALGER présente une meilleure corrélation avec les stations de l'est et de l'ouest, que celles-ci entre elles, tout en étant fortement corrélées avec le groupe "EST". et enfin que les meilleurs coefficients de corrélation sont obtenus entre les stations d'une même région, ce qui permet d'affirmer que le second axe explique l'effet *"REGIONAL"*.

Après avoir montré que seuls les deux premiers axes possèdent une structure spatiale, les autres composantes ne peuvent qu'expliquer des comportements communs à des stations pourtant sans lien spatial ; il est très vraissemblable que ces stations ne se ressemblent que par des comportements exceptionnels à des périodes assez voisines (erreurs accidentelles, simultanées, période de *"détarage"* en grande partie commune).

La condensation notée des variables autour du centre, dans les plans résultants de la combinaison des dernières CP, montre clairement la faible contribution de celles-ci à l'explication des variables qui ont pratiquement des projections nulles sur les axes principaux **Figure III.3.2.**

III.4 VARIATIONS DE L'EVAPOTRANSPIRATION EN FONCTION DES PARAMETRES HYDROMETEROLOGIQUES

III.4.1 DEFINITION DES VARIABLES UTILISEES

L'évaporation, en général est un phénomène hydrologique défini comme étant le passage de la phase liquide à la phase vapeur, il dépend de plusieurs paramètres climatiques. Il peut être mesuré expérimentalement ou calculé directement à partir des formules empiriques.

Pour l'étude de ce phénomène on dispose des données suivantes :

- **L'Insolation : durée pendant laquelle un couvert végétal est exposé aux rayons du soleil, exprimée en heure [h].**
- **La température : température moyenne mensuelle exprimée en degrés Celsius [°C]**

- **La vitesse du vent : exprimée en [m/s]**
- **L'humidité relative : rapport de la pression effective de la vapeur d'eau à la pression maximales exprimée en pourcentage [%]**
- **L'ETP : évaporation transpiration potentiel calculée par la formule de PENMAN représente la demande climatique d'un couvert végétal exprimée en [mm/j].**

Ces paramètres constituent les cinq variables sur lesquelles se base l'étude du phénomène de l'ETP. Pour cela, on dispose de 264 mois (22 années) d'observations des cinq variables étudiées.

III.4.2 RESULTATS DE L'ACP

On effectue une ACP *normée* sur le tableau "264 individus x 5 variables".

Pour une question pratique, chaque variable a été affectée d'un code qui l'identifie. La liste des variables avec leurs codes correspondants est présentée dans le tableau suivant :

Variables	codes affectées
Températures	A0
Vitesse du vent	A1
Insolation	A2
Humidité relative	A3
ETP	A4

Les variances expliquées pour les différentes composantes sont données par :

Composantes k	1	2	3	4	5
Variance expliquée λ(k)	3.160	1.025	0.594	0.193	0.029

En calculant les contributions à la variance totale en terme de pourcentage on obtient

Composante k	1	2	3	4	5
% de variance expliquée	63.2	20.5	11.87	3.85	0.57

Le cumul de la variance expliquée donne :

Composante de 1 à k	1	2	3	4	5
Cumul de variance expliquée %	63.20	83.70	95.58	99.43	100.00

En examinant l'apport de chaque composante (**voir figure III.4.1.a**) on remarque que pour avoir prés de 84% de la variance totale, il suffit de ne considérer que les deux premières CP ; mais vu l'apport assez consistant du 3ème axe (12%) on ne peut négliger ce dernier. Donc on opte pour trois (03) composantes totalisant environ 96% de variance expliquée.

A titre comparatif, on a effectué une ACP sur les données brutes qui a donné les résultats suivants :

Les valeurs propres :

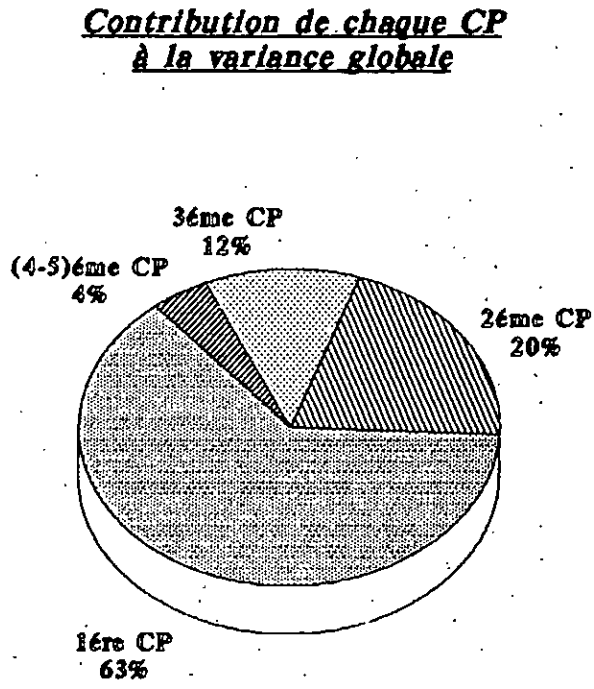
Composante k	1	2	3	4	5
Valeur propre λ(k)	43.00	12.02	2.02	0.48	0.11

Après passage au pourcentage de variance expliquée on obtient :

Composante k	1	2	3	4	5
% de variance expliquée	74.61	20.86	3.53	0.83	0.19

Le pourcentage cumulé est :

Composante de 1 à k	1	2	3	4	5
% Cumulé	74.61	95.47	98.99	99.81	100.00



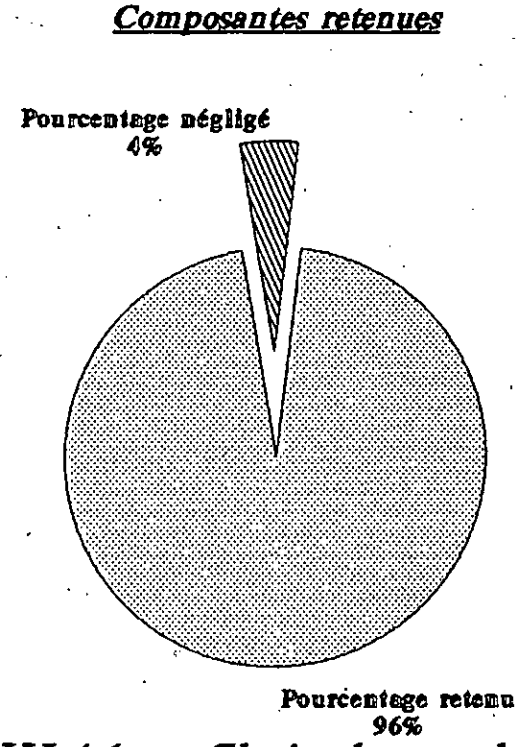
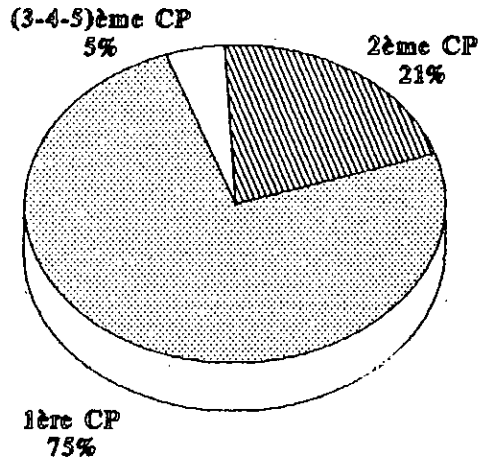


Figure III.4.1.a : Choix du nombre de CP

Contribution de chaque CP à la variance globale



Composantes retenues

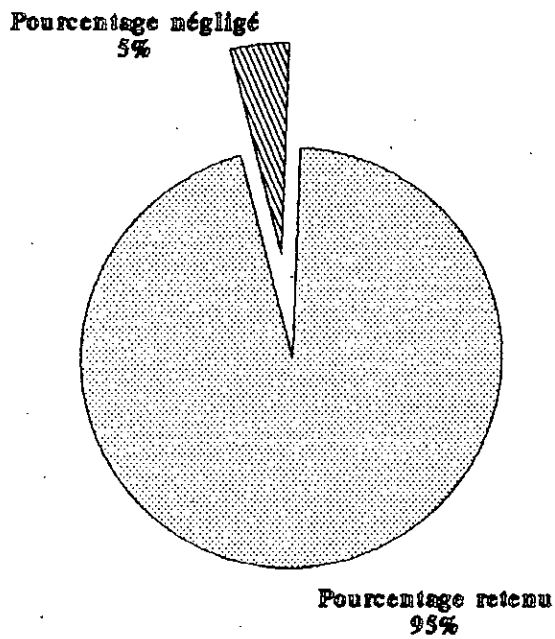


Figure III.4.1.b : Choix du nombre de CP

Une augmentation remarquable est notée pour le premier axe,**(voir figure III.4.1.b)** la différence est évaluée à environ 12% par rapport au cas d'une ACP *normée* ; pour le second axe cette augmentation est insignifiante. La différence, entre les résultats des deux types d'ACP, ne peut être considérée comme étant une perte d'informations pour le cas d'une ACP *normée* car elle est récupérée par les composantes suivantes, soit dans ce cas, la 3 ème composante qui passe de 3.58% (ACP *canonique*) de variance expliquée à 11.87% (ACP *normée*).
Le pourcentage cumulé de variance expliquée peut se limiter dans cette dernière ACP à deux composantes avec un taux de variance expliquée atteignant 95%.

Pour vérifier si le fait de considérer deux (02) composantes principales dans le cas de l'ACP *"brute"* et trois CP dans le cas *"normée"*, ne fausse pas les interprétations, et par conséquent la compréhension du phénomène de l'ETP, on passe aux résultats donnant la contribution de chaque composante à l'explication des variables étudiées.
En se basant sur les **figures (III.4.2.a)** et **(III.4.2.b)**, on obtient le tableau récapitulatif suivant :

Type d'ACP / variable étudiée	ACP brute	ACP normée
A0	(1ère + 2ème) CP	1er CP
A1	4ème CP	2ème CP
A2	1ère CP	1ère CP
A3	(1er + 2ème) CP	(1ère + 3ème) CP
A4	(1+2+3+4+5)ème CP	1ère CP

En conclusion, on dira que le fait de négliger les trois (03) dernières CP dans le cas *"brute"* ne fait intervenir en aucun cas l'influence de la vitesse du vent (A1) et l'ETP (A4) dans la détermination des axes principaux.

Ce problème ne se pose pas dans le cas *"normé"* vu que l'information est concentrée dans les premières CP, donc on peut négliger sans risques les CP à faibles contributions.

Les écarts types des variables initiales sont :

Variables	**températures [°C]**	**Vitesse du vent [m/s]**	**Insolation relative [h]**	**Humidité relative [%]**	**ETP [mm/j]**
Ecart types	5.03	0.66	2.69	4.73	1.82

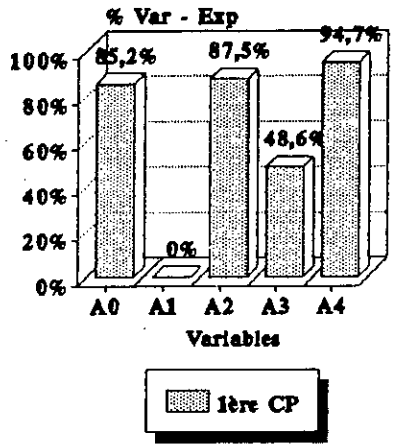
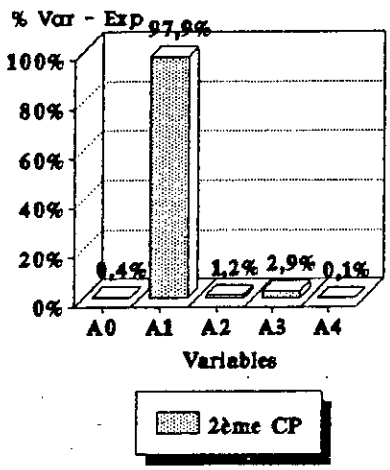
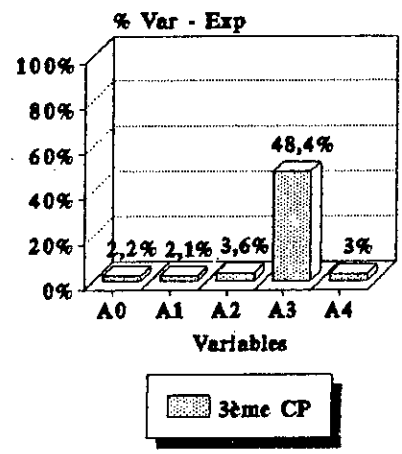
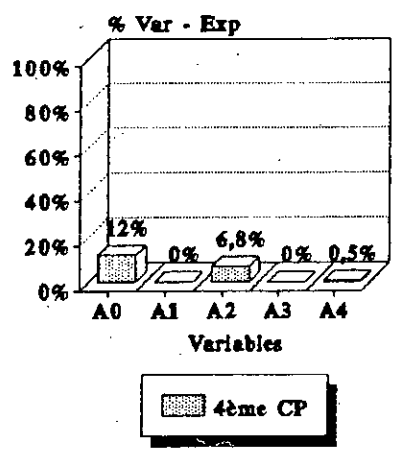
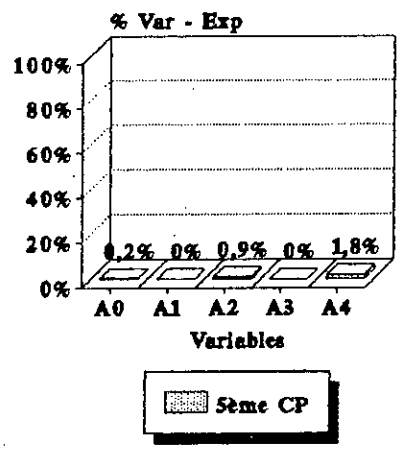


Figure III.4.2.a: Pourcentage de Variance expliqué par chaque composante









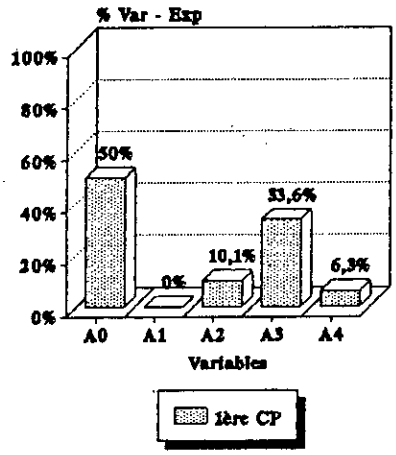
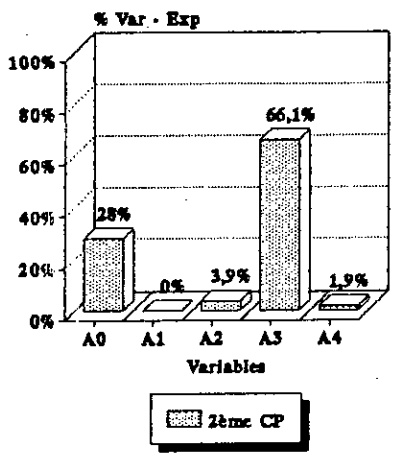
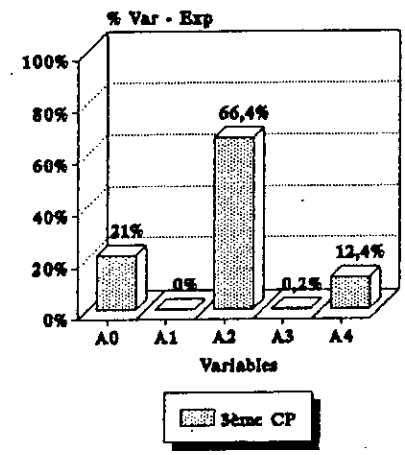
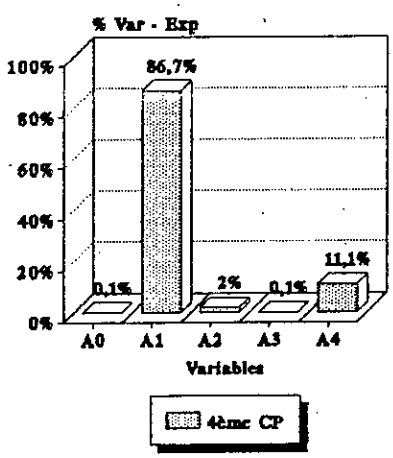
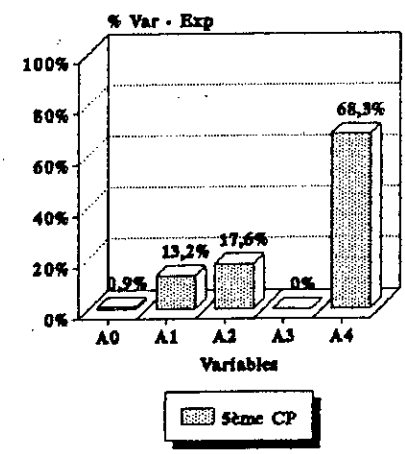


Figure III.4.2.b: Pourcentage de Variance expliqué par chaque composante









On remarque tout de suite que la vitesse du vent et l'ETP possèdent des écarts types relativement faibles par rapport à ceux de la température et l'humidité relative, ainsi que la différence d'unités dans lesquelles sont exprimées les variables étudiées. Effectivement, on passe du degré Celsius, pour la température, mètre par seconde pour la vitesse du vent, heure pour l'insolation, pourcentage pour l'humidité relative, et enfin millimètre par jour pour l'ETP. Toute cette hétérogénéité fait que la vitesse du vent et l'ETP soient expliquées principalement par les composantes à faible contribution qu'on a tendance à négliger.

III.4.3 INTERPRETATIONS GRAPHIQUES

On considère dans le cas *normé* le cercle de corrélation (C1 x C2) (**voir figure III.4.3.a**). On remarque une condensation des variables A0, A2 et A4 autour de l'axe 1, ceci signifie que la température (A0) tout comme l'insolation (A2) sont des paramètres prépondérants dans la détermination de l'ETP. Sachant que ces deux paramètres représentent l'énergie nécessaire pour transformer une goutte d'eau en vapeur, alors l'axe 1 représente [L'ENERGIE], la variable A3 est opposée au groupe formé par A0, A2 et A4 par rapport à l'axe 1 ; le tableau des corrélations (**tableau III.4.1**) montre effectivement une corrélation négative entre l'humidité relative et les autres variables. Physiquement une importante humidité relative est un apport pour le couvert végétal contrairement à l'ETP.

En ce qui concerne la variable A1 elle n'est corrélée qu'avec le second axe et le tableau des corrélations (**tableau III.4.1**) montre presque une indépendance de cette variable avec les autres. Donc l'axe 2 est défini par l'effet du vent jouant le rôle de transport des masses d'air saturées en vapeur d'eau.

En examinant les différents cercles de corrélation (**figure III.4.3.a**) on constate que les variables (A0, A2, A4) sont toujours regroupées. Ceci fait apparaître le phénomène de redondance entre la température (A0) et l'insolation (A2). Dans l'explication de l'ETP (A4). dans le **tableau III.4.1 a** on remarque que la corrélation entre la température et l'insolation est très forte (de l'ordre de 0.94).

	A0	A1	A2	A3	A4
A0	1.000				
A1	-0.031	1.000			
A2	0.813	-0.071	1.000		
A3	-0.525	-0.074	-0.505	1.000	
A4	0.9000	0.064	0.945	-0.565	1.00

Tableau III.4.1 Matrice de corrélation

Figure III-4.3.a : Cercles de corrélation (E.T.P.)

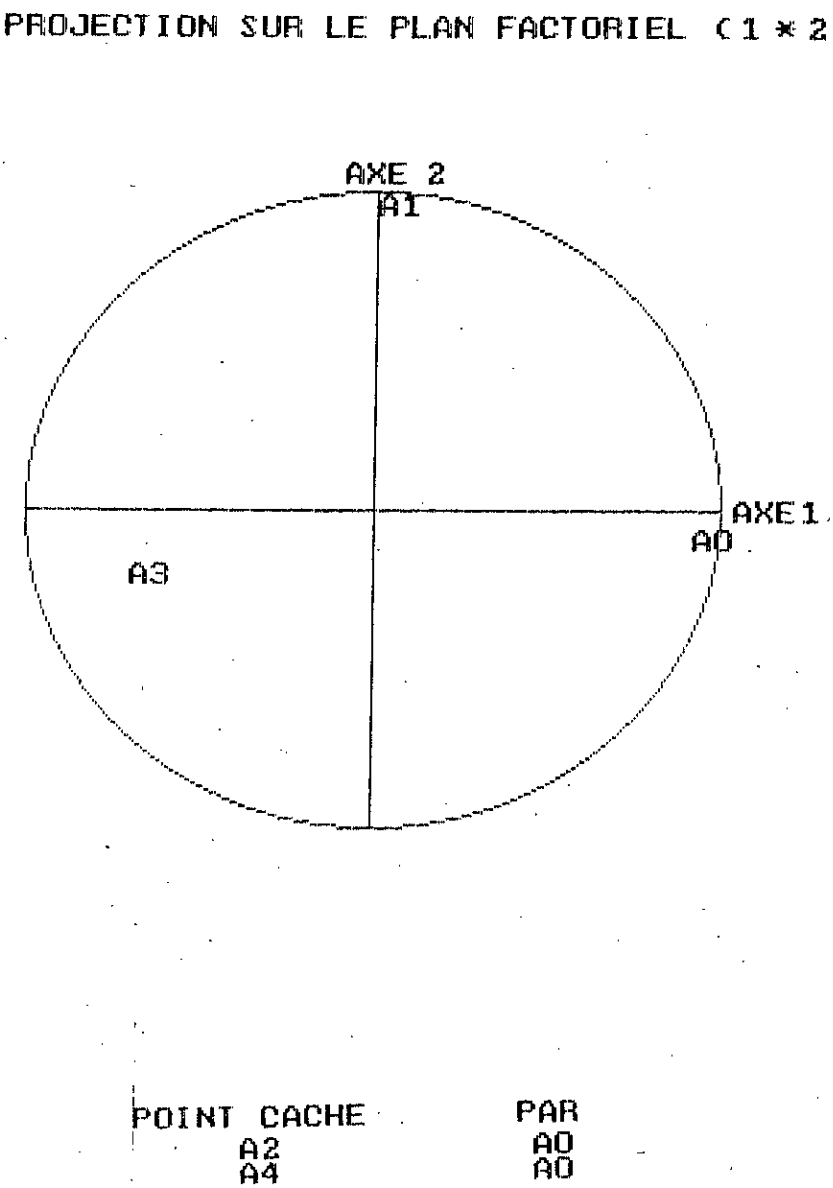


Figure III-4.3.a : Cercles de corrélation (E.T.P.)

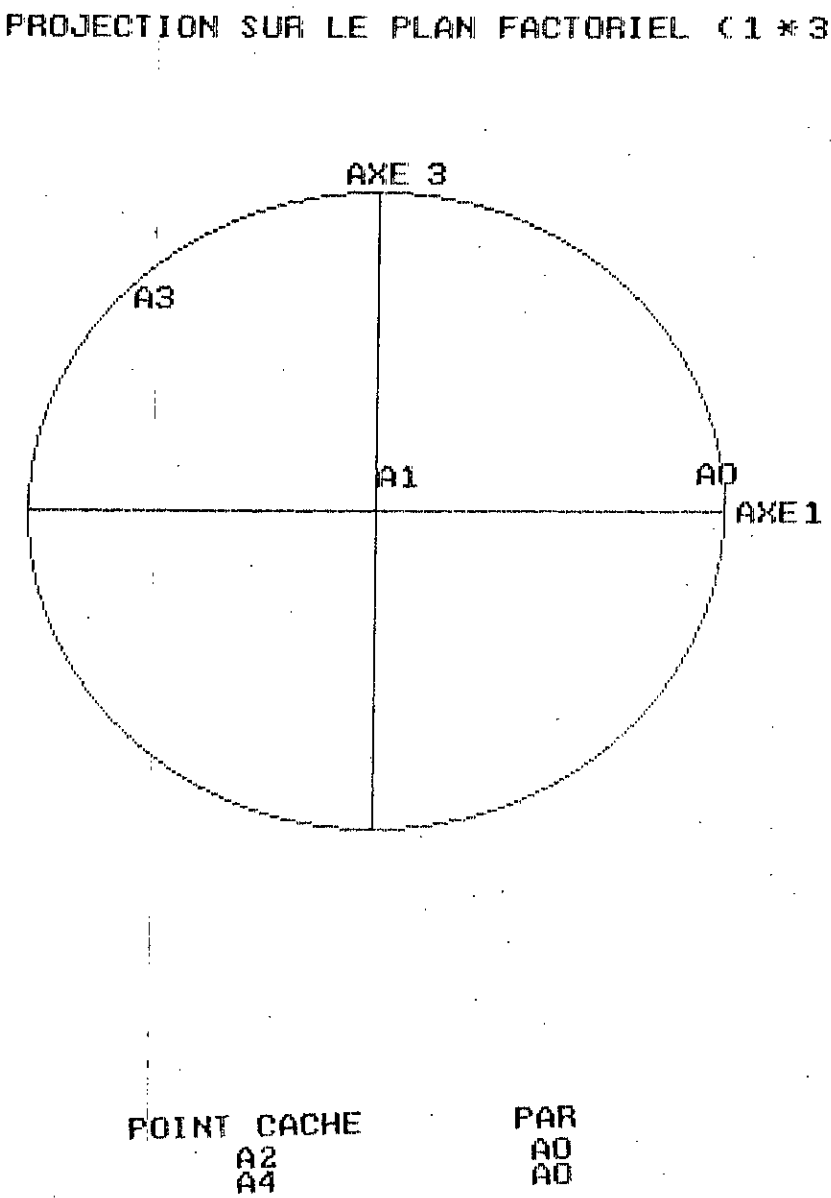


Figure III-4.3.a : Cercles de corrélation (E.T.P.)

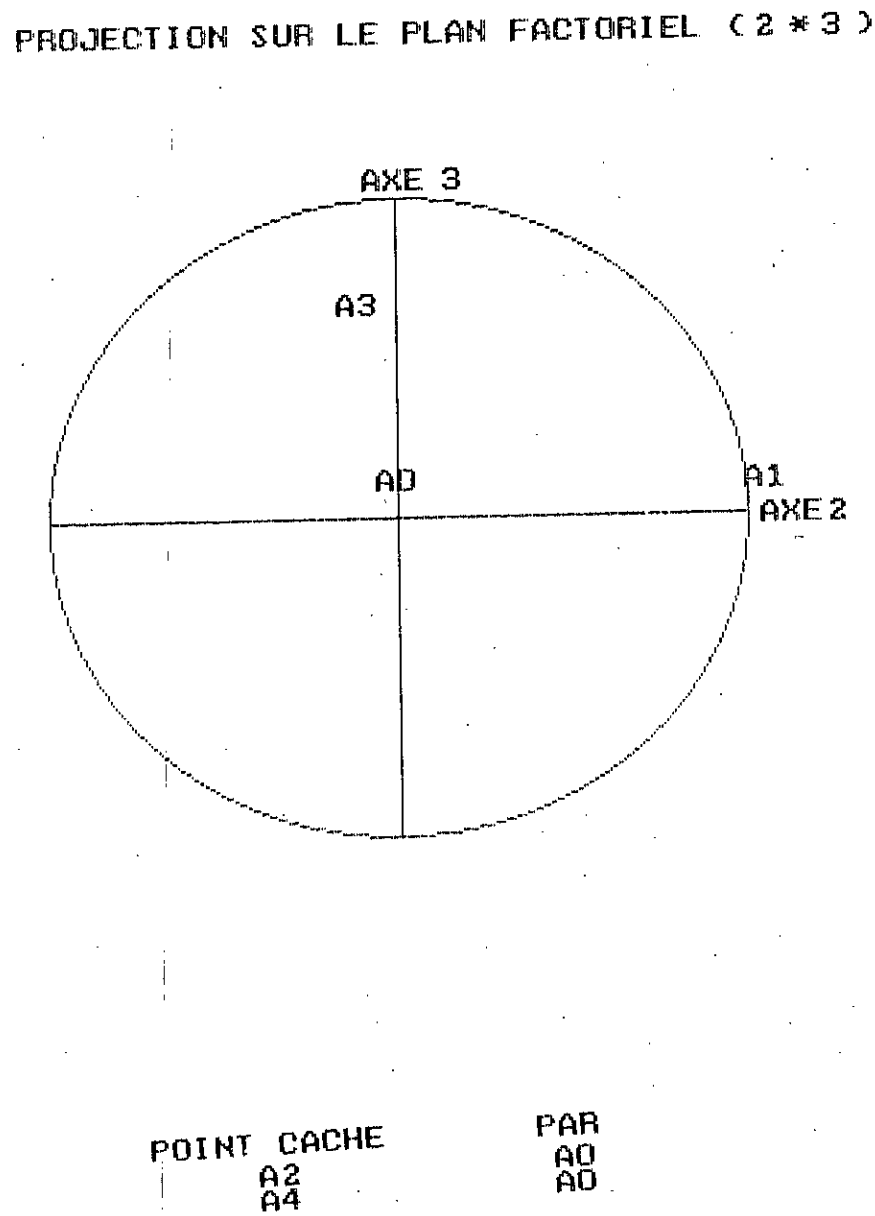
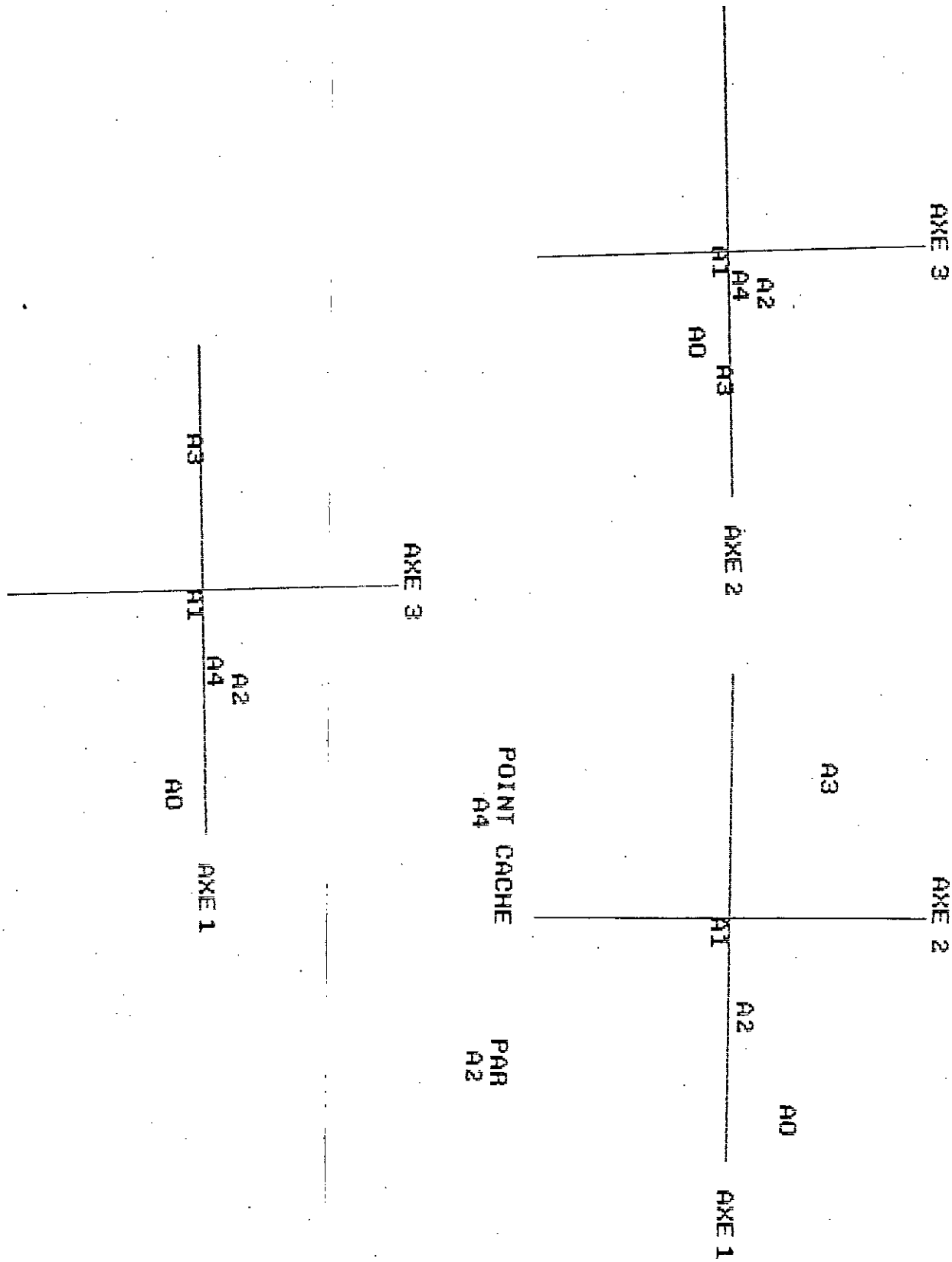


Figure III-4.3.b : Plans factoriels (E.T.P.)



Dans le cas *brute* (**voir figure III.4.3.b**) on remarque que les angles que fait chaque variable avec les axes principaux sont conservés, alors que la distance reliant la variable à l'origine varie en fonction de son écart type ; l'expression analytique de ce dernier est :

$$\sigma_{xj}^2 = (1/N) \sum_{i=1}^{N} (X_{ij} - \overline{X}_j)^2$$

Cette expression est équivalente à 1/N prés à la norme du vecteur reliant le centre de gravité X_j à la variable X_j.

Les variables A0, A2 et A4 regroupées dans le cercle de corrélation (1 x 2) conservent le même angle avec l'axe 1, cependant, la distance les séparant de l'origine varie. Les variables apparaissent dans un ordre qui coïncident avec l'ordre croissant des écarts types respectivement égaux à : 1.823, 2.488 et 5.028. Il en est de même pour la vitesse du vent (A1) dans le plan factoriel (1x2) (**figure III.4.3.b**) qui paraît non expliquée ni par l'axe 1, ni par l'axe 2 ; alors que dans le cercle de corrélation (**figure III.4.3.a**), elle définit presque entièrement l'axe 2. La variable A1 ayant un écart type inférieur à 1 (σ (A1) = 0.659) fait que le coefficient de corrélation qui était important dans le cercle (C1 x C2), se trouve réduit dans le plan factoriel vu que la covariance est le résultat du produit du coefficient de corrélation et de l'écart type.

III.5 CONCLUSION

L'Analyse en Composante Principales dépend essentiellement du type de variables utilisées ; de ce fait toute transformation appliquée aux variables, modifie obligatoirement la structure des CP.

On a montré qu'une A.C.P appliquée sur des données brutes est équivalente à une ACP effectué sur des données centrées, idem pour les données réduites et centrées réduites.

A travers les deux exemples étudiés, on a constaté que des variables à faibles écarts types sont souvent expliquées, dans le cas d'une ACP *brute*, par les CP à faible contribution qu'on a tendance à négliger.

Dans le cas du phénomène de l'évapotranspiration étudie, les variables utilisées étaient exprimées dans un certain système d'unités, tout changement d'unités, par exemple si on passe de [h] à [s] et de [°C] en [°K], donne lieu à des interprétations différentes.

Pour apporter une solution à ce problème, on se doit de standardiser les variables étudiées en les réduisant (écart type = 1)

Il n'est donc pas équivalent de calculer les CP sur des variables centrées (réduites) donc avec une matrice de corrélation ou, sur des variables brutes (centrées) donc avec une matrice de covariance, lorsque les paramètres utilisés ont des variables différentes : car les premiers composantes risquent d'être définies uniquement par les variables à forte variance - celles ayant une faible pondération - alors qu'au contraire elles conditionnent presque totalement les composantes correspondant aux plus faibles valeurs propres qui sont celles que l'on élimine généralement.

L'interprétation graphique des cercles de corrélation et des plans factoriels fait ressortir les points suivants :

- **L'axe 1 explique toujours la tendance centrale du nuage de point : on parle aussi de l'effet de taille. Une condensation des variables autour de cet axe est souvent remarquée.**

- **L'apparition successive d'un regroupement de variables dans différents plans fait ressortir le phénomène de redondance.**

- **Un ensemble de variables groupées dans un cercle de corrélation (ACP *normée*) éclate dans le plan factoriel (ACP *brute*) à cause de la différence des écarts types.**

Il est donc indispensable de tenir compte de ces différents aspects et de leurs conséquences lorsqu'on effectue une ACP.

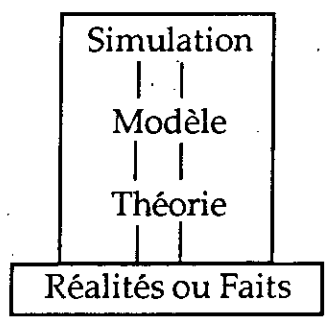
Ces remarques ne diminuent en rien le potentiel du puissant outil d'analyse l'ACP.

CHAPITRE IV
APPLICATION DE L'A.C.P. DANS LE DOMAINE OPERATIONNEL

IV.1. CONCEPT DE SIMULATION

La simulation est une technique qui ressemble à l'expérimentation scientifique utilisant des modèles dont toutes les caractéristiques sont connues.

La figure ci-dessous montre les différentes étapes qui aboutissent à la simulation, à sa base, la figure repose sur la réalité ou les faits.



Selon les situations, il est possible de construire des modèles pour tester ou représenter une théorie. Ainsi, la simulation réside dans l'utilisation d'un modèle qui vise à indentifier ou découvrir le comportement d'un procédé ou d'un système. Elle constitue finalement une approche de résolution de problèmes basée sur l'essai et l'erreur ; elle aide à prendre de meilleures décisions et représente donc un excellent outil de planification.

Les méthodes actuelles de génération présentent une difficulté d'interprétation de *"causes à effets"*, souvent attribuée au manque d'indépendance entre les variables utilisées ; or, les besoins actuels en Hydrologie nous orientent vers la recherche de nouvelles méthodes de génération pour que l'interprétation *"causes à effets"* soit facilitée. Il y aurait avantage à ce que ces méthodes utilisent des variables indépendantes (voir Réf. N°17).

IV.2. MODELE DE SIMULATION

Par définition, les Composantes Principales sont des combinaisons linéaires des variables initiales ; dans le cas d'une A.C.P. *normée*, cela se traduit par :

$$C_l = \sum_{j=1}^{P} a_{lj} \frac{(Xj - \overline{Xj})}{\sigma_{xj}} \qquad \text{(IV .2.1.)}$$

C_l : l^{ième} C.P.
X_j , X_j , σ_{xj} : sont respectivement la variable, moyenne et écart type d'ordre j
a_{lj} : cosinus directeur, élément de la matrice [A] des vecteurs propres.

Soit la composante principale normée C'_l tel que :

$$C'_l = \frac{C_l}{\lambda_l^{1/2}}$$

λ_l : valeur propre correspondant à C_l

En divisant l'équation IV.2.1. par $\lambda^{1/2}_l$, on obtient :

$$C'_l = \frac{C_l}{\lambda_l^{1/2}} \sum_{j=1}^{P} a_{lj} \frac{(Xj - \overline{Xj})}{\sigma_{xj}} \qquad \text{(IV.2.2)}$$

Soit Y la variable centrée réduite telle que :

$$Y_j = \frac{X_j - \overline{X}_j}{\sigma_{xj}} \qquad \text{(IV . 2.3.)}$$

Sous forme matricielle l'expression (IV.2.2.) devient :

$$[C'] = \lambda^{-1/2} [A]^t [Y] \qquad \text{(IV.2.4.)}$$

Avec

λ : Vecteur dont les composantes sont les valeurs propres
[A] : Matrice des cosinus directeurs, formée par les vecteurs propres
$[A]^t$: Matrice transposée de [A]
[Y] : Matrice des variables initiales centrées réduites.

Réciproquement on peut exprimer la variable Y_j en fonction des C.P. *normées*.

En transposant la matrice des vecteurs propres [A] on obtient :

$$[Y] = \lambda^{1/2} [A] [C'] \qquad \text{(IV.2.5.)}$$

Sachant qu'on ne considère que les Q premières C.P. et qu'on ne perd pas de vue la variance non expliquée prise en compte par le terme résiduel noté ε_j.

En explicitant la formule (IV.2.5) on aura :

$$Y_i = \sum_{l=1}^{Q} \lambda_l^{1/2} a_{jl} C'_l + \varepsilon_j \qquad \text{(IV.2.6.)}$$

ε_j : variable résiduelle de moyenne nulle et d'écart type :

$$\sigma_{\epsilon j} = [\, 1 - \sum_{l=1}^{Q} a^2_{jl} \lambda_l \,]^{1/2} \qquad \text{(IV.2.7)}$$

Sachant que :

$$\lambda_j^{1/2} a_{jl} = COV (C'_l, Y_j)$$

$$= COV (C'_l, \frac{X_j - \overline{X_j}}{\sigma_{Xj}})$$

$$= COR (C'_l, X_j)$$

En utilisant ce résultat l'équation (IV.2.6) devient :

$$Y_j = \sum_{l=1}^{Q} COR (C'_l, X_j) . C'_l + \varepsilon_j \qquad \text{(IV. 2.8)}$$

Substituons l'équation (IV.2.3) dans (IV.2.8) :

$$\frac{X_i - \overline{X_i}}{\sigma_{Xj}} = \sum_{l=1}^{Q} COR (C'_l, X_j) . C'_l + \varepsilon_j \qquad \text{(IV.2.8)}$$

$$X_j = \overline{X_j} + \sigma_{xj} . \sum_{l=1}^{Q} COR (C'_l, X_j)\ C'_l + \sigma_{xj}\ \varepsilon_j \qquad \text{(IV.2.9)}$$

$$X_j = \overline{X}_j + \sum_{l=1}^{Q} \sigma_{xj} . COR (C'_l , X_j) \; C'_l + \sigma_{xj} \; \varepsilon_j \qquad (IV.2.9)$$

Etant donné que :

$$\sigma_{C'l} = 1 \qquad (C.P. \text{ "normée"})$$

Alors :

$$\sigma_{Xj} \; COR (C'_l , X_j) = \frac{\sigma_{Xj}}{\sigma_{Xj}} . COR (C'_l , X_j)$$

$$= \beta_{lj}$$

avec :

$\sigma_{C'l}$: écart type de la CP *normée* C'_l, qui est égale à l'unité.
β_{lj} : coefficient de régression entre la variable X_j et la C.P. normée C'_l.

En posant $E_{.j} = \sigma_{Xj} . \varepsilon_{.j}$

l'expression finale du modèle est donnée par :

$$X_{.j} = \beta_{j0} + \sum_{l=1}^{Q} \beta_{jl} C'_l + E_{.j}$$

avec $\beta_{j0} = \overline{X}_j$ (IV .2.10)

En se basant sur la structure du modèle ci-dessus, différentes possibilités de simulation sont à envisager.

On désignera dans ce qui suit par $\hat{X}_j$ la variable simulée et on remarque bien que c'est une combinaison linéaire des C.P. qui sont indépendantes par construction.

IV.3. METHODES DE SIMULATION

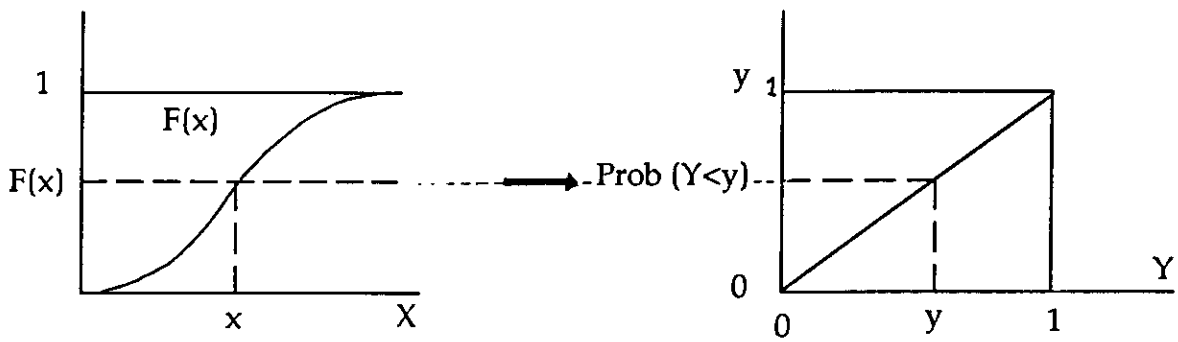
Pour simuler des grandeurs mesurant des phénomènes naturels tels que la pluie ou la température, il ne faut pas perdre de vue que cette notion est basée essesntiellement sur le concept du *hasard*. Ce dernier n'a de sens scientifique que si l'on se réfère au mécanisme de génération.

Toutes les méthodes de simulation se basent sur un tirage de nombres au *hasard*, la différence entre ces méthodes réside dans le choix de l'échantillon sur lequel on opère et la relation entre le paramètre à simuler et le nombre tiré au *hasard*.

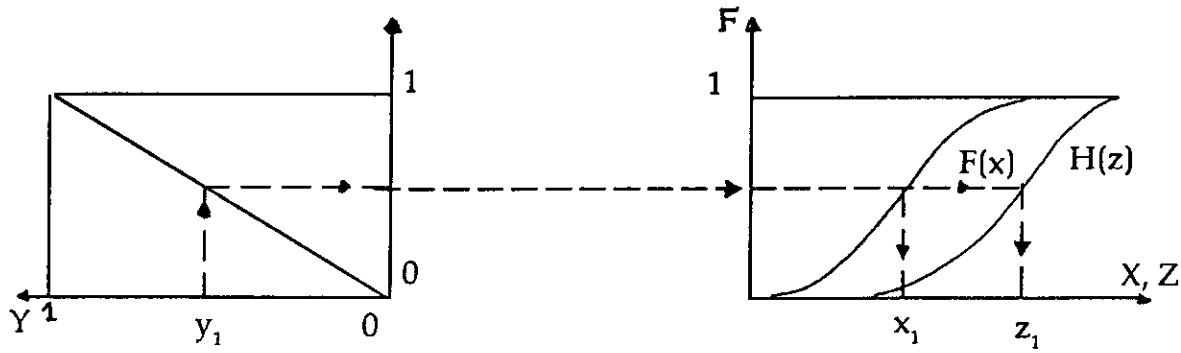
IV. 3.1. SIMULATION PAR LES FONCTIONS DE REPARTITION

Si l'on considère la variable aléatoire X de fonction de répartition F(x), la nouvelle variable Y = F (x) est uniformément répartie sur le segment [0,1]

$$\text{Prob } (Y < y) = \text{Prob } [F(x) < y] = y$$



Quelle que soit la variable aléatoire, on peut toujours la transformer en une variable uniformément distribuée sur le segment [0,1].



La figure ci-dessus représente le tirage au *hasard* :

- **D'une valeur de la variable aléatoire X définie par la loi de probabilité F (x)**
- **D'une valeur de la variable aléatoire Z définie par la loi de probabilité H (z)**

On met la loi de probabilité F ou H dans l'urne des nombres au hasard y_1, qui après transformation fournit la valeur x_1 et /ou z_1.

Remarque

La génération de la fréquence y_1 est remplacée par une loi discrète dont les sauts valent 10^{-k} (K = 3 ou 4) ; on associe des ensembles de K chiffres tirés au *hasard*.

IV. 3.2. SIMULATION PAR LES CHAINES DE MARKOV

Un processus stockastique est un processus dont la valeur à l'instant t dépend des valeurs antérieures, celui-ci est dit MARKOVIEN si la valeur à l'instant t ne dépend en fait que de la valeur la plus récente, c'est-à-dire :

$$P\{X(t) = x / X(t_1) = x_1, X(t_2) = x_2, ..., X(t_n)\} = x_n = \{P\ X(t) = x / X(t_n) = x_n\}$$

Pour simuler une série donnée par les chaînes de MARKOV, il existe deux types de modèles : les modèles "**binaires**" dont le principe est basé sur l'existence de deux états et des modèles "**multiclasses**" où la série est subdivisée en différentes classes.

On utilisera pour notre étude un modèle binaire, d'ordre zéro (0) (Probabilités inconditionnelles) **caractérisé par les deux états zéro (0) et un (1).**

IV.3.2.1. CONSTRUCTION DE LA CHAINE DE MARKOV

On définit en premier lieu **l'état 0 et l'état 1**, sur lesquels on se base pour transformer la série donnée en système binaire (00111010..). Pour obtenir une telle suite, il suffit d'effectuer un comptage des éléments appartenant respectivement à l'état 0 et 1. On calcule ensuite les probabilités inconditionnelles P0 et P1.

$$P0 = \frac{\text{Nombre d'éléments appartenant à l'état 0}}{\text{Nombre total d'éléments}}$$

$$P1 = \frac{\text{Nombre d'éléments appartenant à l'état 1}}{\text{Nombre total d'éléments}}$$

IV.3.2.2. GENERATION PAR LES SERIES DE MARKOV

Après avoir construit la matrice de passage d'ordre zéro (0), on détermine les fonctions de répartition des valeurs appartenant aux états 0 et 1.

La simulation se fait en générant uniformément un nombre aléatoire y dans l'intervalle [0,1], qu'on compare aux probabilités inconditionnelles.

Sachant que $P0 + P1 = 1$

alors $P1 = 1 - P0$

Donc si y est inférieur à P0, la projection de y se fait sur la fonction de répartition correspondant à l'état 0, sinon elle se fera dans celle de l'état 1.

IV .3.3. SIMULATION PAR LES LOIS D'AJUSTEMENT

Les valeurs d'une série de données peuvent avoir une distribution bien spécifique qui est déterminée par l'ajustement des données historiques et vérifiée par les test d'adéquations.

La génération dans la loi uniforme est la base de la simulation dans la plupart des lois d'ajustements statistiques.

Presque tous les calculateurs possèdent une procédure permettant de générer des nombres aléatoires ayant une densité uniforme.

$$f(x) = 1 \quad \text{pour } 0 < x < 1$$

Un tirage aléatoire dans une loi LAPLACE - GAUSS centrée réduite peut être effectuée selon la méthode de BOX et MULLER (1958) : On tire deux nombres U_1 et U_2 grâce à la fonction RANDOM qui effectue un tirage aléatoire dans la loi uniforme sur [0,1].

Et on calcule :

$$Z_1 = (-2 \ln U_1)^{1/2} \cdot \sin (2 \P U_2)$$

$$Z_2 = (-2 \ln U_2)^{1/2} \cdot \sin (2 \P U_1)$$

Ainsi définies, Z_1 et Z_2 sont des variables aléatoires de LAPLACE - GAUSS centrées réduites et indépendantes.

Si la loi d'ajustement est une normale (m, σ) on peut toujours se ramener à une normale de moyenne nulle et d'écart type unité N (0,1) et faire une transformation inverse par la suite.

$$X = m + \sigma . Z$$

Z : variable de GAUSS
X : variable initiale
m, σ : moyenne et écart type de la variable X.

IV.4. OUTIL INFORMATIQUE

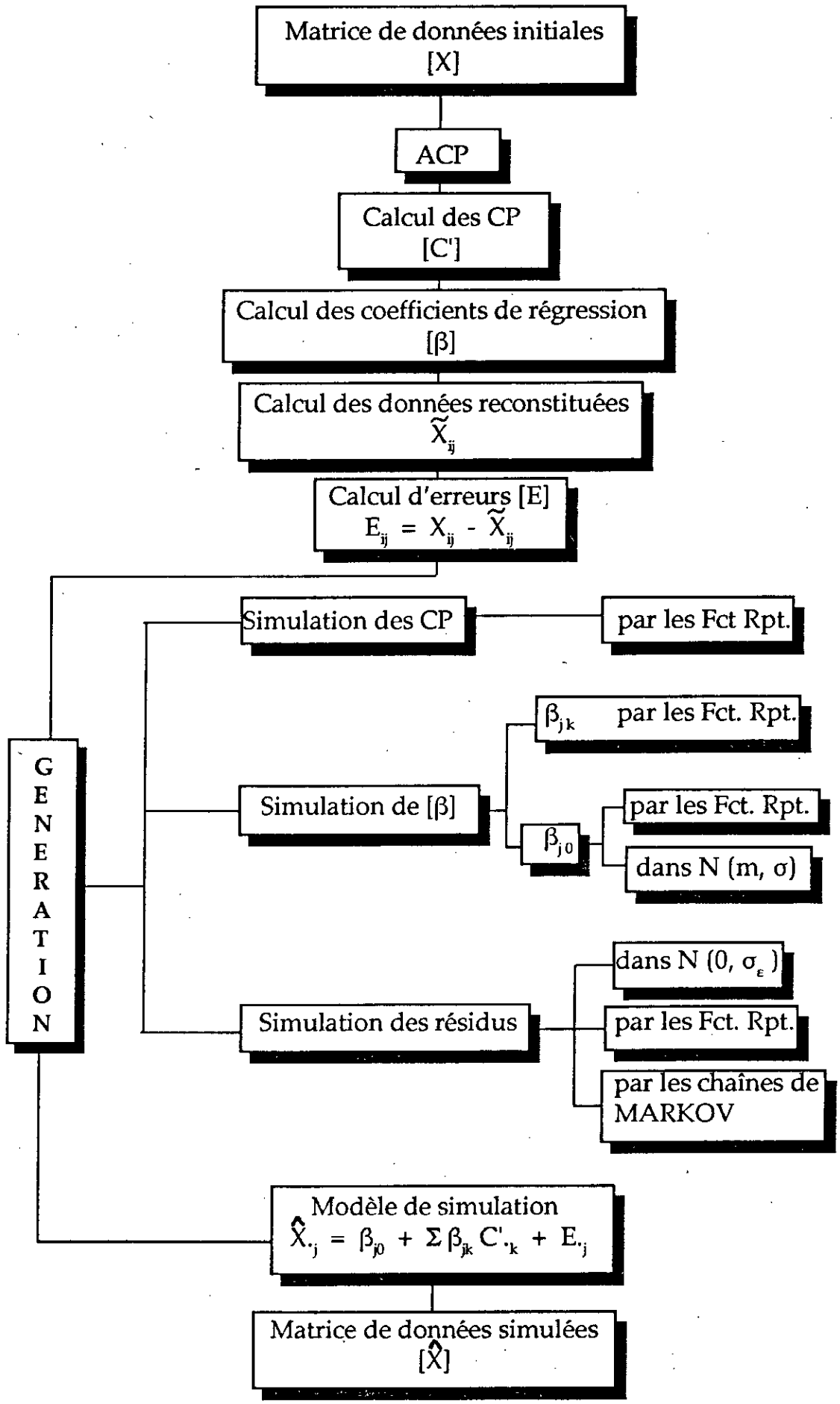
Pour la génération, un programme a été mis au point tenant compte de toutes les méthodes de simulation relatives à chacun des paramètres simulés.

Le menu présenté dans le programme est le suivant :

- **1 - Génération dans la loi normale N (0, 1)**
- **2 - Génération dans la loi normale N (0, sigma)**
- **3 - Génération par les fonctions de répartition**
- **4 - Génération par les chaînes de MARKOV**

On utilisera pour la génération uniforme des nombres aléatoires la fonction RANDOM dont disposent la majorité des calculateurs. Pour la simulation par les fonctions de répartition, la projection de la fréquence simulée sur cette dernière ne peut se faire si on ne dispose pas de l'expression analytique de la fonction elle-même ; pour cela on a fait appel aux méthodes de lissage et en particulier celle du **CUBIC SPLINE.**

IV . 4.1. ORGANIGRAMME DE SIMULATION



IV.4.2. LISSAGE PAR LE CUBIC SPLINE

Le **CUBIC SPLINE** est une méthode de lissage par intervalles qui utilise un polynôme du 3ème degré, c'est l'une des techniques les plus couramment utilisées.

Un polynôme du 3ème degré possède quatre constantes lui conférant une flexibilité suffisante pour assurer non seulement la continuité de la dérivée première sur l'intervalle, mais aussi la continuité de la dérivée seconde sur le même intervalle.

Soit une fonction f à approximer, définie sur l'intervalle [a, b] et un ensemble de nombres dits noeuds tel que :

$$a = x_0 < x_1 < x_N = b$$

On désigne par S la fonction d'interpolation **SPLINE** satisfaisant les conditions suivantes :

1°) S est un polynome du 3ème degré noté S_j dans le sous- intervalle

$[X_j , X_{j+1}]$ $\forall j$ j = 0 à N-1

2°) $S(X_j) = f(X_j)$ $\forall j$ j = 0 à N

3°) $S_{j+1}(X_{j+1}) = S_j(X_{j+1})$ $\forall j$ j = 0 à N-2

4°) $S'_{j+1}(X_{j+1}) = S'_j(X_{j+1})$ $\forall j$ j = 0 à N-2

5°) $S''_{j+1}(X_{j+1}) = S''_j(X_{j+1})$ $\forall j$ j = 0 à N-2

6°) L'une des deux conditions limites suivantes doit être satisfaite :

- $S''(X_0) = S''(X_n) = 0$ extrémité libre
- $S'(X_0) = f'(X_0)$ et $S'(X_N) = f'(X_N)$

Quand la condition limite de l'extrémité libre est satisfaite, le **SPLINE** est dit **SPLINE NATUREL** et le graphe résultant est une courbe flexible forcée de passer par tous les points

$$\{ (X_0 , f(X_0)) ; ((X_1 , f(X_1)) ;.....((X_N . f(X_N))\}$$

Pour construire la fonction d'interpolation du **CUBIC SPLINE** par une fonction f donnée, les conditions citées auparavant peuvent être appliquées au polynôme de degré trois

$$S_j(X) = a_j + b_j(X - X_j) + c_j(X - X_j)^2 + d_j(X - X_j)^3$$

$$\forall j \quad j = 0 \text{ à } N-1$$

IV.5. SIMULATION DES PHENOMENES CYCLIQUE ET ALEATOIRE

IV.5.1. INTRODUCTION

Pour pouvoir appliquer sans distinction le modèle de simulation élaboré, dans le domaine de l'hydrologie, on considère les deux types de phénomènes : **aléatoire** et **cyclique**.

On choisit deux exemples d'application : **les apports** au niveau du barrage de *BENI BAHDEL* pour le premier cas, **le graphe (IV.a)** fait ressortir le caractère aléatoire du phénomène, et **les E.T.P.** dans le second dont **le graphe (IV.b)** montre l'aspect cyclique.

IV.5.2. PHENOMENE ALEATOIRE

IV.5.2.1. Données utilisées

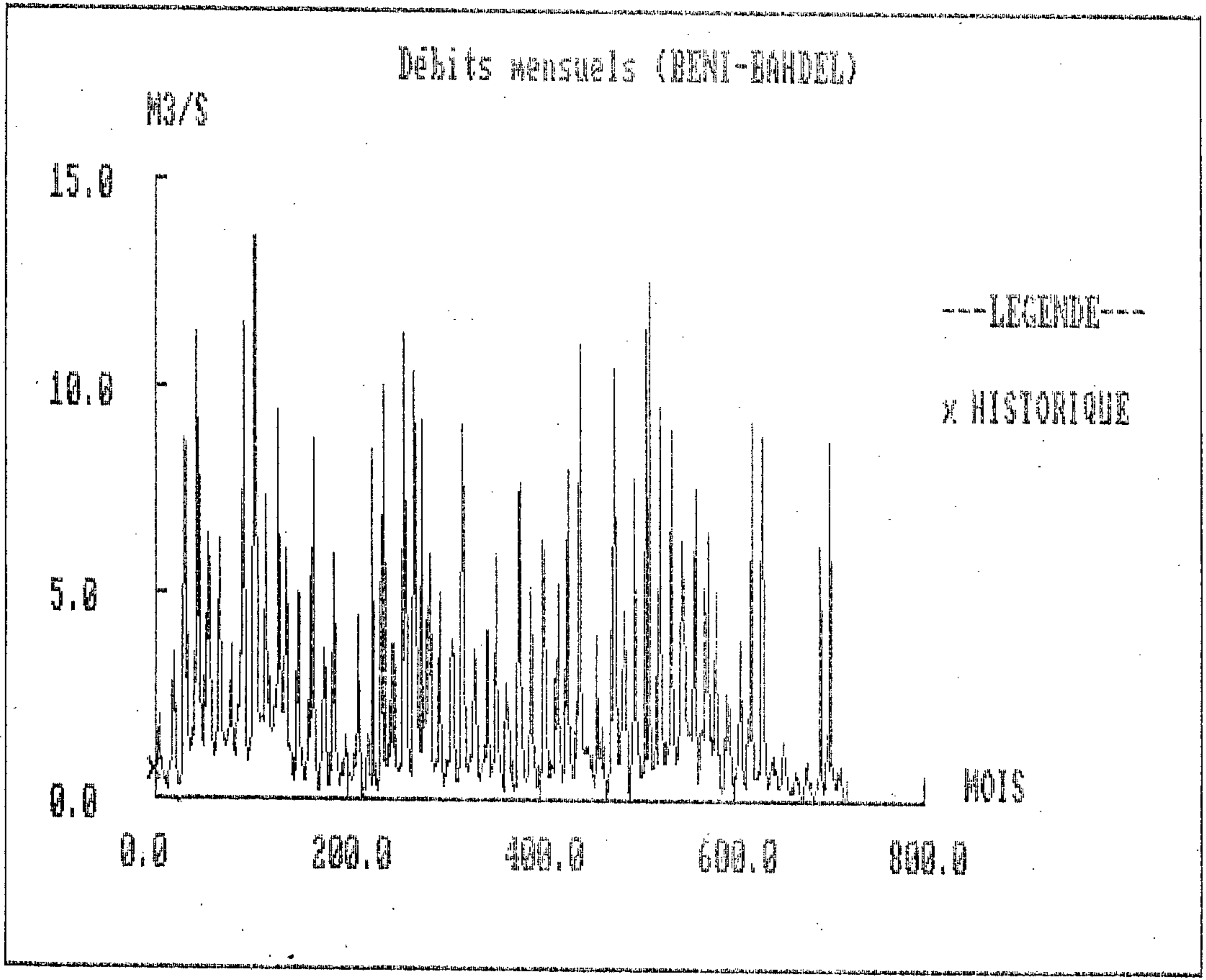
On se propose d'utiliser les apports au niveau du Barrage de *"BENI BAHDEL"* (wilaya de Tlemcen). Pour cela, on dispose des données - tronquées de trois années (1945, 1946 et 1947) manquantes - pour la période allant de 1925 à 1988. On utilisera donc 60 années d'observations. Les calculs seront faits avec la série de SEPTEMBRE 1925 à AOUT 1989 (donc sur l'année hydrologique).

La matrice [X] des données de base est calculée en prenant le cumul des débits journaliers pour les jours du mois considéré, qu'on divise par l'équivalent en secondes du même mois ; ainsi on obtient l'apport moyen mensuel en [m^3/s]. Les paramètres statistiques sont résumés dans le **tableau (IV.1)**, la matrice [60 x 12] est centrée réduite à l'aide de **l'équation (III.2.3.)** pour obtenir la matrice [Y]. Les variables ainsi transformées sont homogènes quant à leurs moyennes et leurs dispersions. En effectuant une A.C.P. *normée*, on obtient le **tableau (IV.2.)** qui donne les valeurs propres ainsi que le pourcentage cumulé. L'étude de ce tableau montre l'importance décroissante en terme de variance expliquée des composantes.

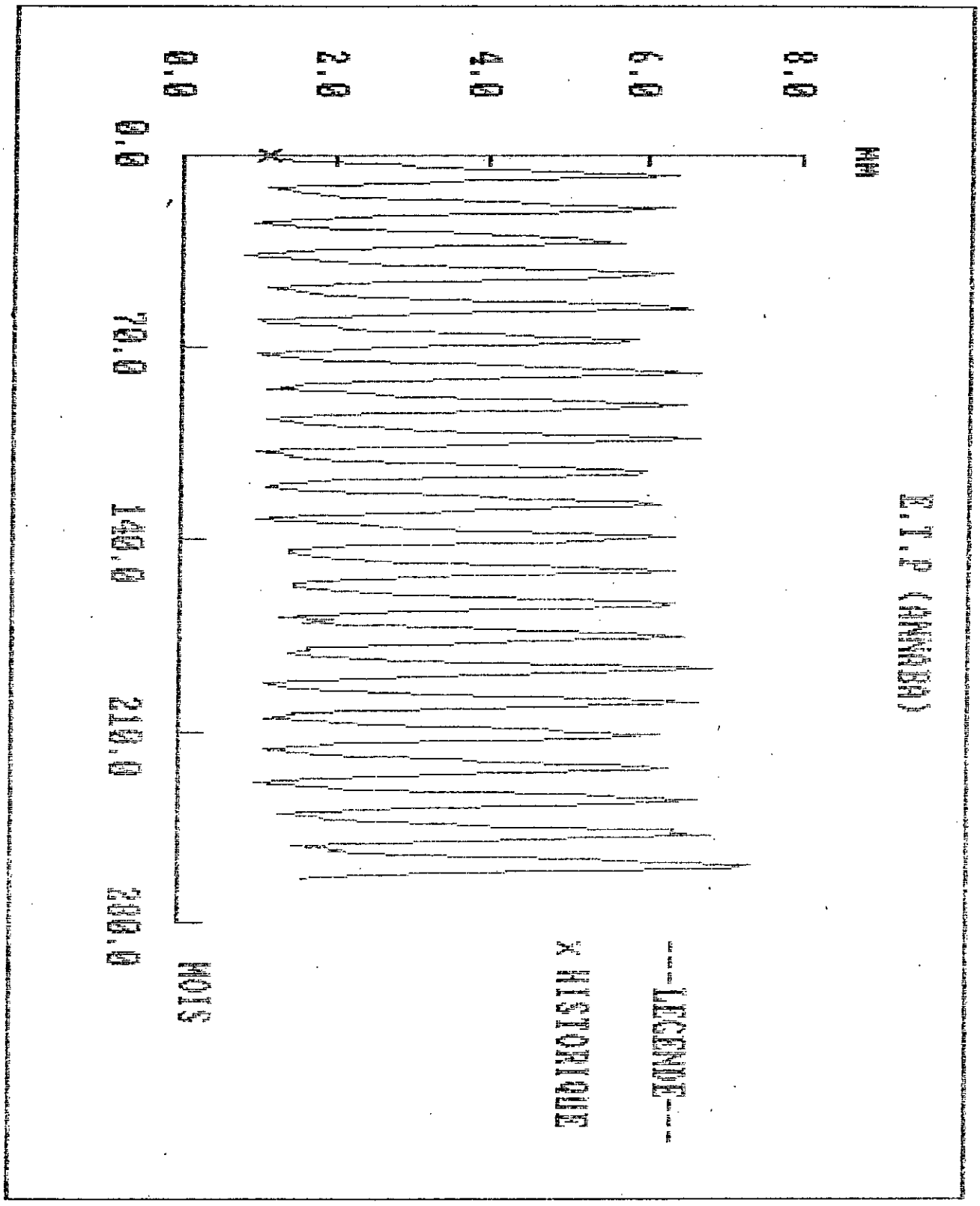
IV. 5.2.2. Choix du nombre de composantes

La **figure (IV.1)** montre l'utilité de réduire le nombre de variables à considérer, le choix de celui-ci doit se faire en fonction du pourcentage de la variance expliquée.

Graphe IV a : Evolution des débits historiques

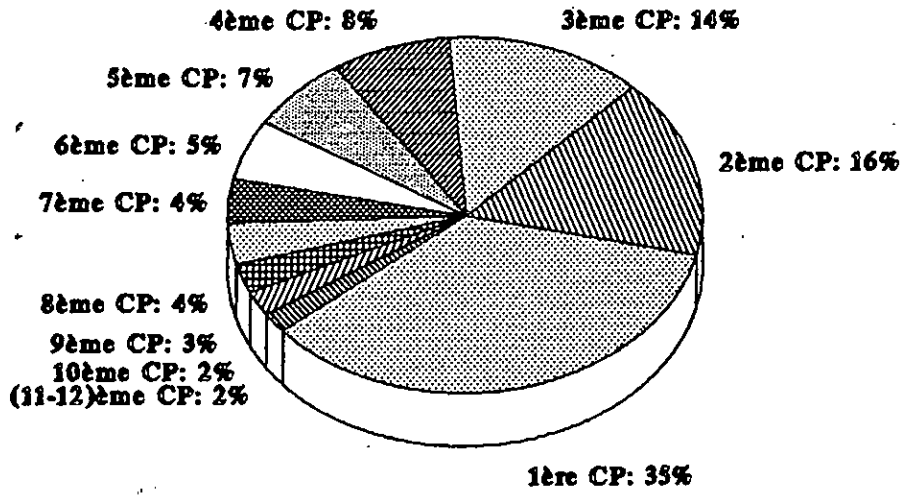


Graphe IV b : Evolution des E.T.P. historiques



Contribution de chaque CP à la Variance totale

Figure IV.1.a



Composantes Retenues

Figure IV.1.b

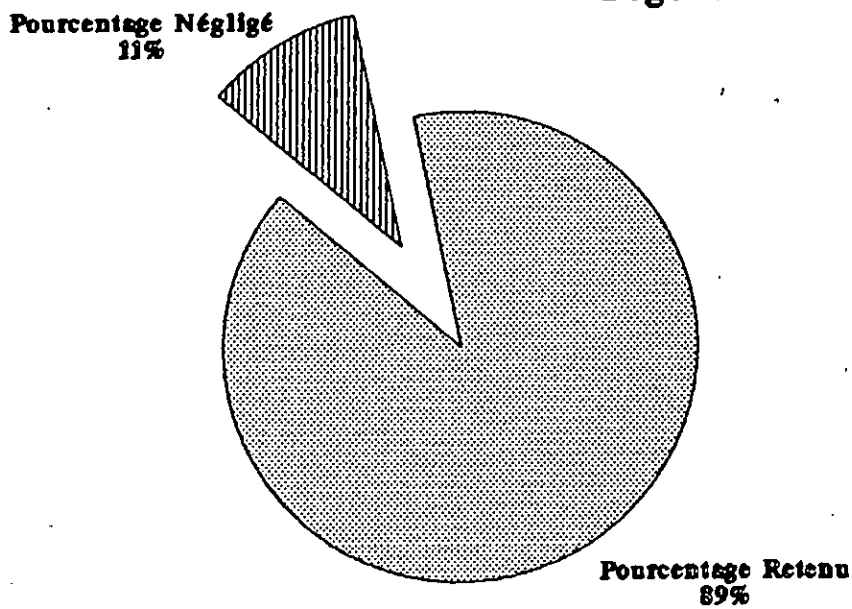


Figure IV.1 : Choix du nombre de CP

Si on prend **cinq (05) C.P.**, on a environ **80 % de variance expliquée** alors que pour **sept (07)** et **dix (10) C.P.** on a respectivement **89 %** et **98 %.**

Pour montrer l'utilité d'inclure une composante, il serait avantageux de voir sur quelles variables agit la composante. Dans ce tout, les graphiques de la variance expliquée de chaque composante sur toutes les variables sont très utiles.

La variance totale relative à une composante K est répartie sur les P variables de la manière suivante :

$$\lambda_k = \sum_{j=1}^{P} a^2_{jk} \lambda_k$$

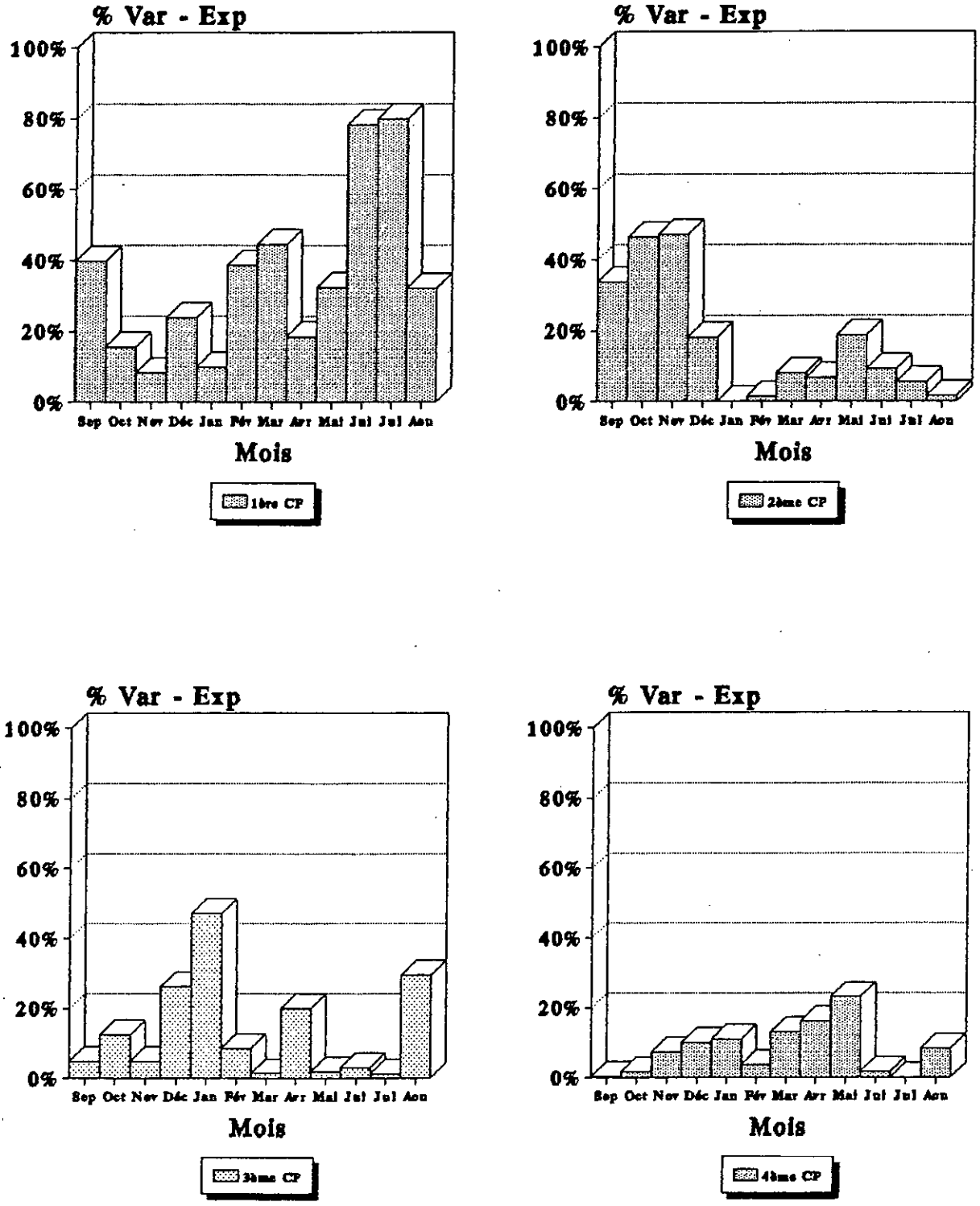
On appelle la quantité $a^2_{jk} \lambda_k$ le pourcentage de variance expliquée de la variable j par la $K^{ième}$ CP.

La **figure (IV.2)** donne les résultats pour les 12 composantes. On constate l'influence de la première composante sur les mois de JUIN et JUILLET et celle de la deuxième composante sur les mois d'OCTOBRE et NOVEMBRE. Donc, même les composantes d'ordre élevé n'expliquent que quelques variables.

Variable / Paramètres statistiques	1	2	3	4	5	6	7	8	9	10	11	12
Moyenne (m^3/s)	0,82	1,044	1,308	2,083	3,461	3,917	4,472	4,065	2,688	1,32	0,838	0,846
Ecart type (m^3/s)	0,483	0,660	0,919	1,684	2,879	3,568	4,477	5,821	2,535	0,850	0,549	1,182

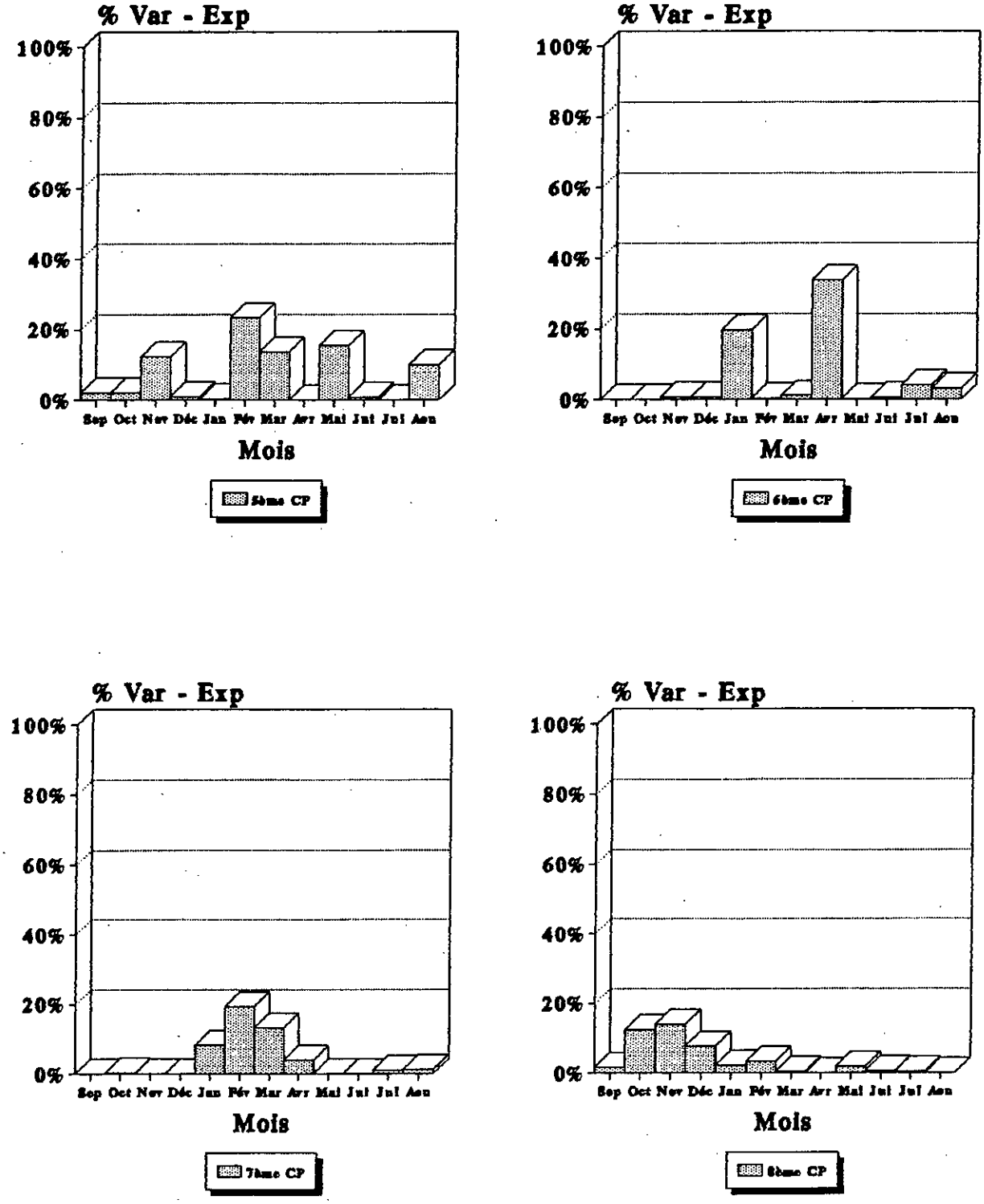
Tableau IV.1 - Paramètres statistiques de la série des apports du Barrage de BENI BAHDEL

Figure IV.2:
Pourcentage de Variance relative à chaque variable



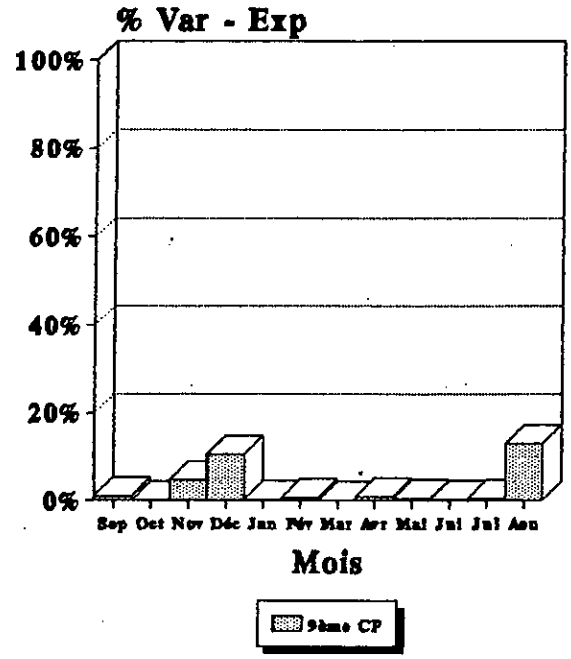
(1/3)

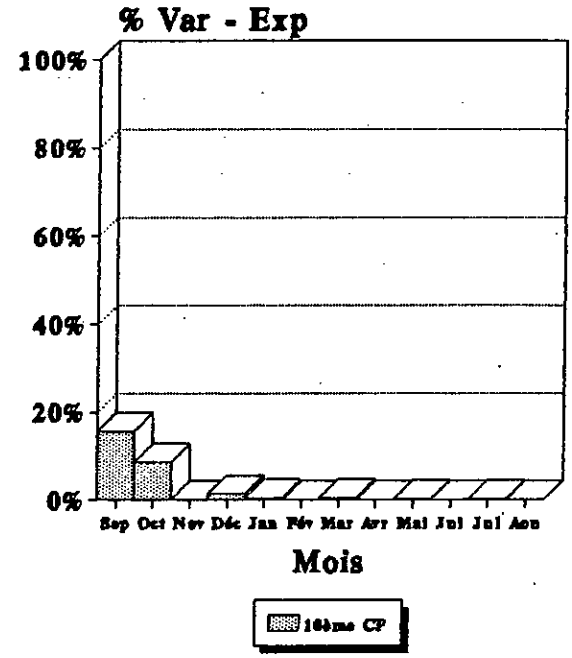
Figure IV.2:
Pourcentage de Variance relative à chaque variable

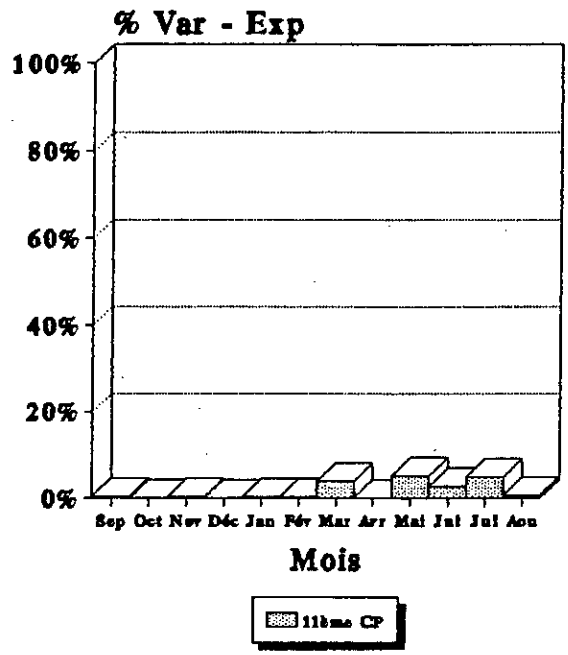


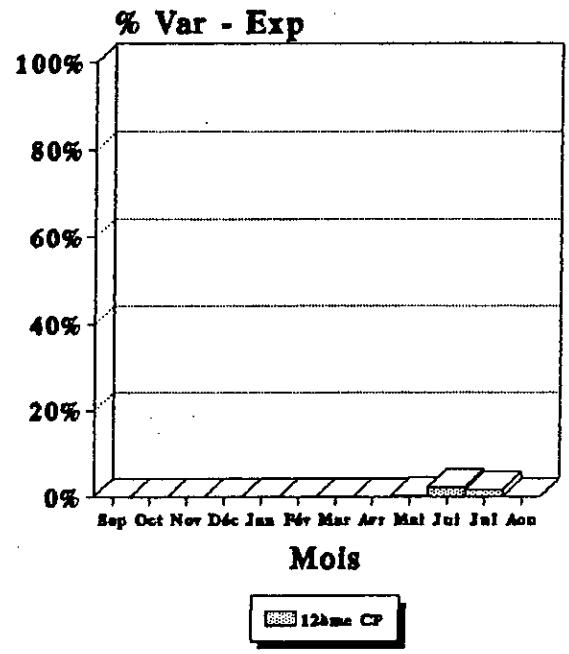
(2/3)

Figure IV.2:
Pourcentage de Variance relative à chaque variable









(3 / 3)

Composante N°	Valeur propre	% de variance expliquée	Pourcentage cumulé
1	4,222	35,18	35,18
2	1,974	16,45	51,63
3	1,617	13,48	65,11
4	0,966	08,05	73,16
5	0,815	06,79	79,95
6	0,651	05,43	85,38
7	0,476	03,96	89,34
8	0,455	03,79	93,14
9	0,324	02,70	95,84
10	0,272	02,27	98,11
11	0,185	01,54	99,65
12	0,042	00,35	100,00

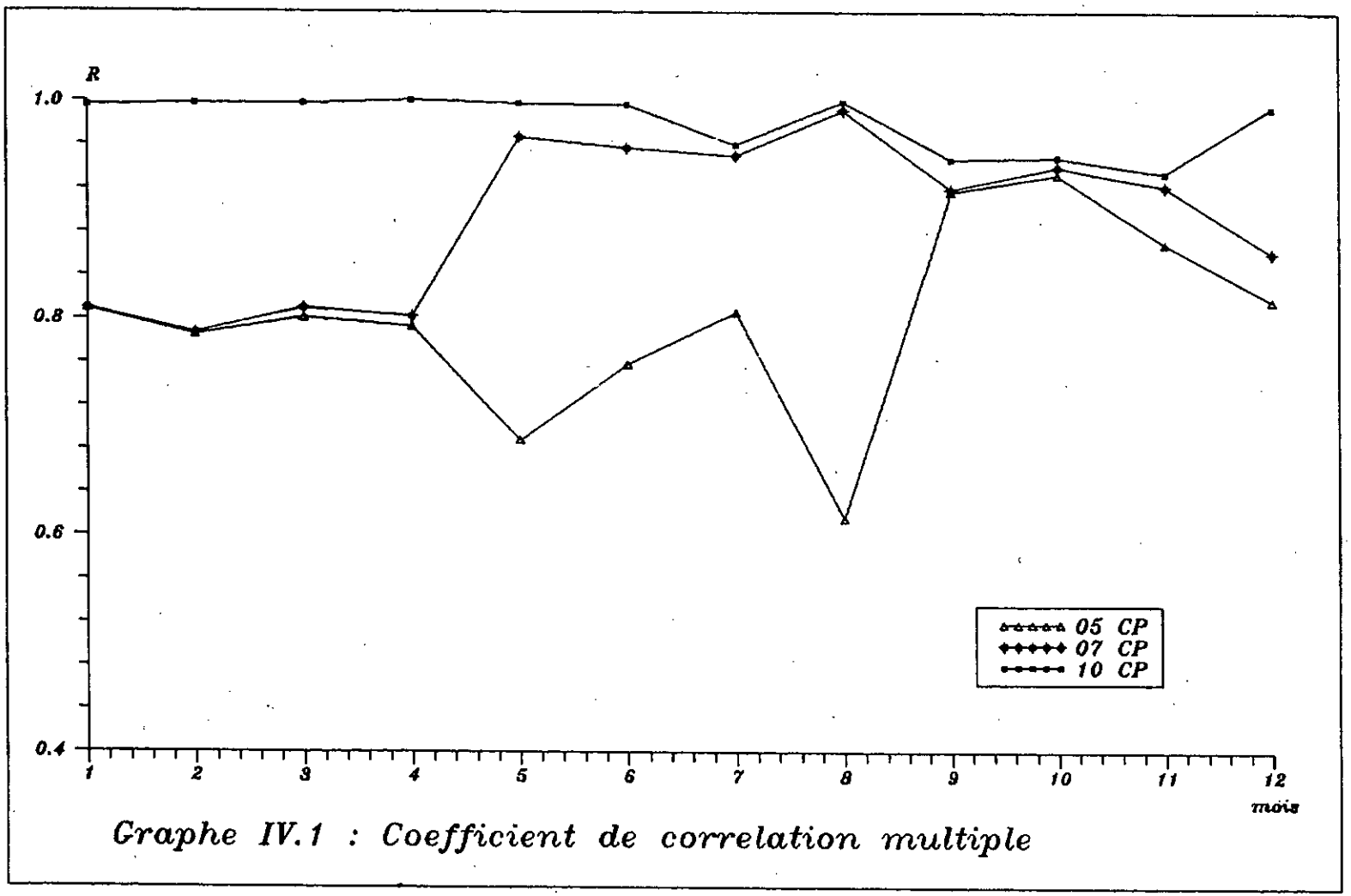
Tableau IV.2 - Pourcentage de variance expliquée

Le **tableau (IV.3)** donne les coefficients de corrélation multiples entre chaque mois et les 5, 7 et 10 premières C.P., il permet de voir l'importance du terme résiduel dans chaque cas.

Le **graphe (IV.1)** représente les résultats du **tableau (IV.3)** et permet de voir rapidement quelles variables sont moins bien expliquées par les 5, 7 ou 10 C.P.

De ces deux dernièrs, on peut déterminer le nombre de composantes à utiliser pour la génération.

En effet, pour 05 C.P. le terme d'erreur est important surtout pour les mois de JANVIER, FEVRIER et AVRIL. Il est pratiquement nul pour 10 C.P., alors que pour 07 C.P., il est d'environ 20 % pour les mois de SEPTEMBRE à DECEMBRE.



Graphe IV.1 : Coefficient de correlation multiple

On prendra donc 07 C.P. pour la génération tout en tenant compte de l'erreur dans le terme résiduel .

Mois	5 composantes	7 composantes	10 composantes
1	0,809	0,810	0,996
2	0,785	0,787	0,998
3	0,801	0,810	0,998
4	0,793	0,802	1,000
5	0,688	0,967	0,998
6	0,758	0,957	0,997
7	0,806	0,950	0,960
8	0,615	0,992	1,000
9	0,917	0,920	0,947
10	0,933	0,940	0,949
11	0,870	0,922	0,934
12	0,817	0,861	0,994

Tableau IV.3 - Coefficients de corrélation multiples

IV. 5.2.3. Calcul des coefficients de régression

La relation entre les coefficients de régression de la variable j et la composante K est donnée par :

$$\beta_{jk} = COR\,(X_j, C'_k) \cdot \frac{\sigma_{xj}}{\sigma_{ck}}$$

Les composantes C'_k étant *normées*, leurs écarts types sont réduits à l'unité :

$$\sigma_{C'k} = 1$$

alors :

$$\beta_{jk} = COR(X_j, C'_k) \cdot \sigma_{xj}$$

Le **tableau (IV.4)** donne ces coefficients pour les 12 composantes. Le terme constant β_{i0} est la moyenne de la variable j.

Les composantes étant orthogonales, les coefficients de régression ne changent pas si l'on utilise trois (03), cinq (05) ou sept (07) C.P.

IV. 5.2.4. Reconstitution des débits

On peut reconstituer les débits observés, au terme résiduel près, par le modèle suivant :

$$\tilde{X}_{ni} = \beta_{i0} + \sum_{j=1}^{Q} \beta_{ij} C'_{ni}$$

$\tilde{X}_{ni}$: débit, correspondant au mois i de l'année n, reconstituée avec les Q premières composantes.

β_{ii} : coefficient de régression entre la i ème régression variable et la j ème composante.

C'_{ni} : composante principale.

β_{i0} : constante égale à la moyenne de la variable.

La reconstitution des débits donne lieu à des valeurs négatives, or les données utilisées sont des apports qui ne peuvent être que positifs ou nuls. Pour cela on procède à la transformation des données initiales ($X_{ij} = \log X_{ij}$). Une fois la matrice reconstituée obtenue, on effectue la transformation inverse ($X_{ij} = EXP(X_{ij})$)

Le **graphe (IV.2)** montre les débits mensuels reconstitués avec cinq (05), sept 07 et dix (10) CP pour l'année 1988-1989. On remarque qu'avec (07) composantes on obtient une très bonne reconstitution.

IV.5.2.5. Calcul et étude des résidus

On définit le résidu comme étant la perte d'informations donnée par la différence entre le débit observé et le débit reconstitué à l'aide de Q composantes principales.

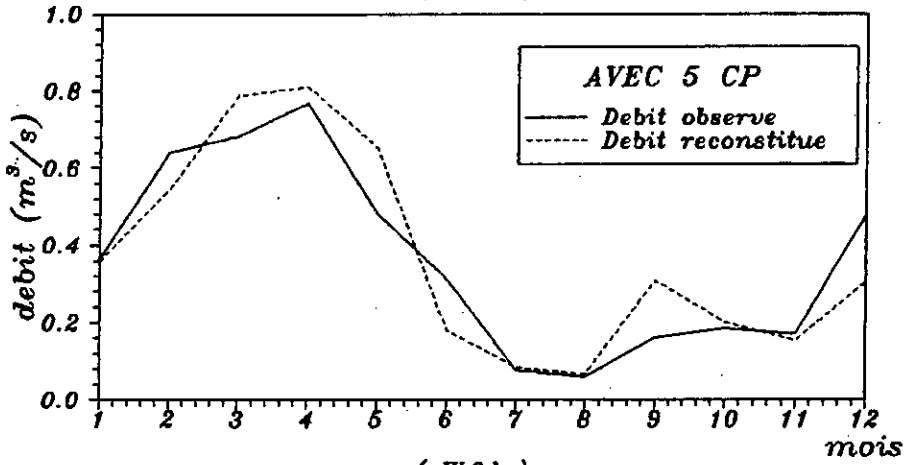
$$E_{.j} = X_{.j} - \tilde{X}_{.j}$$

Mois	$ß_{i0}$	$ß_{i1}$	$ß_{i2}$	$ß_{i3}$	$ß_{i4}$	$ß_{i5}$	$ß_{i6}$	$ß_{i7}$	$ß_{i8}$	$ß_{i9}$	$ß_{i10}$	$ß_{i11}$	$ß_{i12}$
1	0.823	0.305	0.281	-0.107	-0.022	-0.072	-0.014	0.002	-0.064	-0.051	-0.191	0.029	-0.005
2	1.044	0.261	0.450	-0.234	0.089	-0.096	0.009	0.032	-0.232	-0.017	0.193	-0.01	0.011
3	1.308	0.265	0.630	-0.206	0.249	0.322	0.086	0.018	0.344	0.198	0.021	0.044	-0.009
4	2.083	0.822	0.716	0.864	0.532	0.177	0.147	-0.661	0.470	-0.549	0.199	-0.034	0.029
5	3.461	0.910	-0.095	1.978	0.959	0.176	1.278	0.823	-0.449	0.147	0.176	-0.150	-0.009
6	3.917	2.219	-0.0435	1.045	0.696	1.723	-0.257	-1.573	-0.660	-0.271	0.025	0.161	-0.025
7	4.471	2.986	-1.276	-0.533	-0.616	1.652	-0.488	1.625	-0.309	-0.157	0.278	0.895	-0.054
8	4.065	2.509	-1.502	-2.602	-2.334	-0.175	3.382	-1.153	-0.023	0.506	0.041	-0.081	-0.081
9	2.688	1.443	-1.100	-0.350	1.220	-0.999	0.114	-0.094	0.357	-0.177	0.122	0.573	0.101
10	1.324	0.752	-0.260	-0.153	0.111	-0.075	-0.071	0.014	0.067	-0.036	0.032	-0.139	-0.128
11	0.838	0.492	-0.130	-0.063	0.008	0.019	-0.112	0.054	0.038	0.038	-0.025	-0.121	0.071
12	0.0846	0.671	0.0156	0.641	0.0343	-0.374	-0.207	-0.136	-0.044	0.427	0.041	0.090	-0.021

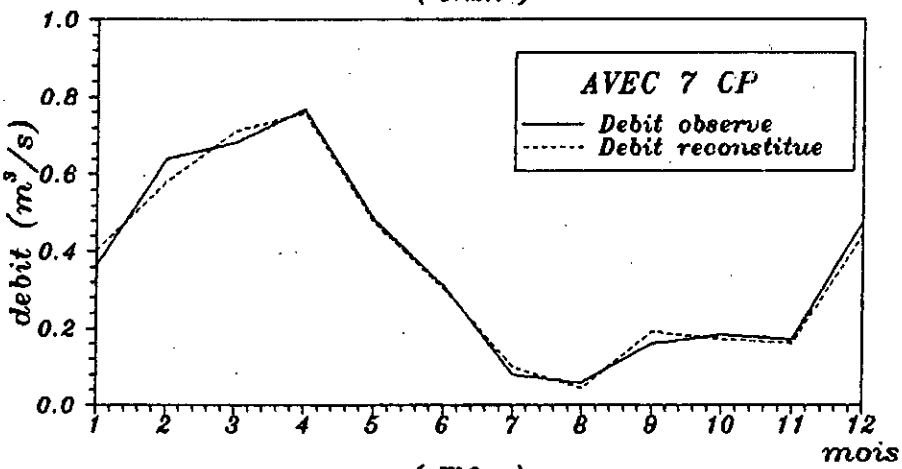
TABLEAU IV.4 : Coéfficients de régression

Graphe IV.2 : Reconstitution de l'annee 1988/1989

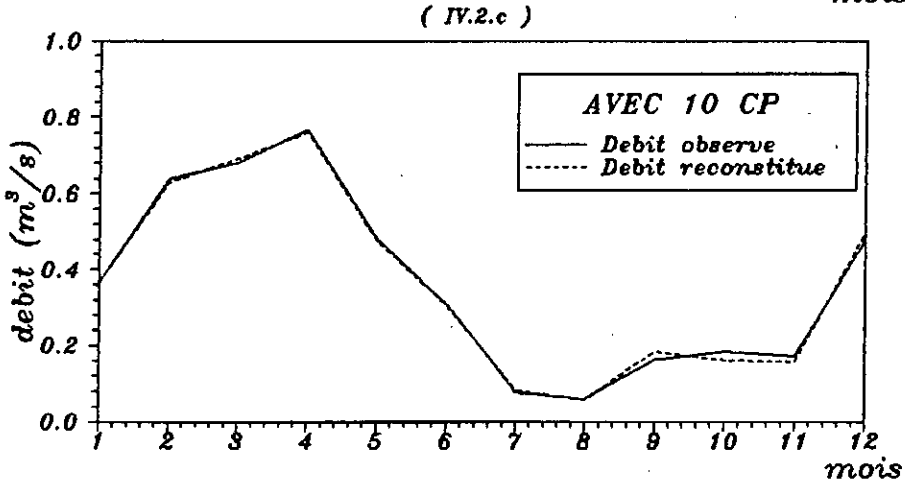
(IV.2.a)



(IV.2.b)



(IV.2.c)



On obtient ainsi une matrice dont les éléments sont les termes résiduels. Chaque colonne de cette matrice suit une loi normale de moyenne nulle et d'écart type σ_j (voir graphe IV.3).

$$\sigma_{ej} = [\, 1 - \sum_{l=1}^{P} a^2_{jl}\, \lambda_l \,]^{1/2}$$

IV. 5.2.6. Simulation des différents paramètres du modèle

Le modèle de simulation est constitué de trois (03) paramètres :

- **Les éléments de la matrice [β]**
  - **$\beta(j,1)$: moyenne inter-individus de la variable j**
  - **$\beta(j,k)_{j=2\,à\,P}$: coefficient de régression.**
- **les éléments de la matrice des CP [C']**
  - **C'_{ik} : projection de l'individu i sur la CP K**
- **Les éléments de la matrice [E]**
  - **ε_{ij} : terme résiduel.**

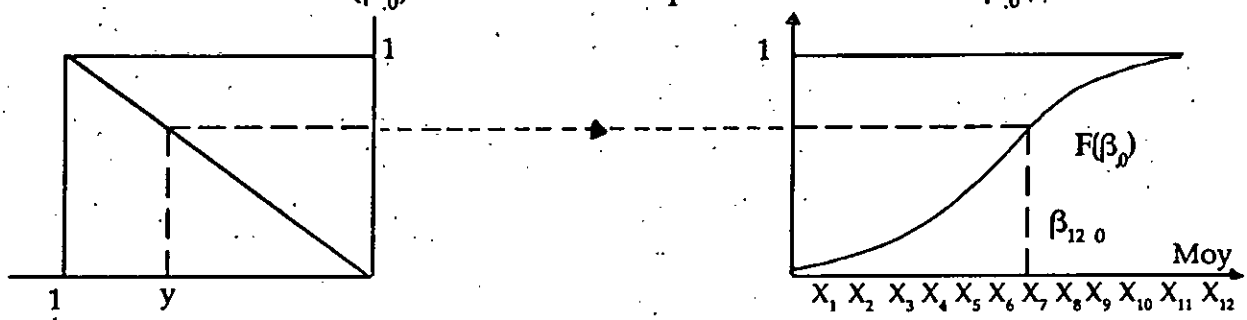
La qualité de la simulation dépendra obligatoirement de la simulation de chacun de ces paramètres.

A partir des différentes possibilités de génération offertes, on pourra établir un menu global des méthodes de simulation, qui fournira par la suite plusieurs combinaisons possibles et donc une variété de scénarios plausibles.

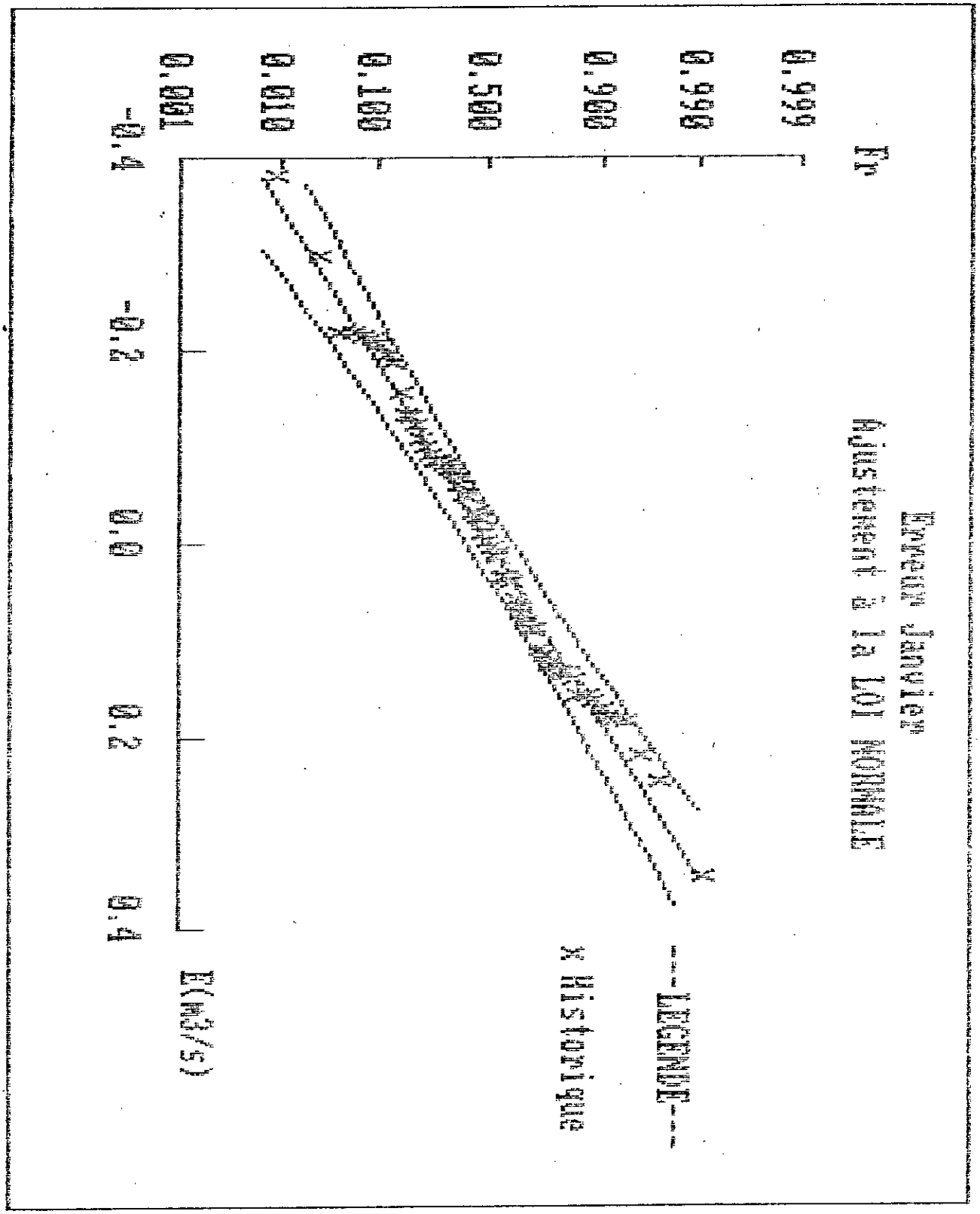
IV. 5.2.6.1. Simulation des éléments de [β]

La matrice [β] est constituée d'un premier vecteur β_0 dont les composantes sont les moyennes calculées sur les données historiques, les autres vecteurs de la matrice ont pour composantes les coefficients de régression entre variables et CP .

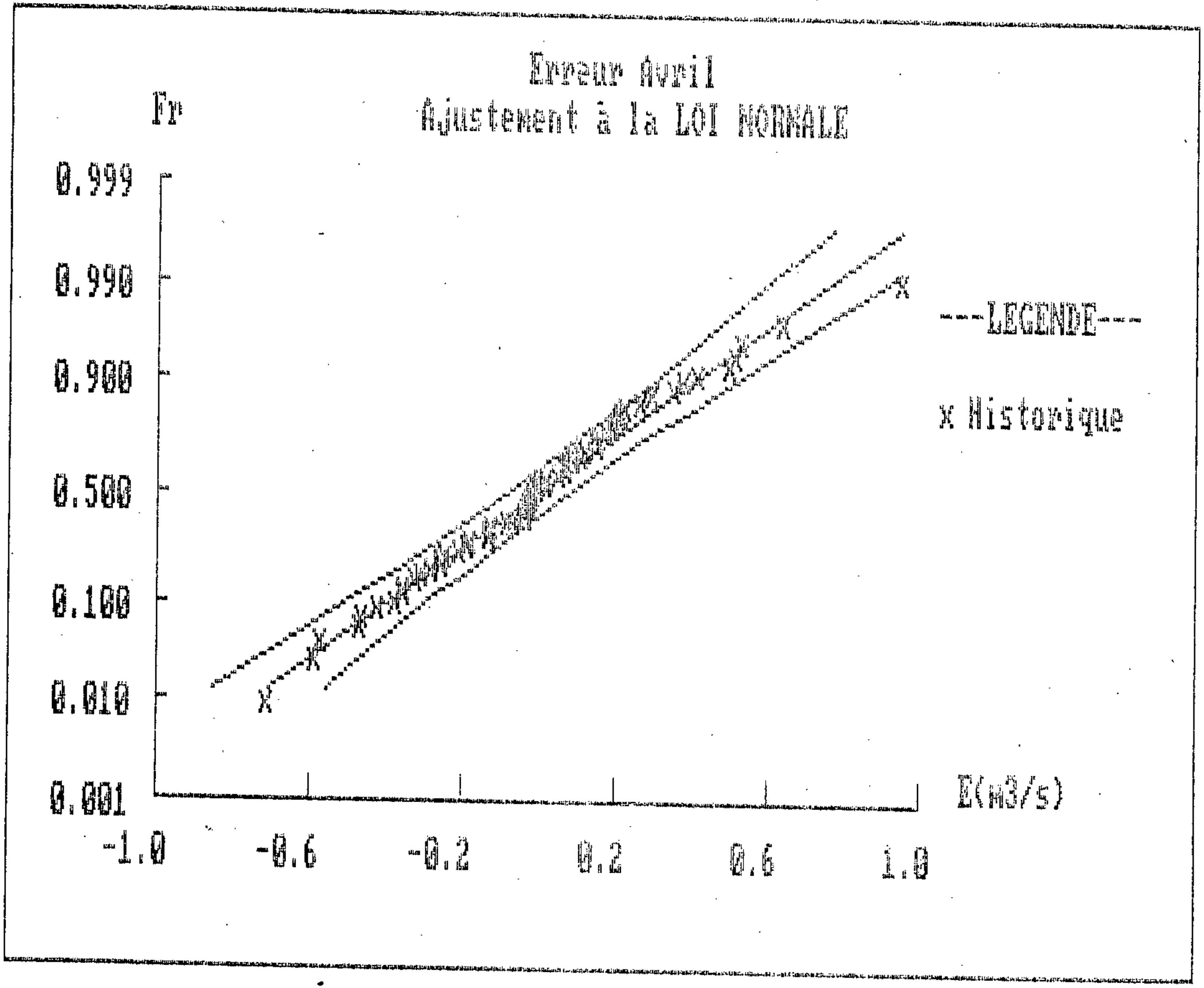
En déterminant la fonction de répartition de chaque vecteur (β_0 , β_1 ,..., β_P) on peut simuler dans ces dernières une nouvelle séquence de vecteur β_j avec j = 0 à P.
La simulation directe du vecteur β_0 dans la fonction de répartition engendre des résultats erronés. Soit F (β_0) la fonction de répartition du vecteur β_0 ;



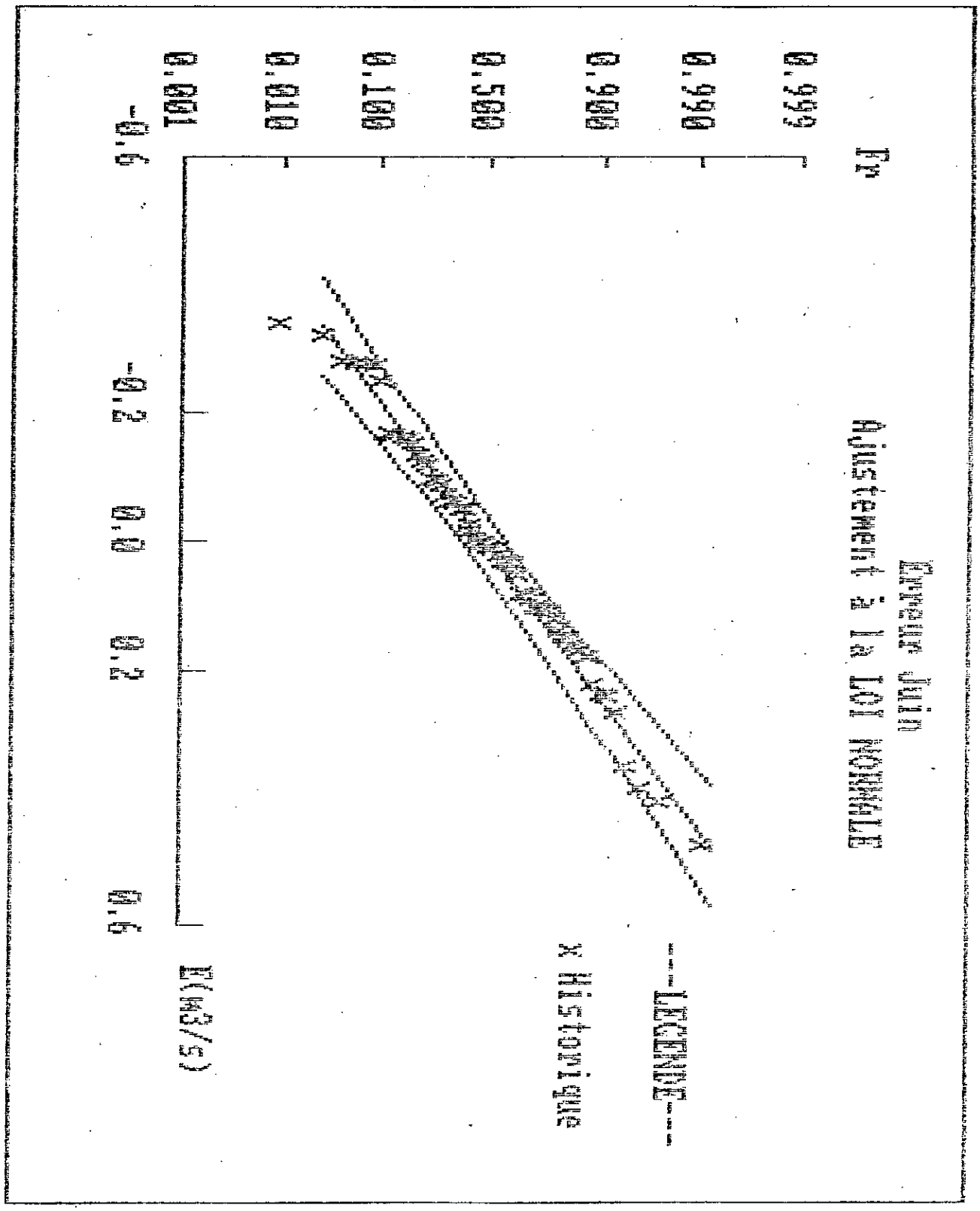
Graphe IV - 3. : Ajustements des résidus



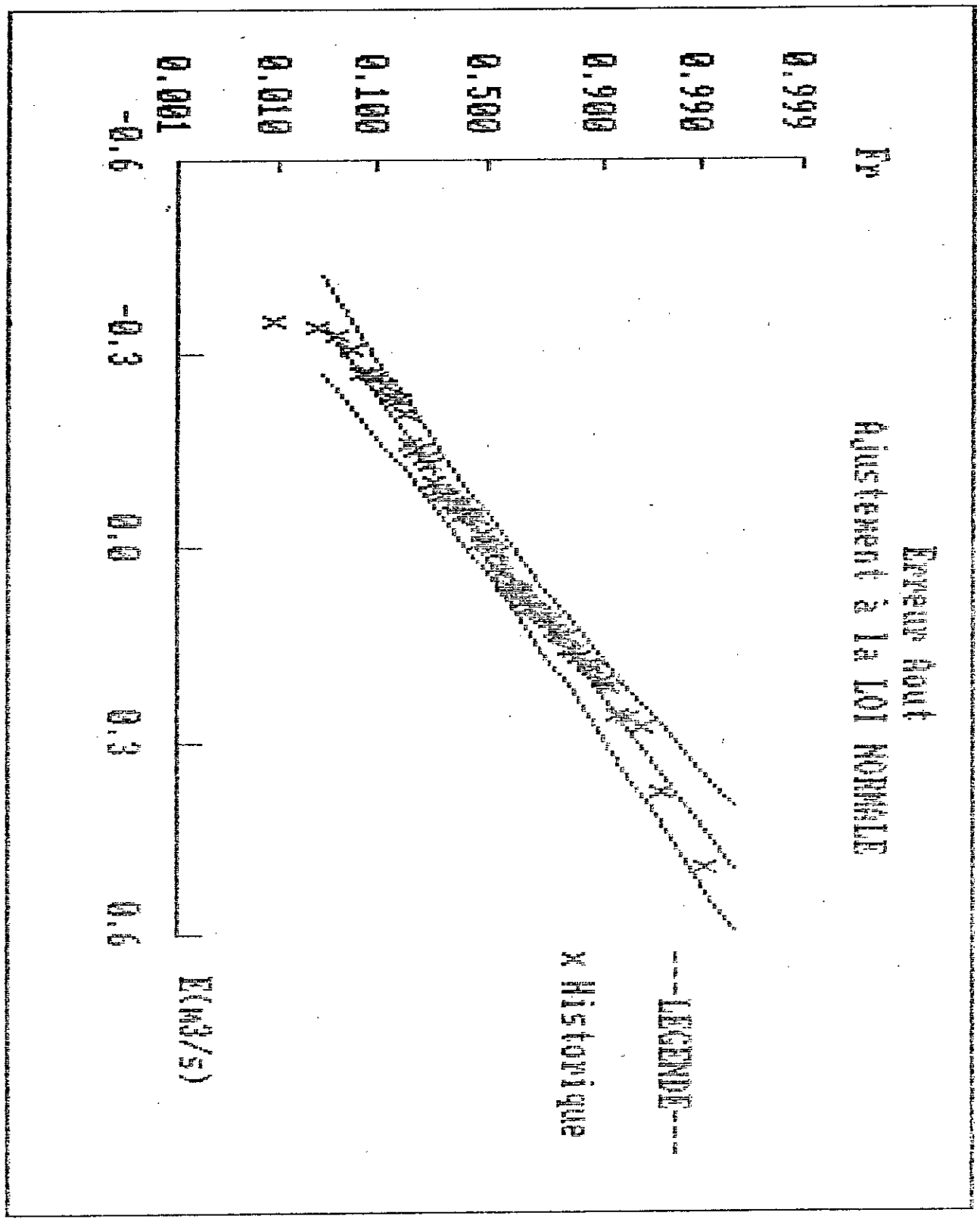
Graphe IV - 3. : Ajustements des résidus



Graphe IV - 3. : Ajustements des résidus



Graphe IV - 3. : Ajustements des résidus



Supposons qu'on veut simuler $\beta_{12\,0}$ (moyenne interannuelle du mois d'AOUT), on voit qu'avec un tirage aléatoire on peut tomber sur une moyenne qui représenterait plutôt un mois humide comme JANVIER ou FEVRIER. Ainsi la simulation du vecteur $\beta_{.0}$ dans sa fonction de répartition ne tient pas compte de l'effet saisonnier.

A cause de l'hétérogénéité dans la structure de la matrice [β], on ne peut faire subir le même traitement à toute la matrice. Pour cela, on propose une autre approche. On simule les vecteurs $\beta_{.j}$ dans leurs fonctions de répartition respectives comme convenu. En ce qui concerne le vecteur $\beta_{.0}$, sachant que chaque élément du vecteur $\beta_{.0}$ est une moyenne mensuelle interannuelle, on a proposé deux méthodes :

- **Par les fonctions de répartition**

On procèdera comme suit : pour simuler la moyenne interannuelle du mois j, on génère N fréquences qu'on projette dans la fonction de répartition du mois considéré, la moyenne de ces N projections sera la moyenne simulée.

- **Par les lois d'ajustement**

Les variables initiales suivent une loi **log normale** vu la transformation effectuée **(voir graphe IV.4.a).**

Quelle que soit la loi suivie la simulation est décrite au **(Chap IV § IV.3.2).**

On désignera par la suite :
La matrice [β] simulée par les lois d'ajustement, le cas où le vecteur $\beta_{.0}$ est simulé dans la loi suivie par les variables initiales (loi normale), et [β] simulé par les fonctions de répartition le cas où $\beta_{.0}$ est simulé dans les fonctions de répartition des variables initiales.

IV.5.2.6.2. Simulation des Composantes Principales [C']

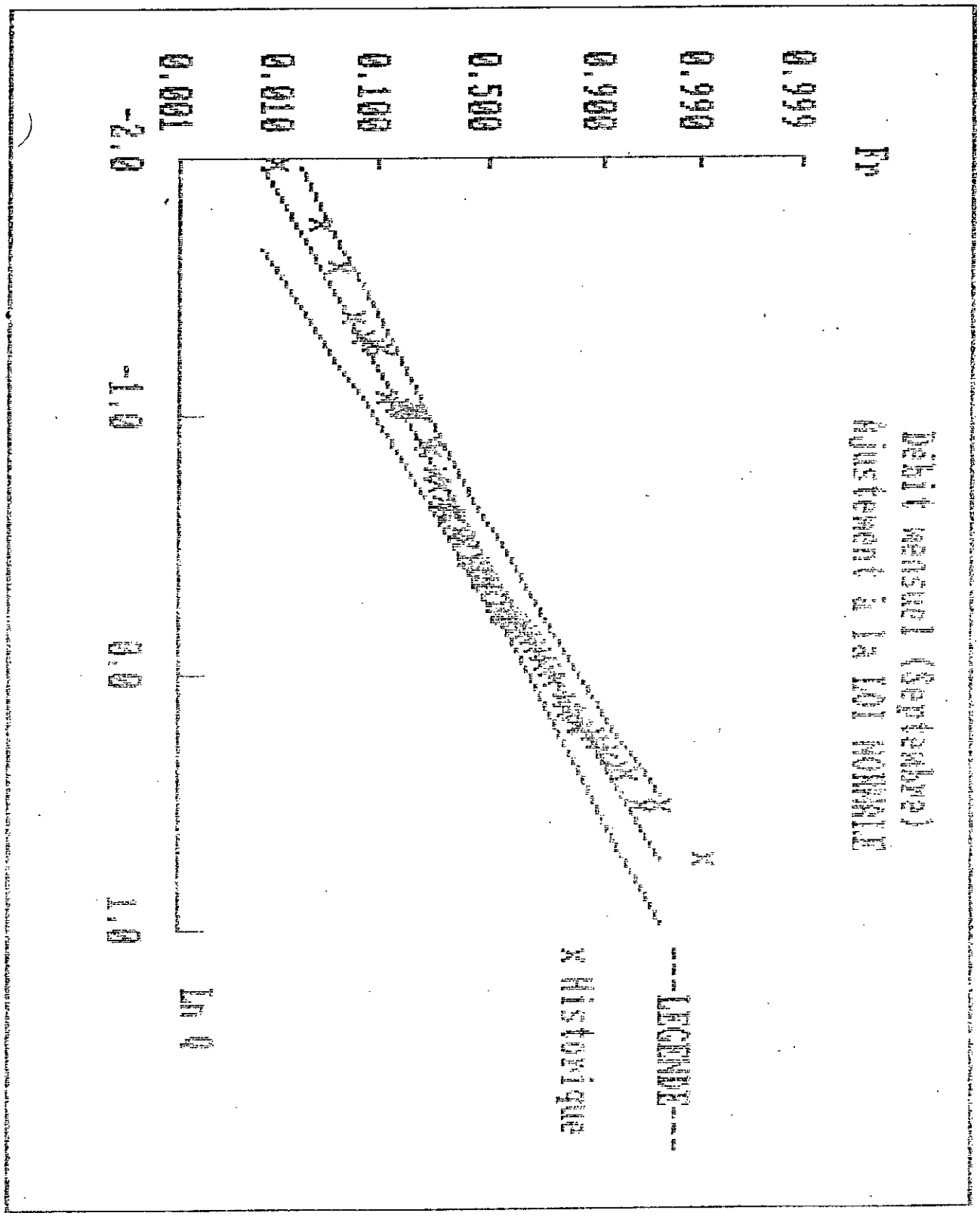
- **Par les fonctions de répartition**

Le procédé est le même, on détermine la fonction de répartition de chaque vecteur $C'_{.j}$ de la matrice [C'] des CP, dans laquelle on générera le nouveau vecteur $C'_{.j}$.

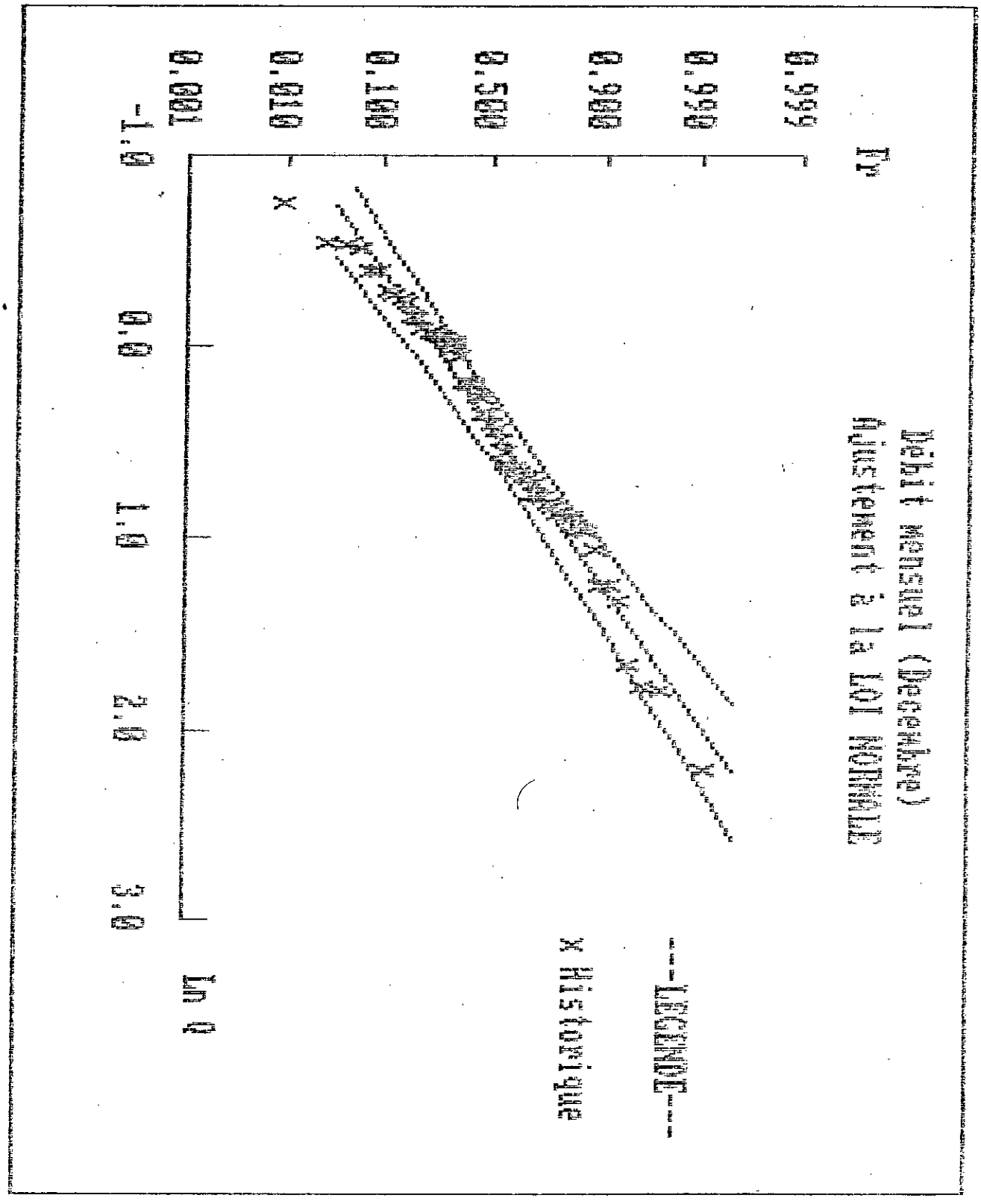
- **Par les lois d'ajustement**

Toute variable peut être simulée dans sa fonction de répartition ou par les chaînes de MARKOV, mais pour simuler dans une certaine loi statistique, il faut vérifier au préalable que cette loi s'ajuste bien à la variable en question.
Pour cela on a procédé à des essais d'ajustement qui ont révélé que les Composantes Principales $C'_{.j}$ suivent une loi normale de moyenne nulle et d'écart type unité (N (0,1)). **(Voir graphe IV.5).**

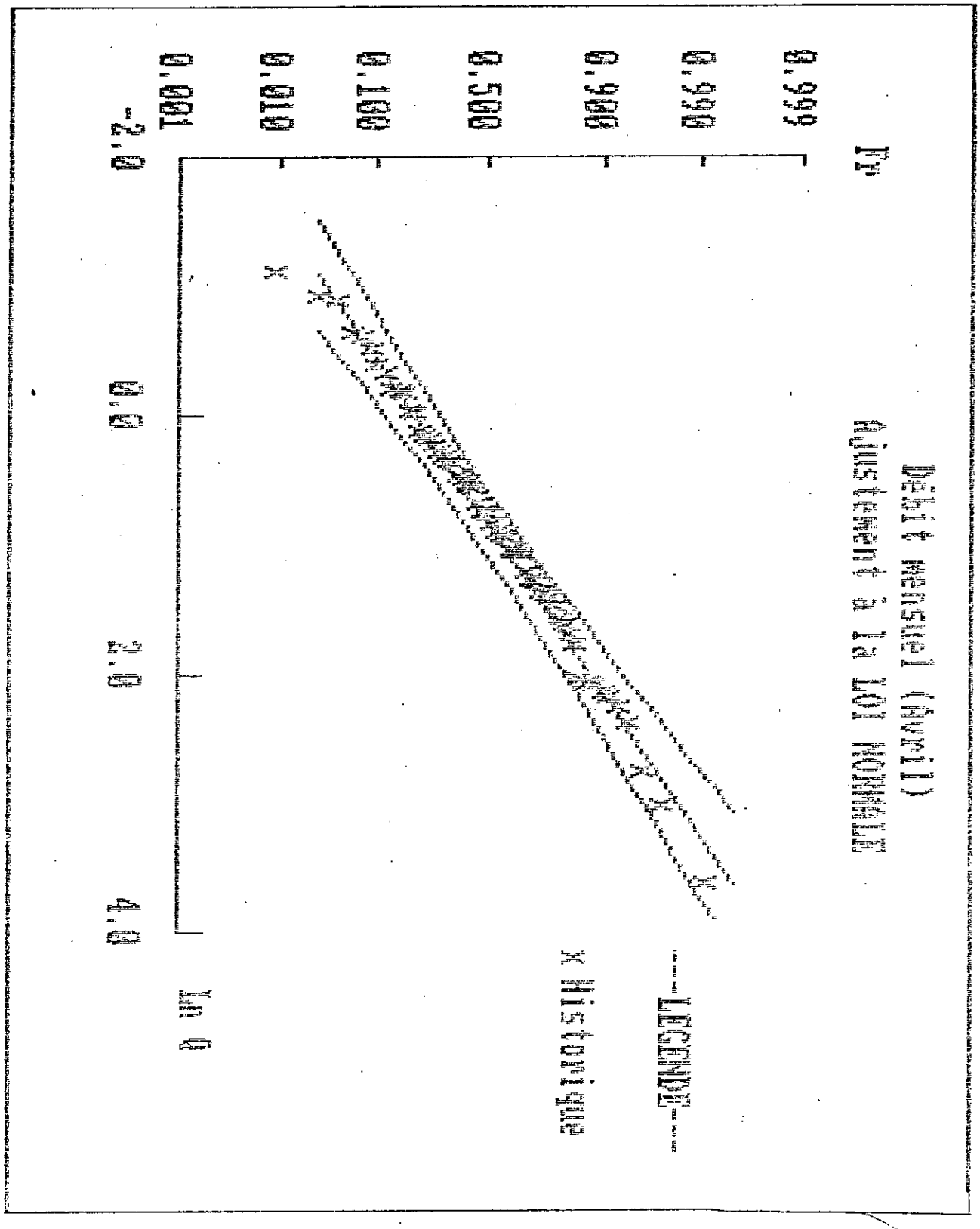
Graphe IV - 4.a : Ajustement des débits



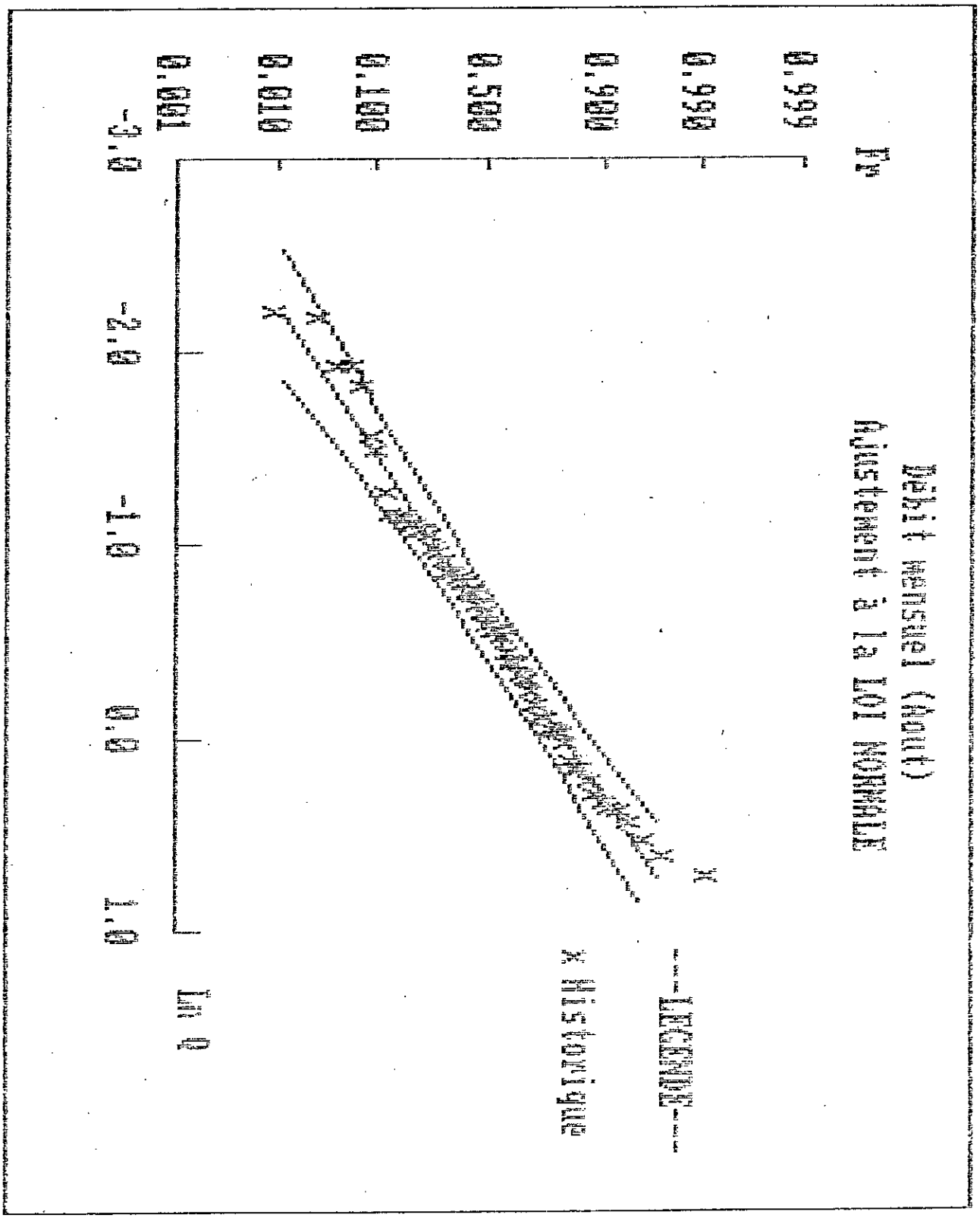
Graphe IV - 4.a : Ajustement des débits



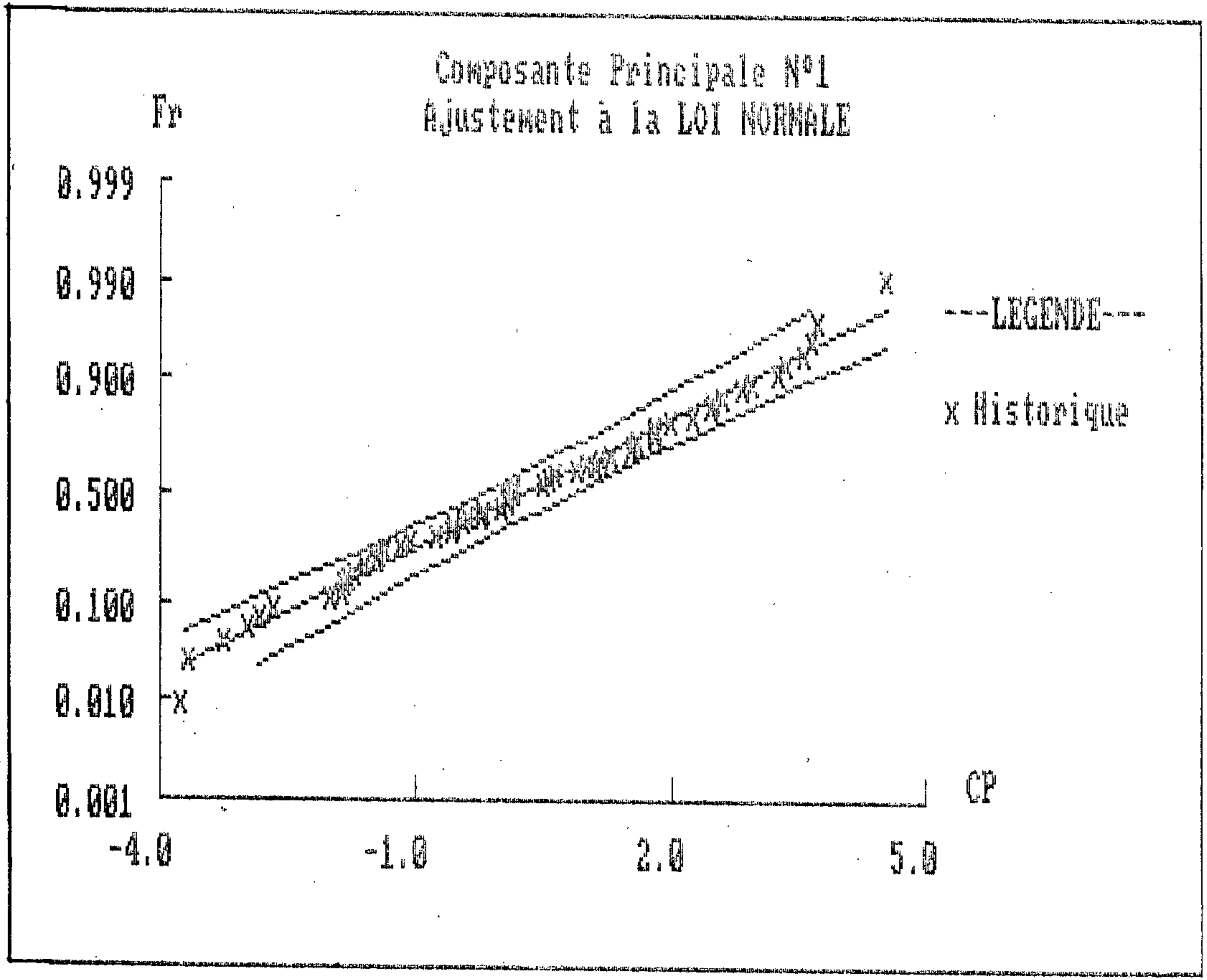
Graphe IV - 4.a : Ajustement des débits



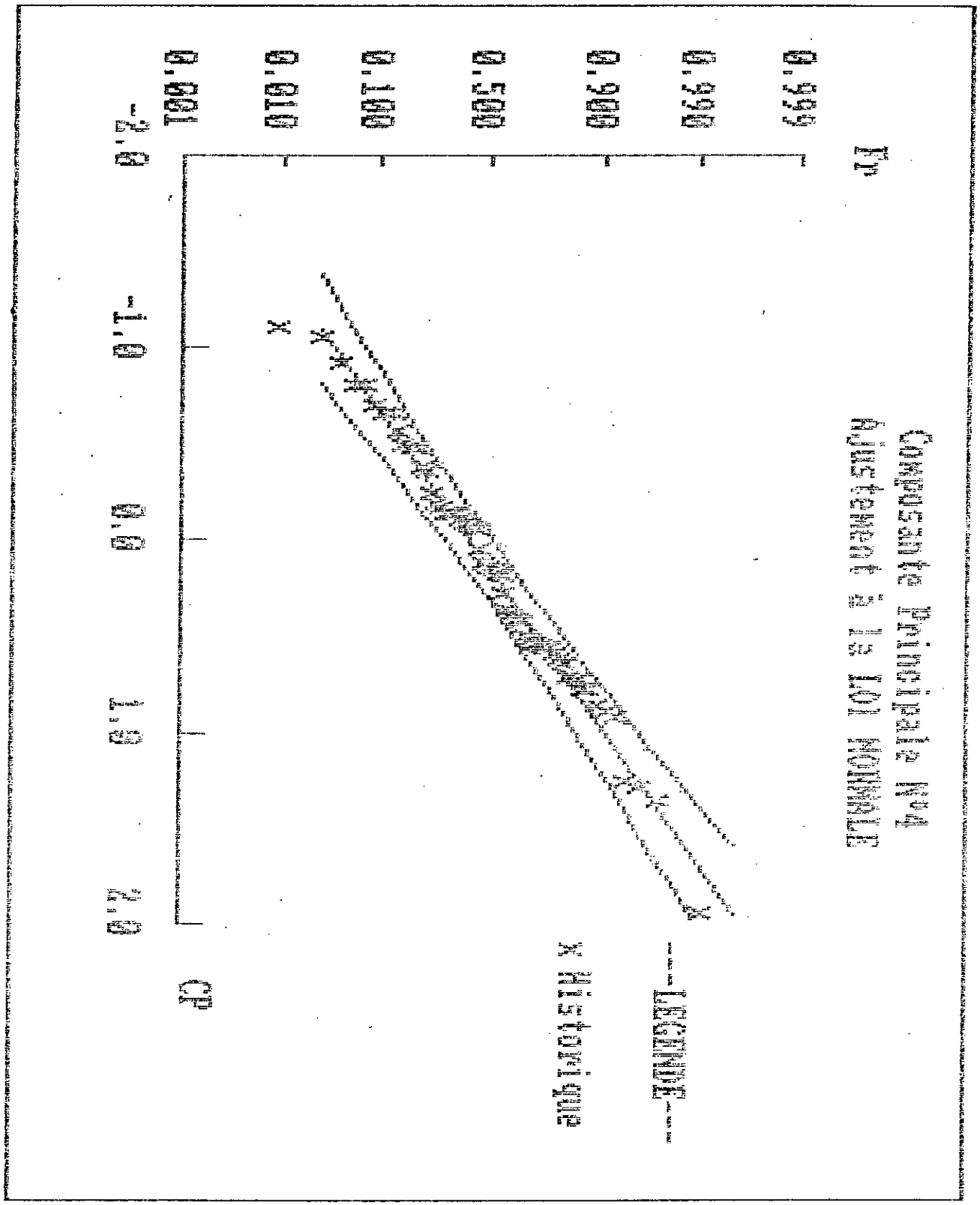
Graphe IV - 4.a : Ajustement des débits



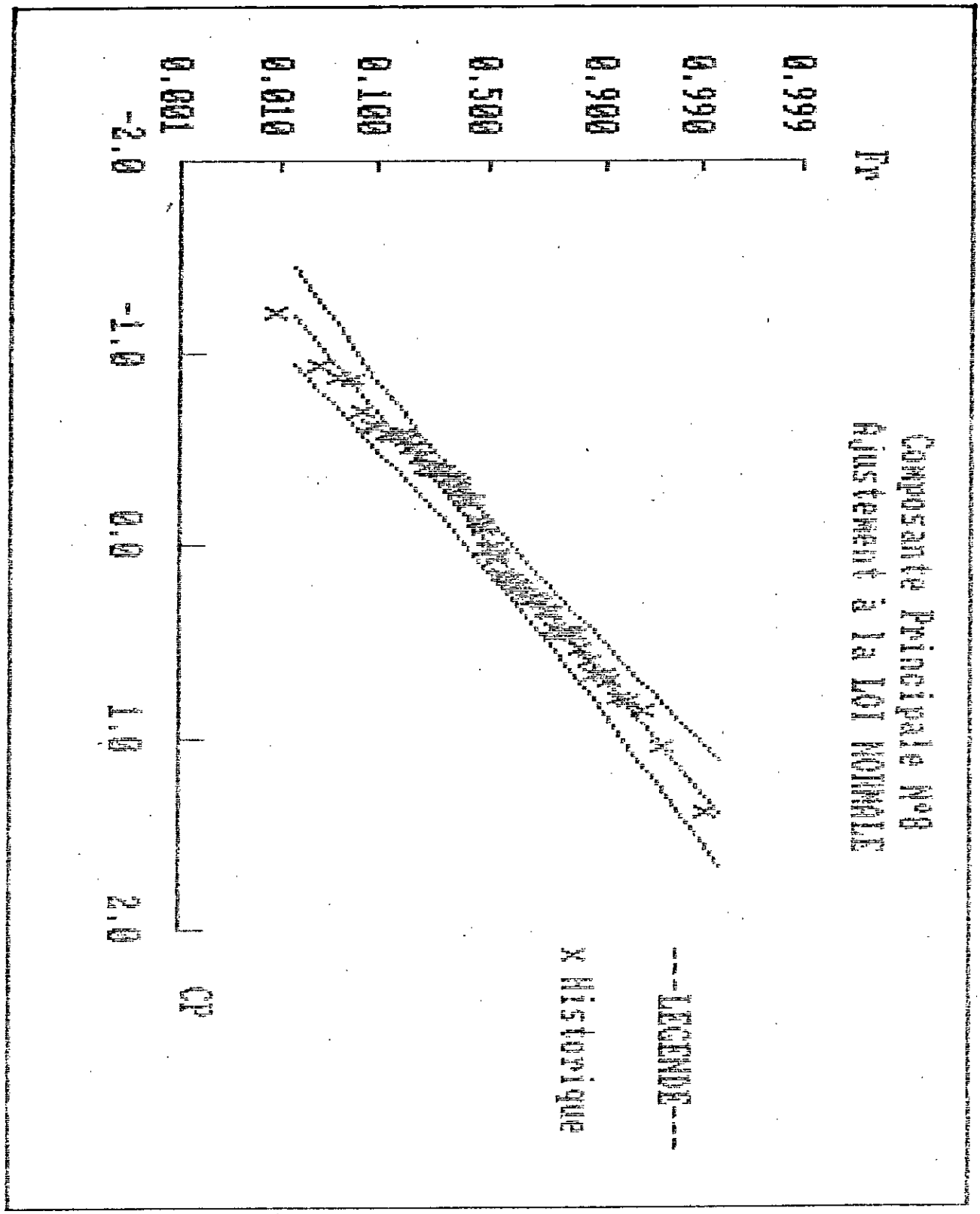
Graphe IV - 5. : Ajustement des C.P.



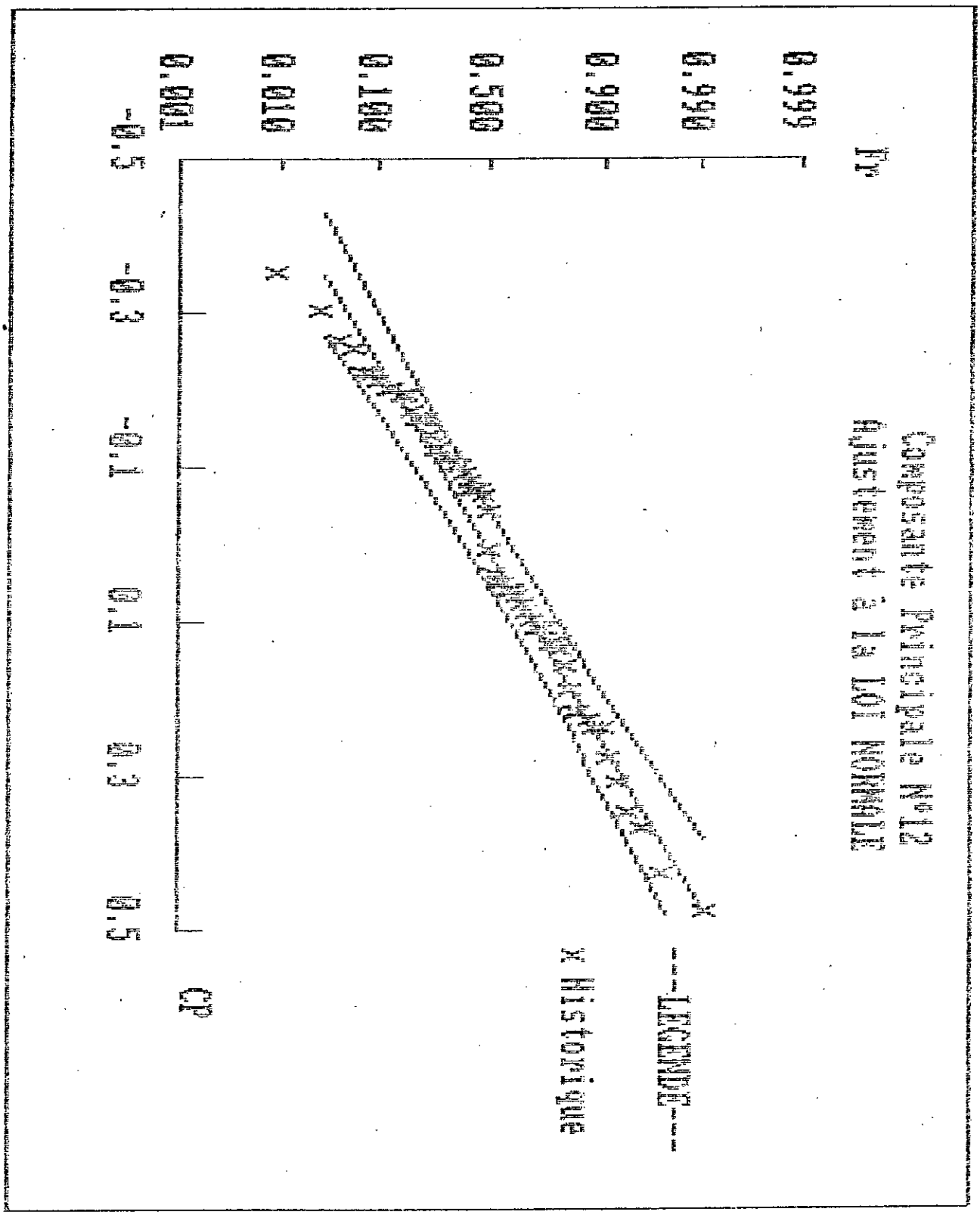
Graphe IV - 5. : Ajustement des C.P.



Graphe IV - 5. : Ajustement des C.P.



Graphe IV - 5. : Ajustement des C.P.



Donc les Composantes Principales peuvent être simulées de deux manières : par les fonctions de répartition ou dans une loi normale N (0,1).

IV.5.2.6.3. Simulation des résidus [E]

- **Par les fonctions de répartition**

De la même manière que pour [β] et [C'], on peut simuler les résidus par les fonctions de répartition, sauf que dans ce cas il n'est pas nécessaire de considérer chaque colonne séparément, on transforme la matrice [E] de dimension (N,P) en un vecteur de dimension (N x P, 1) dont on détermine la fonction de répartition.

- **Par les chaînes de MARKOV**

Cette méthode est utilisée surtout pour affiner la simulation ; elle permet de respecter la structure de la matrice [E] historique en maintenant par exemple le pourcentage de valeurs appartenant au premier état et le pourcentage appartenant au second.

- **Par les lois d'ajustement**

Les résidus calculés au (**IV.5.5**) sont de moyenne nulle et d'écart type σ_ε ; plusieurs essais d'ajustement ont été effectués. La loi qui a donné le meilleur ajustement est la loi normale N (0, σ_ε). **(Voir graphe IV.3)**.

IV.5.3. PHENOMENE CYCLIQUE

IV - 5.3.1. Données utilisées :

On considère la variable E.T.P. calculée sur une période de 22 années à partir des données de la station d'*ANNABA* Les paramètres statistiques sont résumés dans le **tableau (IV.5).**

Variable	JAN	FEV	MAR	AVR	MAI	JUN	JUI	AOU	SEP	OCT	NOV	DEC
Moy [mm]	1.35	1.74	2.38	3.29	4.34	5.58	6.42	5.98	4.23	2.78	1.71	1.22
Ecart type [mm]	0.30	0.20	0.27	0.35	0.38	0.46	0.39	0.42	0.48	0.42	0.23	0.24

Tableau IV.5 : Paramètres statistiques (ETP)

IV - 5.3.2. Choix du nombre de Composantes Principales

Après étude du tableau **(IV.7)**, le nombre de CP retenu est fixé à sept (07) CP correspondant à un taux de variance expliquée de 92%.

N°de CP	Variance expliquée	% de variance expliquée	% Cumulé
1	5.66	47.18	47.18
2	1.26	10.46	57.64
3	1.13	9.39	67.03
4	1.05	8.75	75.78
5	0.74	6.15	81.94
6	0.70	5.86	87.80
7	0.47	3.90	91.69
8	0.35	2.88	94.57
9	0.26	2.18	96.75
10	0.18	1.51	98.26
11	0.13	1.10	99.36
12	0.08	0.64	100.00

Tableau IV.6 : Contribution de chaque CP à la variance totale

Pour étudier l'importance du terme résiduel, on a procédé à une reconstitution, le résultat était proche de l'historique sans présenter de valeurs négatives.

IV.5.3.3. Simulation des différents paramètres

IV. 5.3.3.1. Simulation des résidus et des C.P.

De la même manière que dans le cas du phénomène aléatoire, la génération des CP se fait dans la loi normale N(0, 1) tandis que celle des résidus se fait par les trois méthodes (N (0, σ_ε) ; Fonction de répartition, chaîne de MARKOV)

IV.5.3.3.1. Simulation de [β]

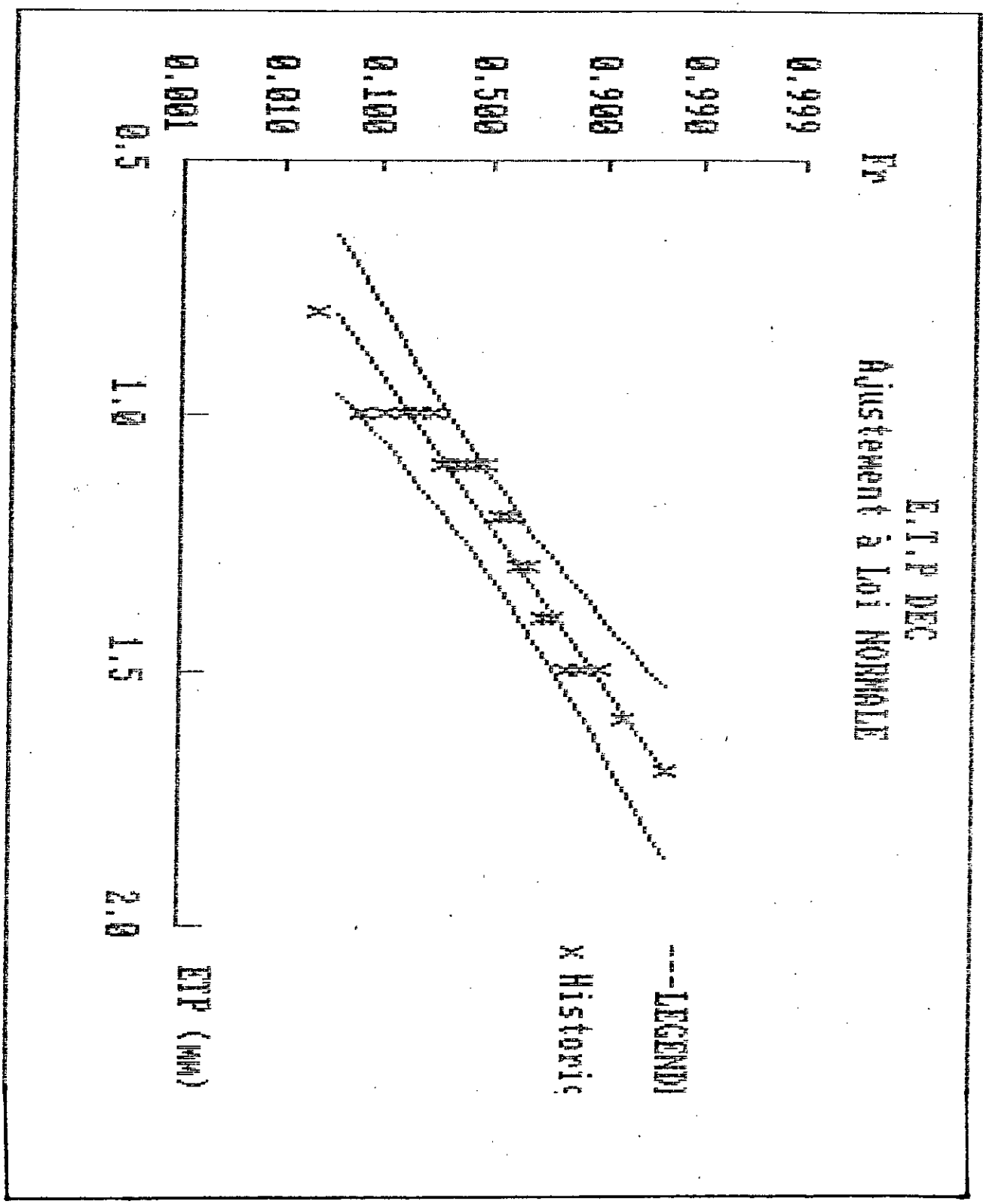
Les vecteurs β_j sont toujours simulés par les fonctions de répartition, β_0 peut être généré par les fonctions de répartition des variables historiques ou dans la loi d'ajustement de ces dernières.
Le **graphe (IV.4.b)** montre que les valeurs mensuelles interannuelles des E.T.P. suivent une loi normale N (m, σ).

IV. 5.4. RESULTATS DE LA GENERATION

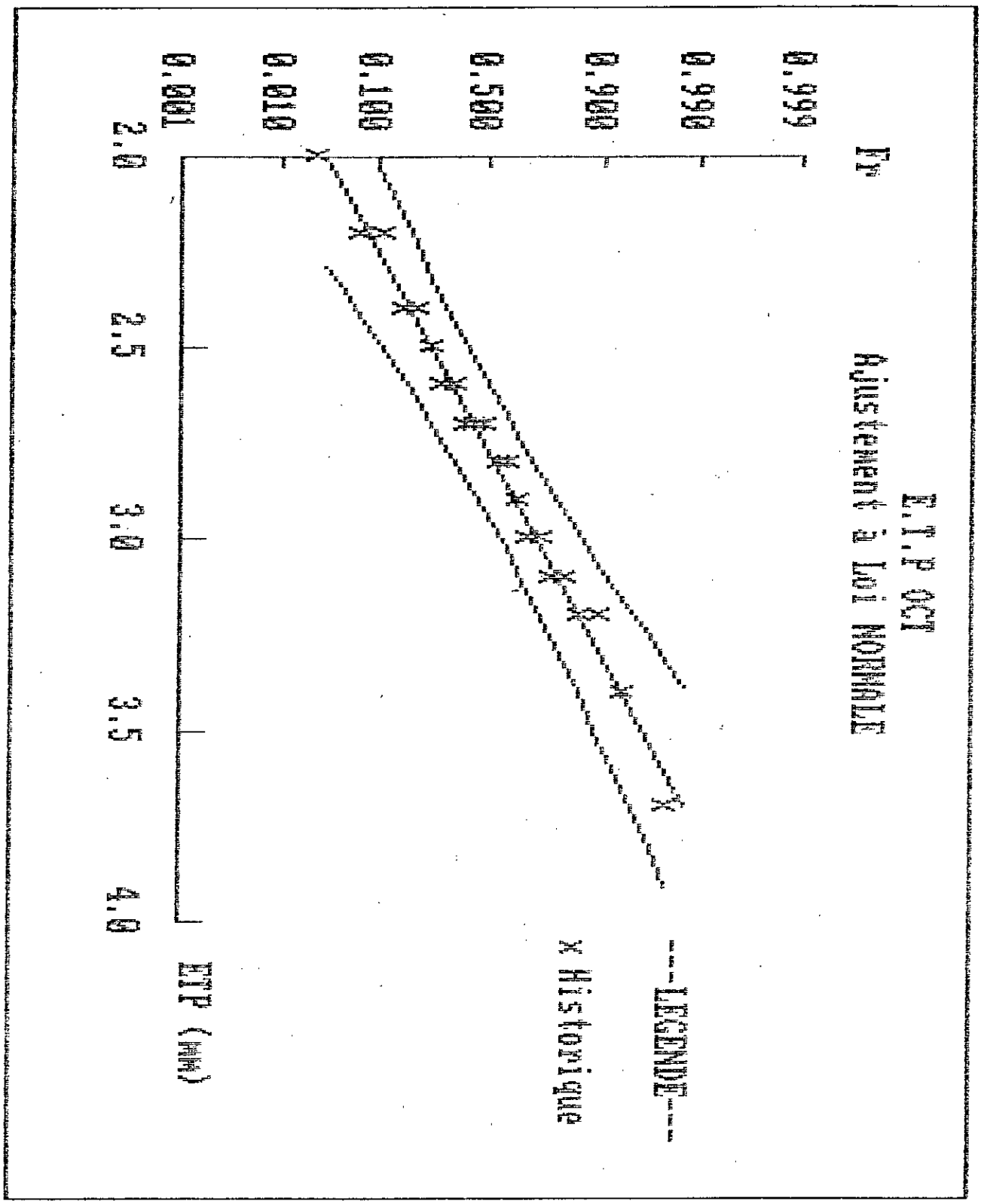
IV.5.4.1. Présentation des différentes catégories de simulation

Vu la variété des méthodes de génération, il est pratique de considérer différentes catégories selon la simulation de la matrice [β]

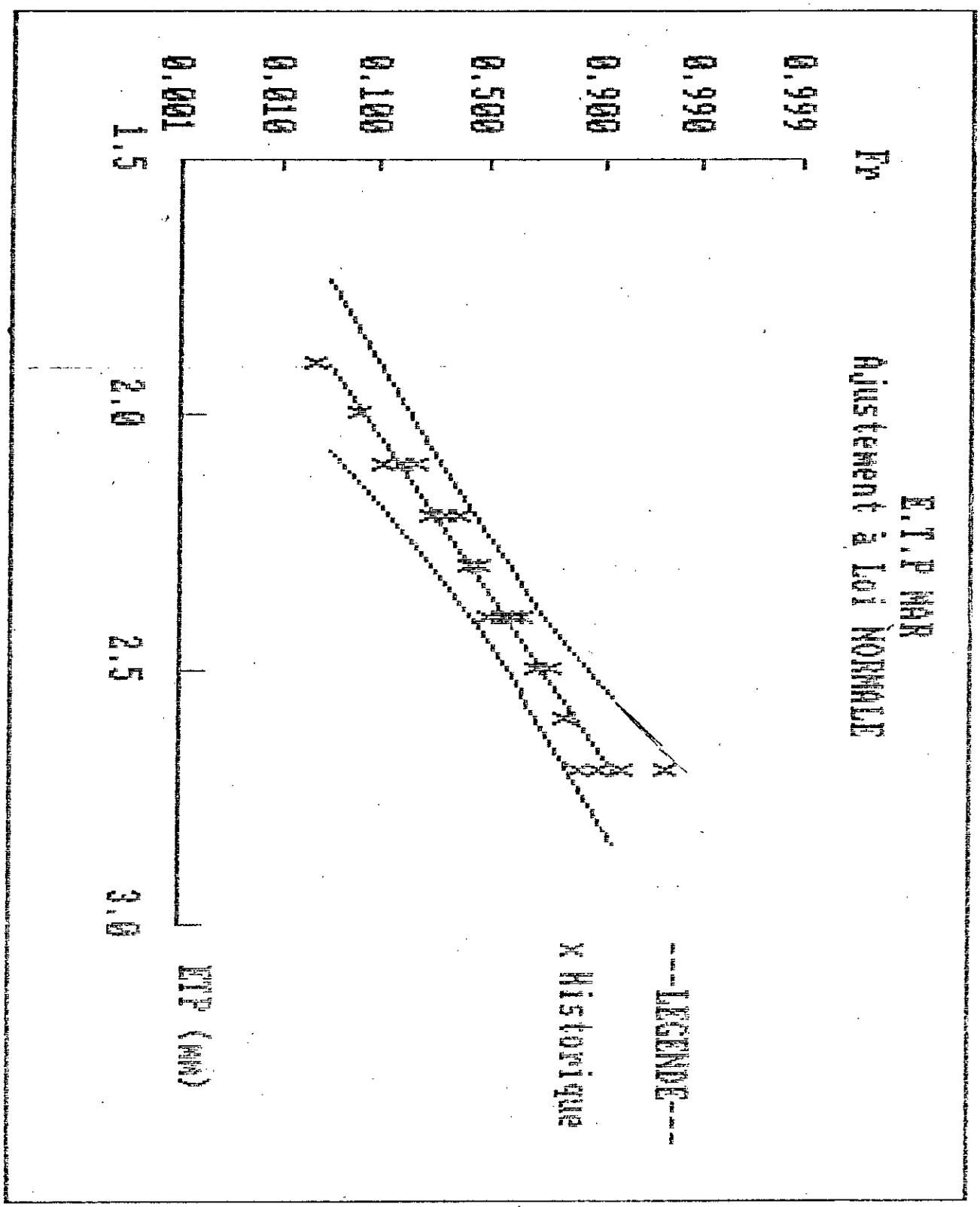
Graphe IV - 4.b : Ajustement des E.T.P.



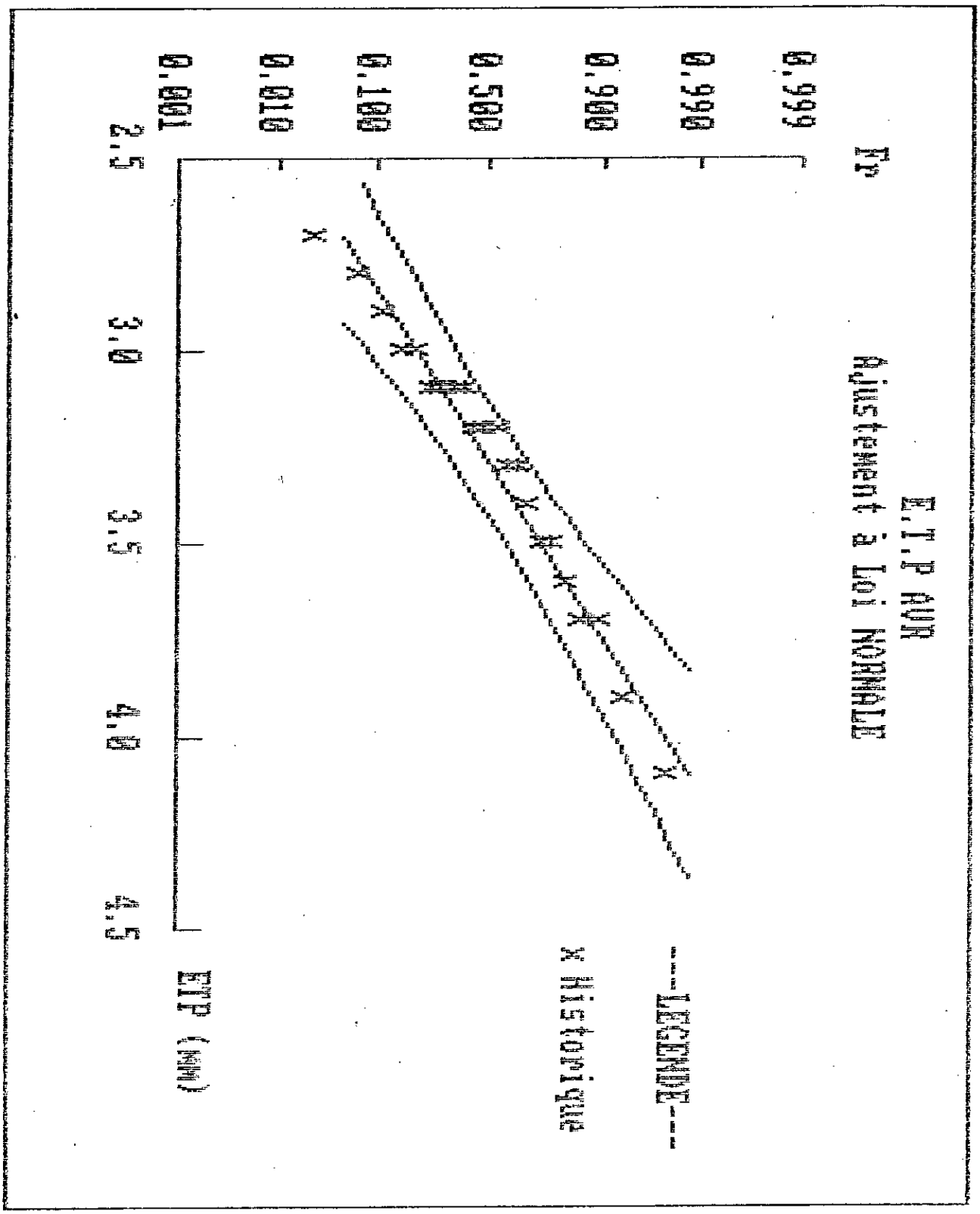
Graphe IV - 4.b : Ajustement des E.T.P.



Graphe IV - 4.b : Ajustement des E.T.P.

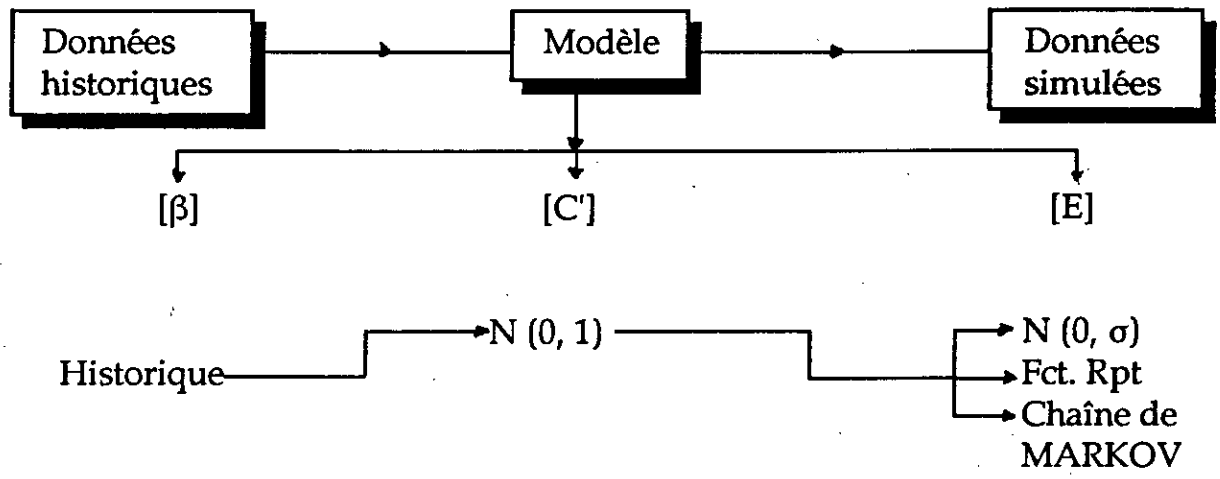


Graphe IV - 4.b : Ajustement des E.T.P.



1ère catégorie

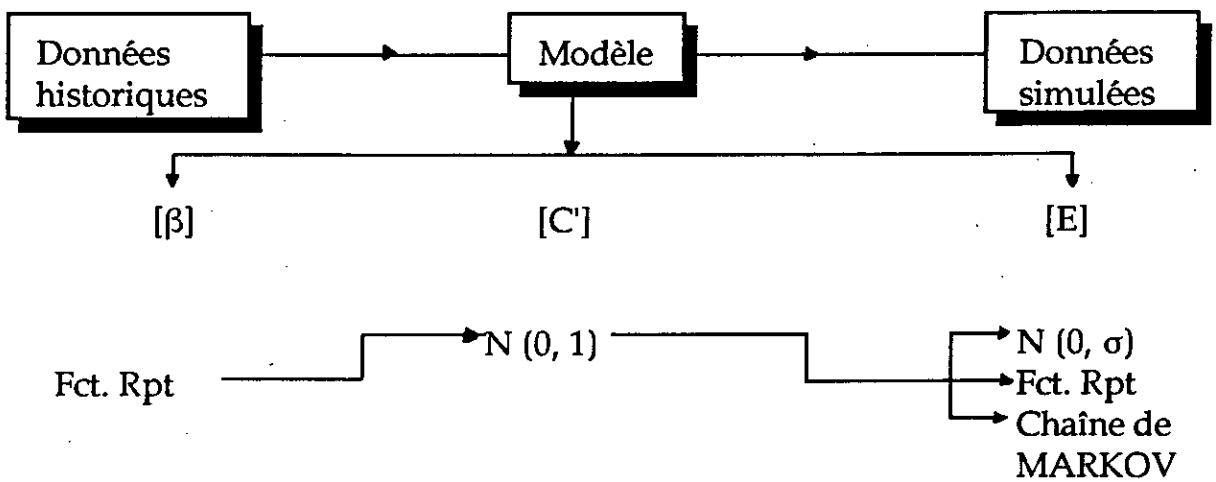
Soit [β] historique, considérons la figure ci-dessous :



Pour [β] non simulées, trois chemins sont possibles.

2ème catégorie

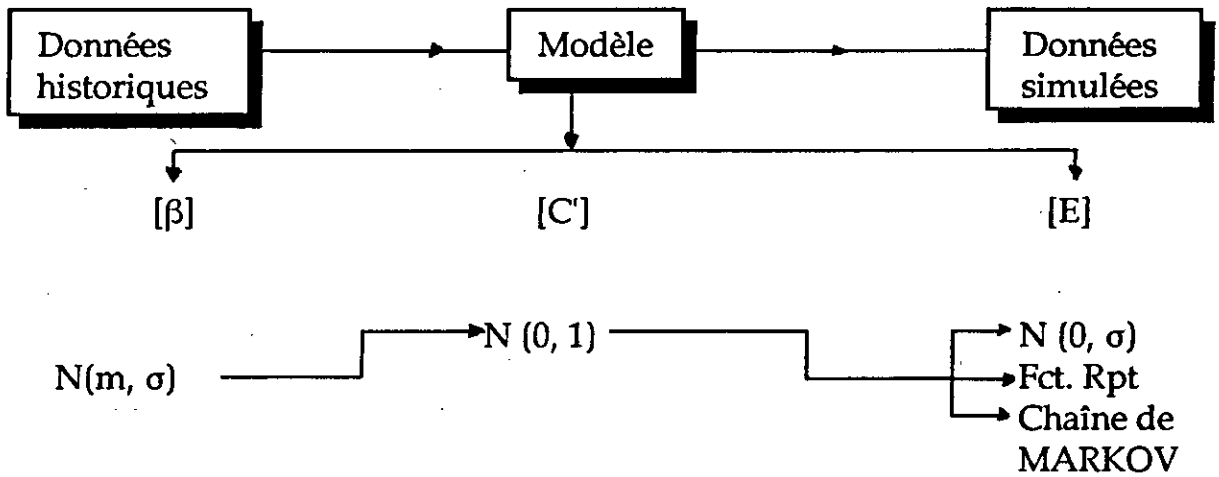
Soit [β] simulée par les fonctions de répartition :



Même dans ce cas, trois chemins sont possibles.

3ème catégorie

Soit le cas où [ß] est simulée dans sa loi d'ajustement :



Comme précédemment, on a trois (03) possibilités de simulation.

On a généré 10 séries de chaque possibilité des différentes catégories pour les deux phénomènes étudiés.

IV.5.4.2. Etude des résultats

La question qui se pose est : ***les séries générées appartiennent -elles à la même population que la série observée*** ? Etant donné qu'on n'a pas les paramètres de la population mais seulement un estimé à l'aide de la série observée ; on va comparer les principaux paramètres statistiques des séries générées et de la série observée et juger sur l'ensemble.

Les paramètres statistiques calculés sur des séries générées par des méthodes satisfaisantes doivent s'approcher des paramètres calculés sur la série observée. Par exemple, si on prend la moyenne des totaux annuels dont on calcule l'intervalle de confiance à partir de l'historique. On détermine par la suite la moyenne des totaux annuels pour les dix séries générées qu'on peut mettre en graphique avec l'intervalle de confiance calculé. Si les valeurs générées n'appartiennent pas à l'intervalle de confiance, on peut douter de la méthode de simulation utilisée.

Mise à part la moyenne des totaux annuels, on a considéré d'autres paramètres statistiques tels que la moyenne mensuelle interannuelle qui doivent appartenir à leurs intervalles de confiance respectifs.

IV.5.4.21.Simulation des ETP

L'Analyse du graphe (**IV.6.1.1.a**) montre que dans le cas où l'on considère la matrice [β] historique (1ere catégorie), l'allure des moyennes des ETP simulées est quasiment linéaire et reste à l'intérieur de l'intervalle de confiance ; ceci est dû à la prépondérance du terme β_0 intervenant sans coefficients de pondération dans le modèle.
Les résultats obtenus dans ce cas sont satisfaisants pour l'ensemble des variantes étudiées. Les mêmes constatations sont faites pour le test sur les moyennes mensuelles (**voir graphes (IV.6.1.2.a,b et c)**).

Mais le fait de garder la matrice [β] *comme apport intégral de l'historique, ne guide-t-il pas les résultats de la simulation ?*

Pour cela on a considéré la seconde et la troisième catégorie où [β] est simulée respectivement par les fonctions de répartition et dans la loi normale N(m,σ) ; l'analyse des graphes (**IV.6.1.1.b,c et IV.6.1.2.b,c**) permet de faire les mêmes remarques que pour la première catégorie.

On peut en conclure que pour un phénomène cyclique tous les modèles proposés donnent de bons et pratiquement mêmes résultats.

IV.5.4.2.2 Simulations des débits

D'Après le graphe (**IV.6.2.1.a**) l'allure des moyennes du débit simulé n'est plus linéaire mais reste dans l'intervalle de confiance, ceci est dû au caractère aléatoire du phénomène étudié.

Pour cette catégorie les trois variantes ne présentent pas de grandes différences, le terme β_0 est toujours prépondérant.

Pour ce qui concerne la question soulevée précédemment, l'analyse du graphe (**IV.6.2.1.b**) concernant la seconde catégorie, montre que le fait de simuler [β] à partir des fonctions de répartition donne d'aussi bons résultats que la première catégorie ; on passe d'une ligne brisée dans le cas [β] historique à une allure curviligne dans le second cas, cela est dû au lissage de la fonction de répartition.

En ce qui concerne la troisième catégorie, (**voir graphe IV.6.2.1.c**) on retrouve une allure légèrement brisée semblable à celle de la première catégorie.
Le bon ajustement de la loi normale aux débits transformés fait que le vecteur β_0 généré dans la loi normale N(m,σ), ne séloigne pas vraiment de celui calculé sur les données observées.
On remarque qu'en général les résultats appartiennent à l'intervalle de confiance tout en étant meilleurs pour la première variante (ε généré dans N(0,σ)).

On peut dire que la simulation du terme moyenne par les fonctions de répartition n'est qu'un lissage des données historiques.

On obtient par cette méthode un double avantage : premièrement on diminue l'erreur des mesures sur les débits ainsi que l'effet de tout caractère exceptionnel de l'échantillon, deuxièmement les débits générés dans ce cas représenteront mieux la population.

Pour dégager le meilleur modèle entre [β] simulée respectivement par les fonctions de répartition et la loi normale N(m,σ). (les combinaisons faisant intervenir [β] historique étant écartées dés le départ), on examine les graphes (**IV.6.2.1 et IV.6.2.2**), on remarque que tous les résultats sont satisfaisants, mais on opte pour le modèle donné par la seconde catégorie deuxième variante ([β] et ε simulés par les fonctions de répartition).

IV.5.5.ETUDE DE L'INFLUENCE DU NOMBRE DE CP

La reconstitution des paramètres débit et évapotranspiration a été faite en tenant compte des sept (07) premières CP qui contribuent à l'explication de la variance totale. Cette reconstitution s'étant faite d'une façon satisfaisante, le nombre de CP a été fixé à 07 dans les différents modèles de simulation.

On s'est demandé si la réduction du nombre de CP a une influence significative sur les résultats obtenus.

Pour prendre en considération la fluctuation des résidus, on a opté pour la simulation de ceux-ci par les chaînes de Markov ; en effet lorsque le nombre de CP est faible, la variance expliquée sera elle aussi faible et le complément d'information sera contenu dans les résidus. On a travaillé alors avec le modèle donné par la seconde catégorie troisième variante et, on a fait varier le nombre des CP (1, 2 et 4 CP), trois (03) séries ont été générées pour chaque cas.

Les résultats du test sur la moyenne des totaux annuels se sont avérés satisfaisants (**voir tableau.IV.7**) même en ne tenant compte que d'une seule CP. Le modèle choisi ne permet pratiquement aucune perte d'informations quelque soit le pourcentage p de variance expliquée apporté par les CP ; le terme résiduel permet de récupérer presque la totalité de la variance résiduelle c'est-à-dire (1-p)%.

	ETP			DEBIT		
	IC $^-$	Moyenne[mm]	IC $^+$	IC $^-$	Moyenne [m^3/s]	IC $^+$
1 CP	53.26	75.48 74.34 70.70	97.81	88.48	164.60 178.35 169.52	180.20
2.CP		72.69 74.25 74.78			131.39 127.90 120.74	
4 CP		74.75 72.07 70.79			152.37 139.84 151.65	

Tableau IV.7 Influence du nombre de CP

IV.5.6 INFLUENCE DE LA TAILLE DE L'ECHANTILLON

Le but de la simulation est de reproduire le plus fidèlement possible les caractéristiques statistiques de la série historique. Autrement dit, *le modèle choisi peut-il donner des séries simulées ayant les mêmes caractéristiques que l'historique, quelque soit la taille de l'échantillon?*
Pour cela, on tronque la série initiale en prenant N=13 pour les deux phénomènes étudiés : L'**ETP** et le **débit** et, on génère trois (03) séries pour chacun, en utilisant le même modèle que précédemment.

- **Variable ETP**

Les résultats obtenus lors de la simulation de l'évapotranspiration avec un échantillon de 13 années sont satisfaisants comme le montre **le tableau IV.8.**
On dira que l'échantillon ainsi réduit, possède l'information nécessaire pour décrire le phénomène étudié ; ceci est dû au caractère cyclique de ce dernier.

- **Variable débit**

Les mêmes constatations sont faites pour le débit : Les résultats présentés par le **tableau IV.8** montre que les différentes moyennes des totaux annuels des séries générées, appartiennent à l'intervalle de confiance.

	IC⁻	Moyenne	IC⁺
ETP [mm]	30.2	41.6 42.1 42.3	56.1
DEBIT [m^3/s]	24.20	33.95 31.83 35.70	50.12

Tableau IV.8 : Influence de la taille de l'échantillon

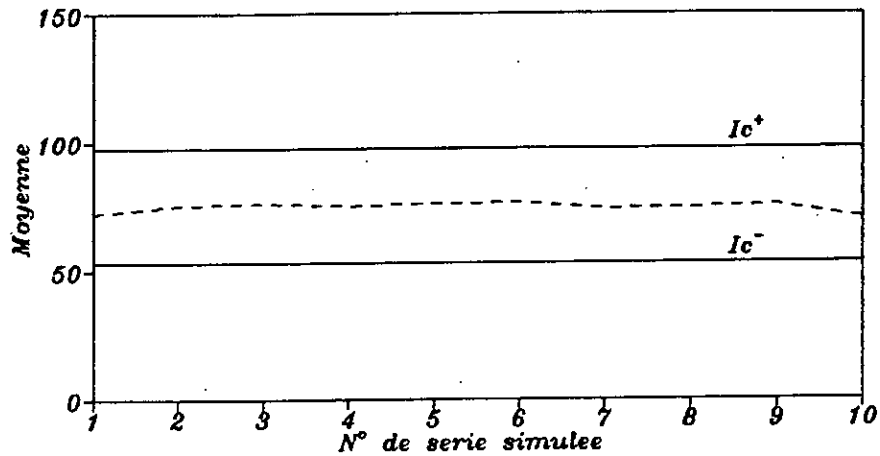
IV.6.CONCLUSION

Le modèle de génération développé à l'aide des composantes principales présente de très grands avantages par rapport à ceux couramment utilisés jusqu'à présent. Ceci est dû au fait que le modèle de simulation posséde plusieurs degrés de liberté, il offre une importante variété de combinaisons de méthodes de génération.

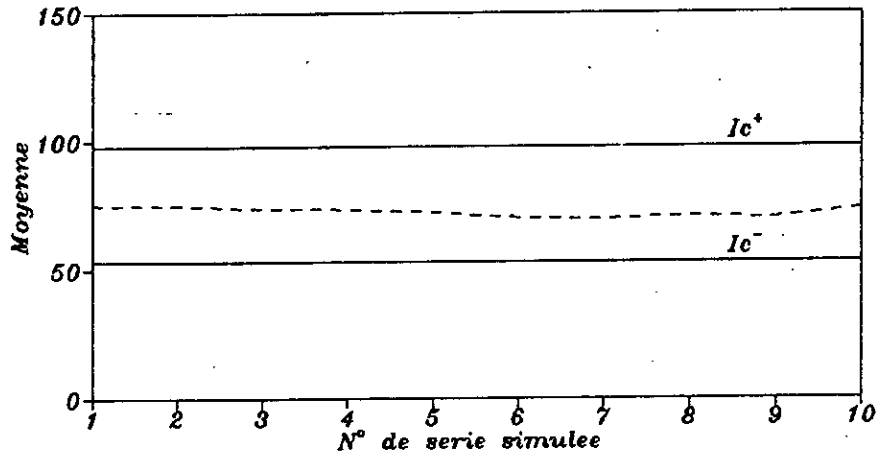
L'étude de ces différentes combinaisons a donné lieu aux constatations suivantes :

- **Le modèle basé sur [β] historique donne des séries simulées pratiquement identiques à celles observée. Alors que si [ß] est générée par les fonctions de répartition, l'information véhiculée est pratiquement dépourvue d'erreurs (erreurs systématiques et/ou accidentelles).**

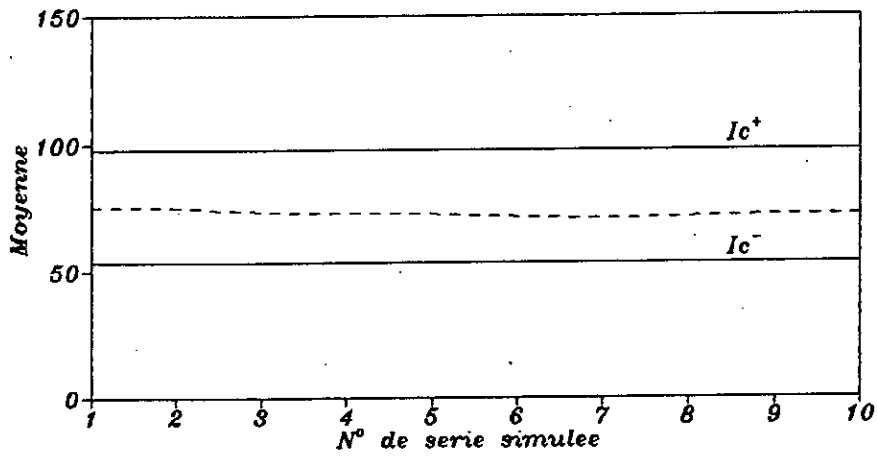
- **Le modèle élaboré est valable aussi bien pour les phénomènes cyclique qu'aléatoire, avec une transformation des données initiales pour ce dernier. Il ne dépend pas du nombre de CP retenu, vu que le taux de variances résiduelle est récupéré par le terme erreur, si ce dernier est simulé par une méthode adéquate.**



CP N(0,1) et ε N(0,σ)



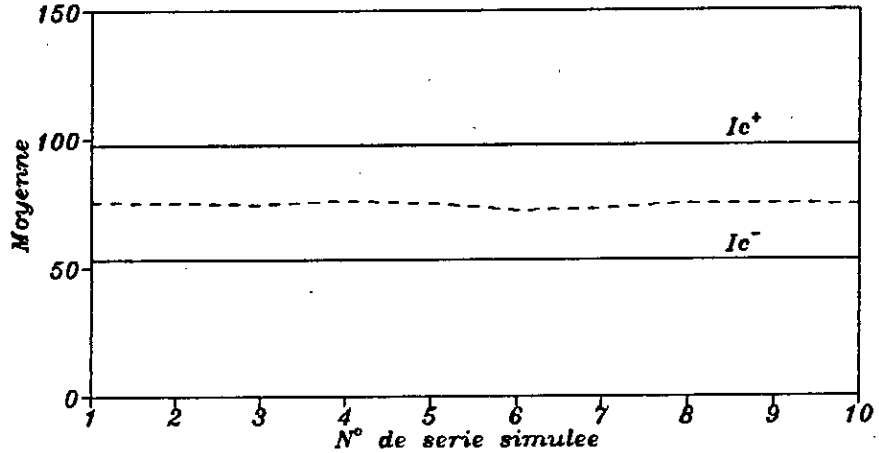
CP N(0,1) et ε fonction de repartition



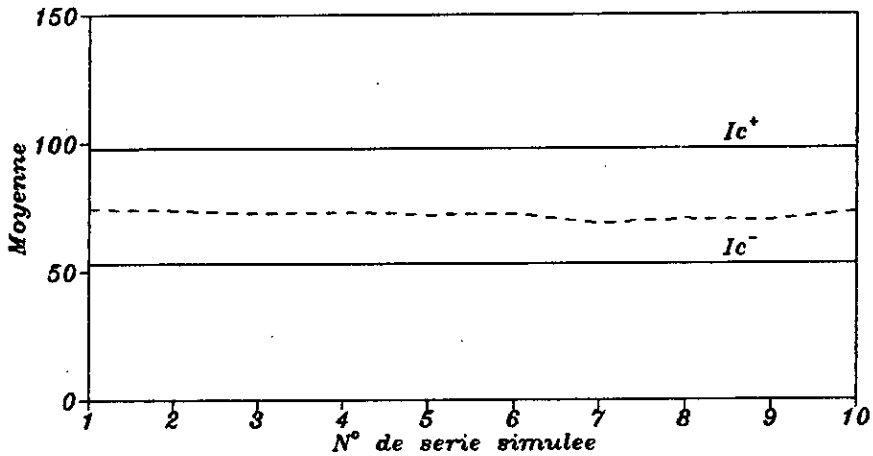
CP N(0,1) et ε chaine de MARKOV

Graphe IV.6.1.1.α : Test sur les Totaux Annuels
- Premiere Categorie -

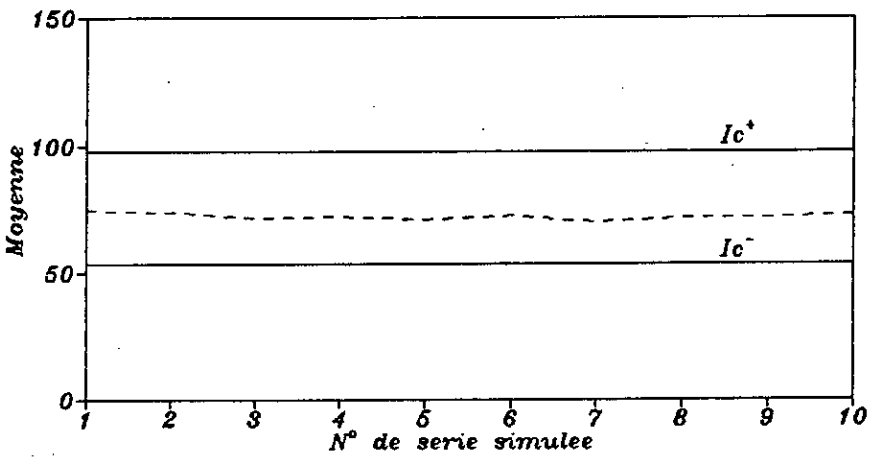
——— Intervalle de Confiance
- - - - Valeur Simulee



CP N(0,1) et ε N(0,σ)



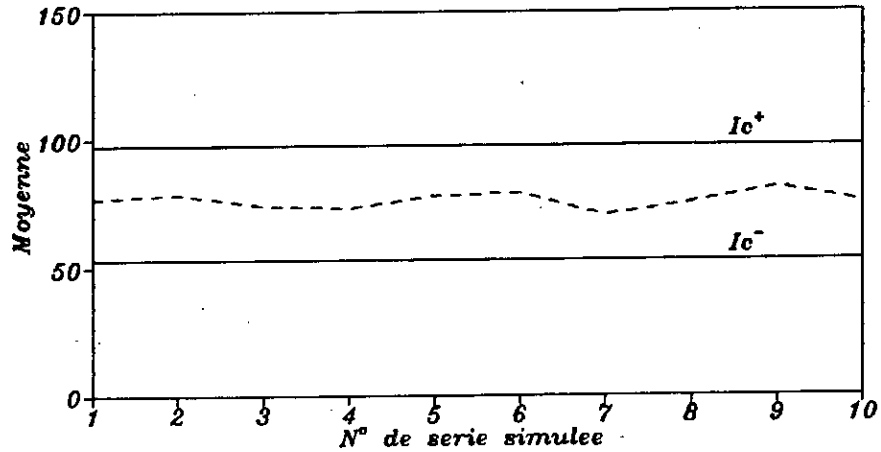
CP N(0,1) et ε fonction de repartition



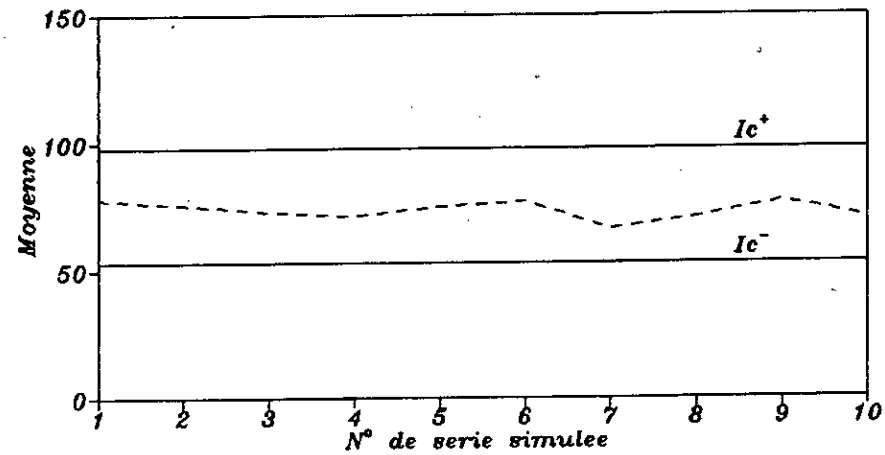
CP N(0,1) et ε chaine de MARKOV

Graphe IV.6.1.1.b : Test sur les Totaux Annuels
– Deuxieme Categorie –

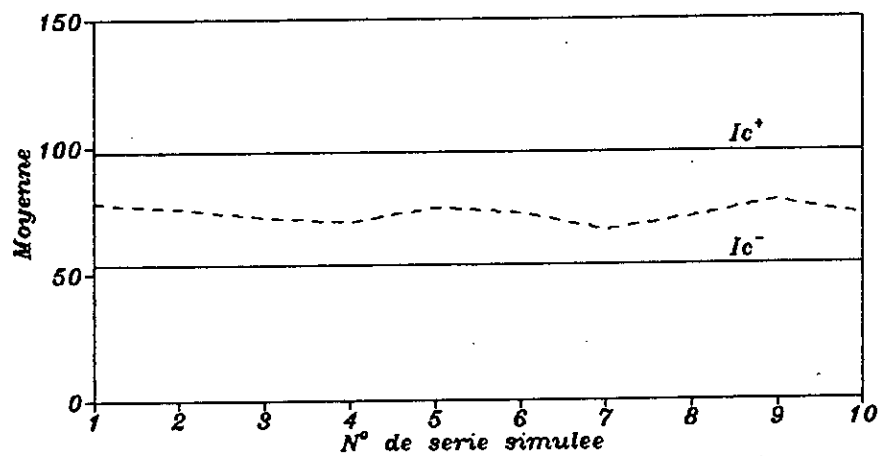
——— Intervalle de Confiance
- - - - Valeur Simulee



CP N(0,1) et ε N(0,σ)



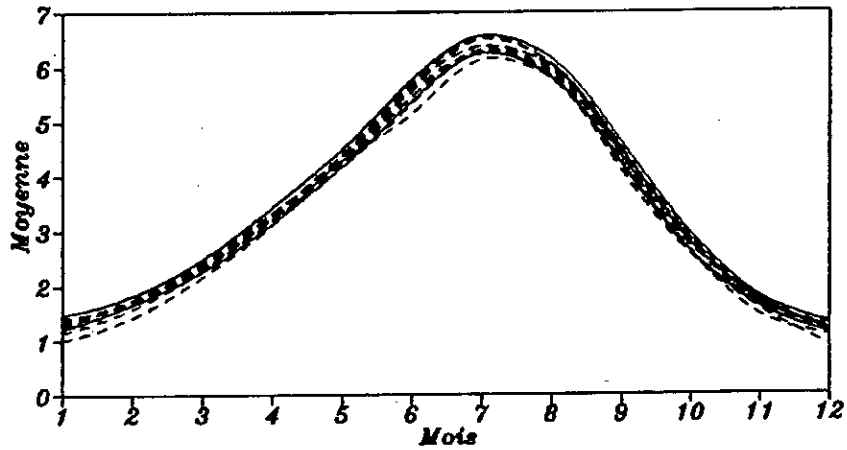
CP N(0,1) et ε fonction de repartition



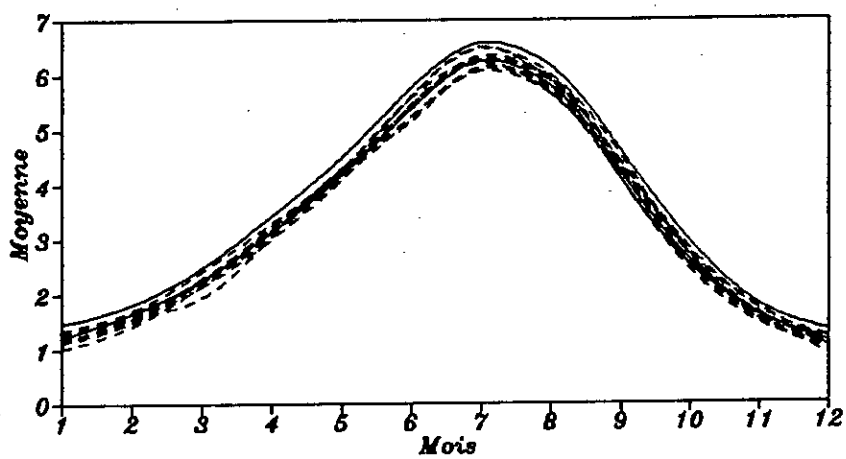
CP N(0,1) et ε chaine de MARKOV

Graphe IV.6.1.1.c : Test sur les Totaux Annuels
- Troisieme Categorie -

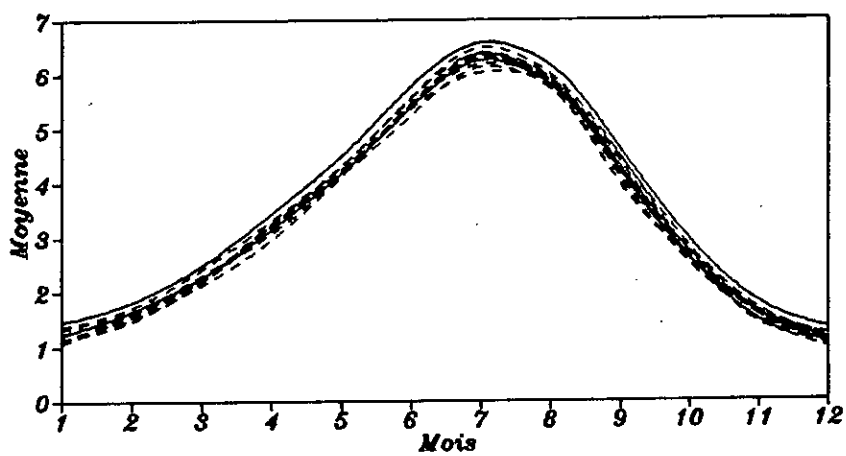
——— Intervalle de Confiance
---- Valeur Simulee



CP N(0,1) et ε N(0,σ)



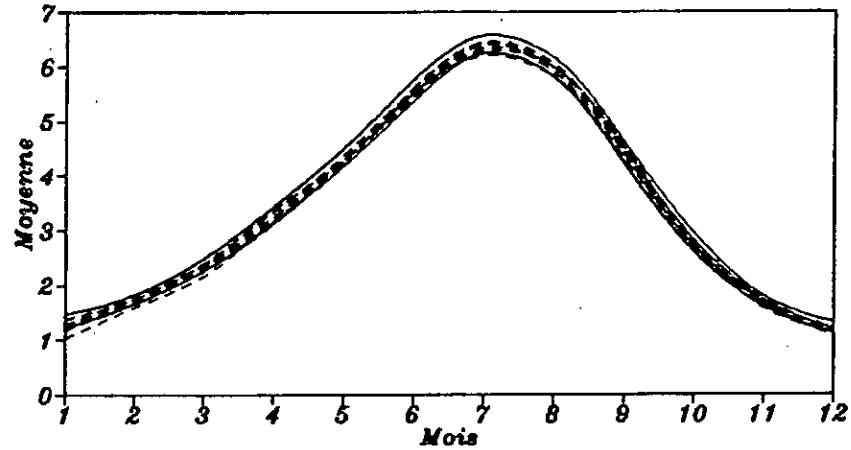
CP N(0,1) et ε fonction de repartition



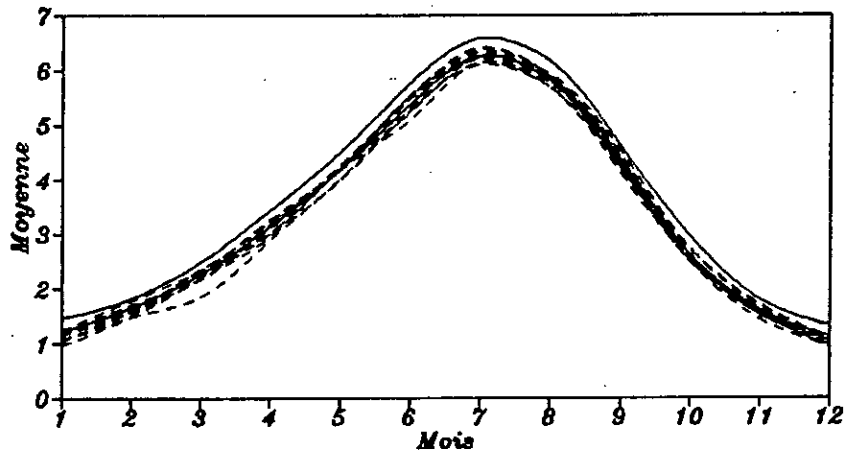
CP N(0,1) et ε chaine de MARKOV

Graphe IV.6.1.2.a : Test sur les Moyennes Mensuelles
– Premiere Categorie –

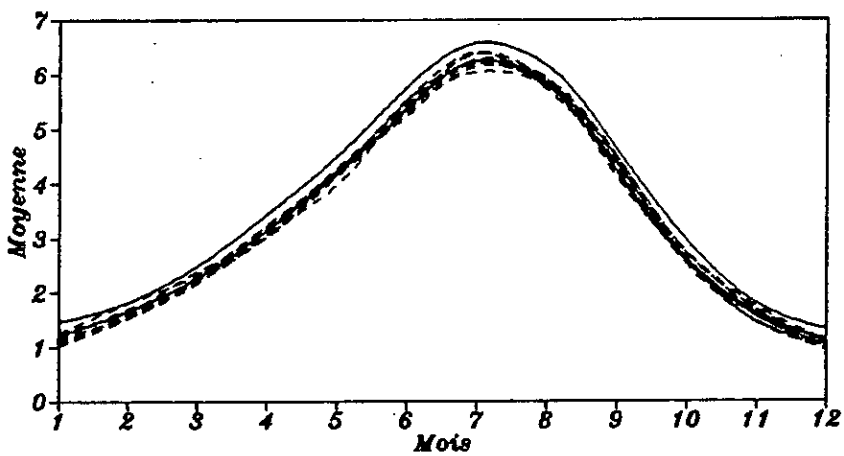
——— Intervalle de Confiance
- - - - - Valeur Simulee



CP N(0,1) et ε N(0,σ)



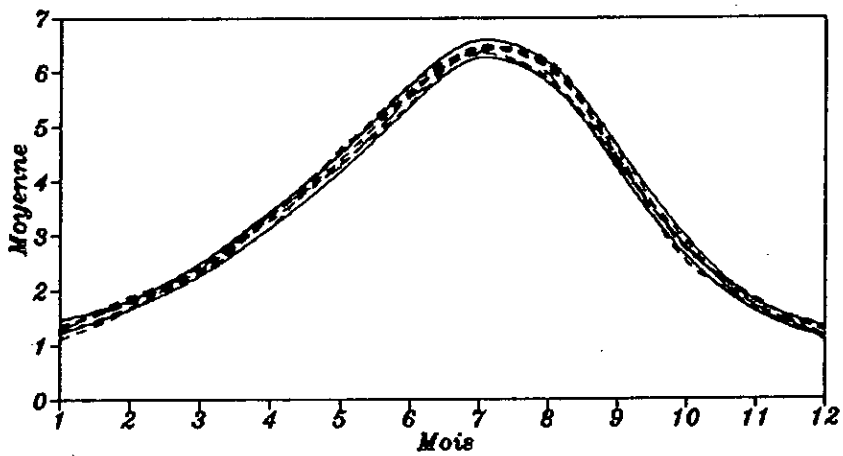
CP N(0,1) et ε fonction de repartition



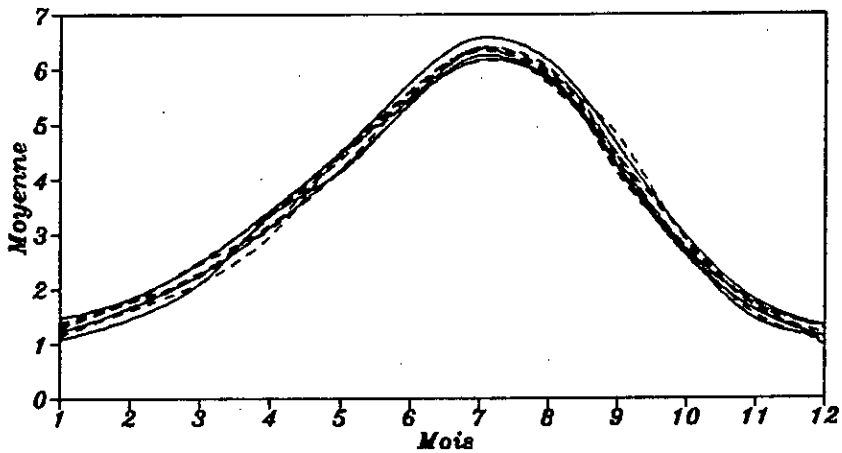
CP N(0,1) et ε chaine de MARKOV

Graphe IV.6.1.2.b : Test sur les Moyennes Mensuelles
– Deuxieme Categorie –

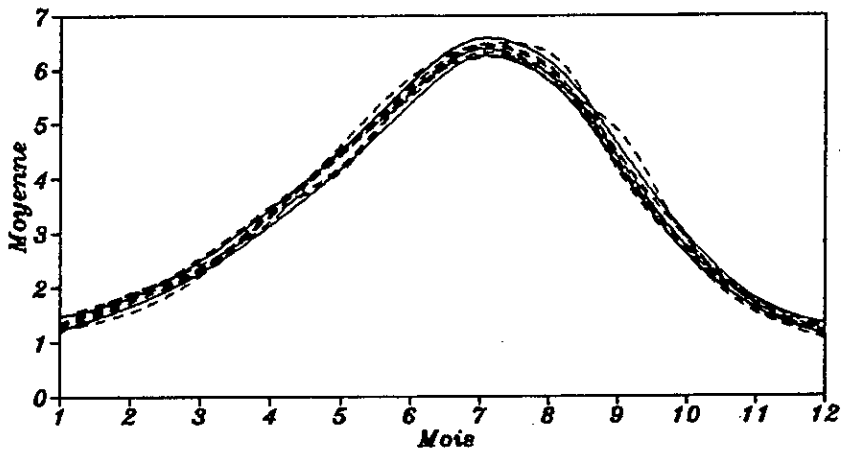
Intervalle de Confiance
Valeur Simulee



CP N(0,1) et ε N(0,σ)



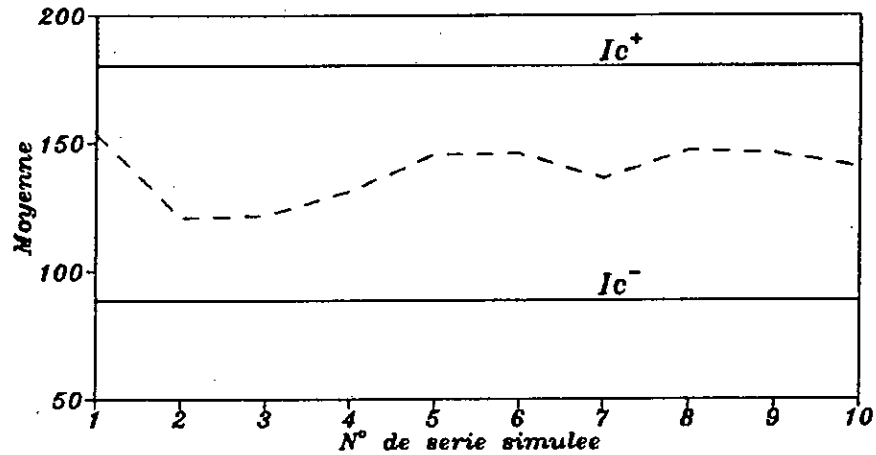
CP N(0,1) et ε fonction de repartition



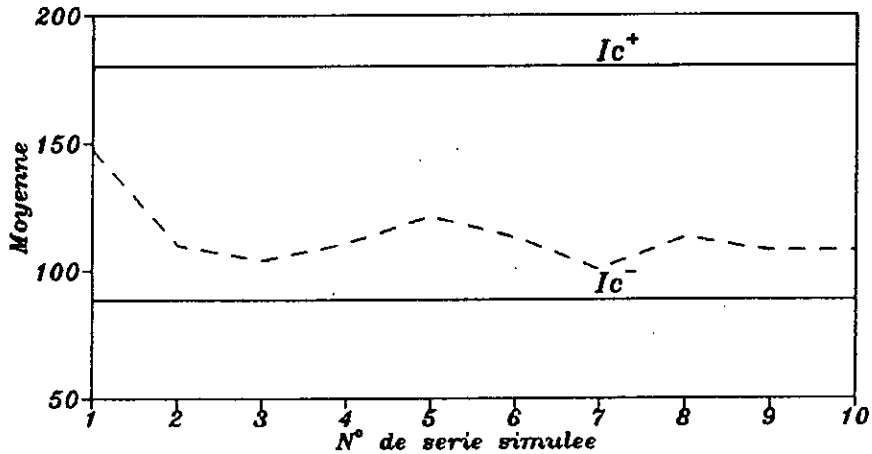
CP N(0,1) et ε chaine de MARKOV

Graphe IV.6.1.2.c : Test sur les Moyennes Mensuelles
-Troisieme Categorie -

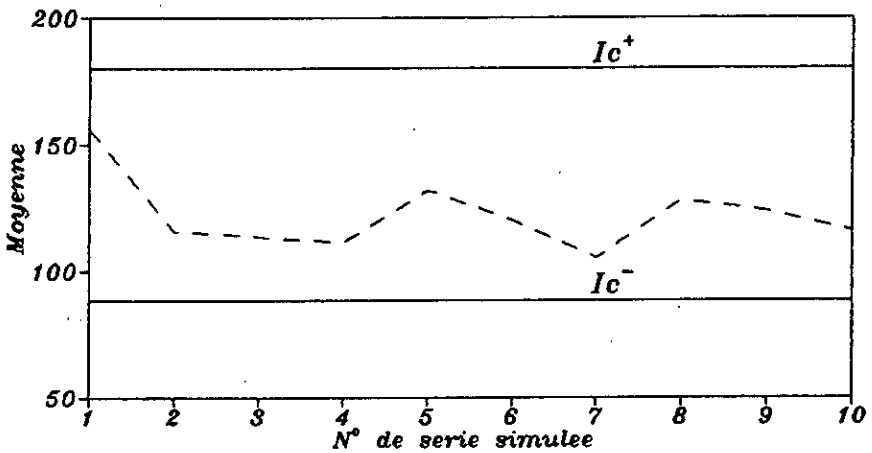
Intervalle de Confiance
Valeur Simulee



CP N(0,1) et ε N(0,σ)



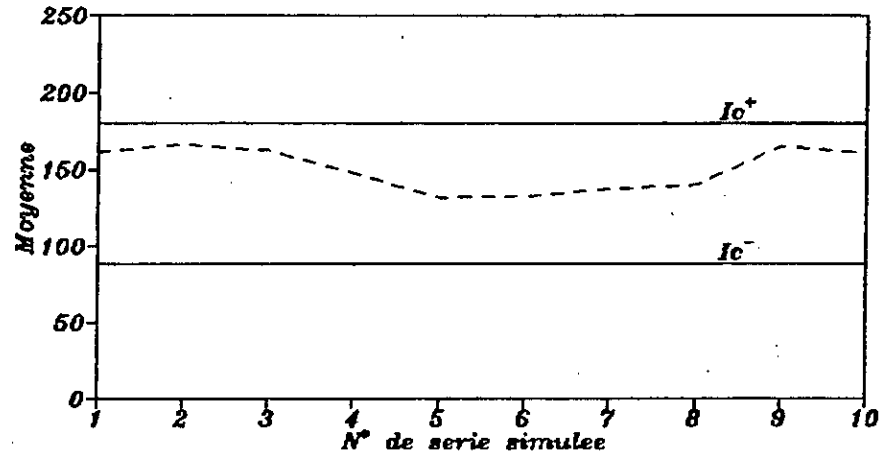
CP N(0,1) et ε fonction de repartition



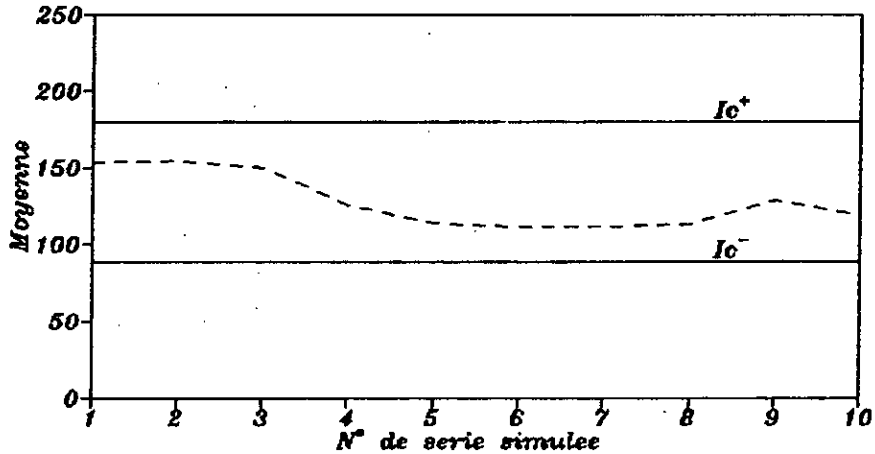
CP N(0,1) et ε chaine de MARKOV

Graphe IV.6.2.1.a : Test sur les Totaux Annuels
– Premiere Categorie –

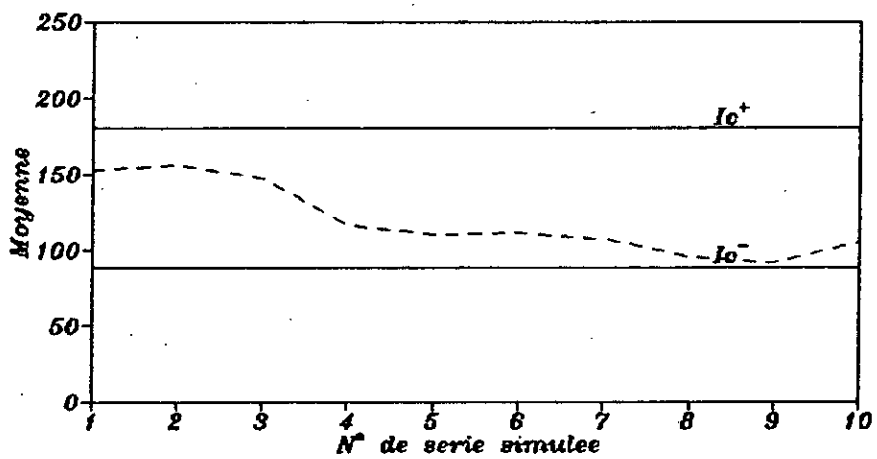
Intervalle de Confiance
Valeur Simulee



CP N(0,1) et ε N(0,σ)

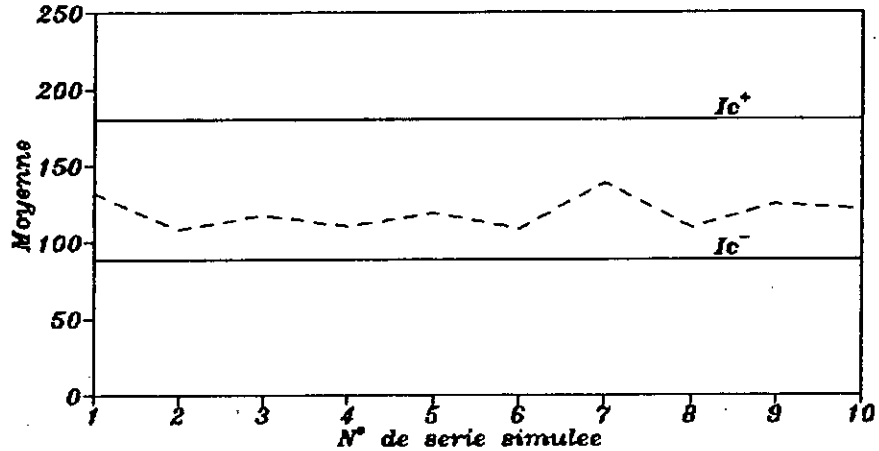


CP N(0,1) et ε fonction de repartition

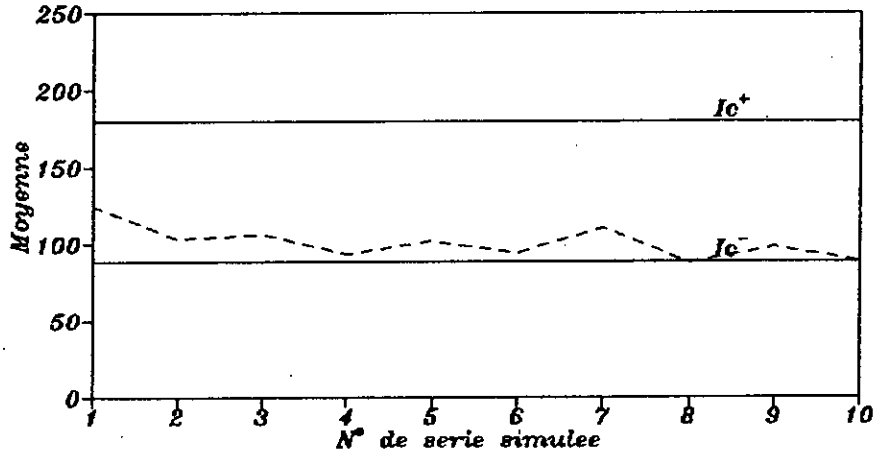


CP N(0,1) et ε chaine de MARKOV

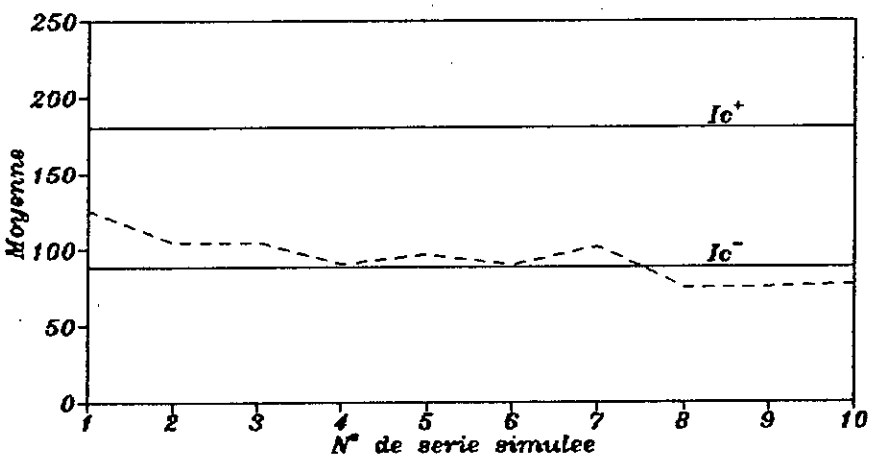
Graphe IV.6.2.1.b : Test sur les Totaux Annuels
– Deuxieme Categorie –



CP N(0,1) et ε N(0,σ)



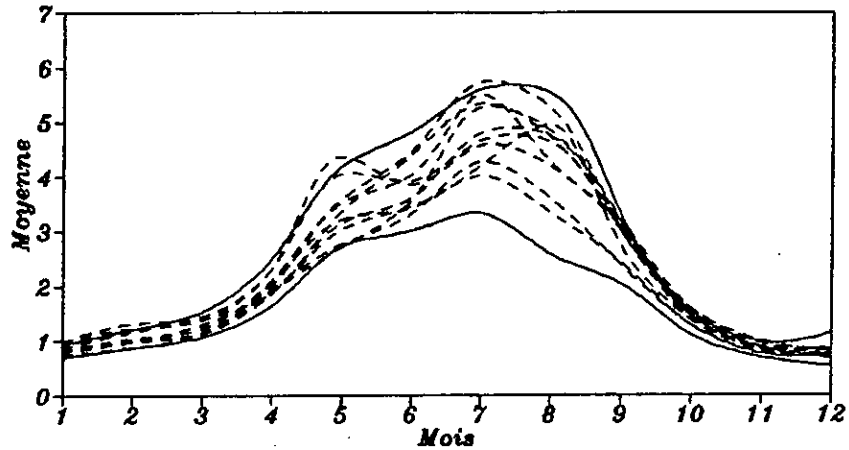
CP N(0,1) et ε fonction de repartition



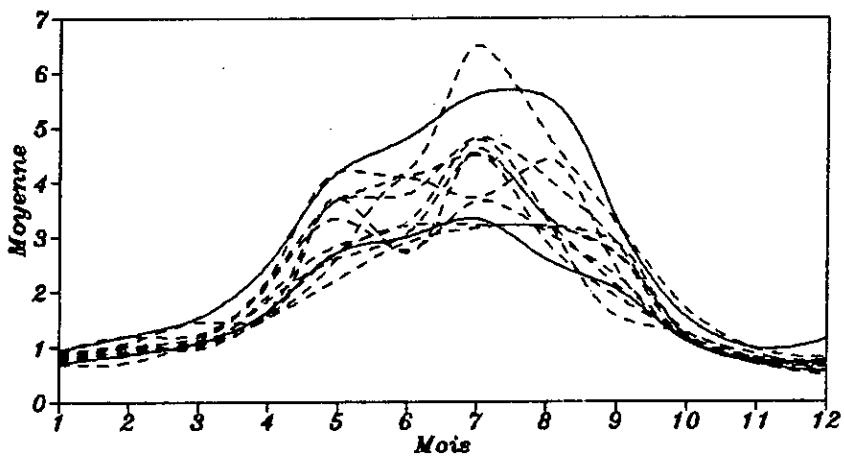
CP N(0,1) et ε chaine de MARKOV

Graphe IV.6.2.1.c : Test sur les Totaux Annuels
– Troisieme Categorie –

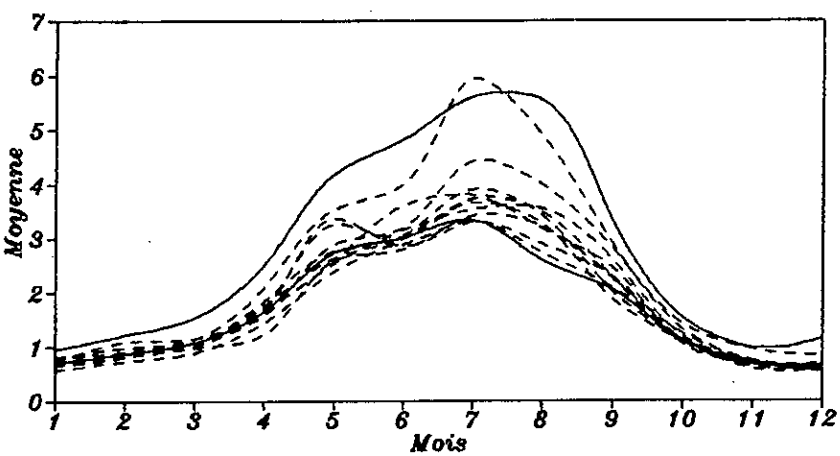
——— Intervalle de Confiance
- - - - Valeur Simulee



CP N(0,1) et ε N(0,σ)



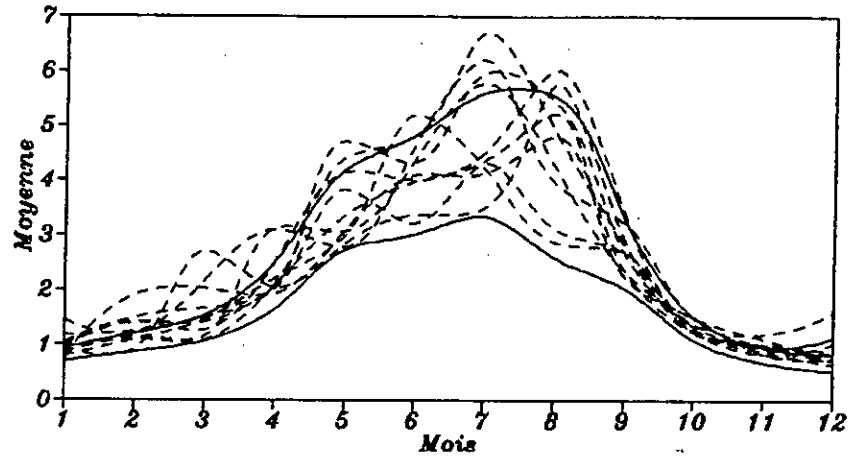
CP N(0,1) et ε fonction de repartition



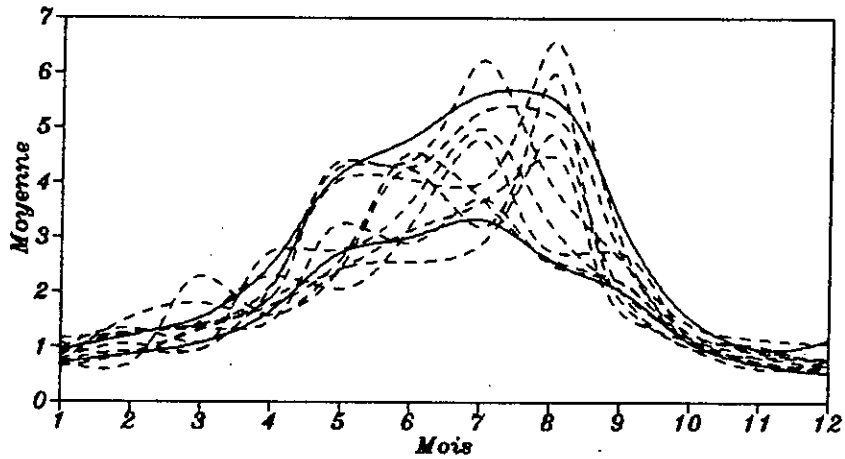
CP N(0,1) et ε chaine de MARKOV

Graphe IV.6.2.2.a : Test sur les Moyennes Mensuelles
- Premiere Categorie -

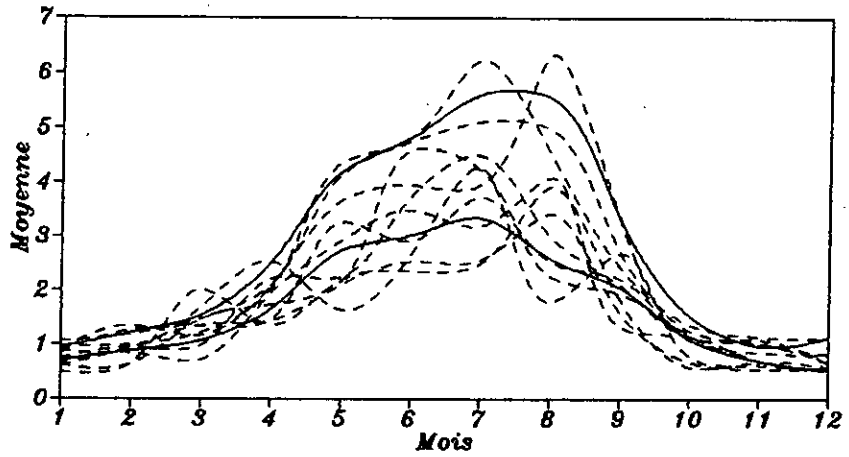
——— *Intervalle de Confiance*
- - - - *Valeur Simulee*



CP N(0,1) et ε N(0,σ)



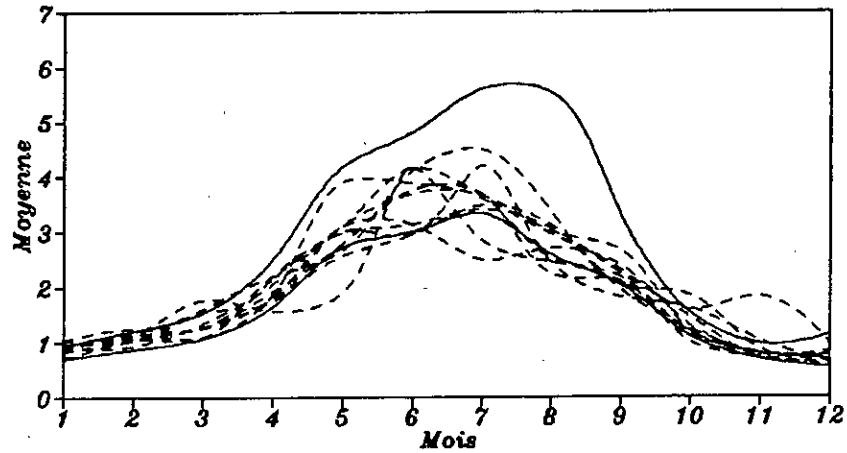
CP N(0,1) et ε fonction de repartition



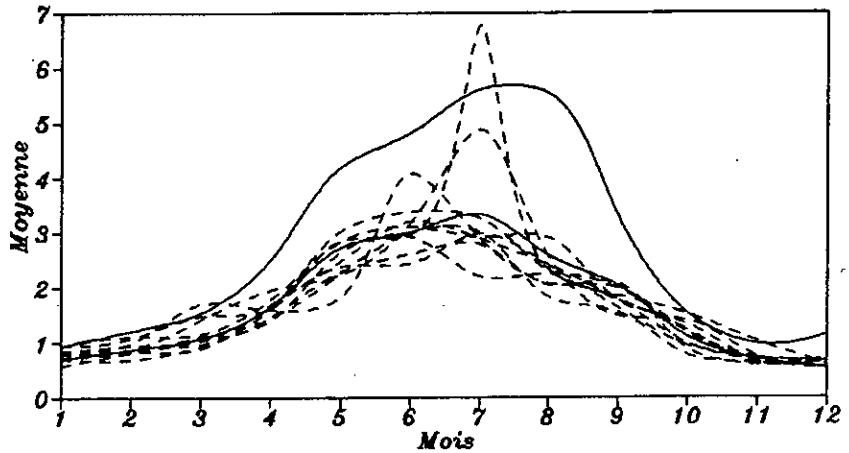
CP N(0,1) et ε chaine de MARKOV

Graphe IV.6.2.2.b : Test sur les Moyennes Mensuelles
– Deuxieme Categorie –

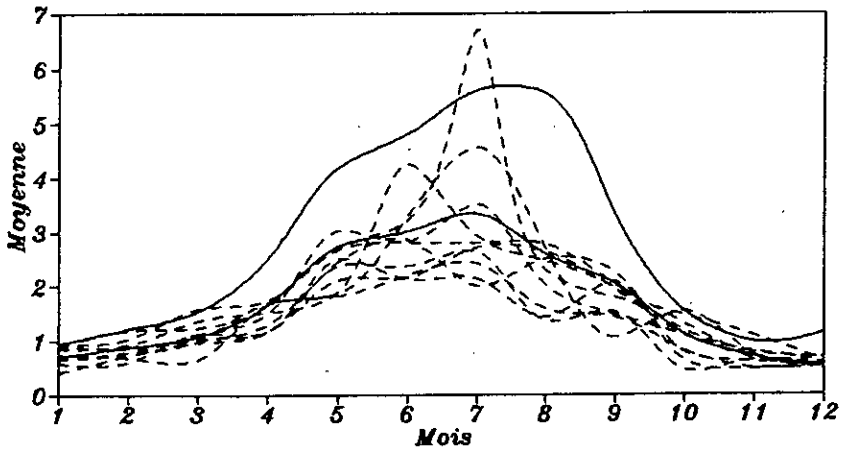
Intervalle de Confiance
Valeur Simulee



CP N(0,1) et ε N(0,σ)



CP N(0,1) et ε fonction de repartition



CP N(0,1) et ε chaine de MARKOV

Graphe IV.6.2.2.c : Test sur les Moyennes Mensuelles
- Troisieme Categorie -

____ *Intervalle de Confiance*
_ _ _ _ *Valeur Simulee*

CONCLUSION GENERALE

L'Analyse en Composantes Principales permet de construire des variables statistiques artificielles qui expliquent l'ensemble de celles prises en compte.

Les représentations obtenues par une A.C.P. ne sont qu'une description des données que l'on possède, par conséquent elles ne peuvent en aucun cas servir de preuve pour une quelconque théorie, l'approche est purement descriptive. La description ainsi obtenue facilite la recherche d'une structure de l'ensemble des données analysées par cette méthode de représentation.

La simulation des paramètres hydrométérologiques permet d'avoir des séries synthétiques qui sont utilisées dans la gestion des ressources en eau. A cet effet, plusieurs modèles de simulation sont proposés (ARMA, Désagrégation, Broken line, MARKOV, etc...). Ceux-ci ne peuvent être appliqués que sous certaines condtions : type de phénomène, chronologie, stationnarité...

Les modèles élaborés basés sur les Composantes Principales, permettent de s'affranchir de ces contraintes.

En effet, les différents tests menés sur des variables cycliques (E.T.P., température) et aléatoires (débit, pluie) ont donné d'excellents résultats. La stationnarité du phénomène et la chronologie des séries ne constituent pas un handicap dans l'application de ces modèles ; cela est dû respectivement au fait de normer les variables brutes est de travailler avec la matrice de corrélation inter-variables.

L'étude de l'influence du nombre de Composantes Principales sur la qualité de la simulation a révélé que celui-ci peut être réduit tout en aboutissant à d'aussi bons résultats. Cela s'explique par le fait que les variations des erreurs sont prises en compte dans la simulation.

Les essais réalisés sur des échantillons de petite taille (13 - 30 années) ont montré que l'effet de la taille n'est pas significatif dans les modèles développés, car l'information contenue dans l'échantillon est valorisée par les C.P.

Pour la suite de cette recherche, on recommande :

- d'approfondir la modélisation par l'A.C.P. pour la critique des données et la prévision.
- de calculer les C.P. sur un grand nombre de bassins et trouver les relations entre les paramètres physiques de ces derniers et les coefficients de régression.
- d'affiner les modèles de simulation en réduisant le pas de temps (décadaire, journalier, ...).
- de simuler les variables étudiées en corrélation (Exemple : pluie - E.T.P., pluie - débit).

Enfin, on peut dire que l'Analyse en Composantes Principales est une technique qui allie **simplicité** et **puissance**.

LISTE DES FIGURES

LISTE DES GRAPHES

LISTE DES TABLEAUX

BIBLIOGRAPHIE

1/AURAY.J.P (1990) : "Analyse des données multidimensionnelles". pp 19 -75

2/BENZECRI .J.P (1982) : "Histoire et préhistoire de l'Analyse des données " DUNOD.

3/BOUROCHE J.M. (1983) :" L'Analyse des données" 2eme Ed Que sais-je? pp 5 - 45.

4/ BURDEN .R.L (1981) : "Numerical Analysis" 2eme Ed BOSTON - MASSACHUSETTES- pp 73 - 107.

5/ CHEVALIER .P.(1985) : "Pluie et crue au SAHEL.Hydrologie de la mare d'Oursi (BURKINA-FASO)"- Ed L'ORSTON - pp 79 -88.

6/CHTCHERBATSKI (1989) : "Analyse Numérique - Cours et Problèmes- O.P.U.

7/ DE LAGARDE.J.(1983) : " Initiation à l'Analyse des données". DUNOD- pp 1-91.

8/ DIPAY.E. (1989) : "Eléments d'Analyse des Données" - DUNOD - pp 168 - 221.

9/ DUBAND.D (1989) : "Hydrologie statistique Approfondie" I.N.P.G - pp VI.1 - VI.43

10/ DUBREUIL.P. (1974) : "Initiation à l'Analyse hydrologique" ORSTON (PARIS)

11/ MCGRAW. HIll : "Numérical Analysis".

12/ KOTTEGODA (1980) : "Stochastic Water Ressources technology".- The MACMILLAN PRESS LTD.

13/ LAADOUA ABDELHAK (1987) : "Les variations spatio-temporelles des précipitations au Maroc septentrionale" Thèse DEA - Université des sciences et techniques de LILLE Flandres . Artois

14/ LABORDE.J.P.: "Rapport d'activité concernant la période allant du 1er Février 1991 au 31 Octobre 1991". ANRH.

15/ LEBART.L. (1982) : "Technique de la description statistique multidimensionnelle" - MASSON - pp 119 - 148.

17/ MORIN .G. (1974) : "Génération de chronique de débits Conception nouvelle utilisant des fonctions orthogonales". - Thèse de DEA- Grenoble.

18/ NAKACHE.J.P (1970) :"Analyse en Composantes Principales" - (PARIS).

19/ ROBERT.J. (1984) : "Les statistiques : une approche nouvelle"
MC GRAW - HILL- pp : 409 - 418.

20/ TESSIER DANIEL .(1983) : "Analyse statistique des caractéristiques
physiographiques de petits bassins versants
en relation avec le débit de crue - thèse (M.SC.A) -
MONTREAL.

21/ TROUDE CECILE (1992) : "Analyse des données" - Cycle S.E.A - pp 1 - 8.

22/ VOLLE.M. (1985) " Analyse des données". Economica - p.p 1 - 25.