ÉCOLE NATIONALE SUPÉRIEURE
D'HYDRAULIQUE DE GRENOBLE

D. DUBAND

HYDROLOGIE
STATISTIQUE APPROFONDIE

1982

INSTITUT NATIONAL POLYTECHNIQUE
DE GRENOBLE

Ce cours n'a pas la prétention de constituer un traité exhaustif de tout ce qui concerne la technique statistique appliquée à l'hydrologie ou étude du cycle de l'eau dans la nature.

Les exposés qui suivent ont été rédigés dans un cadre bien défini qui est celui de la gestion des ressources en eau, au plan opérationnel. Cette projection dans l'avenir, donc dans l'incertain, implique nécessairement :

- la prise en compte de la nature aléatoire du cycle hydrométéorologique qui conditionne le renouvellement spatio-temporel de la ressource en eau ;

- la schématisation des facteurs statiques (topographie, géologie, pédologie, végétation, etc...) et dynamiques (précipitation liquide et solide, rayonnement, température, saturation en humidité de l'air, etc...), ainsi que celle des processus complexes d'évolution des systèmes aquifères en réponse à ces impulsions d'origine atmosphérique, dans un environnement.

L'optimisation de la gestion des ressources en eau, quelles que soient les performances des algorithmes de calcul, n'a de sens que par référence à un <u>état donné</u> de l'information hydrologique : il faut donc extraire le maximum de l'information contenue dans les échantillons de mesures <u>disponibles</u> effectuées sur le terrain. S'appuyant sur des bases physiques simples et éprouvées, l'analyse statistique des données observées se révèle alors utile et pratique.

S O M M A I R E

Introduction

Comment définir la statistique moderne : un ensemble de méthodes pour prendre des décisions raisonnables en présence d'incertitudes.

Cette définition est loin de celle qu'on attribuait à la statistique au siècle dernier et antérieurement, car on considérait alors la statistique comme la science du dénombrement. Il s'agissait d'établir des statistiques, une statistique étant un tableau de chiffres - relevé systématique d'observations concernant un phénomène quelconque : un "état".

Or, depuis le début du siècle, il y a une différence de conception énorme entre cet aspect et la statistique mathématique utilisée comme méthode d'investigation scientifique. Car, alors que cela n'avait pas été le cas avant 1900, malgré la naissance et les développements du calcul des probabilités, ce sont les théoriciens anglo-saxons, dont les chefs de file furent K. PEARSON et FISHER, qui ont mis l'accent sur l'induction, c'est-à-dire : à partir de résultats d'expérience, utiliser des modes de raisonnement permettant de connaître quelque chose de la structure interne des phénomènes. On peut dire que l'hydrologie, c'est-à-dire la connaissance des phénomènes d'écoulement, constitue le champ d'application idéal d'une telle méthode. Il s'agit pour nous d'extraire des inférences valables des séries statistiques de précipitations, de débits et de températures.

o
o o

Quelle démarche suit-on lorsqu'on entreprend une étude statistique ? Il y a 3 phases principales :

1°- Tout d'abord description - consiste à effectuer une mise en ordre de la série de chiffres considérés. Cette mise en ordre, ou classement, qui permet de réduire le tableau des données à un volume de chiffres beaucoup moins important et plus maniable; on condense l'information fournie par ces données à l'aide de quelques graphiques et valeurs types (moyenne, écart type, etc.), techniques multidimensionnelles.

2°- Il s'agit ensuite d'effectuer l'analyse de ces résultats. On s'efforce de les habiller en leur appliquant un modèle probabiliste, c'est-à-dire que l'on essaye de formaliser l'information contenue dans la série par une expression mathématique. Toute théorie ne s'ajustant pas aux faits doit être rejetée.

3°- La Prévision - on projette dans l'avenir le modèle choisi pour pouvoir organiser l'avenir de la façon la plus avantageuse dans le contexte économique considéré, faire des choix rationnels et ainsi guider la décision.

o
o o

Après ces quelques généralités qui n'ont pas la prétention de définir et décrire exhaustivement tout ce en quoi consiste la statistique, mais de donner un aperçu, je vous propose le plan suivant pour la série d'exposés qui constituent l'introduction à la statistique appliquée :

I/ - Les données : leur mise en forme (la statistique descriptive).

II/ - Les modèles probabilistes unidimensionnels

III/ - Application des modèles aux faits observés :
- . échantillonnage
- . estimation
- . les tests

IV/ - L'ajustement d'une relation - les erreurs

V/ - Les liaisons stochastiques :
- . la corrélation entre variables aléatoires
- . l'autocorrélation : corrélation de variables dépendant du temps (la mémoire ou persistance)

VI/ - Analyse factorielle :
- . recherche des composantes principales spatiales et temporelles.

VII/- Analyse spectrale

Dans tout cela on ne peut se passer du calcul des probabilités car la statistique en constitue l'application et le prolongement naturel. La

frontière entre ces deux disciplines n'est pas facile à apprécier. Les théorèmes asymptotiques sont très importants (loi des grands nombres - théorème de Bernouilli - théorème central limite).

Pour terminer cette introduction, je voudrais définir en quelques mots dans quel esprit nous utilisons ces méthodes statistiques pour résoudre les problèmes hydrologiques qui se posent à E.d.F. et dans quel but. Il s'agit d'extraire, des données dont on dispose, la meilleure information et la plus complète possible, dans un temps limité.

Toutes ces études ont en effet pour but une utilisation économique des ressources en eau destinées essentiellement à la production d'énergie électrique : le problème des valeurs extrêmes de crue étant également un problème économique (relatif au génie civil) mais non prévisionnel.

Ainsi, 90 % de l'information accessible par une méthode simple et opérationnelle vaut mieux que 92 % ou même 95 % par une technique apparemment plus raffinée mais souvent hypothétique et surtout plus coûteuse.

Cette approche globale des phénomènes fait ressortir l'ordre de grandeur des influences : les termes négligés ont une influence notablement inférieure à celle des termes retenus, résultats qui peuvent cependant orienter des recherches ultérieures plus poussées vers une connaissance plus exacte des faits.

La statistique mathématique ne peut en aucun cas suppléer au manque ou à l'insuffisance des données d'observation : "on ne peut faire sortir un lapin d'un chapeau"; ce n'est pas une magie.

Dans ce domaine moins qu'ailleurs, il n'y a pas de recettes mais une approche intelligente des problèmes, il faut du bon sens, une certaine largeur d'esprit pour suppléer aux lacunes des outils mathématiques. Il faut se défier des solutions standards; le culte de la formule est nocif.

Enfin, comme toute science expérimentale ou même science dite exacte, la statistique n'est pas neutre : on peut soit induire un résultat en agissant

sur les hypothèses ou données initiales, soit suggérer une interprétation des résultats, sans les tronquer, mais par une présentation appropriée et par certaines omissions.

Il est donc indispensable de bien définir les données (conditions de leur acquisition et mesure), les transformations qu'on leur fait subir : le choix des éléments significatifs d'une réalité et leur organisation selon le champ des décisions à appliquer.

I - LES DONNEES NUMERIQUES - MISE EN FORME

1.1 - Définition du hasard

Le hasard est défini, dans deux dictionnaires classiques, comme "la cause fictive" des évènements apparemment :

- inexplicables, souvent personnifiée (fortune, sort, destin),
- soumis à la seule loi des probabilités.

Les phénomènes naturels observables tels que débits, précipitations, températures de l'air, résultent d'une telle complexité de causes que les grandeurs qui les mesurent en un point peuvent être considérées comme variables aléatoires (variables pouvant prendre un ensemble de valeurs, à chacune desquelles est associée une probabilité).

La théorie des variables aléatoires et des distributions ou "théorie de la probabilité", doit être considérée comme un ensemble de propositions mathématiques établies pour former un modèle des régularités statistiques observées en relation avec des suites de tirages au hasard.

Or, l'expression "au hasard" ne signifie pas n'importe comment : ainsi les chiffres 5, 8, 2, 2, 7, 1, 3, sont-ils au hasard ? cela n'a pas de sens. "Au hasard" n'a de sens scientifique que si l'on se réfère au mécanisme d'obtention de quelque chose et non au résultat : ce mécanisme est-il susceptible d'une schématisation définie par une loi de probabilité (uni ou multidimensionnelle) ?

Le concept de hasard, tel qu'il est défini ici et dans la suite, est lui-même plus restrictif que le concept d'incertitude. Ceci est à l'origine de deux grands courants chez les statisticiens, quant au contenu concret de la théorie de la probabilité :

- pour les empiristes, dans tout phénomène au hasard on peut réaliser ou concevoir une suite d'épreuves, et les fréquences constatées pour les réalisations tendent vers des limites (loi des grands nombres); ainsi, la probabilité est une limite de fréquence ;

- pour les intuitionnistes, les degrés de croyance rationnelle que l'on peut avoir à l'égard de propositions incertaines se combinent entre eux d'une façon dont le calcul des probabilités rend compte, c'est la notion de vraisemblance.

Disposant de l'ensemble des valeurs d'un phénomène : tirer au hasard un échantillon de n valeurs parmi cette population signifie que chaque individu a même chance d'être choisi, et que tous les échantillons possibles d'effectif "n" ont même chance d'être sélectionnés. Pour réaliser cette opération, indépendamment de toute influence humaine ou physique, on utilise les nombres au hasard. Plus généralement, cette branche des mathématiques expérimentales englobe les méthodes de Monte-Carlo, appliquées aussi bien au domaine probabiliste que déterministe.

1.2 - Nombres au hasard

Les tables de nombres au hasard constituent de grands échantillons dérivant de lois de probabilité particulièrement simples. En effet, une suite de chiffres au hasard est engendrée par un mécanisme probabiliste conduisant à l'obtention de chiffres successifs tels, qu'à chaque instant la probabilité qu'un chiffre soit égal à 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 est de une chance sur dix.

Pour fabriquer ces tables de nombres au hasard, on a utilisé divers procédés (à présent au musée) tels que :

- annuaires téléphoniques,
- résultats de tirages de la Loterie Nationale,
- tables de logarithmes (on a extrait les décimales de rang p à p+q), etc.

Des procédés automatiques permettent de générer des nombres au hasard :

- la méthode du carré médian, qui consiste à créer une séquence de nombres en extrayant la partie centrale du nombre précédent élevé au carré, exemple :

$N_1 = 47$ $\qquad (47)^2 = 2209$

$\Downarrow$

$N_2 = 20$ $\qquad (20)^2 = 0400$

$\Downarrow$

$N_3 = 40$ $\qquad \ldots\ldots\ldots\ldots$

Cette technique est maintenant abandonnée pour insuffisance.

- les méthodes de congruence, actuellement les plus utilisées parmi les algorithmes de génération, de nombres pseudo-aléatoires, qui créent une séquence de nombres d'après la relation de récurrence suivante,

$$x_i = a_1\, x_{i-1} + a_2\, x_{i-2} + \ldots + a_p\, x_{i-p} + a_{p+1} \text{ (modulo m)},$$

le choix du paramètre m définit la longueur de la période de cette suite de nombres, les valeurs de a_j et m conditionnent la qualité du "hasard", ce choix dépend aussi du type de calculateur.

Deux méthodes sont plus particulièrement programmées dans les bibliothèques ordinateurs :

. la méthode congruentielle mixte $x_i \equiv a\, x_{i-1} + c$ (modulo m)

. la méthode congruentielle multiplicative $x_i \equiv a\, x_{i-1}$ (modulo m)

où x_i est le reste de la division de $a\, x_{i-1}$ par m, a étant une puissance de 5 par exemple et m une puissance élevée de 2; cette méthode a été testée comme la plus performante (la période est de $\frac{m}{4}$ si "a" diffère de 3 du plus proche multiple de 8 et x_o est impair).

En fait, il est très difficile d'imiter le hasard, bien que ces tables de nombres au hasard soient soumises à une série de tests, exemples :

- la fréquence n_i du chiffre i (0, 1, ..., 9) doit être peu différente de 1/10 sur l'ensemble des N chiffres; on teste la quantité $\sum_{i=0}^{9} \frac{(n_i - N/10)^2}{N/10}$ qui suit une loi du (X^2) à 9 degrés de liberté ;

- test des paires, deux chiffres successifs doivent être indépendants ;

- test des lacunes qui est la fréquence des cas où l'on a 0, 1, 2, ..., n chiffres entre 2 chiffres identiques, etc.

Un biais important peut provenir de périodicités induites dans les séries de chiffres au hasard; pour les détecter, on calcule les coefficients d'autocorrélation. En fait, il y a souvent une période, mais l'important est qu'elle soit très grande (tous les 500 000 chiffres ou plus). Pour les expériences spatiales, la NASA a dû construire des calculateurs spéciaux générateurs de nombres au hasard, dans le but de simuler la fiabilité des appareillages (les probabilités utiles étant très petites).

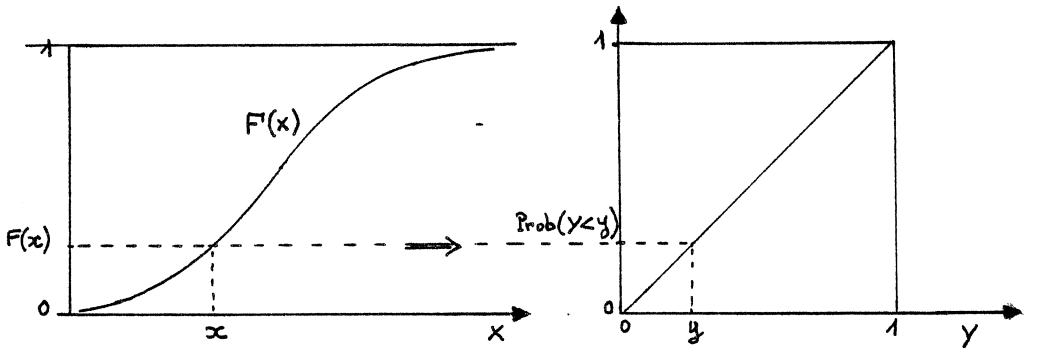
Pour les usages courants, la méthode de congruence suffit.

1.3 - Utilisation des nombres au hasard

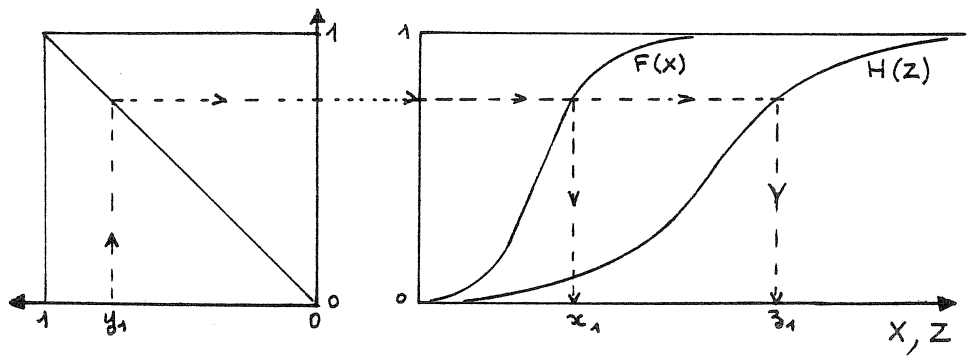
Si l'on considère la variable aléatoire X de fonction de répartition F(x), la nouvelle variable Y = F(x) est uniformément répartie sur le segment [0, 1] :

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

on a réalisé ainsi une anamorphose rectangulaire.



Quelle que soit la variable aléatoire donnée, nous pouvons toujours la ramener à une variable uniformément distribuée sur le segment [0, 1].



Le dessin 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 "plonge" la loi de probabilité F ou H dans l'urne des nombres au hasard, c'est une sorte de filtre, et l'on tire un nombre au hasard y_1 qui, après transformation, fournit la valeur x_1 ou z_1.

En pratique, on remplace la distribution rectangulaire recherchée par une loi discrète dont les sauts valent 10^{-k} (k = 3 ou 4); on associe des ensembles de k chiffres extraits des tables de nombres au hasard.

Exemple -

La fonction de répartition des débits moyens annuels de la LOIRE à Blois étant une loi normale de moyenne 360 et d'écart type 100, on veut générer un échantillon de 10 valeurs de débits moyens annuels. On prendra, par exemple, les 10 derniers nombres de 4 chiffres dans les 4 premières colonnes :

0347	1622
9774	8442
1676	6301
1256	3321
5559	5760

Si les sauts valent 10^{-3}, les valeurs des nombres au hasard sur l'intervalle [0,1] seront :

y_1 = .034	y_6 = .162
y_2 = .977	y_7 = .844
y_3 = .167	y_8 = .630
y_4 = .125	y_9 = .332
y_5 = .555	y_{10} = .576

on cherchera dans la table de la fonction normale $[F(u), u]$ les valeurs centrées réduites de u ($u = \frac{Q - 360}{100}$) qui correspondent aux dix valeurs $F(u) = y$; exemple $y_1 = .977 \Longrightarrow F(u) = .977 \Longrightarrow u = 2 \Longrightarrow Q = 360 + 200 = 560$

1.4 - Réflexion sur les procédés de simulation - intérêt et limites de cette technique

Dans l'exemple ci-dessus, il est évident que si l'on génère un échantillon de milliers de valeurs de débits moyens annuels de la LOIRE à Blois, la moyenne et l'écart type de cette série fictive seront peu différents de 360 et 100, de même la distribution empirique des valeurs classées sera quasiment gaussienne.

Il n'y a pas plus d'information dans l'échantillon simulé que dans l'échantillon des 100 valeurs observées : cette information est simplement présentée avec un fort grossissement. Les conclusions tirées de l'échantillon fictif ne sont légitimes qu'à condition que l'on puisse, théoriquement, extraire les mêmes de la série réelle.

Cependant une telle simulation apporte une information sur les séquences des débits : en particulier cela permet d'étudier comment se comporterait un réservoir d'accumulation annuelle face à une série de 4 ou 5 ans d'apports annuels déficitaires, par exemple.

Dans un cas aussi simple, on peut se poser les questions suivantes :

- la fonction de répartition ajustée à l'échantillon de valeurs annuelles représente-t-elle correctement les valeurs de faibles probabilités ?
- l'échantillon est-il suffisant pour estimer avec une bonne confiance les paramètres de cette fonction ?
- existe-t-il ou non une dépendance entre années successives (mémoire courte) ?
- existe-t-il ou non des accumulations de valeurs fortes ou faibles (mémoire longue) ?
- y a-t-il stationnarité ou non du phénomène dans le temps (influence de l'homme) ?

On peut d'ailleurs s'appuyer sur un phénomène directeur dont on connaît mieux la fonction de répartition (longue période d'observation) et la liaison avec le phénomène que l'on cherche à simuler, comme c'est le cas dans la relation pluie-débit.

Les techniques de simulation apportent toute leur utilité lorsque les phénomènes sont complexes, c'est le cas lorsqu'on cherche à créer artificiellement des débits-pluies-températures, hebdomadaires-journalières-horaires, ponctuellement et régionalement, car l'on doit assurer l'homogénéité et la cohérence des données simulées par référence aux observations réelles, non seulement dans le temps mais aussi dans l'espace. Ce qu'il est alors impossible de calculer analytiquement, lois de probabilité conditionnelles (spatio-temporelles) compliquées et non stationnaires, devient relativement aisé par la méthode de Monte-Carlo.

En résumé, la génération à l'aide de nombres au hasard permet :

- d'étudier les propriétés statistiques (dispersion, distribution) de paramètres de lois de probabilité complexes ;

- de simuler des séries de données hydrologiques donnant le moyen de :

  . caractériser les structures d'équipement, déterminer le dimensionnement optimal des ouvrages de protection et de régulation des eaux (d'après les caractéristiques du régime d'écoulement) ,

  . contrôler les ressources en eau sur un bassin, en particulier de tester les règles d'exploitation d'un réservoir ou d'un système de réservoirs (en tenant compte éventuellement des prévisions calculées) face à l'occurrence d'évènements extrêmes ou à la conjonction d'évènement particuliers dans le temps et l'espace, avec différentes hypothèses économiques. La simulation de plusieurs milliers de données permet de lisser les résultats que l'on aurait obtenus avec de courtes séries d'observations. Cette technique est ainsi très utilisée en recherche opérationnelle.

1.5 - Distribution empirique

La série des débits moyens annuels de la LOIRE à Blois forme une suite d'observations qui peuvent être considérées comme des tirages dans l'urne des modules de la LOIRE : F(Q).

Implicitement on admet que l'une est stable - il n'y a pas d'évolution de climat (à l'échelle séculaire) - et formée d'une population infinie de débits annuels.

On classe les n valeurs de débit observé dans l'ordre croissant, par exemple : i = 1, 2, ... n.

On peut considérer les n valeurs comme n points sur un axe. La distribution de l'échantillon sera définie en affectant une masse de $\frac{1}{n}$ à chaque point.

C'est une fonction en escalier. $F^{*}(x_i) = \frac{c}{n}$ pourcentage des valeurs $< x$.

Dans la pratique, on affecte $\frac{2i - 1}{2n}$ à l'observation de rang i qui est le centre de la marche de hauteur $\frac{1}{n}$. L'avantage de cette représentation est qu'il permet un lissage de la représentation en "escalier".

En fait il existe plusieurs modes de représentation découlant de la formule générale :

$$\frac{i - a}{n + b}$$

avec a = 0 ; b = 0
a = 0 ; b = 1 (Gumbel)
a = .5 ; b = 0
a = .375 ; b = .25 (Gauss)
a = .44 ; b = .12 (Gringorten pour les valeurs extrêmes)
a = -.38 ; b = -.31 (Chegodaxev - valeurs extrêmes)

Ces formules étant plus ou moins justifiées par des considérations théoriques, lorsque n est de l'ordre de quelques dizaines elles sont très voisines. Pour notre part nous avons choisi la représentation a = .5 ; b = 0.

Avec 30 ou 50 valeurs il est peu réaliste de tracer un histogramme de fréquence (on aurait des classes d'effectif 0, 1, 2, 3), le choix de l'amplitude de classes est arbitraire ainsi que celui de l'origine de celles-ci; en hydrologie on utilise généralement les fréquences cumulées (exemple de la LOIRE à Blois), excepté dans le cas des précipitations journalières pour lesquelles on dispose d'un grand nombre d'observations indépendantes, 1000 à 4000.

Exercice : Construire la distribution empirique du débit moyen annuel de la LOIRE à Blois de 1863 à 1887.

On passe intuitivement à la notion de courbe de fréquence, dans le cas présent de fonction de répartition, en imaginant des intervalles de classes de plus en plus étroits, n augmentant.

Le théorème de Bernouilli est le pont entre fréquence et probabilité. $F^{*}(x)$ est l'image statistique de $F(x)$; ε et δ étant aussi petits que l'on veut, il y a une probabilité δ pour que: $\text{Prob}\left\{ \left| F^{*}(x) - F(x) \right| > \varepsilon \right\} < \delta$

Exemple de la LOIRE à Blois : série de 100 ans.

1.6 - Synthèse des distributions statistiques

On cherche à résumer et réduire par un, deux, trois paramètres simples l'information contenue dans les n valeurs de l'échantillon $(x_1, x_2, \ldots x_n)$.

On utilise habituellement pour caractériser :

- la tendance centrale

  . la moyenne arithmétique simple $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \Longrightarrow$ moment d'ordre 1,

  $E(x) = \int_{\alpha 1}^{\alpha 2} x\, f(x)\, dx$

. la moyenne arithmétique pondérée $\bar{\bar{x}} = \frac{\Sigma \lambda_i x_i}{\Sigma \lambda_i}$, λ_i étant le poids affecté à x_i

. la moyenne géométrique g : $\log g = \frac{1}{n} \sum_{i=1}^{n} \log x_i$

. la moyenne harmonique h : $\frac{1}{h} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i}$

. la moyenne quadratique q : $q = \sqrt{\frac{1}{n} \Sigma x_i^2}$

. le mode x_{M_o}, valeur pour laquelle la fréquence est maximale

. la médiane x_{50}, telle que $F^{\#}(x_{50}) < 50\ \%$

- la dispersion ou variabilité du phénomène

. la variance $s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \Longrightarrow V(x) = \int_{\alpha 1}^{\alpha 2} x^2 f(x)\, dx - \left[\int_{\alpha 1}^{\alpha 2} x f(x)\, dx \right]^2$ ou l'écart type s

. une différence de quantiles $\frac{x_{90} - x_{10}}{2}$

Rappelons que le quantile x_p de la fonction de répartition $F(x)$ est défini par $F(x_p) = P\ \%$

Nous n'utilisons guère les moments centrés d'ordre supérieur à 2 car le poids des valeurs extrêmes devient alors prépondérant surtout dans le cas de petits échantillons.

Ces moments servent à calculer :

- le coefficient de dissymétrie : $\beta_1 = \frac{\left\{ \sum_{i=1}^{n} \left[(x_i - \bar{x})^3 \right] \right\}^2}{\left\{ \sum_{i=1}^{n} \left[(x_i - \bar{x})^2 \right] \right\}^3}$ ou $\sqrt{\beta_1}$

qui s'annule lorsqu'il y a symétrie; on peut également utiliser le rapport des quantiles : $\frac{x_{90} - x_{50}}{x_{50} - x_{10}}$, qui est :

$\left\{ \begin{array}{l} > 1 \text{ si la dissymétrie est positive} \\ = 1 \text{ s'il y a symétrie} \\ < 1 \text{ si la dissymétrie est négative} \end{array} \right.$

- le coefficient d'aplatissement : $\beta_2 = \dfrac{\sum_{i=1}^{n} (x_i - \overline{x})^4}{\left\{ \sum_{i=1}^{n} \left[(x_i - \overline{x})^2 \right] \right\}^2} - 3$

qui est égal à 0 dans le cas d'une distribution normale.

Pour les débits on utilise parfois :

$\text{Log} \dfrac{x_{90}}{x_{10}}$ comme paramètre de dispersion

$\text{Log} \dfrac{x_{90} \cdot x_{10}}{x_{50}}$ comme paramètre de dissymétrie.

Le coefficient de variation $c_v = \dfrac{s}{\overline{x}}$ est indépendant des unités et permet de comparer deux distributions, exemple :

. pour les débits moyens annuels de la LOIRE : $c_v = .30$

. pour les débits moyens du 25 octobre de la LOIRE : $c_v = 1.43$

. pour les débits moyens d'octobre de la LOIRE : $c_v = .9$

on pourra aussi utiliser : $\dfrac{x_{90} - x_{10}}{x_{90} + x_{10}}$.

A l'aide des distributions empiriques et sans faire d'hypothèse mathématique, on peut déjà disposer d'une information fort utile.

Les exemples de la répartition des débits moyens journaliers pour la LOIRE à Blois - la ROMANCHE au Chambon - le DRAC au Sautet, l'illustrent bien.

Sans parler de l'intérêt hydrologique qui permet de caractériser de façon simple trois régimes différents : pluvial - nival - pluvio-nival, ces renseignements peuvent être fort utiles, par exemple, à un Service d'Exploitation qui projette d'effectuer des travaux en rivière un an à l'avance, donc sans possibilité d'utiliser de prévision. Il est possible de choisir la période durant laquelle les débits ont une certaine fréquence d'être ou de ne pas être dépassés.

1.7 - Le contrôle et la critique des données

C'est la partie ingrate et la phase préliminaire de toute étude statistique. On peut affirmer sans exagération que 30 à 50 % du travail consiste à critiquer les séries de données que l'on utilisera dans le calcul.

On n'insistera jamais assez sur cet aspect, car ce n'est pas un travail "noble"; de plus, la compilation de chiffres est rebutante, fastidieuse. Il n'y a pourtant pas de meilleur moyen d'acquérir la notion des ordres de grandeur. De ce contrôle dépend toute la suite de l'étude.

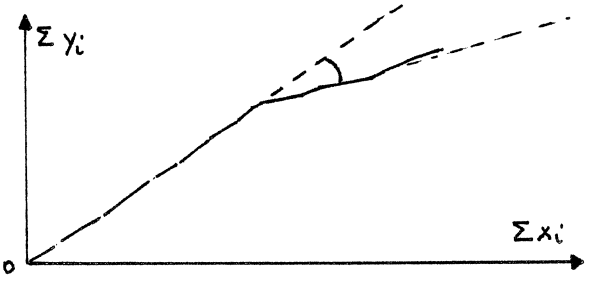
Types d'erreurs fréquentes :

- erreurs aléatoires : erreurs de mesure, erreurs de transcription de données,
- erreurs systématiques : changement d'appareil ou de station à partir d'une date.

On détecte ces erreurs, soit en étudiant la continuité des mesures, soit par comparaison avec des stations voisines.

Dans ce dernier cas on utilisera le contrôle par corrélation.

Une méthode rapide, si l'on dispose d'une bonne série de référence et d'une série dont on veut tester l'homogénéité, consiste à tracer la ligne des valeurs cumulées : $\sum_{1}^{n} x_i$ en fonction de $\sum_{1}^{n} y_i$, s'il y a une hétérogénéité systématique on constate une cassure très nette :



Nous reviendrons sur ce contrôle lors du cours sur la corrélation.

1.8 - Graphiques à échelles fonctionnelles

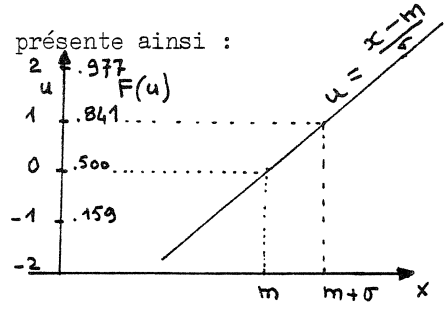
Soit x une variable aléatoire et F(x) sa fonction de répartition, la représentation graphique $\left[x, F(x)\right]$ en coordonnées naturelles est généralement une courbe en S continue.

On a construit des graphiques pour les fonctions les plus courantes : normale, log normale, doublement exponentielle (Gumbel), tels que la représentation $\left[x, F(x)\right]$ soit linéaire.

Quand on étudie la distribution empirique des n valeurs d'un échantillon, on peut ainsi s'assurer visuellement que la fonction choisie pour représenter cet échantillon, convient ou non, et éventuellement l'ajuster (il suffit de 2 points) et estimer ainsi graphiquement les paramètres de cette fonction de répartition.

Le "papier" gausso-arithmétique se présente ainsi :

en ordonnée, l'échelle est arithmétique en variable normale centrée réduite et, en abscisse, l'échelle est arithmétique (exemple de la LOIRE).



Le papier gausso-logarithme a la même échelle des ordonnées, et l'échelle des abscisses a une graduation logarithmique.

Cette présentation dilate la zone centrale (entre 10 % et 90 %).

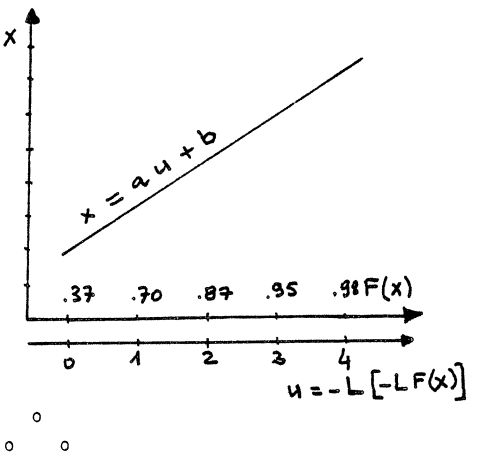
Le "papier" Gumbel se présente ainsi : en abscisse l'échelle est arithmétique en $-L\left[-L\ F(x)\right]$, en ordonnée l'échelle est arithmétique.

Sur le graphique, la fonction $F(x) = e^{-e^{-(\alpha x + \beta)}}$ est une droite (toute fonction en exponentielle simple a une représentation linéaire dans la zone des fortes probabilités).

Cette représentation a pour avantage de dilater l'échelle dans les fortes probabilités, on peut noter que pour F(x) voisin de 1 :

$-L\left[-L\,F(x)\right] \simeq LT$ avec $T = \frac{1}{1-F}$

(exemple des précipitations journalières à Orcière)



BIBLIOGRAPHIE

J. BASS - Eléments de calcul des probabilités (Masson)

GNEDENKO et KHINTCHINE - Introduction à la théorie des probabilités (monographie Dunod)

MORICE et CHARTIER - Méthode statistique, tome I, INSEE (Imprimerie Nationale)

KENDALL and STUART - The advanced theory of statistics, tome I (Griffin)

TORTRAT - Principes de Statistique Mathématique (Dunod)

J.M. HAMMERSLEY and D.C. HANDSCOMB - Monte-Carlo Methods (Methuen's Monographs).

TABLEAUX ET FIGURES DU CHAPITRE I (Pour travaux dirigés)

TABLE XXXIII. RANDOM NUMBERS (I)

03	47	43	73	86	36	96	47	36	61	46	98	63	71	62	33	26	16	80	45	60	11	14	10	95
97	74	24	67	62	42	81	14	57	20	42	53	32	37	32	27	07	36	07	51	24	51	79	89	73
16	76	62	27	66	56	50	26	71	07	32	90	79	78	53	13	55	38	58	59	88	97	54	14	10
12	56	85	99	26	96	96	68	27	31	05	03	72	93	15	57	12	10	14	21	88	26	49	81	76
55	59	56	35	64	38	54	82	46	22	31	62	43	09	90	06	18	44	32	53	23	83	01	30	30
16	22	77	94	39	49	54	43	54	82	17	37	93	23	78	87	35	20	96	43	84	26	34	91	64
84	42	17	53	31	57	24	55	06	88	77	04	74	47	67	21	76	33	50	25	83	92	12	06	76
63	01	63	78	59	16	95	55	67	19	98	10	50	71	75	12	86	73	58	07	44	39	52	38	79
33	21	12	34	29	78	64	56	07	82	52	42	07	44	38	15	51	00	13	42	99	66	02	79	54
57	60	86	32	44	09	47	27	96	54	49	17	46	09	62	90	52	84	77	27	08	02	73	43	28
18	18	07	92	46	44	17	16	58	09	79	83	86	19	62	06	76	50	03	10	55	23	64	05	05
26	62	38	97	75	84	16	07	44	99	83	11	46	32	24	20	14	85	88	45	10	93	72	88	71
23	42	40	64	74	82	97	77	77	81	07	45	32	14	08	32	98	94	07	72	93	85	79	10	75
52	36	28	19	95	50	92	26	11	97	00	56	76	31	38	80	22	02	53	53	86	60	42	04	53
37	85	94	35	12	83	39	50	08	30	42	34	07	96	88	54	42	06	87	98	35	85	29	48	39
70	29	17	12	13	40	33	20	38	26	13	89	51	03	74	17	76	37	13	04	07	74	21	19	30
56	62	18	37	35	96	83	50	87	75	97	12	25	93	47	70	33	24	03	54	97	77	46	44	80
99	49	57	22	77	88	42	95	45	72	16	64	36	16	00	04	43	18	66	79	94	77	24	21	90
16	08	15	04	72	33	27	14	34	09	45	59	34	68	49	12	72	07	34	45	99	27	72	95	14
31	16	93	32	43	50	27	89	87	19	20	15	37	00	49	52	85	66	60	44	38	68	88	11	80
68	34	30	13	70	55	74	30	77	40	44	22	78	84	26	04	33	46	09	52	68	07	97	06	57
74	57	25	65	76	59	29	97	68	60	71	91	38	67	54	13	58	18	24	76	15	54	55	95	52
27	42	37	86	53	48	55	90	65	72	96	57	69	36	10	96	46	92	42	45	97	60	49	04	91
00	39	68	29	61	66	37	32	20	30	77	84	57	03	29	10	45	65	04	26	11	04	96	67	24
29	94	98	94	24	68	49	69	10	82	53	75	91	93	30	34	25	20	57	27	40	48	73	51	92
16	90	82	66	59	83	62	64	11	12	67	19	00	71	74	60	47	21	29	68	02	02	37	03	31
11	27	94	75	06	06	09	19	74	66	02	94	37	34	02	76	70	90	30	86	38	45	94	30	38
35	24	10	16	20	33	32	51	26	38	79	78	45	04	91	16	92	53	56	16	02	75	50	95	98
38	23	16	86	38	42	38	97	01	50	87	75	66	81	41	40	01	74	91	62	48	51	84	08	32
31	96	25	91	47	96	44	33	49	13	34	86	82	53	91	00	52	43	48	85	27	55	26	89	62
66	67	40	67	14	64	05	71	95	86	11	05	65	09	68	76	83	20	37	90	57	16	00	11	66
14	90	84	45	11	75	73	88	05	90	52	27	41	14	86	22	98	12	22	08	07	52	74	95	80
68	05	51	18	00	33	96	02	75	19	07	60	62	93	55	59	33	82	43	90	49	37	38	44	59
20	46	78	73	90	97	51	40	14	02	04	02	33	31	08	39	54	16	49	36	47	95	93	13	30
64	19	58	97	79	15	06	15	93	20	01	90	10	75	06	40	78	78	89	62	02	67	74	17	33
05	26	93	70	60	22	35	85	15	13	92	03	51	59	77	59	56	78	06	83	52	91	05	70	74
07	97	10	88	23	09	98	42	99	64	61	71	62	99	15	06	51	29	16	93	58	05	77	09	51
68	71	86	85	85	54	87	66	47	54	73	32	08	11	12	44	95	92	63	16	29	56	24	29	48
26	99	61	65	53	58	37	78	80	70	42	10	50	67	42	32	17	55	85	74	94	44	67	16	94
14	65	52	68	75	87	59	36	22	41	26	78	63	06	55	13	08	27	01	50	15	29	39	39	43
17	53	77	58	71	71	41	61	50	72	12	41	94	96	26	44	95	27	36	99	02	96	74	30	83
90	26	59	21	19	23	52	23	33	12	96	93	02	18	39	07	02	18	36	07	25	99	32	70	23
41	23	52	55	99	31	04	49	69	96	10	47	48	45	88	13	41	43	89	20	97	17	14	49	17
60	20	50	81	69	31	99	73	68	68	35	81	33	03	76	24	30	12	48	60	18	99	10	72	34
91	25	38	05	90	94	58	28	41	36	45	37	59	03	09	90	35	57	29	12	82	62	54	65	60
34	50	57	74	37	98	80	33	00	91	09	77	93	19	82	74	94	80	04	04	45	07	31	66	49
85	22	04	39	43	73	81	53	94	79	33	62	46	86	28	08	31	54	46	31	53	94	13	38	47
09	79	13	77	48	73	82	97	22	21	05	03	27	24	83	72	89	44	05	60	35	80	39	94	88
88	75	80	18	14	22	95	75	42	49	39	32	82	22	49	02	48	07	70	37	16	04	61	67	87
90	96	23	70	00	39	00	03	06	90	55	85	78	38	36	94	37	30	69	32	90	89	00	76	33

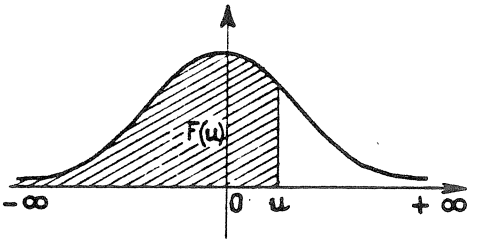
TABLE XXXIII. RANDOM NUMBERS (II)

53	74	23	99	67	61	32	28	69	84	94	62	67	86	24	98	33	41	19	95	47	53	53	38	09
63	38	06	86	54	99	00	65	26	94	02	82	90	23	07	79	62	67	80	60	75	91	12	81	19
35	30	58	21	46	06	72	17	10	94	25	21	31	75	96	49	28	24	00	49	55	65	79	78	07
63	43	36	82	69	65	51	18	37	88	61	38	44	12	45	32	92	85	88	65	54	34	81	85	35
98	25	37	55	26	01	91	82	81	46	74	71	12	94	97	24	02	71	37	07	03	92	18	66	75
02	63	21	17	69	71	50	80	89	56	38	15	70	11	48	43	40	45	86	98	00	83	26	91	03
64	55	22	21	82	48	22	28	06	00	61	54	13	43	91	82	78	12	23	29	06	66	24	12	27
85	07	26	13	89	01	10	07	82	04	59	63	69	36	03	69	11	15	83	80	13	29	54	19	28
58	54	16	24	15	51	54	44	82	00	62	61	65	04	69	38	18	65	18	97	85	72	13	49	21
34	85	27	84	87	61	48	64	56	26	90	18	48	13	26	37	70	15	42	57	65	65	80	39	07
03	92	18	27	46	57	99	16	96	36	30	33	72	85	22	84	64	38	56	98	99	01	30	98	64
62	95	30	27	59	37	75	41	66	48	86	97	80	61	45	23	53	04	01	63	45	76	08	64	27
08	45	93	15	22	60	21	75	46	91	98	77	27	85	42	28	88	61	08	84	69	62	03	42	73
07	08	55	18	40	45	44	75	13	90	24	94	96	61	02	57	55	66	83	15	73	42	37	11	61
01	85	89	95	66	51	10	19	34	88	15	84	97	19	75	12	76	39	43	78	64	63	91	08	25
72	84	71	14	35	19	11	58	49	26	50	11	17	17	76	86	31	57	20	18	95	60	78	46	75
88	78	28	16	84	13	52	53	94	53	75	45	69	30	96	73	89	65	70	31	99	17	43	48	70
45	17	75	65	57	28	40	19	72	12	25	12	74	75	67	60	40	60	81	19	24	62	01	61	16
96	76	28	12	54	22	01	11	94	25	71	96	16	16	88	68	64	36	74	45	19	59	50	88	92
43	31	67	72	30	24	02	94	08	63	38	32	36	66	02	69	36	38	25	39	48	03	45	15	22
50	44	66	44	21	66	06	58	05	62	68	15	54	35	02	42	35	48	96	32	14	52	41	52	48
22	66	22	15	86	26	63	75	41	99	58	42	36	72	24	58	37	52	18	51	03	37	18	39	11
96	24	40	14	51	23	22	30	88	57	95	67	47	29	83	94	69	40	06	07	18	16	36	78	86
31	73	91	61	19	60	20	72	93	48	98	57	07	23	69	65	95	39	69	58	56	80	30	19	44
78	60	73	99	84	43	89	94	36	45	56	69	47	07	41	90	22	91	07	12	78	35	34	08	72
84	37	90	61	56	70	10	23	98	05	85	11	34	76	60	76	48	45	34	60	01	64	18	39	96
36	67	10	08	23	98	93	35	08	86	99	29	76	29	81	33	34	91	58	93	63	14	52	32	52
07	28	59	07	48	89	64	58	89	75	83	85	62	27	89	30	14	78	56	27	86	63	59	80	02
10	15	83	87	60	79	24	31	66	56	21	48	24	06	93	91	98	94	05	49	01	47	59	38	00
55	19	68	97	65	03	73	52	16	56	00	53	55	90	27	33	42	29	38	87	22	13	88	83	34
53	81	29	13	39	35	01	20	71	34	62	33	74	82	14	53	73	19	09	03	56	54	29	56	93
51	86	32	68	92	33	98	74	66	99	40	14	71	94	58	45	94	19	38	81	14	44	99	81	07
35	91	70	29	13	80	03	54	07	27	96	94	78	32	66	50	95	52	74	33	13	80	55	62	54
37	71	67	95	13	20	02	44	95	94	64	85	04	05	72	01	32	90	76	14	53	89	74	60	41
93	66	13	83	27	92	79	64	64	72	28	54	96	53	84	48	14	52	98	94	56	07	93	89	30
02	96	08	45	65	13	05	00	41	84	93	07	54	72	59	21	45	57	09	77	19	48	56	27	44
49	83	43	48	35	82	88	33	69	96	72	36	04	19	76	47	45	15	18	60	82	11	08	95	97
84	60	71	62	46	40	80	81	30	37	34	39	23	05	38	25	15	35	71	30	88	12	57	21	77
18	17	30	88	71	44	91	14	88	47	89	23	30	63	15	56	34	20	47	89	99	82	93	24	98
79	69	10	61	78	71	32	76	95	62	87	00	22	58	40	92	54	01	75	25	43	11	71	99	31
75	93	36	57	83	56	20	14	82	11	74	21	97	90	65	96	42	68	63	86	74	54	13	26	94
38	30	92	29	03	06	28	81	39	38	62	25	06	84	63	61	29	08	93	67	04	32	92	08	00
51	29	50	10	34	31	57	75	95	80	51	97	02	74	77	76	15	48	49	44	18	55	63	77	09
21	31	38	86	24	37	79	81	53	74	73	24	16	10	33	52	83	90	94	76	70	47	14	54	36
29	01	23	87	88	58	02	39	37	67	42	10	14	20	92	16	55	23	42	45	54	96	09	11	06
95	33	95	22	00	18	74	72	00	18	38	79	58	69	32	81	76	80	26	92	82	80	84	25	39
90	84	60	79	80	24	36	59	87	38	82	07	53	89	35	96	35	23	79	18	05	98	90	07	35
46	40	62	98	82	54	97	20	56	95	15	74	80	08	32	16	46	70	50	80	67	72	16	42	79
20	31	89	03	43	38	46	82	68	72	32	14	82	99	70	80	60	47	18	97	63	49	30	21	30
71	59	73	05	50	08	22	23	71	77	91	01	93	20	49	82	96	59	26	94	66	10	67	08	60

TABLE 2-1

FONCTION DE REPARTITION DE LA LOI NORMALE REDUITE

(Probabilité de trouver une valeur inférieure à u)



u	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7290	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9779	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

Table pour les grandes valeurs de u

u	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,8	4,0	4,5
F(u)	0,99865	0,99904	0,99931	0,99952	0,99966	0,99976	0,999841	0,999928	0,999968	0,999997

Nota :- La table donne les valeurs de F(u) pour u positif. Lorsque u est négatif il faut prendre le complément à l'unité de la valeur lue dans la table.

Exemple . pour u = 1,37 F(u) = 0,9147
pour u = - 1,37 F(u) = 0,0853

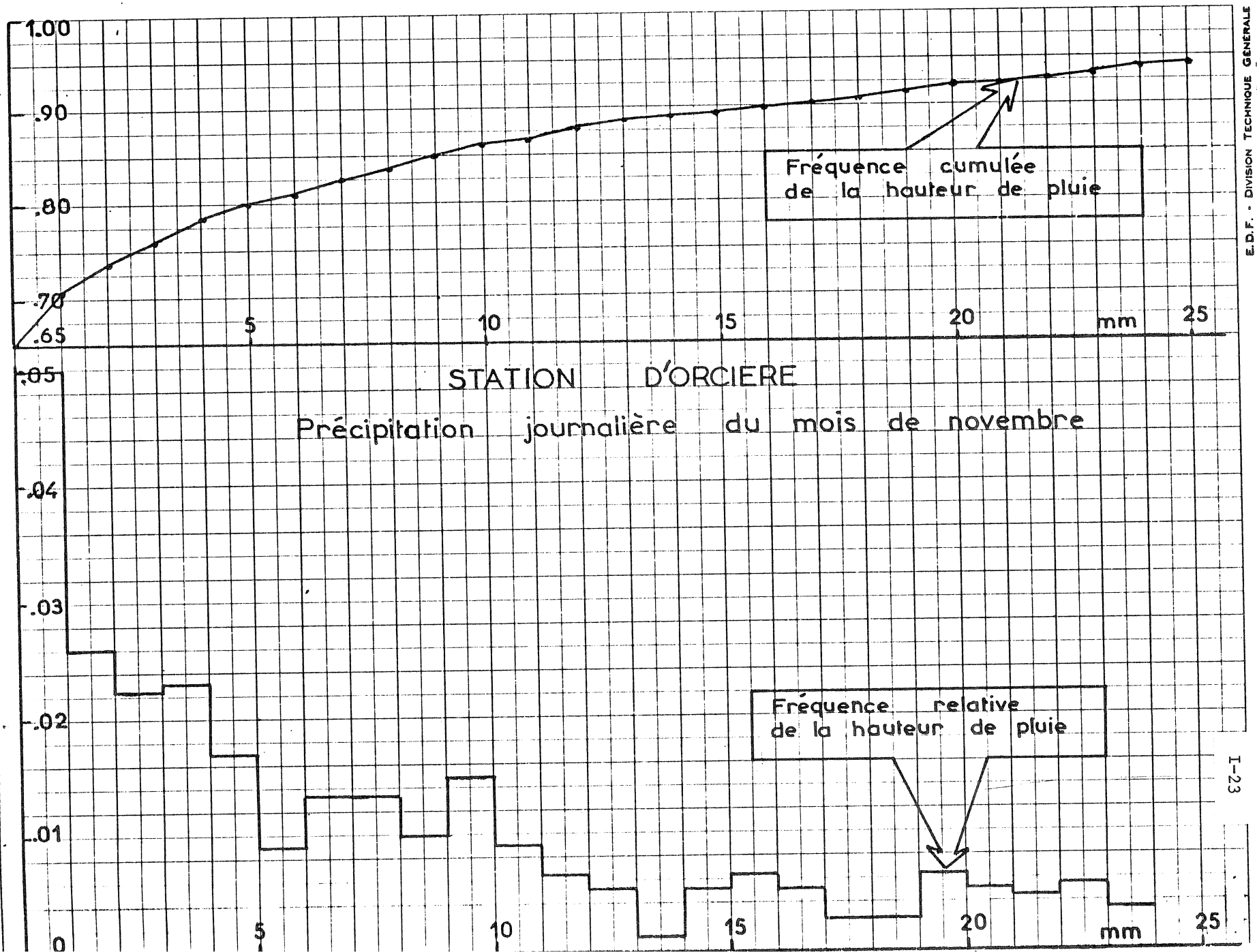
Précipitations journalières du mois de NOVEMBRE (1931 - 1967)

ORCIERES

MOYENNE 4.713 ECART TYPE 11.501

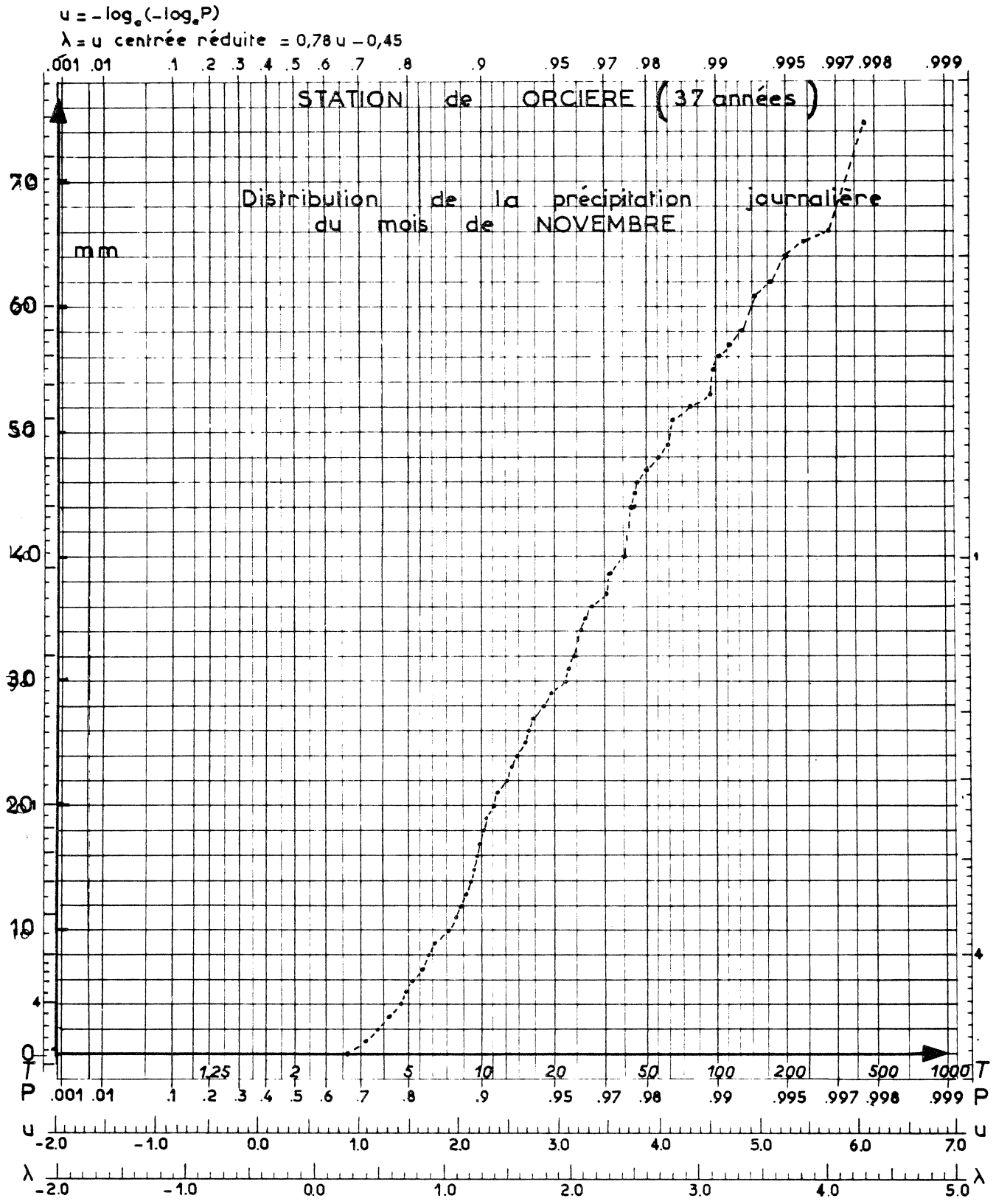
NO	LIM.SUP	N	FREQUENCE	F.CUMULEE		
1	.00	732	65.95	65.95		
2	1.00	56	5.05	70.99		
3	2.00	29	2.61	73.60		
4	3.00	25	2.25	75.86		
5	4.00	26	2.34	78.20		
6	5.00	19	1.71	79.91		
7	6.00	10	.90	80.81		
8	7.00	15	1.35	82.16		
9	8.00	15	1.35	83.51		
10	9.00	11	.99	84.50		
11	10.00	17	1.53	86.04		
12	11.00	10	.90	86.94		
13	12.00	7	.63	87.57		
14	13.00	6	.54	88.11		
15	14.00	1	.09	88.20		
16	15.00	6	.54	88.74		
17	16.00	7	.63	89.37		
18	17.00	6	.54	89.91		
19	18.00	3	.27	90.18		
20	19.00	3	.27	90.45		
21	20.00	7	.63	91.08		
22	21.00	6	.54	91.62		
23	22.00	5	.45	92.07		
24	23.00	6	.54	92.61		
25	24.00	4	.36	92.97		
26	25.00	6	.54	93.51		
27	26.00	2	.18	93.69		
28	27.00	3	.27	93.96	99.64	
29	28.00	5	.45	94.41	99.73	
30	29.00	6	.54	94.95	99.82	
31	30.00	6	.54	95.50	99.91	
32	31.00	3	.27	95.77	100.00	
33	32.00	2	.18	95.95		
35	34.00	1	.09	96.04	.09	
36	35.00	4	.36	96.40	.09	
37	36.00	3	.27	96.67	.09	
38	37.00	5	.45	97.12	.09	
40	39.00	1	.09	97.21	.09	
41	40.00	4	.36	97.57		
45	44.00	1	.09	97.66	1	
46	45.00	1	.09	97.75	1	
47	46.00	1	.09	97.84	1	
48	47.00	2	.18	98.02	1	
49	48.00	2	.18	98.20		
50	49.00	2	.18	98.38		
52	51.00	1	.09	98.47		
53	52.00	3	.27	98.74		
54	53.00	2	.18	98.92		
56	55.00	1	.09	99.01		
57	56.00	1	.09	99.10		
58	57.00	1	.09	99.19		
59	58.00	1	.09	99.28		
62	61.00	1	.09	99.37		
63	62.00	1	.09	99.46		
65	64.00	1	.09	99.55		

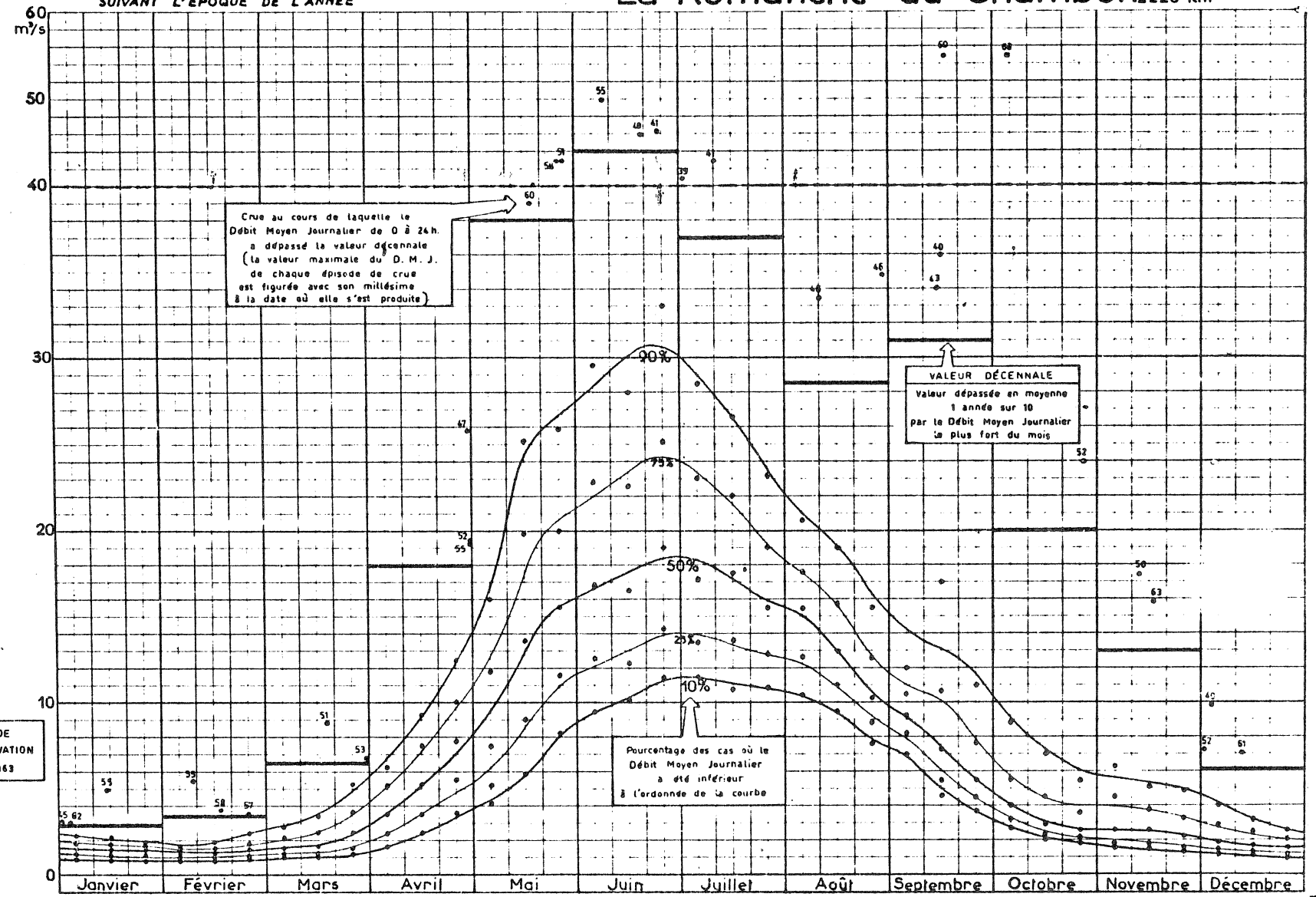
66	65.00
67	66.00
71	75.00
73	85.00
78	120.00



E.D.F. - DIVISION TECHNIQUE GÉNÉRALE

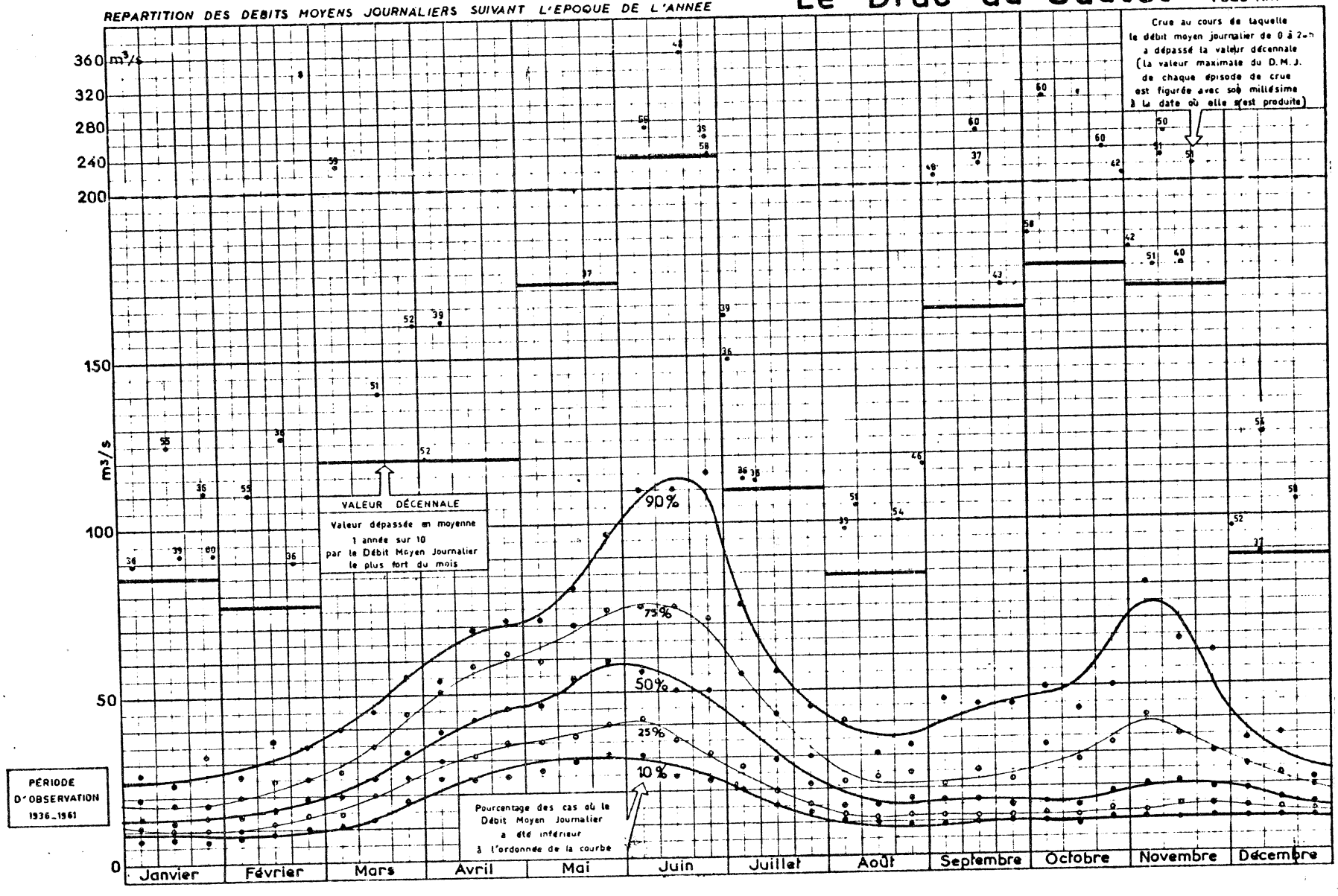
I-23


u = - log. (- log. P)
λ = u centrée réduite = 0,78 u - 0,45
STATION de ORCIERE (37 années)
Distribution de la précipitation journalière du mois de NOVEMBRE
mm
T
P
u
λ


RÉPARTITION DES DÉBITS MOYENS JOURNALIERS
SUIVANT L'ÉPOQUE DE L'ANNÉE
La Romanche au Chambon_220 km²
60 m³/s
50
40
30
20
10
0
Janvier
Février
Mars
Avril
Mai
Juin
Juillet
Août
Septembre
Octobre
Novembre
Décembre
90%
75%
50%
25%
10%
Crue au cours de laquelle le Débit Moyen Journalier de 0 à 24 h. a dépassé la valeur décennale (la valeur maximale du D. M. J. de chaque épisode de crue est figurée avec son millésime à la date où elle s'est produite)
VALEUR DÉCENNALE
Valeur dépassée en moyenne 1 année sur 10 par le Débit Moyen Journalier le plus fort du mois
Pourcentage des cas où le Débit Moyen Journalier a été inférieur à l'ordonnée de la courbe
PÉRIODE D'OBSERVATION 1939_1963

Le Drac au Sautet 1020 km²

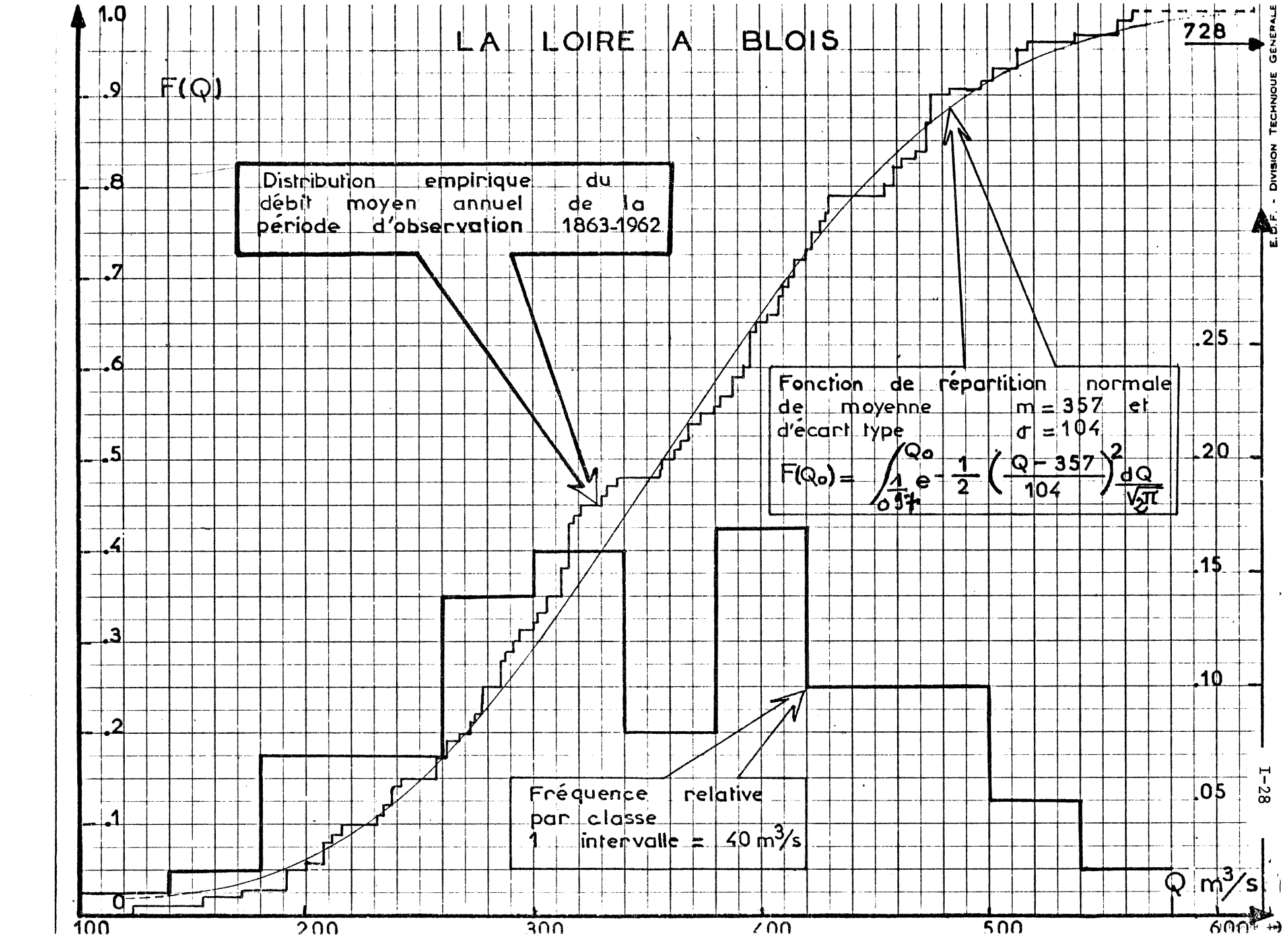
REPARTITION DES DEBITS MOYENS JOURNALIERS SUIVANT L'EPOQUE DE L'ANNEE

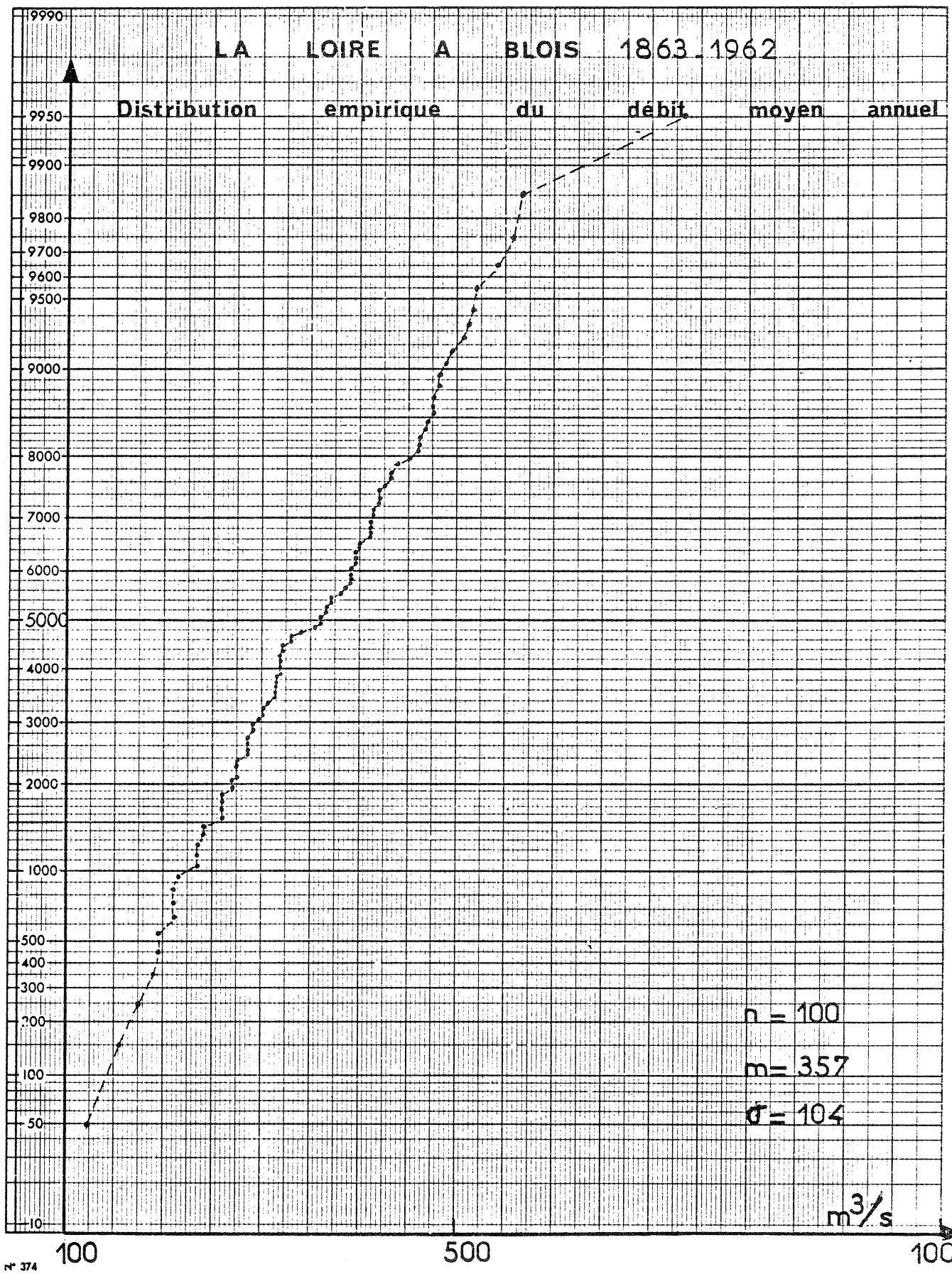


DEBITS MOYENS ANNUELS

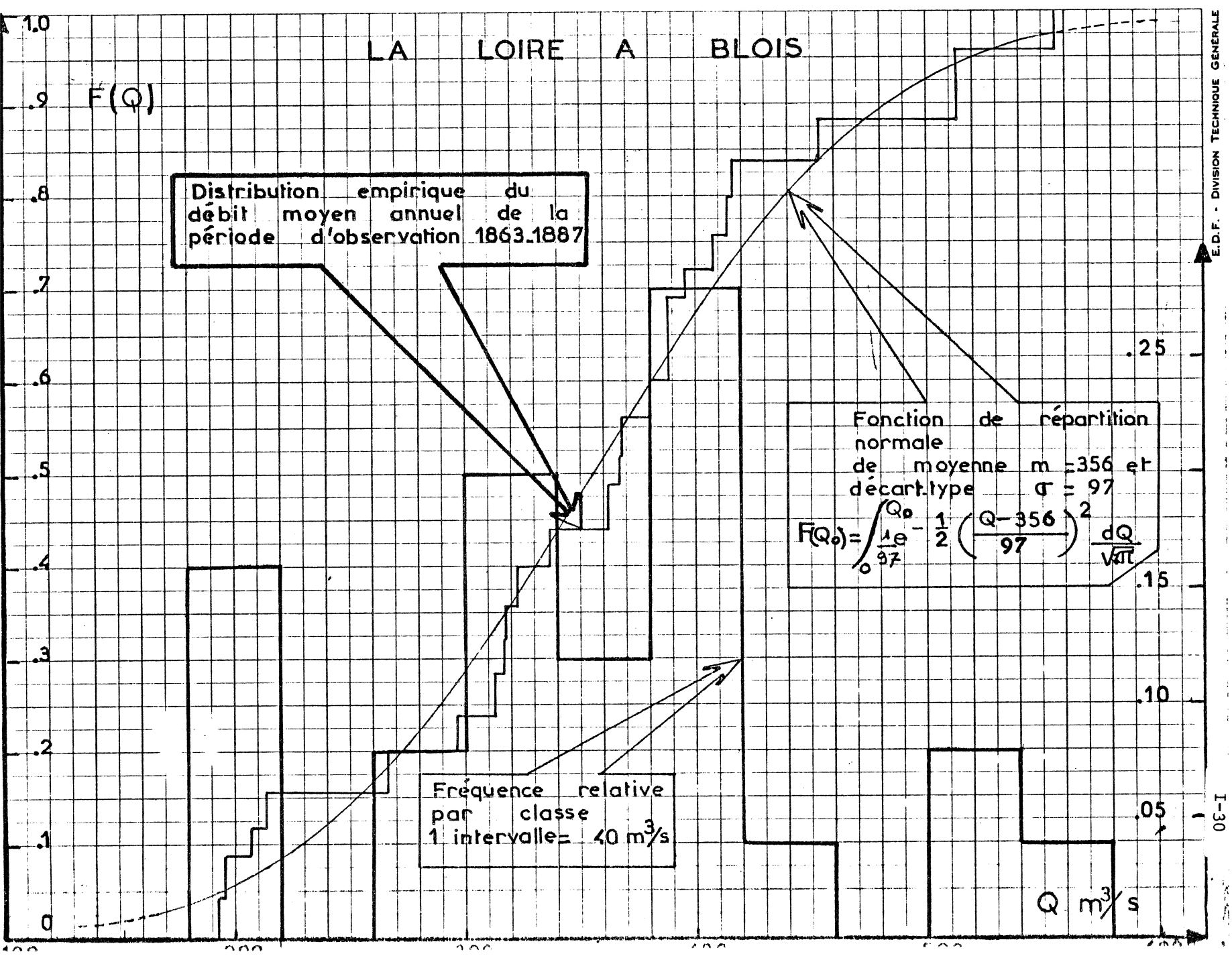
La LOIRE à BLOIS

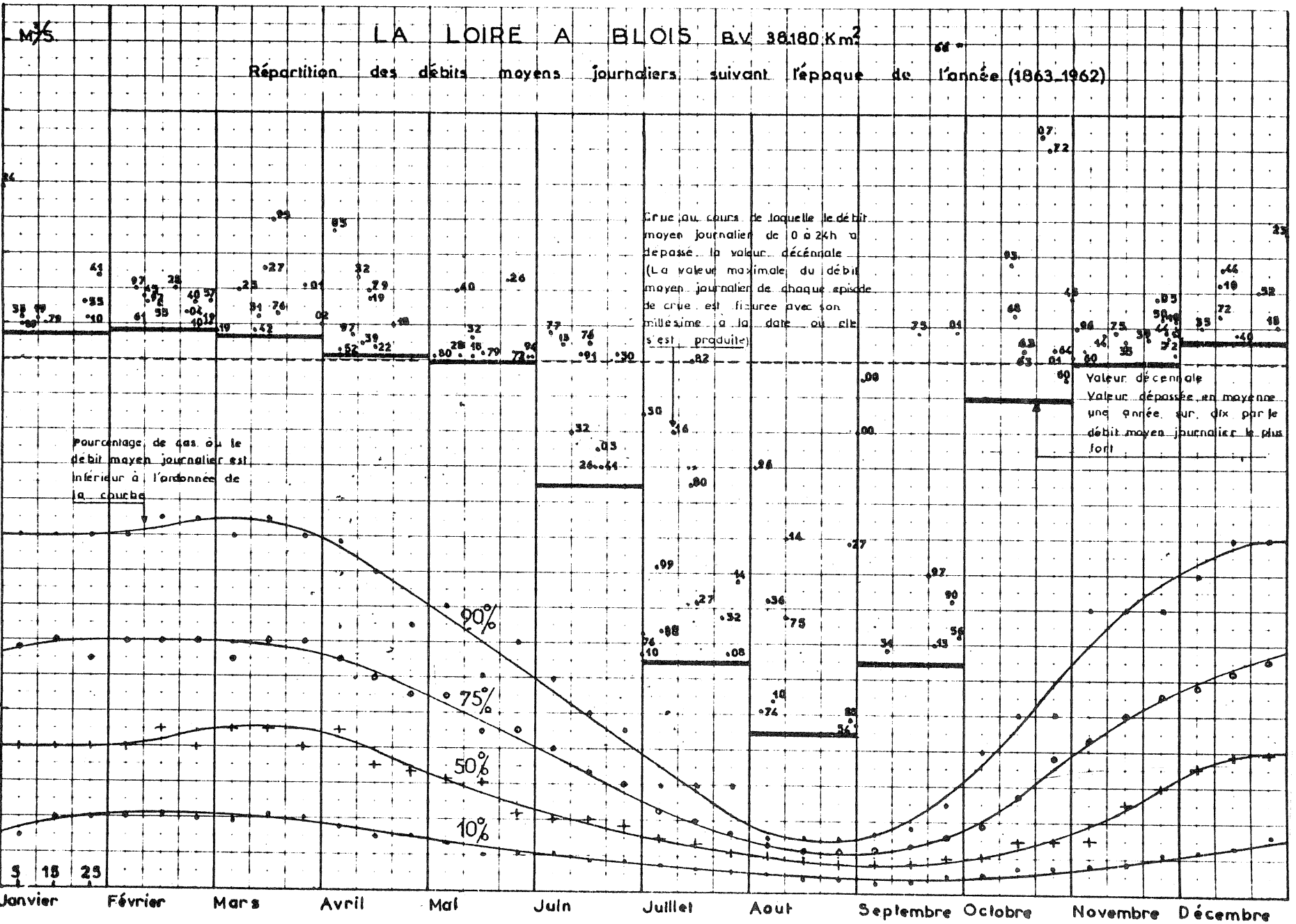
	m³		m³		m³		m³
1863	317	1888	426	1913	484	1938	233
64	268	89	420	14	422	39	473
65	322	1890	305	15	396	1940	503
66	512	91	370	16	468	41	537
67	453	92	398	17	430	42	306
68	412	93	263	18	394	43	279
69	311	94	215	19	497	44	430
1870	197	95	315	1920	289	45	235
71	208	96	382	21	156	46	199
72	511	97	410	22	356	47	208
73	364	98	276	23	408	48	239
74	192	99	237	24	313	49	124
75	414	1900	392	25	259	1950	241
76	409	01	474	26	403	51	464
77	380	02	472	27	422	52	472
78	388	03	314	28	429	53	173
79	557	04	275	29	271	54	286
1880	338	05	311	1930	564	55	312
81	299	06	314	31	459	56	302
82	387	07	459	32	471	57	264
83	367	08	286	33	296	58	396
84	211	09	333	34	330	59	287
85	394	1910	728	35	519	1960	369
86	361	11	278	36	471	61	259
87	316	12	356	37	414	62	291


LA LOIRE A BLOIS
F(Q)
728
Distribution empirique du débit moyen annuel de la période d'observation 1863-1962
Fonction de répartition normale de moyenne m = 357 et d'écart type σ = 104
$F(Q_0) = \int_{0}^{Q_0} \frac{1}{97} e^{-\frac{1}{2}\left(\frac{Q-357}{104}\right)^2} \frac{dQ}{\sqrt{2\pi}}$
Fréquence relative par classe
1 intervalle = 40 m³/s
Q m³/s
1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
0
.25
.20
.15
.10
.05
100
200
300
400
500
600
E.D.F. - DIVISION TECHNIQUE GÉNÉRALE
I-28


LA LOIRE A BLOIS 1863-1962
Distribution empirique du débit moyen annuel
9990
9950
9900
9800
9700
9600
9500
9000
8000
7000
6000
5000
4000
3000
2000
1000
500
400
300
200
100
50
10
100
500
1000
n = 100
m= 357
σ= 104
m3/s

N° 374


LA LOIRE A BLOIS
F(Q)
1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
0
.25
.15
.10
.05
Q m³/s
Distribution empirique du débit moyen annuel de la période d'observation 1863-1887
Fonction de répartition normale de moyenne m = 356 et décart-type σ = 97
$F(Q_0)=\int_0^{Q_0}\frac{1}{97}e^{-\frac{1}{2}\left(\frac{Q-356}{97}\right)^2}\frac{dQ}{\sqrt{2\pi}}$
Fréquence relative par classe
1 intervalle = 40 m³/s
E.D.F. - DIVISION TECHNIQUE GÉNÉRALE
I-30


M³/S
LA LOIRE A BLOIS B.V. 38.180 Km²
Répartition des débits moyens journaliers suivant l'époque de l'année (1863-1962)
Crue au cours de laquelle le débit moyen journalier de 0 à 24h a dépassé la valeur décennale (La valeur maximale du débit moyen journalier de chaque épisode de crue est figurée avec son millésime à la date où elle s'est produite)
Valeur décennale
Valeur dépassée, en moyenne une année sur dix par le débit moyen journalier le plus fort
Pourcentage de cas où le débit moyen journalier est inférieur à l'ordonnée de la courbe
90%
75%
50%
10%
5
15
25
Janvier
Février
Mars
Avril
Mai
Juin
Juillet
Août
Septembre
Octobre
Novembre
Décembre

II - LES MODELES PROBABILISTES

Il existe en hydrologie un arsenal fort important de fonctions de répartition. J'évite à dessein l'expression loi de probabilité qui suggère implicitement une justification physique et peut faire croire que les considérations théoriques imposent le choix d'un modèle spécifique pour représenter tel phénomène hydrométéorologique (pluie, débit, température).

Dans la pratique, la seule justification à l'emploi d'une fonction de répartition est en général purement empirique : on constate la cohérence des résultats dans un grand nombre d'applications comparables.

Parfois plusieurs fonctions de répartition peuvent être pratiquement confondues dans un domaine de l'intervalle $[0,1]$, pour caractériser un phénomène; si l'on ne dispose pas d'éléments complémentaires permettant de décider du choix, la règle générale consiste à utiliser la fonction la plus simple qui contient le moins de paramètres. La justification sera donnée lors du 3ème exposé.

Ces propos n'ont pas pour but de diminuer l'intérêt des modèles probabilistes mais de présenter honnêtement la réalité.

Certains pratiquent le lissage des distributions empiriques, sur graphique $x_i, \left[\frac{2i-1}{2n}\right]$ et estiment que cela suffit pour déterminer la probabilité d'un évènement, évitant ainsi de s'encombrer d'hypothèses mathématiques qu'il est impossible de vérifier avec certitude. Cette procédure est sans doute acceptable et sans grand risque entre les quantiles 20 % et 80 % pour des échantillons de 30 à 50 observations.

A l'extérieur de cet intervalle cela devient dangereux, on s'expose à de graves mécomptes en étant trop tributaires des aléas de l'échantillonnage et de certaines valeurs extrêmes. De plus ce type de lissage est trop subjectif, deux personnes effectueront rarement un ajustement graphique identique sur le même échantillon - des expériences de ce genre, faites sur un grand nombre de cas, montrent la grande variabilité des résultats. Aussi, bien que cette méthode paraisse inoffensive, on ne peut la déconseiller.

En fait, la réalité est moins défavorable, et il existe heureusement une continuité temporelle et spatiale des phénomènes hydrométéorologiques, continuité qui a permis de mettre à l'épreuve les fonctions de répartition traitées dans ce chapitre.

2.1 - La fonction gamma incomplète

$$f(x) = \frac{1}{\Gamma(\lambda)} e^{-\frac{x}{\rho}} \left(\frac{x}{\rho}\right)^{\lambda-1} \frac{1}{\rho} \quad \text{définie entre 0 et } \infty$$

$$\int_0^{\infty} f(x)\, dx = 1$$

ρ : est le paramètre d'échelle

λ : est le paramètre de forme

Pour faciliter les calculs on posera : $y = \frac{x}{\rho}$

$$f(y) = \frac{1}{\Gamma(\lambda)} e^{-y} y^{\lambda-1}$$

Rappelons que $\Gamma(\lambda) = \int_0^{\infty} f(y)\, dy = (\lambda-1)\ !$

Les moments se calculent simplement :

- espérance mathématique $E(y) = \int_0^{\infty} y\, f(y)\, dy = \lambda$

- variance $V(y) = \int_0^{\infty} y^2 f(y)\, dy - \lambda^2 = \lambda\,(\lambda+1) - \lambda^2 = \lambda$

- moment centré d'ordre 3 : $\mu_3(y) = E(y^3) - 3\,E(y^2)\,E(y) + 2\left[E(y)\right]^3 = 2\,\lambda$

- le moment centré d'ordre 4 :

$$\mu_4(y) = E(y^4) - 4\,E(y^3)\,E(y) + 6\,E(y^2)\left[E(y)\right]^2 - 3\left[E(y)\right]^4 = 3\,(\lambda+2)\,\lambda$$

En x on obtient

$$m = E(x) = \lambda \rho$$

$$\sigma^2 = V(x) = \lambda \rho^2$$

$$\mu_3(x) = 2 \lambda \rho^3$$

$$\mu_4(x) = 3 \lambda (\lambda + 2) \rho^4$$

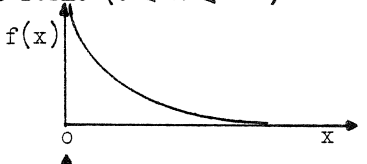
On détermine aisément les relations suivantes :

$$\lambda = \frac{m^2}{\sigma^2}$$

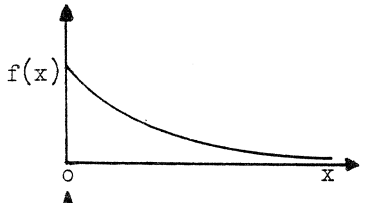
$$\rho = \frac{\sigma^2}{m}$$

Cette fonction est entièrement définie à l'aide de la moyenne et de la variance. Selon les valeurs du paramètre de forme ($0 < \lambda < \infty$)

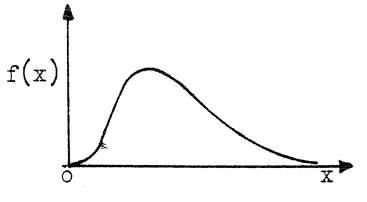
. $0 < \lambda < 1$ on obtient une courbe en j avec branche infinie à l'origine



. $\lambda = 1$ on obtient la fonction exponentielle



. $\lambda > 1$ définit les fonctions à allure en cloche. Leur dissymétrie sera d'autant plus importante que λ est faible (quelques unités) pour $\lambda = 25$ la dissymétrie est déjà très atténuée et pour $\lambda > 60$ on obtient une courbe pratiquement symétrique et approximable par la fonction gaussienne.



Remarque : on peut établir la fonction Γ incomplète pour les valeurs entières de λ à partir d'une somme de variables aléatoires dont la fonction de répartition est exponentielle.

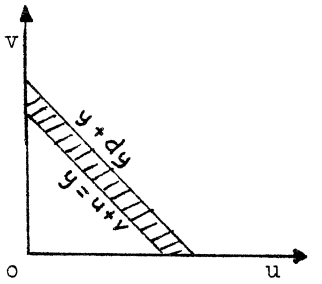
On considèrera des variables réduites :

$$u \text{ de fonction } f(u) = e^{-u}$$

$$v \text{ de fonction } g(v) = e^{-v}$$

La densité de probabilité du couple de ces 2 variates (contraction de variable aléatoire) indépendantes s'écrit :

$$dH\,(u,v) = e^{-u}\, e^{-v}\, du\, dv$$



posons $y = u + v$ et calculons l'intégrale sur u

$$v = y - u$$

$$\int_{u=0}^{u=y} e^{-y}\, du\, dy$$

soit :

$$e^{-y}\, y\, dy$$

Pour 3 variates $u + v + w = y + w$ on utilisera le même processus; de proche en proche on trouve la densité de répartition de la somme de k variates exponentielles :

$$\frac{1}{(k-1)!}\, e^{-y}\, y^{k-1}$$

Cette fonction a été tabulée par Karl Pearson.

Remarque : la fonction de répartition gamma incomplète appliquée au logarithme de la variate x, est recommandée aux Etats-Unis pour représenter la distribution des valeurs extrêmes de débits de crue.

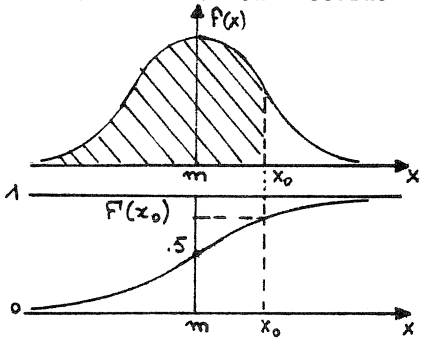
En fait cela revient à un moyen terme entre la fonction de Gumbel et les fonctions de Fréchet et log normale. Ce genre de compromis ne repose sur aucune justification.

2.2 - Fonction de répartition normale ou gaussienne

On l'appelle parfois abusivement la "loi" du hasard.

A ce propos, il est instructif de citer l'anecdote suivante : tout le monde croit à la "loi" normale des erreurs, les expérimentateurs parce qu'ils pensent qu'elle peut être prouvée par les mathématiciens, et les mathématiciens parce qu'ils croient qu'elle a été établie par l'observation. Les deux thèses étant d'ailleurs correctes dans la mesure où on ne leur accorde pas la valeur d'un postulat.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \; e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$$



Elle est définie pour x variant entre $-\infty$ et $+\infty$.

Comment caractériser physiquement cette fonction :

la valeur x de la variable résulte de l'action d'un grand nombre de facteurs dont les effets sont additifs - dont les fluctuations sont indépendantes, distribuées suivant des lois de probabilité quelconques mais dont les premiers moments existent - les fluctuations sont du même ordre de grandeur - la fluctuation d'un facteur particulier est petite par rapport à la fluctuation totale due à l'ensemble des facteurs.

Cette fonction est extrêmement utile et a des propriétés importantes, mais on lui a trop souvent attribué une valeur quasi métaphysique.

Ses moments :

$$E(x) = m$$
$$V(x) = \sigma^2$$

les moments d'ordre impair sont nuls, du fait de la symétrie; les moments d'ordre pair s'obtiennent à partir de la relation :

$$\mu_m = (m-1)\;\sigma^2\;\mu_{m-2}$$

que l'on établit simplement à l'aide de la relation différentielle :

$$d\left[y^{m-1}\, e^{-\frac{y^2}{2}}\right] = (m-1)\, y^{m-2}\, e^{-\frac{y^2}{2}}\, dy - y^m\, e^{-\frac{y^2}{2}}\, dy$$

on peut obtenir l'expression de la loi normale par addition de variates uniformes.

2.3 - <u>La fonction de répartition log normale</u> (Galton-Gibrat)

$$\text{La densité : } f(x) = \frac{1}{S\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{\log x - M}{S}\right)^2}$$

ce n'est autre que la fonction normale appliquée au logarithme de x.

C'est la "loi" dite de l'effet proportionnel : la variable x est la résultante d'effets multiplicatifs en valeurs naturelles. Ces effets deviennent additifs en logarithme, on retombe ainsi sur le cas de la fonction normale.

Les paramètres :

M est le logarithme de la moyenne géométrique des observations,
S^2 est la variance des logarithmes des observations.

Relations qui existent entre la moyenne, la variance de x et la moyenne et variance de log x :

posons $y = a \log_e x + b$; y étant une variable log normale centrée réduite

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a \log_e x + b} e^{-\frac{y^2}{2}}\, dy$$

$$m_x = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x(y)\, e^{-\frac{y^2}{2}}\, dy$$

or $x = e^{\left(\frac{y-b}{a}\right)}$ ou $\exp\left(\frac{y-b}{a}\right)$

d'où $m_x = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{y^2}{2} + \frac{y-b}{a}} dy$

$$m_x = \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2a^2} - \frac{b}{a}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}\left(y - \frac{1}{a}\right)^2} dy$$

$$m_x = e^{\frac{1}{2a^2} - \frac{b}{a}} = m_1 \qquad (1)$$

de même les moments d'ordre k s'obtiennent à l'aide de :

$$m_k = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} x^k (y)\, e^{\frac{-y^2}{2}} dy$$

soit après transformation

$$m_k = e^{\frac{k^2}{2a^2} - \frac{kb}{a}}$$

en particulier $m_2 = e^{\frac{2}{a^2} - \frac{2b}{a}}$

$$\sigma^2 = m_2 - m_1^2 = e^{\frac{2}{a^2} - \frac{2b}{a}} - e^{\frac{1}{a^2} - \frac{2b}{a}}$$

$$= e^{\frac{1}{a^2} - \frac{2b}{a}} \left(e^{\frac{1}{a^2}} - 1\right) \qquad (2)$$

M étant la moyenne des logarithmes : $M = \frac{1}{n} \Sigma \log_e x = \log_e g$

si on note g la moyenne géométrique;

S étant l'écart type de logarithme, en remplaçant a et b par leur valeur en fonction de M et S on a :

$$a = \frac{1}{S}$$

$$b = -\frac{M}{S} = -\frac{\log_e g}{S}$$

D'après (1) $\log_e m_x = \frac{S^2}{2} + \log_e g$ (3)

D'après (2) $\log_e \sigma = \log_e g + \frac{S^2}{2} + \log_e (e^{S^2} - 1)$ (4)

Rappelons que le coefficient de variation de x : $C_v = \frac{\sigma_x}{m_x}$

$$\text{d'où } \log_e C_v = \log_e \sigma_x - \log_e m_x = \frac{1}{2} \log_e (e^{\frac{1}{a^2}} - 1)$$

$$C_v^2 = e^{\frac{1}{a^2}} - 1 = e^{S^2} - 1$$

$$\text{d'où} \qquad S = \sqrt{\log_e (1 + C_v^2)}$$

$$\text{on obtient alors} \quad g = \frac{m_x}{\sqrt{1 + C_v^2}}$$

2.4 - Application de la fonction de répartition normale à des variables transformées - Mélange de lois normales

D'une façon générale on peut toujours appliquer la fonction de répartition gaussienne à des transformées de la variate initiale x, soit $y = \varphi(x)$, y étant une fonction monotone de x. C'est le cas pour $y = a \log_e x + b$; un autre exemple d'application consiste à effectuer $y = \alpha x^{\frac{1}{m}}$ avec $m > 1$, généralement $m = 2$ ou 3.

Il est évident que l'on dispose ainsi d'un large éventail de fonctions de répartition, mais il faut se garder d'abuser de cette trop grande souplesse d'adaptation aux distributions empiriques d'observations, sans autre justification.

On dispose également d'un autre procédé pour reproduire une gamme de fonctions de répartition très variée, en mélangeant deux fonctions normales.

Soit la fonction de répartition normale $H(x)$ de moyenne m_1, écart type σ_1, et la fonction de répartition $G(x)$ de moyenne m_2 et d'écart type σ_2, le mélange de ces 2 fonctions est une fonction de répartition de la forme :

$$F(x) = p\,H(x) + q\,G(x) \qquad \text{avec } p + q = 1, \quad p \text{ et } q > 0$$

Cette fonction dépend de 5 paramètres p, m_1, σ_1, m_2, σ_2 qu'il faudra calculer d'après les moments de $F(x)$.

Cette représentation a été proposée pour représenter la distribution des débits d'une rivière, en considérant que $H(x)$ représente la fonction de répartition des débits "ordinaires" et $G(x)$ la fonction de répartition des débits extrêmes supérieurs (crues), la proportion p étant au moins égale à .9 .

Ce modèle n'est pratiquement pas utilisé en hydrologie actuellement.

On peut d'ailleurs imaginer de mélanger K fonctions de répartitions normales :

$$F(x) = \sum_{i=1}^{K} p_i\,H_i(x) \qquad \text{avec } \sum_{i=1}^{K} p_i = 1$$

modèle défini par $3K-1$ paramètres (m_i, σ_i, p_i).

On voit que le nombre de paramètres à calculer devient vite imposant, eu égard au nombre d'observations; bien que la souplesse d'adaptation

d'une telle fonction de répartition puisse être séduisante, on ne peut que déconseiller son utilisation pratique.

Remarques

- Il y a parfois confusion entre la fonction de répartition d'une variate résultant du mélange de deux fonctions de répartition normales (cf. ci-dessus), et, la fonction de répartition de la somme de 2 variates gaussiennes de paramètres m_1, σ_1 et m_2, σ_2. Dans ce dernier cas il s'agit d'une fonction de répartition gaussienne de moyenne $m = m_1 + m_2$ et d'écart type $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$, moments tout à fait différents de ceux de $F(x) = p\,H(x) + (1 - p)\,G(x)$.

- Ainsi, sans recourir à des fonctions de répartition sophistiquées dont l'expression analytique est mathématiquement complexe, par le seul jeu de simples transformations de variables ou mélange de 2 fonctions de répartition, avec la fonction gaussienne on peut obtenir une panoplie variée et très large des modèles probabilistes.

2.5 - Fonction de répartition harmonique (loi de Halphen)

Il s'agit d'un cas particulier des lois de Halphen du type A

$$f(x) = \frac{1}{2\,\mu^{\nu}\ K_{\nu}\ (\alpha)}\ e^{-\frac{\alpha}{2}\left(\frac{x}{\mu} + \frac{\mu}{x}\right)}\ x^{\nu - 1}$$

pour $\nu = 0$ on obtient la fonction harmonique :

$$f(x) = \frac{1}{2x\ k_o\ (\alpha)}\ e^{-\frac{\alpha}{2}\left(\frac{x}{\mu} + \frac{\mu}{x}\right)}$$

k_o est la fonction de Bessel Basset d'ordre zéro.

Pour la caractériser, il suffit de deux paramètres : un paramètre d'échelle : $\mu = \sqrt{mh}$ qui est aussi la médiane de la distribution (h est la moyenne harmonique : $\frac{1}{h} = \frac{1}{n} \Sigma \frac{1}{x_i}$) et d'un paramètre de forme soit $C_v = \frac{\sigma}{m}$, soit $\lambda = \sqrt{\frac{m}{h}}$.

A l'aide de l'abaque ci-joint, on peut construire la fonction de répartition $\int_0^u f(x)\,dx$; elle est définie par une droite passant par l'origine et le point correspondant au C_v sur l'axe de gauche. Cette droite coupe des courbes verticales qui définissent la probabilité et des droites horizontales qui définissent la valeur réduite correspondant à la probabilité; on multiplie cette dernière par μ pour obtenir la valeur naturelle.

Cette fonction peut rendre des services pour des débits dont la distribution est dissymétrique; on l'a appliquée sur $\sqrt{x}$ plutôt que sur x.

Je m'étendrai peu sur ces fonctions, il existe également les lois du type B : $f(x) = k \exp\left[-\frac{x^2}{v} + b\frac{x}{v}\right] x^{-2\alpha-1}$, qui dépendent de trois paramètres et peuvent faire illusion par leur mathématique complexe. Il s'agit là d'un outil beaucoup trop raffiné pour être utilisé en hydrologie à l'heure actuelle. Et il ne faudrait pas croire que la complexité mathématique augmente la précision des calculs.

2.6 - <u>Fonction de répartition de Gumbel</u>

$$F(x) = e^{-e^{-\left(\frac{x-\theta}{a}\right)}}$$

avec $\begin{cases} a = \frac{\sqrt{6}}{\pi}\sigma \\ \theta = m - \gamma a \end{cases}$ (γ ou nombre d'Euler $\# .577$)

C'est la "loi" dite des valeurs extrêmes; on retrouve, là encore, le même abus d'usage que pour la fonction normale, on attribue à cette fonction des propriétés qu'elle ne satisfait qu'avec une approximation plus ou moins étroite.

Soit un échantillon de n valeurs indépendantes, la fonction de répartition associée à l'échantillon G(x), la loi de probabilité de la plus forte de ces n valeurs s'écrit :

$$G(x)^n$$

Nombreux sont ceux qui sont persuadés que $G(x)^n = F(x)$. Cette égalité n'est rigoureuse que si G(x) est une expression en exponentielle simple; dans tous les autres cas il ne s'agit que d'une approximation.

En effectuant une transformation logarithmique sur x ($y = \log_e x$), la nouvelle variate "y" ayant pour fonction de répartition la fonction de Gumbel, on obtient pour x la fonction de répartition de Fréchet.

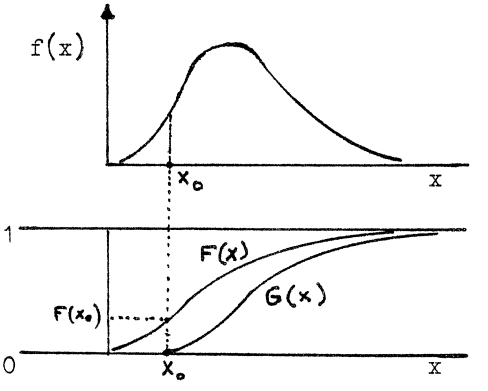
Ce type de transformation a été utilisé précédemment pour les fonctions de répartition de Galton-Gibrat et log Pearson III. Cette fonction, utilisée parfois dans l'étude des valeurs extrêmes de débits, conduit à des valeurs considérables lorsqu'on l'extrapole au-delà de la probabilité .99 .

2.7 - Fonction de répartition tronquée - censurée

2.7.1 - Fonction de répartition tronquée

Soit une fonction de répartition F(x), de densité f(x), si on effectue une troncature à la valeur x_o, en ne s'intéressant qu'aux valeurs supérieures à x_o, cela revient à créer une nouvelle fonction de répartition pour $x \geqslant x_o$:

$$G(x) = \frac{F(x) - F(xo)}{1 - F(xo)}$$



2.7.2 - Fonction de répartition censurée

Si on effectue une coupure au point x_o sur une fonction de répartition initiale $F(x)$, cela revient à accumuler en x_o toutes les valeurs qui lui sont inférieures $(0 < x < x_o)$.

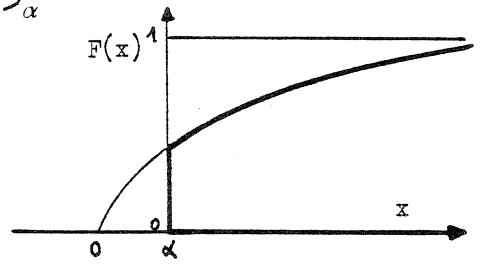
Prenons comme exemple la fonction de répartition exponentielle :

$$F(x) = 1 - e^{-\frac{x}{a}}$$, définie pour x variant de 0 à $+\infty$.

On effectue une censure au point $x_o = \alpha$, on notera $\hat{F}(x)$ la nouvelle fonction de répartition.

Calculons le moment d'ordre 1 de $F(x)$:

$$E(x) = \int_0^{\infty} x\, e^{-\frac{x}{a}}\, dx = \int_0^{\alpha} x\, e^{-\frac{x}{a}}\, dx + \int_{\alpha}^{\infty} x\, e^{-\frac{x}{a}}\, dx = a$$



soit :

$$E(x) = a\left[1 - e^{\frac{\alpha}{a}}\left(1+\frac{\alpha}{a}\right)\right] + a\, e^{-\frac{\alpha}{a}}\left(1+\frac{\alpha}{a}\right)$$

Effectuer une censure consiste à rendre nulles les valeurs de x pour $0 < x < \alpha$ donc le premier moment de $\hat{F}(x)$ sera :

(1) $$\hat{E}(x) = m = a\,(1-\theta)\left(1+\frac{\alpha}{a}\right) \qquad \text{avec } \theta = F(\alpha)$$

on calculera également le moment d'ordre 2 :

$$E(x^2) = \int_0^{\alpha} x^2\, e^{-\frac{x}{a}}\, dx + \int_{\alpha}^{\infty} x^2\, e^{-\frac{x}{a}}\, dx = 2a^2$$

$$E(x^2) = a^2\left\{2\left(1-e^{-\frac{\alpha}{a}}\right) - 2\,e^{-\frac{\alpha}{a}}\frac{\alpha}{a} - e^{-\frac{\alpha}{a}}\frac{\alpha^2}{a^2}\right\} + a^2\left\{2\,e^{-\frac{\alpha}{a}} + 2\,e^{-\frac{\alpha}{a}}\frac{\alpha}{a} + e^{-\frac{\alpha}{a}}\frac{\alpha^2}{a^2}\right\}$$

du fait que $x = 0$ pour $0 < x < \alpha$ on obtient le moment d'ordre 2 de $\hat{F}(x)$

$$\hat{E}(x^2) = e^{-\frac{\alpha}{a}}\left(2 + 2\frac{\alpha}{a} + \frac{\alpha^2}{a^2}\right)a^2$$

d'où la variance de $\hat{F}(x)$:

$$\hat{V}(x) = \hat{E}(x^2) - \left[\hat{E}(x)\right]^2$$

(2) soit $\sigma^2 = a^2 (1 - \theta) \left[(1 + \frac{\alpha}{a})^2 \theta + 1\right]$

Si l'on effectue le changement d'origine $X = x - \alpha$, notons :

m' la moyelle des valeurs de X

σ'^2 la variance des valeurs de X

on trouve alors que $\begin{cases} m' = m - \alpha (1 - \theta) \\ \sigma'^2 = \sigma^2 - 2 \alpha m (1 - \theta) + \alpha^2 (1 - \theta) \theta \end{cases}$

d'après (1) et (2) on obtient alors :

$$\begin{cases} m' = (1 - \theta) a \\ \sigma'^2 = (1 - \theta^2) a^2 \end{cases} \qquad \text{et } C'_v = \sqrt{\frac{1 + \theta}{1 - \theta}}$$

2.8 - Fonction de répartition de Poisson

C'est un exemple de loi discontinue, la variate y peut prendre les valeurs 0, 1, 2, ... avec la probabilité :

$$\text{Prob } (Y=y) = e^{-m} \cdot \frac{m^y}{y!} \quad ; \quad (y = 0, 1, 2, \ldots)$$

on démontre facilement que la moyenne et la variance sont égales à m. C'est la loi des évènements rares.

Exemple d'application: les crues extrêmes :

Si on fait un relevé des statistiques de crues maximales annuelles x, on définit la crue centenaire par x_p (la valeur x_p a p % chances d'être dépassée) p = .01, on note $T = \frac{1}{p} = 100$ ans la durée de retour de cet évènement - en moyenne on observe une crue tous les 100 ans.

Indépendamment de la définition fréquentiste contestable de la crue centenaire ou millénaire, si on admet cette hypothèse on peut montrer que la probabilité d'observer 2 crues centenaires en 10, 50, 100 ans n'est pas négligeable. Notons Y la variate nombre d'observations supérieures à x_p relevées au cours d'une période fixe de N années. Y obéit à une loi de Poisson de moyenne Np.

$$\text{Probabilité } (Y \geqslant y) = \sum_{k=y}^{\infty} e^{-Np} \frac{(Np)^k}{k!}$$

y \ N	10	50	100
1	.095	.393	.593
2	.005	.090	.227
3	-	.014	.063

2.9 - Fonctions de répartition utilisées dans les tests

2.9.1 - Loi du X^2

Soit une variate normale centrée réduite (moyenne 0, écart type 1)

$$f(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

Effectuons le changement de variable $\begin{cases} Z = u^2 \\ dZ = 2\,udu \end{cases}$

La densité de probabilité de Z s'écrit :

$$g(Z)\,dZ = \frac{1}{\sqrt{2\pi Z}} e^{-\frac{Z}{2}} dZ \qquad \text{pour } Z > 0$$

$$= 0 \qquad \text{pour } Z \leqslant 0$$

Il s'agit d'une fonction de répartition Γ incomplète.

Il suffit de poser $a = \lambda = \frac{1}{2}$ dans l'expression :

$$\frac{1}{\Gamma(\lambda)} a^{\lambda} x^{\lambda-1} e^{-ax}$$ et de se rappeler que $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

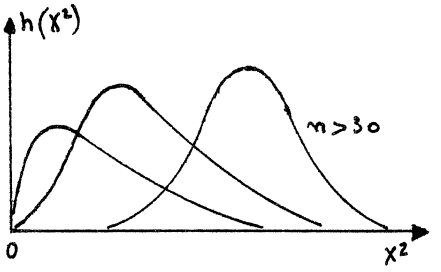
Si $u_1 \ldots\ldots u_n$ sont n variates normales centrées réduites, la variate $X^2 = \sum_{i=1}^{i=n} u_i^2$ suit une loi Γ incomplète de paramètres : $a = \frac{1}{2}$

$$\lambda = n\left(\frac{1}{2}\right)$$

d'où la densité : $h(X^2) = \frac{1}{2^{\frac{n}{2}}\ \Gamma\left(\frac{n}{2}\right)}\ x^{\frac{n}{2}-1}\ e^{-\frac{x}{2}}$ (avec $X^2 = x$)

n est le nombre des degrés de liberté, nous verrons sa signification dans la suite. La démonstration est évidente à l'aide des fonctions caractéristiques : on sait que la fonction caractéristique d'une somme de variates de même loi de probabilité est égale au produit des fonctions caractéristiques élémentaires, soit :

$$\left[\frac{1}{(1 - 2it)^{\frac{1}{2}}}\right]^n$$



2.9.2 - Loi de Student

Soit n+1 variates normales : u, u_1, u_2 ... u_n centrées et réduites (m = 0, σ=1);

posons $v = \sqrt{\frac{1}{n}\ \Sigma\ u_i^2}$ (racine > 0)

Considérons la variate $\frac{u}{v} = s$,

on veut calculer Probabilité (s < x).

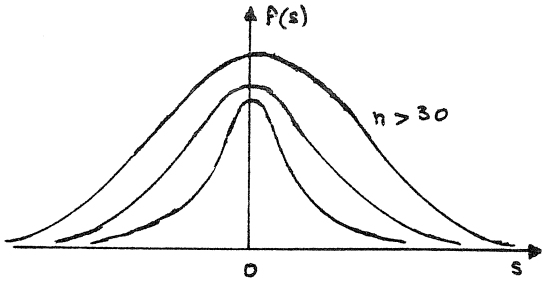
Sachant que la densité de répartition d'un produit de variates indépendantes est égale au produit des densités respectives, en posant

$$\begin{cases} v = w \\ u = sw \end{cases}$$

et en intégrant sur w on trouve :

$$f(s) = \frac{1}{\sqrt{n\pi}}\ \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\ \left(1 + \frac{s^2}{n}\right)^{-\frac{n+1}{2}}$$

expression dans laquelle n représente le nombre de degrés de liberté. (Rappelons que Student était le pseudonyme de l'anglais Gosset).



CONCLUSION

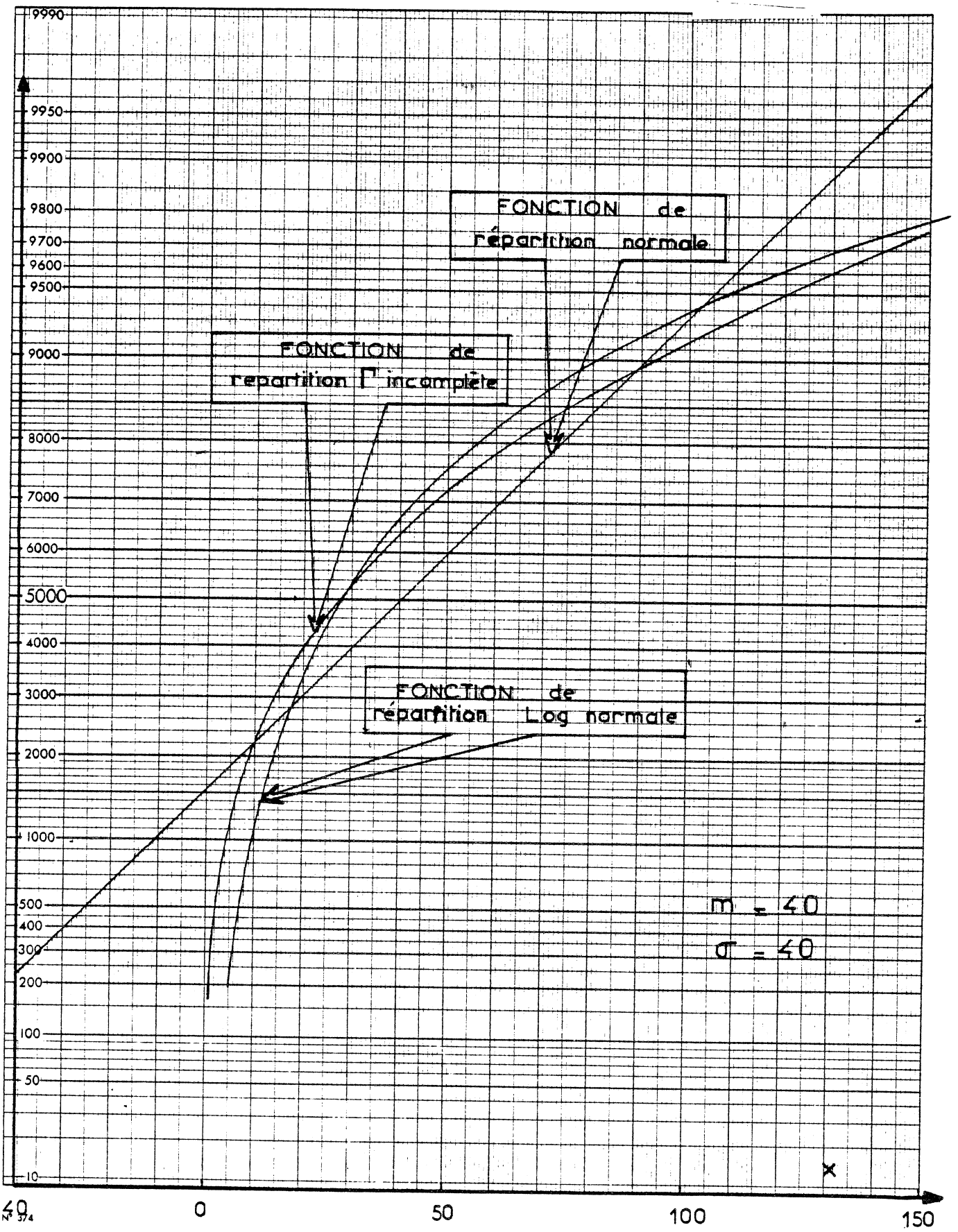
Il est bon d'insister sur les deux formes d'approximation importantes que sont les fonctions de Gauss et de Gumbel :

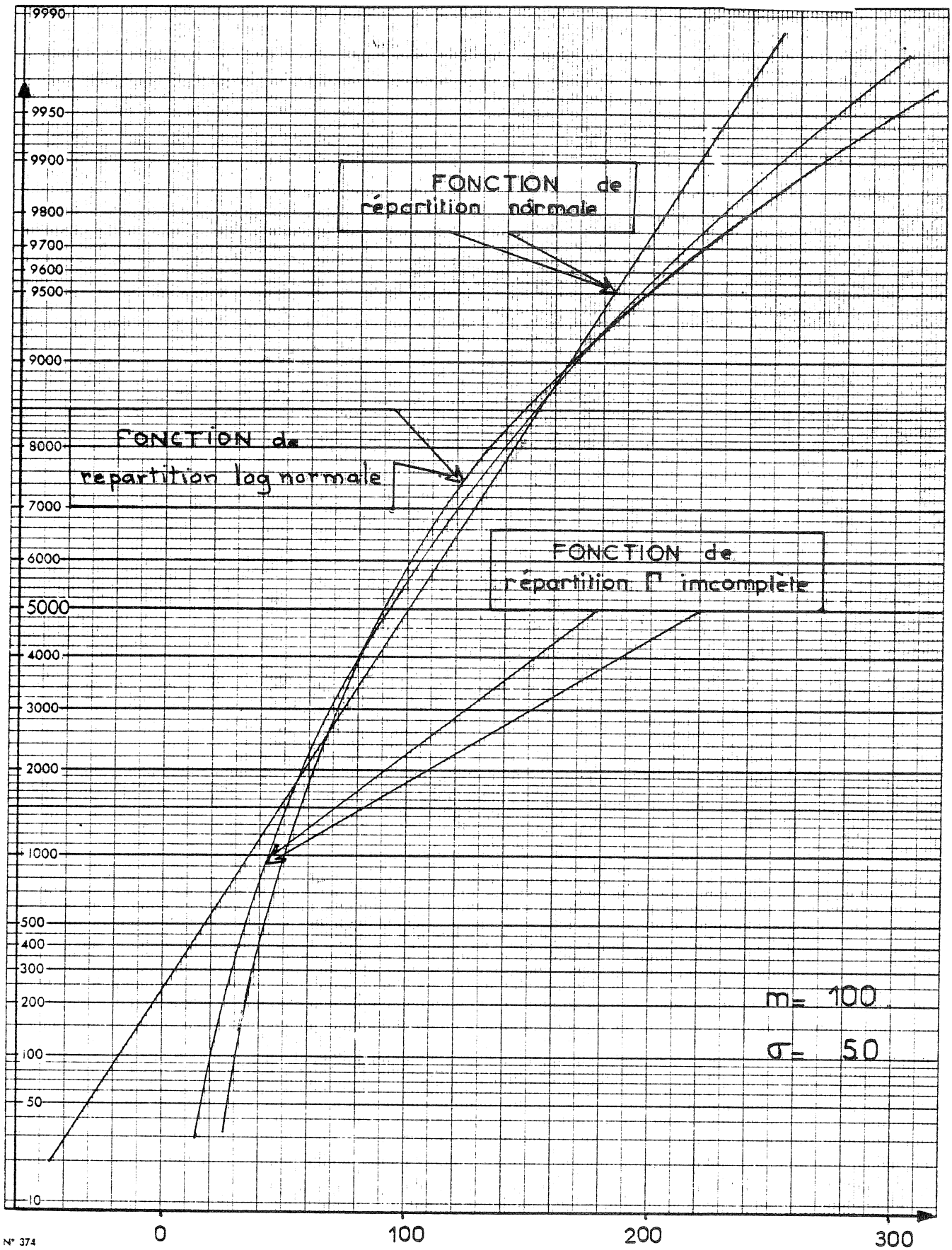
- la première est l'approximation en $e^{-\frac{x^2}{2}}$, la partie centrale des densités de répartition peut souvent être approximée par un arc de parabole; sur le graphique gausso-arithmétique on peut remplacer la partie centrale du graphe de la fonction de répartition par un segment de droite (entre 15 % et 85 % par exemple) ;

- la seconde est l'approximation en e^{-x} et intéresse les queues des courbes de densité sur graphique $\left[x, -L\ (-LF(x)\right]$ on peut remplacer une courbe par un segment de droite, entre les probabilités 95 % et 99.9 %, par exemple.

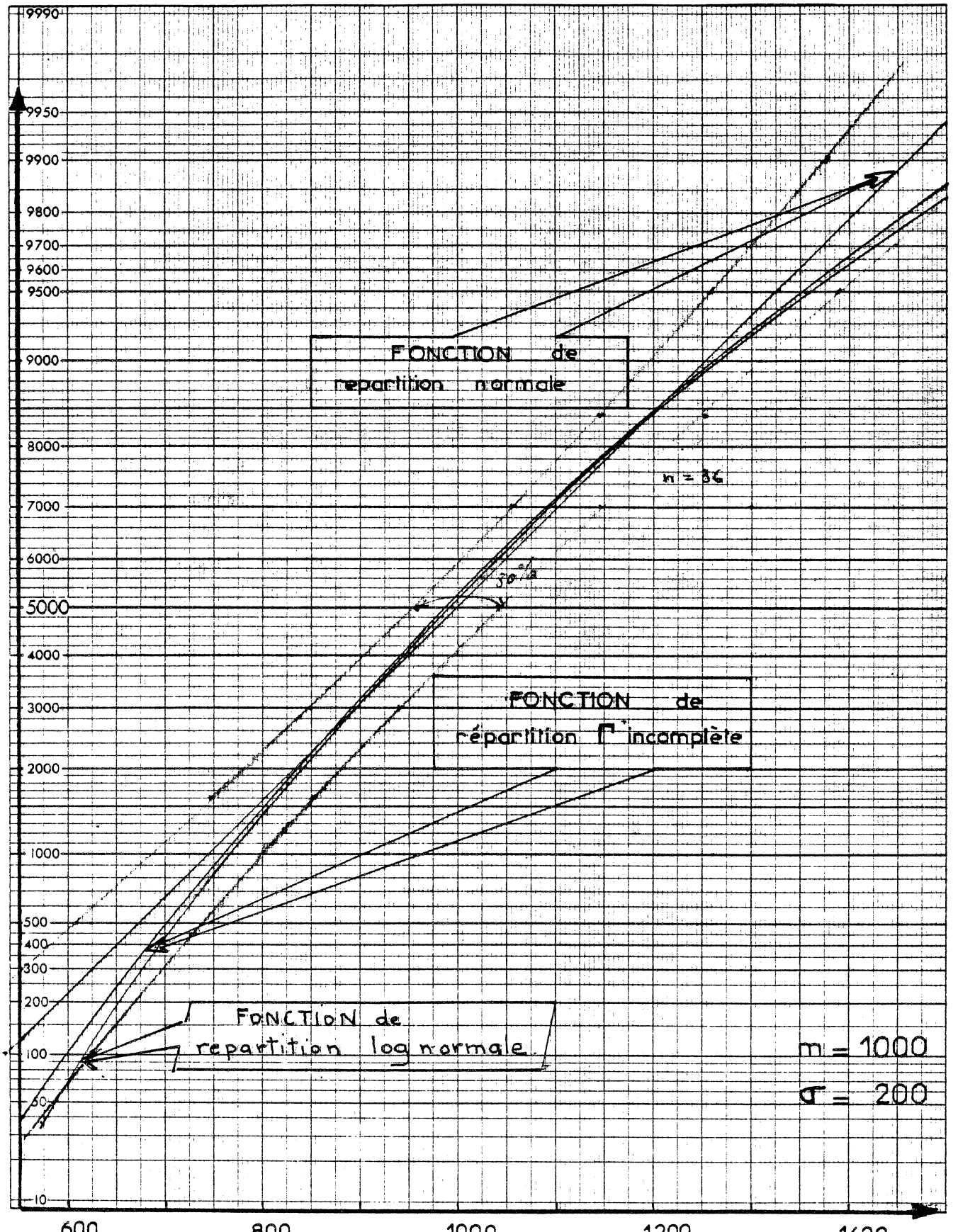
Bien entendu, ces approximations sont valables dans un domaine limité (qui peut être très large d'ailleurs) de l'intervalle $\left[0,1\right]$.

Les graphiques suivants donnent une idée de ces approximations pour différentes fonctions.

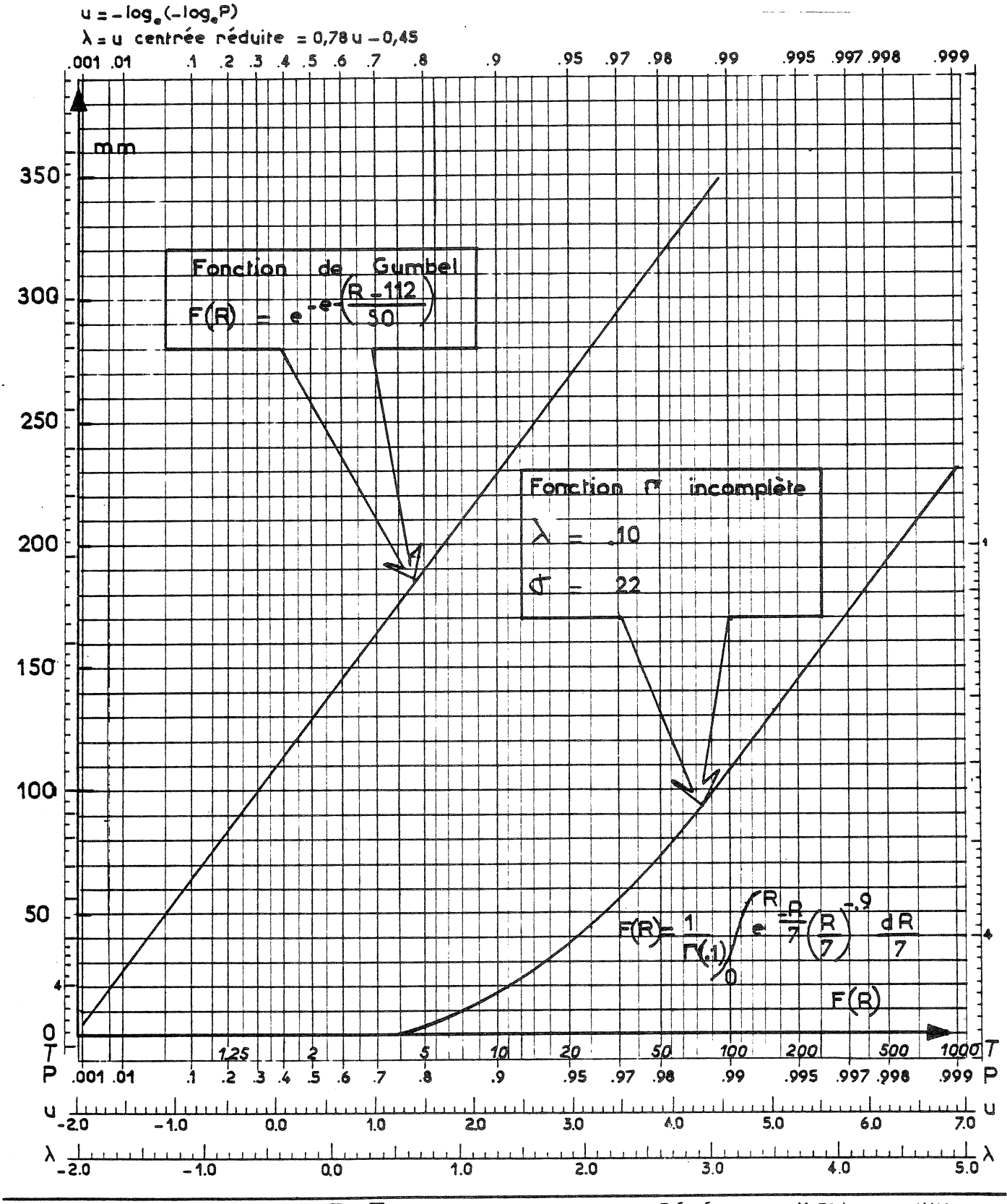

9990
9950
9900
9800
9700
9600
9500
9000
8000
7000
6000
5000
4000
3000
2000
1000
500
400
300
200
100
50
10
FONCTION de répartition normale
FONCTION de répartition Γ incomplète
FONCTION de répartition Log normale
m = 40
σ = 40
-40
0
50
100
150


9990
9950
9900
9800
9700
9600
9500
9000
8000
7000
6000
5000
4000
3000
2000
1000
500
400
300
200
100
50
10
FONCTION de répartition normale
FONCTION de répartition log normale
FONCTION de répartition Γ incomplète
m= 100
σ= 50
0
100
200
300

N° 374


9990
9950
9900
9800
9700
9600
9500
9000
8000
7000
6000
5000
4000
3000
2000
1000
500
400
300
200
100
50
10
600
800
1000
1200
1400
FONCTION de
repartition normale
n = 36
30°
FONCTION de
répartition Γ incomplète
FONCTION de
repartition log normale
m = 1000
σ = 200

N° 374


u = − log_e(− log_e P)
λ = u centrée réduite = 0,78 u − 0,45
.001 .01 .1 .2 .3 .4 .5 .6 .7 .8 .9 .95 .97 .98 .99 .995 .997 .998 .999
mm
350
300
250
200
150
100
50
0
Fonction de Gumbel
F(R) = e^{-e^{-(R-112)/50}}
Fonction Γ incomplète
λ = .10
α = 22
F(R) = 1/Γ(.1) ∫_0^R e^{-R/7} (R/7)^{-.9} dR/7
F(R)
T
1.25 2 5 10 20 50 100 200 500 1000
P
u
-2.0 -1.0 0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0
λ
-2.0 -1.0 0.0 1.0 2.0 3.0 4.0 5.0

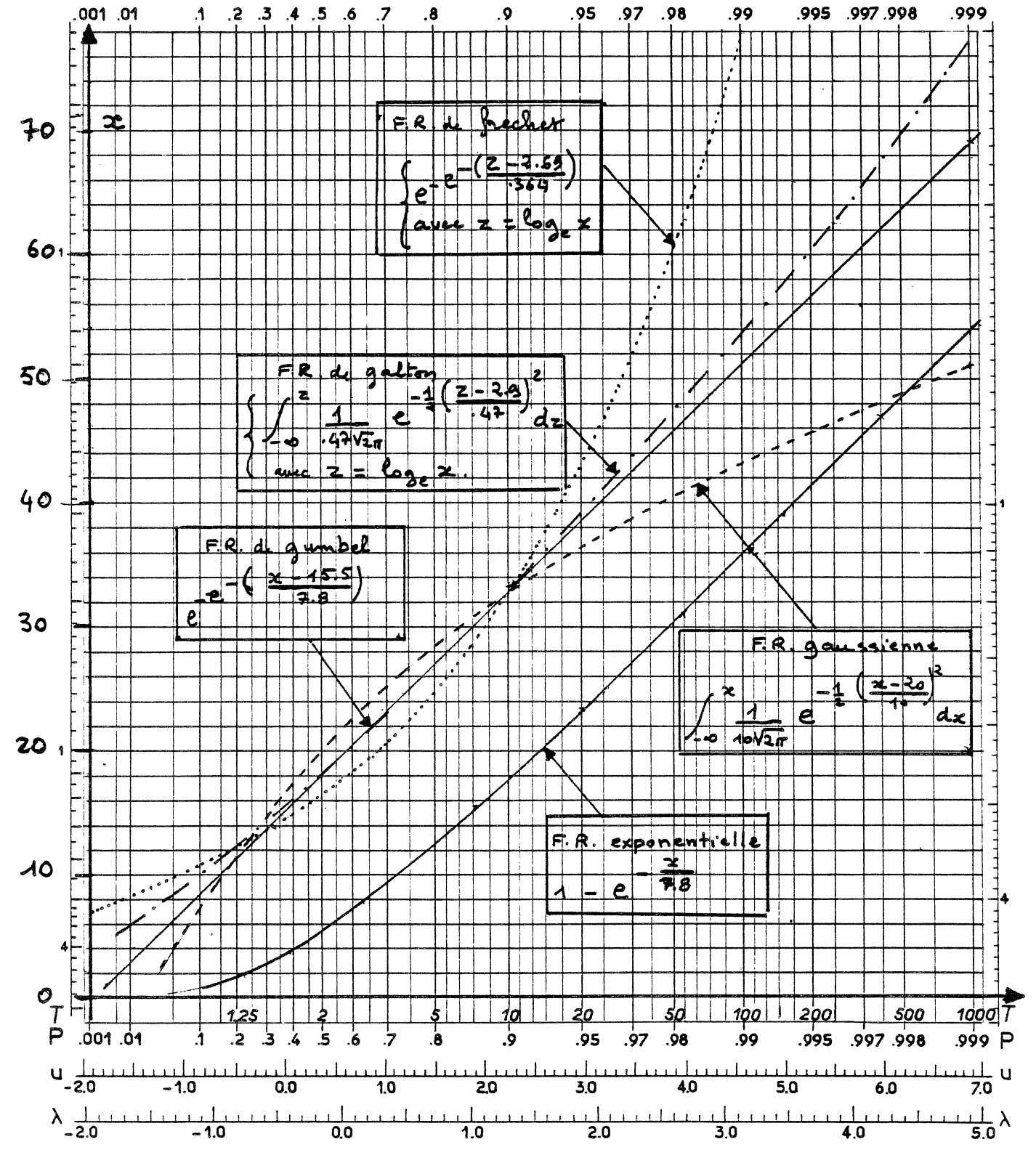
REPRESENTATION GRAPHIQUE DE 5 FONCTIONS DE REPARTITION

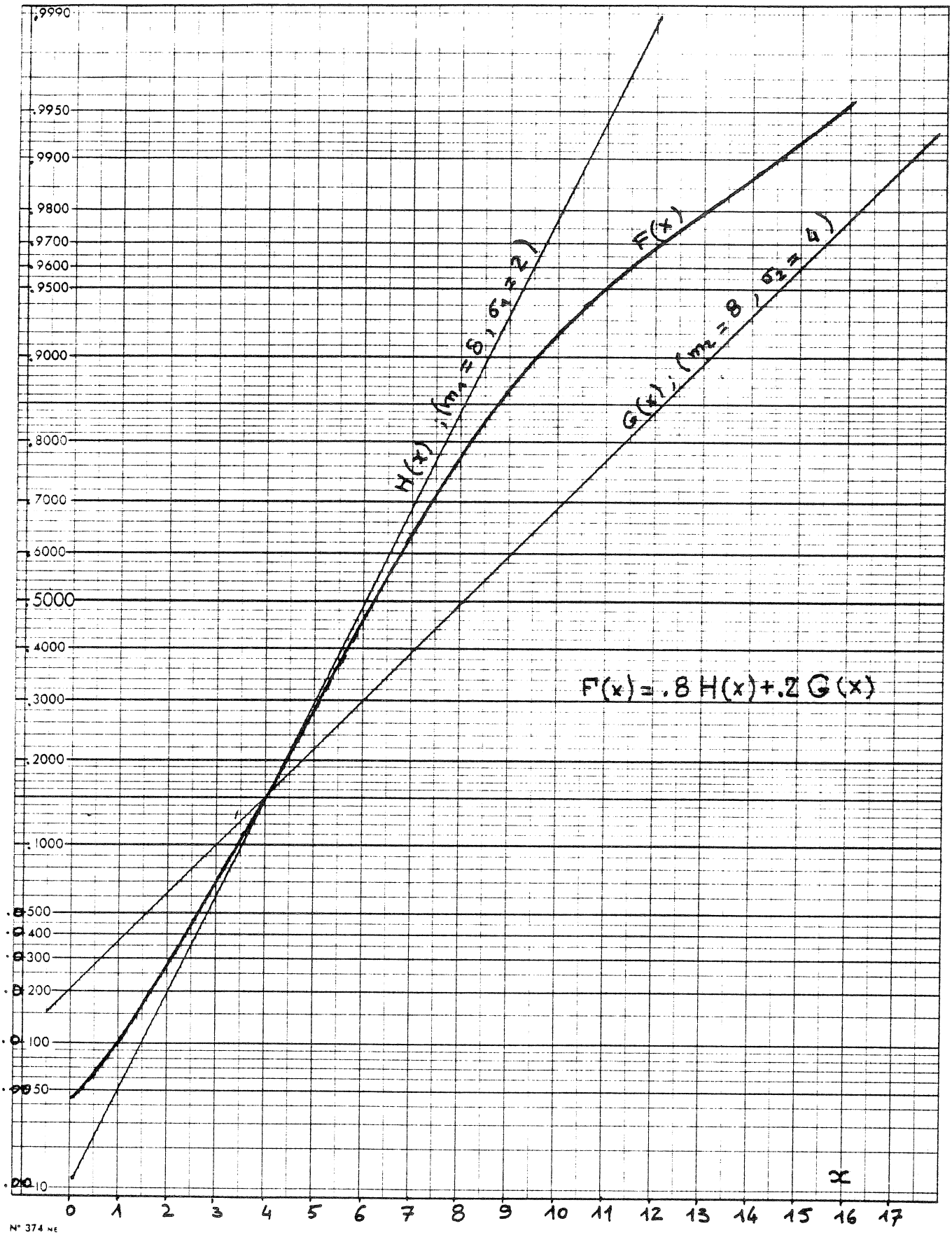
$m_x = 20$

$\sigma_x = 10$

$u = -\log_e(-\log_e P)$

$\lambda = u$ centrée réduite $= 0{,}78\,u - 0{,}45$




.9990
.9950
.9900
.9800
.9700
.9600
.9500
.9000
.8000
.7000
.6000
.5000
.4000
.3000
.2000
.1000
500
400
300
200
100
50
10
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
x
H(x) ; (m1 = 6 ; b2 = 2)
F(x)
G(x) ; (m2 = 8 ; b2 = 4)
F(x) = .8 H(x) + .2 G(x)

UTILISATION DE L'ABAQUE

Cet abaque permet d'ajuster rapidement la fonction Gamma incomplète aux distributions empiriques de précipitations et de débits journaliers mensuels, pluri-mensuels et annuels.

Description de l'abaque :

- l'axe des abscisses est gradué en valeurs de C_v et $\lambda = \frac{1}{C_v^2}$;
- les courbes sont cotées en variable réduite $u = \frac{R}{\sigma}$, R étant la quantité dont on veut étudier la répartition en probabilité ;
- l'axe des ordonnées est gradué en probabilité F(u) d'avoir une valeur inférieure ou égale à $u = \frac{R}{\sigma}$.

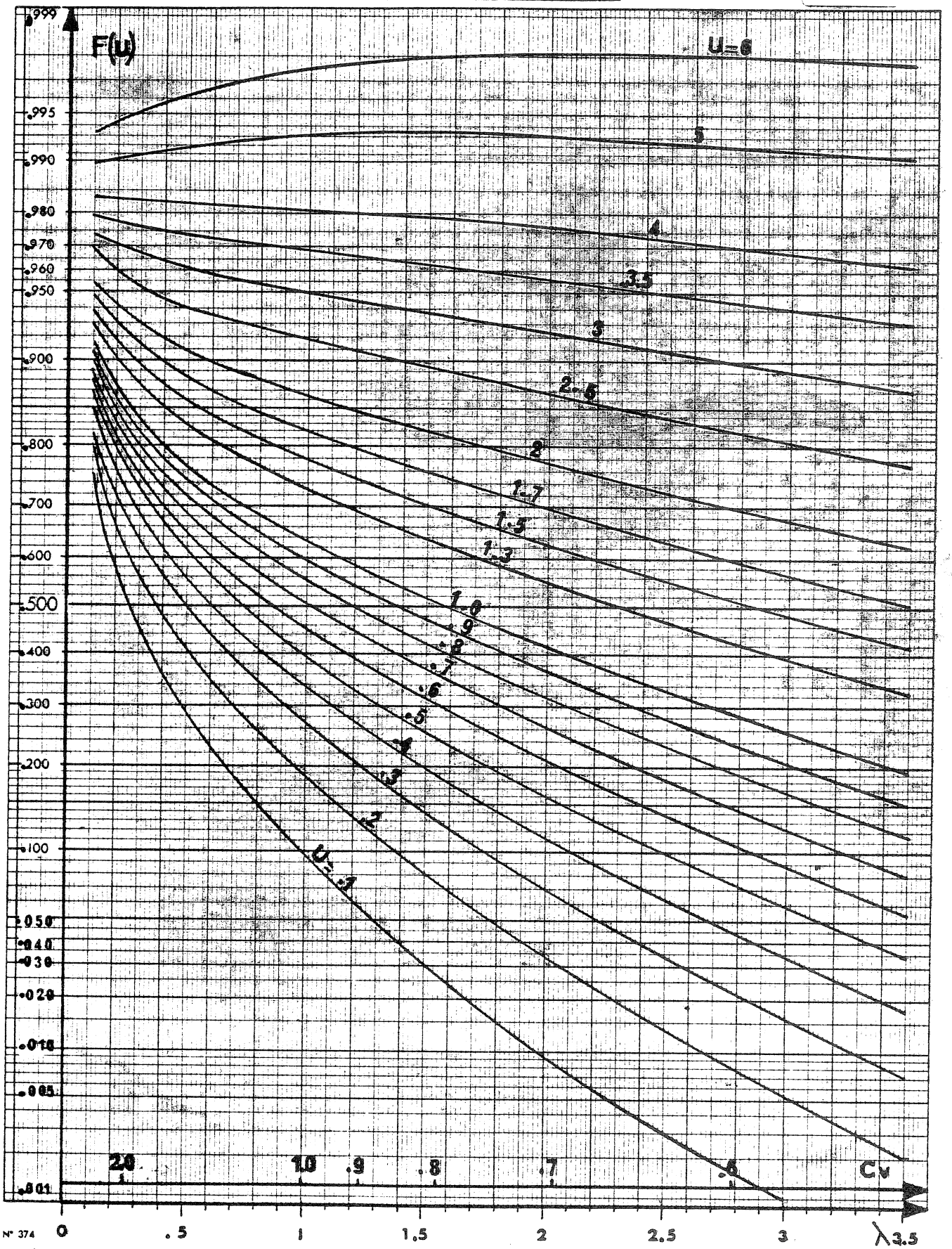
Utilisation :

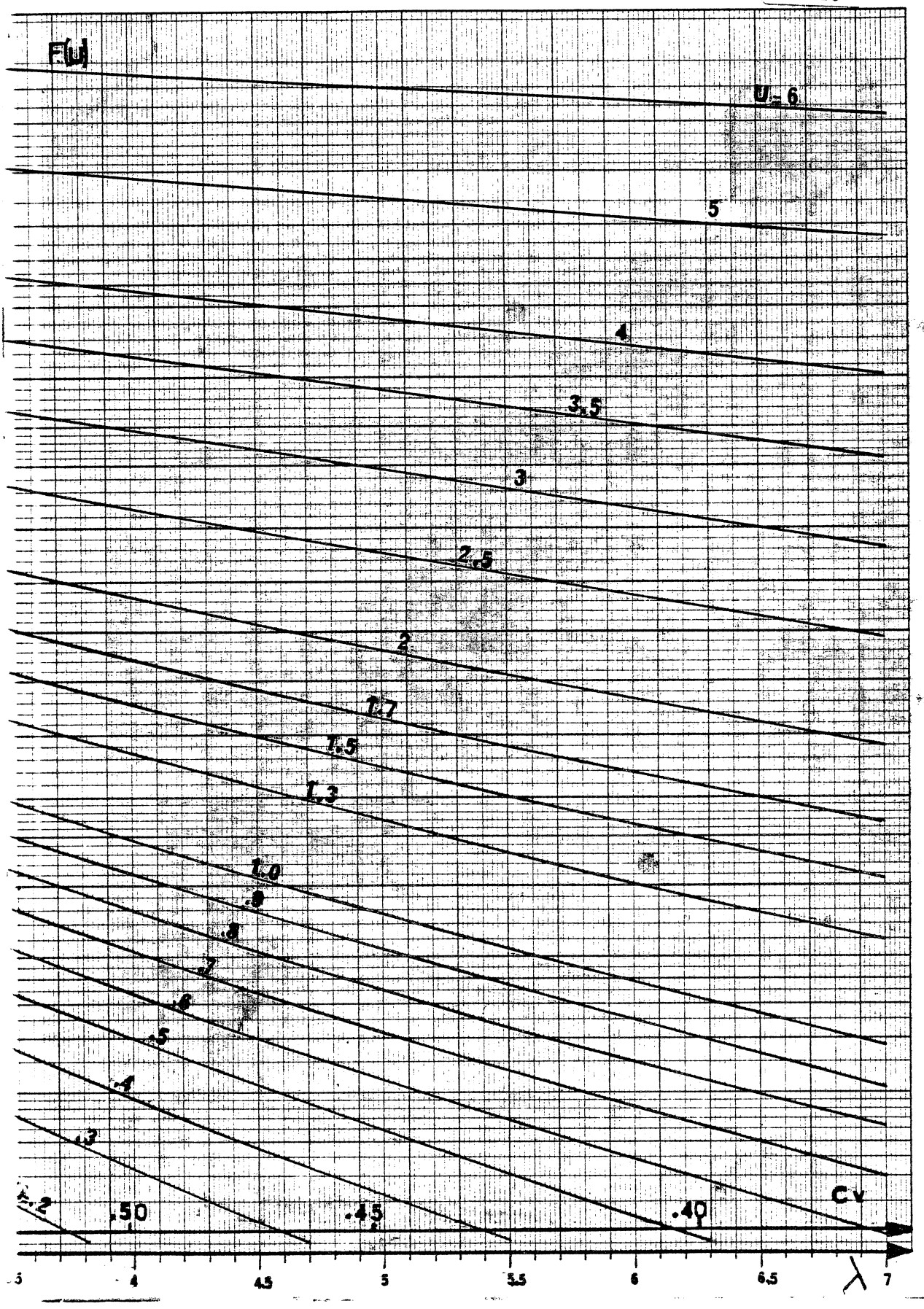
A partir des données de la série de n années d'observations d'une station pluviométrique $(R_1 \ldots R_n)$ on calcule :

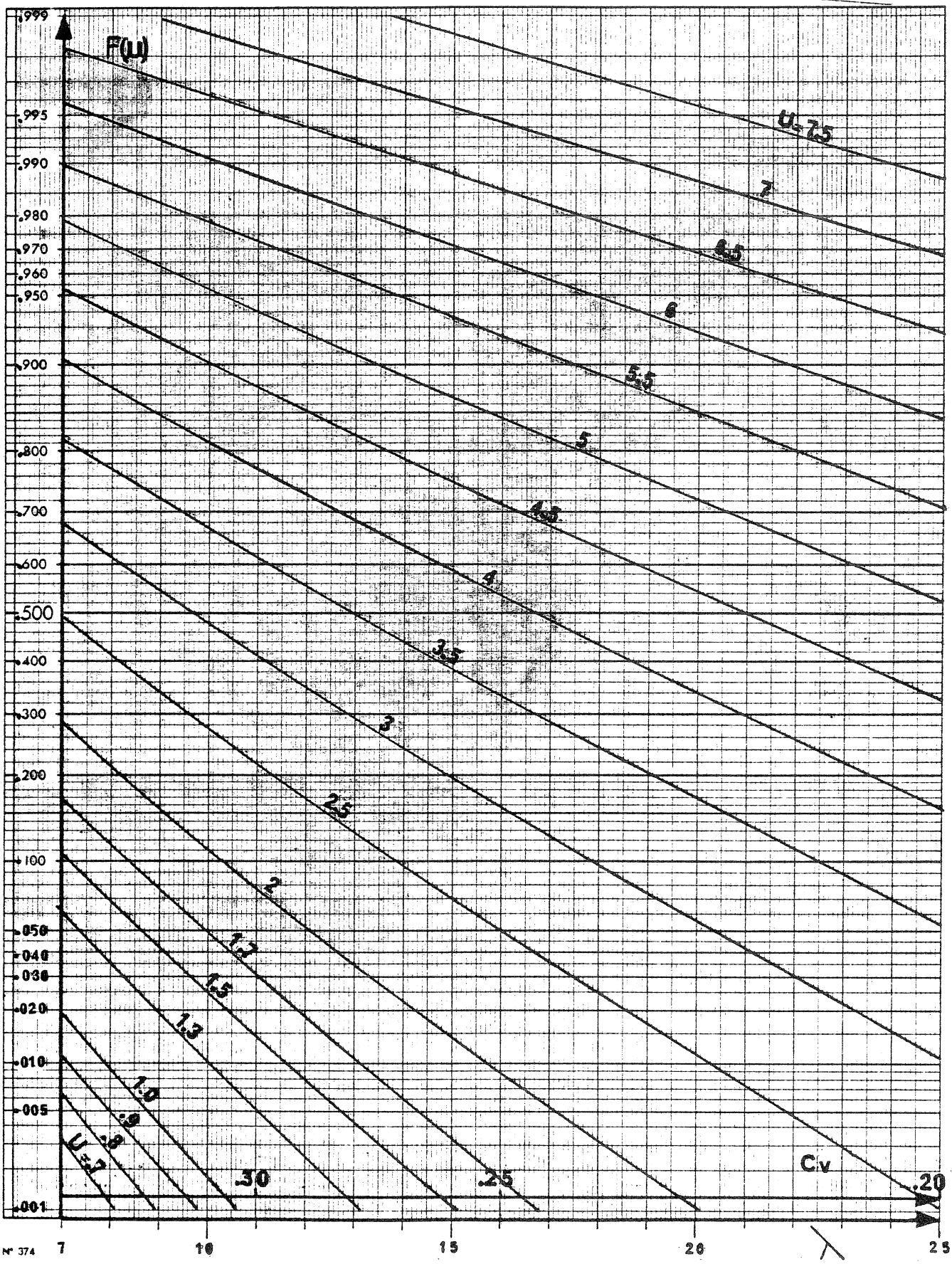
- la moyenne $\overline{R} = \frac{\sum_{i=1}^{i=n} R_i}{n}$; l'écart type $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{i=n} (R_i - \overline{R})^2}$
- le coefficient de variation $C_v = \frac{s}{\overline{R}}$ ou le paramètre de forme $\lambda = \frac{\overline{R}^2}{s^2}$

La probabilité d'avoir une précipitation inférieure à $R_i = s \cdot u_i$ est définie à l'intersection de la verticale d'abscisse λ et de la courbe $u_i = \frac{R_i}{s}$

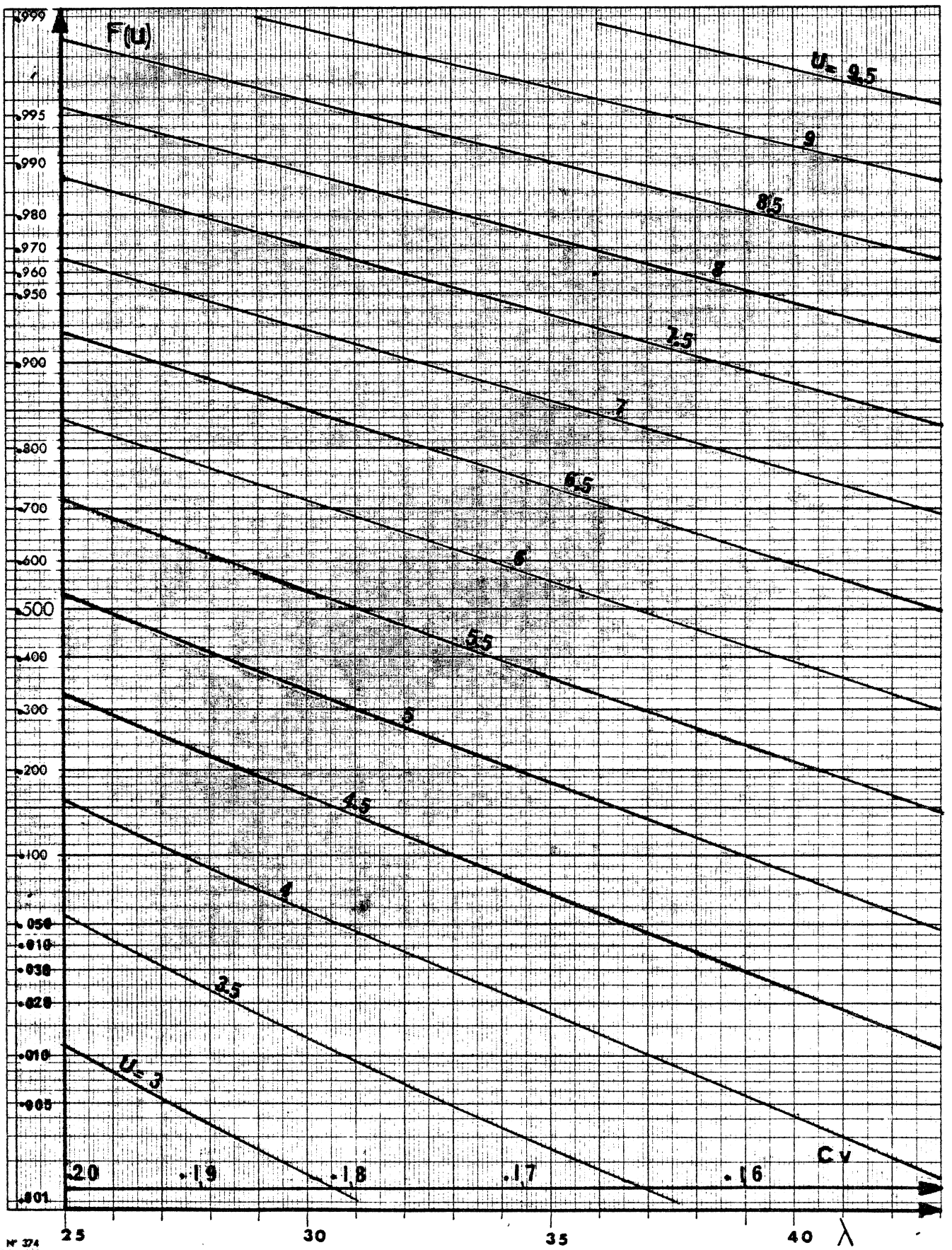
ABAQUE DE LA FONCTION Γ INCOMPLETE



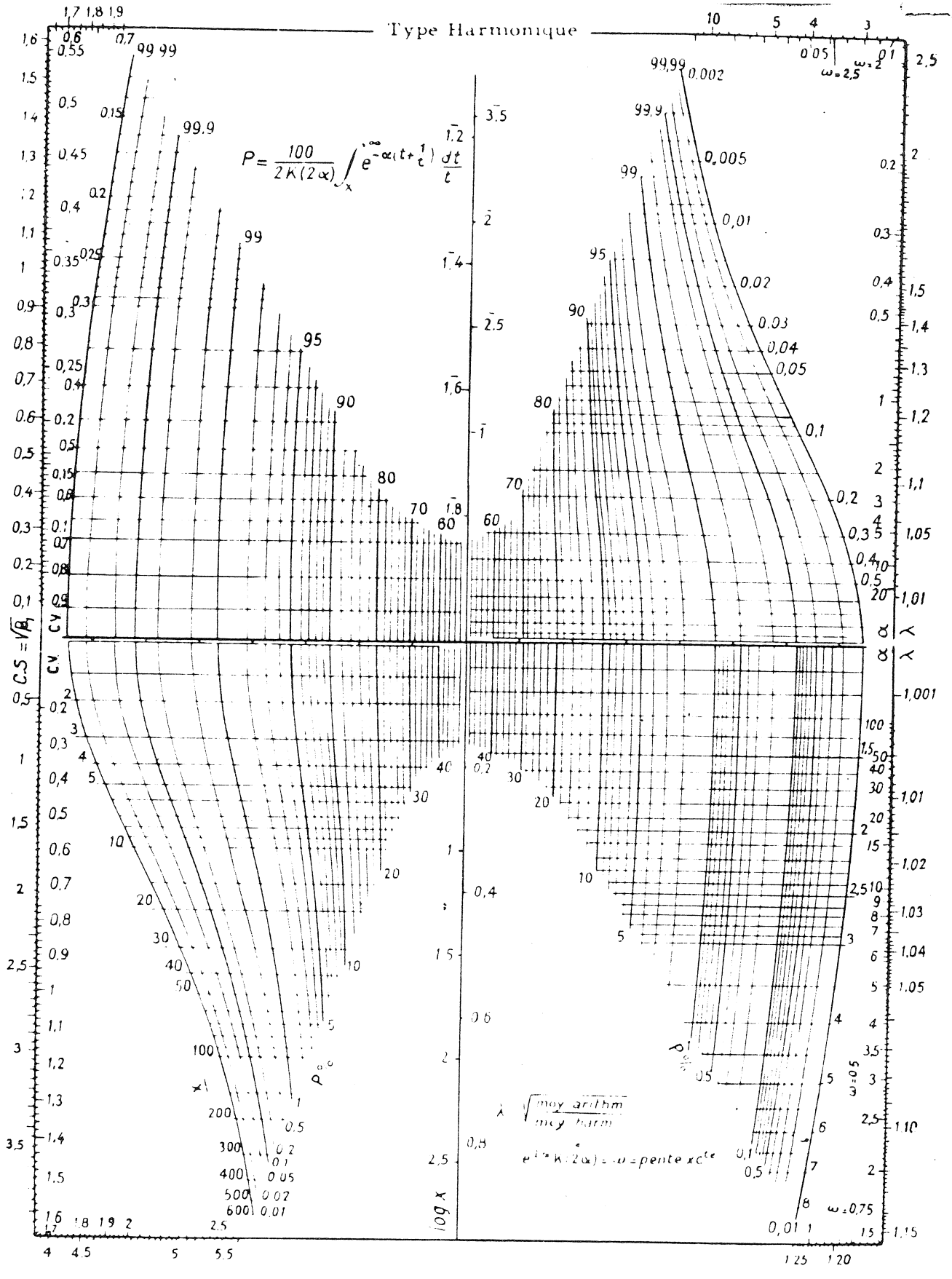

FU
U=6
5
4
3,5
3
2,5
2
1,7
1,5
1,3
1,0
,9
,8
,7
,6
,5
,4
,3
,2
,50
,45
,40
Cv
4
4,5
5
5,5
6
6,5
λ
7


F(u)
.999
.995
.990
.980
.970
.960
.950
.900
.800
.700
.600
.500
.400
.300
.200
.100
.050
.040
.030
.020
.010
.005
.001
U=7.5
7
6.5
6
5.5
5
4.5
4
3.5
3
2.5
2
1.7
1.5
1.3
1.0
.9
.8
U=.7
.30
.25
.20
Civ
7
10
15
20
25

N° 374


F(u)
999
995
990
980
970
960
950
900
800
700
600
500
400
300
200
100
050
010
030
020
010
005
001
U= 9,5
9
8,5
8
7,5
7
6,5
6
5,5
5
4,5
4
3,5
U= 3
Cv
,20
,19
,18
,17
,16
25
30
35
40
λ

Type Harmonique



III - L'APPLICATION DES MODELES PROBABILISTES AUX FAITS OBSERVES

Ayant établi la distribution empirique d'une série d'observations on est tenté de rechercher la fonction de répartition qui épouse au mieux celle-ci, de substituer à une représentation discontinue une représentation continue. Il est en effet fort possible de trouver une telle fonction avec suffisamment de paramètres, de même que l'on peut toujours faire passer par n points du plan un polynôme de degré n-1. Or "un cambrioleur n'ouvre pas une serrure avec une clé de cire mais avec un dispositif métallique auquel il s'efforce de donner la forme la plus probable convenant à la serrure". Cette phrase d'Etienne Halphen me paraît tout à fait adaptée pour caractériser le comportement que l'on doit avoir lors du choix de la fonction de répartition. Une fois ce choix effectué, restera le problème de l'estimation des paramètres de cette fonction à l'aide des observations. En fait, ces deux problèmes ne sont pas indépendants, ils interviennent lorsqu'on teste l'écart entre l'échantillon d'observations et le schéma probabiliste de référence.

o
o o

Dans l'exposé qui suit, nous insisterons essentiellement sur les raisons physiques qui permettent de "justifier" ou plutôt d'assurer la vraisemblance du choix d'une fonction de répartition pour représenter les débits, précipitations et températures, en ne traitant que brièvement l'utilisation des tests d'adéquation classiques dont l'intérêt, primordial dans d'autres domaines, est extrêmement limité en hydrologie, nous verrons pourquoi.

3.1 - Le choix du modèle probabiliste
(il ne s'agira ici que de fonctions à une variable)

3.1.1 - Les débits

Remarque :

L'ajustement d'une courbe analytique à l'ensemble des fréquences du débit journalier d'une rivière (365 jours x n années) ou courbe des débits

classés n'est pas une opération statistique au sens habituel, c'est-à-dire définition d'une fonction de répartition : elle permet seulement de calculer l'énergie productible en année moyenne.

Dans ce qui suit, nous considérons essentiellement un même phénomène, c'est-à-dire une urne de composition unique.

Dans tous ces problèmes d'ajustements, il y a 4 facteurs physiques fondamentaux à ne pas perdre de vue et qui assurent une garantie infiniment supérieure à l'utilisation scolaire de tests statistiques classiques :

- l'unité de temps 2, 4, ... 24 heures, 1 mois, 1 an
- l'unité de surface 10 km², 1 000 km², 10 000 km², 1000 000 km²
- le relief plaine-montagne
- la saison.

Ainsi la fonction de répartition normale peut être une excellente approximation dans les cas suivants :

- un bassin situé à haute altitude, dont l'alimentation est purement nivale, dans les Alpes essentiellement, pour les débits journaliers de printemps et a fortiori pour les débits moyens mensuels et moyens annuels, les apports sont régularisés par la fusion nivale ;

- un grand bassin (plusieurs dizaines de milliers de km²) pour les débits moyens mensuels et annuels, pluriannuels, du fait de la dimension du bassin, le débit à l'exutoire est la résultante d'un grand nombre d'effets liés à la pédologie, végétation, alimentation en eau, déphasage des débits des sous-bassins, nappes souterraines ;

- un bassin pluvial de quelques milliers de km² pour le débit moyen annuel ou bi-annuel.

Dans tous ces cas, on peut admettre que le débit est la résultante d'un très grand nombre d'effets additifs, d'importance équivalente.

L'approximation par la fonction gausso-normale n'est plus valable si l'on s'intéresse au débit journalier d'un bassin à alimentation pluviale. On constate alors que l'effet de la pluie sur le bassin, donc sur le débit, sera d'autant plus important que les conditions initiales de saturation seront importantes; on est tenté, et l'expérience le confirme, d'ajuster la fonction log-normale (Galton-Gibrat). Mais il ne faut pas oublier que si l'hypothèse d'effets multiplicatifs est satisfaisante pour les débits non extrêmes, à la limite elle conduit à deux absurdités :

- s'il n'y a plus d'eau dans la rivière, quelle que soit la quantité de pluie qui va tomber, il ne coulera rien ;

- si au contraire le terrain est complètement saturé, il coulera plus d'eau qu'il n'en tombe.

Cette fonction convient remarquablement pour représenter les distributions empiriques de débits moyens journaliers de bassins du Massif Central (500 à 30 000 km²), de bassins tels que le Rhône au Teil et le Rhin à Bâle en hiver. Parfois la dissymétrie de la fonction log-normale n'est pas suffisante et on peut utiliser alors non plus la fonction normale appliquée au logarithme du débit mais la fonction Gamma incomplète appliquée au logarithme du débit.

Une répartition dissymétrique s'observe aussi pour un bassin tel que la Loire à Blois (38 000 km²), lorsqu'on s'intéresse au débit moyen journalier maximum de l'automne (mois de septembre et octobre), c'est le cas de la plupart des débits extrêmes de crue de bassins à alimentation pluviale; dans ce cas, la méthode du Gradex permet d'obtenir la forme de fonction de répartition de ces débits dans la zone des faibles probabilités à l'aide d'un raisonnement physico-statistique: le comportement asymptotique de la distribution des débits extrêmes est le même que celui de la distribution des pluies extrêmes sur le bassin.

3.1.2 - Les précipitations

La pluie est un phénomène d'autant plus discontinu que l'on considère un intervalle de temps court.

Ainsi, dans le Massif Central en hiver (de novembre à mars), si on étudie les statistiques de précipitations en 2 heures, en effectuant le dépouillement de pluviogrammes, il ne pleut que dans 20 à 25 % des cas, ce chiffre peut avoisiner 5 à 10 % des cas dans les régions méditerranéennes en été. Si l'on s'intéresse à la précipitation journalière (08 - 08 h), la fréquence des pluies atteint 50 % dans l'ouest du Massif Central et 30 % dans les Alpes du Sud.

Les précipitations sont mesurées ponctuellement mais c'est un phénomène organisé dans l'espace et qui présente une certaine cohérence et homogénéité, compte tenu de l'effet du relief, alors que ce n'est pas le cas de débits à l'exutoire de bassins versants voisins.

Deux fonctions de répartition fournissent une excellente approximation de la distribution empirique des pluies pour des intervalles de temps compris entre 2 heures et 5 jours :

- la fonction Γ incomplète avec un paramètre de forme $\lambda < 1$ donc de forme en J ;
- la fonction $F(R) = 1 - \alpha e^{-\frac{R}{a}} - \beta e^{-\frac{R}{c}}$ dans laquelle $1 - (\alpha + \beta) = 1 - \theta$ représente la fréquence des pluies nulles $\alpha = \frac{m^2}{\sigma^2}$ et "a" le Gradex, c'est-à-dire le gradient des valeurs extrêmes; $F(o)$ et "a" sont les paramètres caractéristiques de la fonction de répartition des précipitations, le paramètre c étant secondaire :

$$\begin{cases} a = \dfrac{m\,\alpha + \sqrt{\alpha\,\beta\,(K\,\theta - m^2)}}{\alpha\,\theta} \\ c = \dfrac{m - \alpha a}{\beta} \end{cases} \qquad \text{avec } K = \frac{m^2 + \sigma^2}{2}$$

Cette dernière fonction est plus proche de la réalité que la fonction Γ incomplète à l'origine et dans les valeurs extrêmes.

Toutefois, pour déterminer la distribution des précipitations décadaire, mensuelle, plurimensuelle à une station, on obtient une excellente approximation à l'aide de la fonction Γ calculée par composition des

lois de probabilité de pluies journalières, en tenant compte de l'autocorrélation entre pluies successives si faible soit-elle : cette fonction est définie pour n jours par :

$$m_n = n \cdot m$$

$$\sigma_n^2 = \sigma^2 \left[n + \frac{2\, r_1}{1 - \theta\, r_1} \left(n - \frac{1 - \theta^n\, r_1^n}{1 - \theta r_1} \right) \right] \quad ; \; r_1 \neq \frac{1}{3} \quad ; \; \theta = \alpha + \beta$$

en notant m la moyenne de la précipitation journalière et σ^2 sa variance.

Remarque :

On a parfois utilisé la fonction normale appliquée à la racine carrée ou cubique ou racine $n^{ième}$ de la pluie, et même au logarithme. L'inconvénient de ces transformations est leur trop grande souplesse : on peut toujours trouver une transformation qui "collera" à l'échantillon considéré.

3.1.3 - Les températures

Il s'agit de la température de l'air mesurée à 1,50 mètre au-dessus du sol, sous abri. C'est sans doute le seul phénomène physique, avec la pression atmosphérique, dont la distribution soit remarquablement gaussienne quels que soient le lieu, l'unité de temps considérée, qu'il s'agisse des valeurs extrêmes en maximum, minimum ou de moyennes journalières, mensuelles, etc.

3.2 - Estimation des paramètres d'une fonction de répartition - Leur dispersion d'échantillonnage

Une distribution à une ou plusieurs dimensions étant connue dans sa forme analytique, sa fonction de répartition dépend d'un ou plusieurs paramètres de valeurs inconnues qu'il s'agit d'estimer au mieux à partir d'observations, exemples :

$$\frac{1}{\sigma\sqrt{2\pi}} \; e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$$

$$\frac{1}{\rho} \cdot \frac{1}{(\lambda-1)} \; e^{\frac{x}{\rho}} \cdot \left(\frac{x}{\rho}\right)^{\lambda-1}$$

Il y a deux méthodes d'estimation : l'une ponctuelle, l'autre par intervalles. Pour comparer les estimations d'un paramètre, on compare leur distribution d'échantillonnage.

Un <u>estimateur</u> est une fonction de n observations de l'échantillon qui fournit une valeur aléatoire dite <u>estimation</u>.

Quelques définitions :

- une estimation $t(x_1, x_2, \ldots, x_n)$ consistante, donc converge vers le paramètre estimé T si lorsque $n \longrightarrow \infty$ il existe ε et η petits tels que :

  probabilité $\left\{ |t - T| < \eta \right\} > 1 - \varepsilon$

- une estimation est sans distorsion ou biais si :

  $E(t) = T$

  exemples :

  . la moyenne de n observations $x_1, \ldots x_n$ est :

  $$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

  on peut considérer x_i comme des variables aléatoires indépendantes extraites de l'urne de composition F(x)

  $$E(\overline{x}) = \frac{1}{n} \Sigma \, E(x_i) = m$$

  on notera que $V(\overline{x}) = \frac{1}{n^2} \Sigma \, V(x_i) = \frac{\sigma^2}{n}$

  . la variance de n observations

  $$E(s^2) = E\left\{ \frac{1}{n} \Sigma (x_i - \overline{x})^2 \right\} = E\left\{ \frac{1}{n} \Sigma \left[(x_i - m) + m - \overline{x} \right]^2 \right\}$$

  plaçant l'origine à la moyenne de la population on a m = 0

$$E(s^2) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$$

c'est un estimateur avec biais; pour éliminer ce biais on écrit :

$$s^2 = \frac{1}{n-1} \sum (x_i - \overline{x})^2$$

- un estimateur sans distorsion est dit efficace si sa variance est minimum; on compare l'efficience de deux estimateurs à l'aide du rapport de leurs variances ;

- l'estimateur du paramètre γ est exhaustif, (c'est-à-dire qu'il n'y a pas perte d'informations) si la probabilité conditionnelle de $x_1, x_2, \ldots, x_n$, lorsqu'on connaît la valeur $t(x_1, \ldots, x_n)$, est indépendante de γ .

3.2.1 - Méthodes d'estimation ponctuelle

a) - la méthode des moments :

Elle consiste à égaler les moments théoriques de la fonction de répartition aux moments empiriques obtenus à partir des observations : ainsi pour la fonction Γ incomplète :

$$\hat{\lambda} = \frac{\overline{x}^2}{s^2} \qquad \text{avec} \qquad \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\hat{\rho} = \frac{s^2}{\overline{x}} \qquad\qquad s^2 = \frac{1}{n} \sum (x_i - \overline{x})^2$$

Elle est extrêmement pratique à mettre en application :

b) - La méthode du maximum de vraisemblance :

Considérons une densité de répartition $f(x, \gamma)$, qui caractérise la composition d'une urne. Les valeurs $x_1, x_2, \ldots, x_n$ obtenues à partir de n tirages dans cette urne sont des variables aléatoires indépendantes de même $f(x, \gamma)$, la probabilité élémentaire d'une telle suite est :

$$L(x_1, \ldots, x_n, \gamma)\, dx_1, dx_2, \ldots, dx_n = f(x_1, \gamma) \cdot f(x_2 \gamma), \ldots, f(x_n \gamma) \cdot dx_1, \ldots, dx_n$$

L est appelée la fonction de vraisemblance de l'échantillon.

La méthode du maximum de vraisemblance consiste à choisir un estimateur qui rende maximum L, soit pour $\log_e L$ ce qui revient au même :

$$\frac{\delta \log_e L}{\delta \gamma} = 0 \qquad (1)$$

Moyennant certaines conditions, existence des densités $\frac{\delta \log_e L}{\delta \gamma}$ et $\frac{\delta^2 \log_e L}{\delta \gamma^2}$ et $\frac{\delta^3 \log_e L}{\delta \gamma^3} < M(x)$, on démontre qu'il y a une probabilité $\longrightarrow 1$ pour que l'équation (1) ait une solution $\hat{\gamma}$ qui converge en probabilité vers la vraie valeur γ . Lorsque $n \longrightarrow \infty$.

Exemple de la fonction Γ incomplète :

$$L(x_i) = \frac{1}{\rho^{n\lambda} \left[\Gamma(\lambda)\right]^n} \cdot \exp\left(-\frac{\Sigma x_i}{\rho}\right) \cdot \sum_i x_i^{\lambda-1}$$

$$\log_e L = -n\lambda \log_e \rho - n \log_e \Gamma(\lambda) - \frac{\Sigma x_i}{\rho} + (\lambda - 1) \sum_1 \log_e x_i$$

La solution du maximum de vraisemblance consiste à écrire :

$$\frac{\delta \log_e L}{\delta \lambda} = 0$$

$$\frac{\delta \log_e L}{\delta \rho} = 0$$

soit
$$\begin{cases} \lambda\rho - \dfrac{\Sigma x_i}{n} = 0 \\ -\log_e \rho - \dfrac{d\log_e \Gamma(\lambda)}{d\lambda} + \dfrac{1}{n} \Sigma \log_e x_i = 0 \end{cases}$$

après élimination de $\rho = \frac{\overline{x}}{\lambda}$, on doit résoudre :

$$I(\lambda) = \log_e \lambda - \frac{d\log_e \Gamma(\lambda)}{d\lambda} = \log_e \overline{x} - \log_e g$$

(g étant la moyenne géométrique des x_i).

Une estimation approchée est fournie par :

$$\lambda' = \frac{1 + \sqrt{\frac{4}{3} z + 1}}{4 z} \text{ , avec } z = \log \bar{x} - \log g$$

Il existe des tables de la fonction $\frac{d \log_e \Gamma(\lambda)}{d\lambda}$

Calcul des variances de ces deux estimateurs :

1°- $\hat{\lambda} = \frac{\bar{x}^2}{s^2}$

On utilise l'expression classique pour une fonction de moments :

$$V(\hat{\lambda}) = (\frac{\delta\lambda}{\delta\sigma})^2 V(s) + \frac{2\,\delta\lambda}{\delta\sigma} \cdot \frac{\delta\lambda}{\delta m} \text{ cov } (\bar{x} \cdot s) + (\frac{\delta\lambda}{\delta m})^2 V(\bar{x})$$

sachant que $V(s) = \frac{1}{2n} \rho^2 (\lambda + 3)$

$$\text{Cov } (\bar{x}\, s) = \rho^2 \frac{\sqrt{\lambda}}{n}$$

$$V(\bar{x}) = \frac{\rho^2}{n}$$

d'où : $$V(\lambda) = \frac{2\lambda(\lambda + 1)}{n}$$

2°- $$\log_e \hat{\lambda} - \frac{d \log_e \Gamma(\hat{\lambda})}{d\hat{\lambda}} = \log_e \bar{x} - \log e\, g$$

On obtient la variance de cet estimateur λ en inversant la matrice d'information :

$$I(\gamma) = - E \left(\frac{\delta^2 \log_e L}{\delta \gamma_i \, \delta \gamma_j} \right) ; \quad \gamma_i \text{ sont les paramètres}$$

soit : $$V(\lambda) = \frac{\lambda}{n \left[\frac{d^2 \log_e \Gamma(\lambda)}{d\lambda^2} - 1 \right]}$$

Or la fonction $\frac{d^2 \log (\lambda - 1)!}{d\lambda^2}$ est tabulée

Notons :

H_o l'hypothèse de travail,

D une fonction des observations : on convient de rejeter H_o si d $> D_o$ correspondant à une probabilité fixée a priori,

$D > D_o$ est la zone de rejet ou région critique.

On teste l'hypothèse H_o contre une hypothèse alternative H_1 :

. erreur de première espèce : rejeter H_o quand H_o est vraie
erreur de probabilité α

. erreur de deuxième espèce : accepter H_o quand H_o est fausse
erreur de probabilité β

Le choix de la région critique s'effectue ainsi : il faut minimiser β ou maximiser $1 - \beta$ = puissance du test.

Exemple d'une fonction normale de variance égale à 1 et dont on cherche à tester pour la moyenne l'hypothèse nulle $m = h_o$, l'hypothèse alternative $m = h_1$, la moyenne arithmétique $\overline{x}$ étant calculée sur un échantillon de taille n :

La probabilité $(\overline{x} > c/H_o) = \alpha$, sachant que $\overline{x}$ est une variable aléatoire de densité de répartition :

$$f(\overline{x}/H_o) = \sqrt{\frac{n}{2\pi}}\, e^{-\frac{1}{2}(\overline{x} - h_o)^2 n}$$

d'où $u(\alpha) = \sqrt{n}\,(c - h_o)$ d'après la table de Gauss $\left[u, F(u)\right]$; d'autre part $(1 - \quad)$ = probabilité $(\overline{x} \in W/H_1)$, W étant la région critique, soit :

$$(1 - \beta) = \int_c^{\infty} \sqrt{\frac{n}{2\pi}}\, e^{-\frac{1}{2}(\overline{x} - h_1)^2 n}\, d\overline{x}$$

On peut voir que le risque de deuxième espèce diminue, ou $1 - \beta \longrightarrow 1$, lorsque n augmente.

3.3.1 - Le test du X^2 utilisé comme test d'adéquation :

Problème général : une fonction de répartition F(x) étant donnée, on veut lui confronter un échantillon de n valeurs $x_1, \ldots, x_n$.

Le principe consiste à associer à F(x) par un découpage en N points $\alpha_1, \ldots, \alpha_N$ une urne à N catégories.

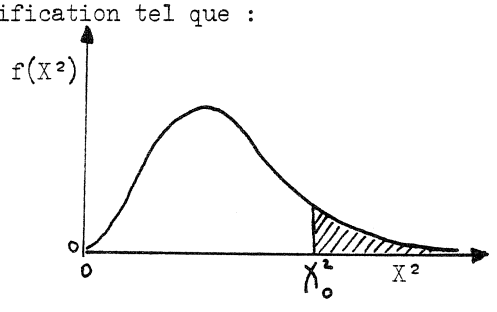
Une épreuve x constituera une épreuve de la $j^{ième}$ catégorie si $\alpha_{j-1} < x \leq \alpha_j$

Les probabilités attachées à cette urne seront données par :

$$p_j = \int_{\alpha_{j-1}}^{\alpha_j} d\,F(x)$$

on calcule alors la quantité $X^2 = \sum_{j=1}^{N} \frac{(n_j - np_j)^2}{np_j}$ qui est une "distance" entre la composition de l'urne fictive et la distribution empirique définissant l'échantillon : on teste si cette valeur est significative ou non. Pour cela, on se définit a priori un seuil de signification tel que :

Prob $(X^2 > X_o^2) = 1 - F(X_o^2) = \alpha$
c'est-àdire si le $X^2 > X_o$ on met en doute la représentativité de l'échantillon par F(x).



Ici, on peut ouvrir une parenthèse pour définir la loi multinomiale caractérisant la composition d'urnes à N catégories.

Soit une population dont chaque individu est affecté d'un caractère pouvant prendre N valeurs $A_1, A_2, \ldots A_N$ dont les probabilités respectives sont : $p_1, .., p_N$.

Quand on extrait de la population un échantillon de taille n, la probabilité d'obtenir n_1 individus possédant le caractère A_1 n_2, le caractère A_2 ... est :

$$\text{Prob}\,(n_1, \ldots, n_N) = \frac{n!}{n_1!, \ldots, n_N!}\; p_1^{n_1}, \ldots p_N^{n_N}\;; \text{avec } n = \sum_{i=1}^{N} n_i \qquad (1)$$

à l'aide de la formule de Stirling ($n! \# n^{n+\frac{1}{2}} e^{-n} \sqrt{2\pi}$) on peut montrer que les variables centrées réduites $x_1 = \frac{n_i - np_i}{\sqrt{n}}$, lorsque n croît $\longrightarrow \infty$, tendent vers la loi de Gauss :

$$\frac{1}{(2\pi)^{\frac{N}{2}}} \quad \frac{1}{\sqrt{p_1 \cdot p_2 \cdot \cdots \cdot p_N}} \quad \exp\left[-\frac{1}{2}\left(\frac{x_1^2}{p_1} + \ldots + \frac{x_N^2}{p_N}\right)\right]$$

Cette loi n'est autre que la fonction X^2 a $\nu = N - 1$ degré de liberté (les n_i sont liés par la relation (1) donc on enlève 1 degré de liberté à N). On voit ici le lien avec le problème évoqué en début de paragraphe.

De plus, on remarque que $X^2 = \sum_{j=1}^{k} \frac{(n_j - np_j)^2}{np_j}$ suit approximativement la loi du X^2 à K-1 degrés de liberté, la convergence ayant lieu lorsque $n \longrightarrow \infty$.

On ne sait pas facilement établir la loi de probabilité exacte de X^2, mais on peut comparer les moments des deux distributions, par exemple les deux premiers moments :

X^2 théorique	χ^2 réellement calculé
$m = (K-1)$;	$m' = K-1$
$\sigma^2 = 2(K-1)$;	$\sigma'^2 = 2(K-1) - \frac{1}{n}\left(k^2 + 2k - 2 - \sum \frac{1}{p_j}\right)$
$\mu_3 = 8(K-1)$;	$\mu'_3 = 8(K-1) - \frac{1}{n} k(k, p_j) + \frac{1}{n^2} g(k, p_j, p_j^2)$

Les deux moyennes sont identiques car :

$$E(X^2) = E\left[\frac{\sum (n_j - np_j)^2}{np_j}\right] = \frac{\sum np_j (1 - np_j)}{np_j} = \sum (1 - p_j) = K - 1$$

On voit que σ^2 dépend du nombre de classes k, de la taille de l'échantillon n et des probabilités p_j; les moments centrés de X^2 font inter-

venir en plus $\Sigma \frac{1}{p_j}^2$ pour μ'_3 et $\Sigma \frac{1}{p_j}^3$ pour μ'_4.

Dans le cas particulier où l'on prend tous les p_j égaux et pour un nombre de classes compris entre 2 et 10, on sait alors établir la loi exacte de X^2 dont les moments sont :

$$m' = K-1$$
$$\sigma'^2 = 2\,(K-1)\,(1 - \frac{1}{n})$$
$$\mu'^3 = 8\,(K-1)\,(1 + \frac{K-8}{2n} - \frac{K-6}{2n^2})$$

avec $p_j = \frac{1}{k}$

On peut alors montrer que la courbe continue du X^2 théorique à k-1 degrés de liberté est une bonne approximation de la fonction de répartition discontinue du X^2 calculé. On voit en particulier que les moments ne dépendent plus que de n pour la variance et K et n pour les moments d'ordre supérieur à 2.

Toutefois, même dans ce cas favorable, en prenant n = 35 et k = 7, on a alors p = .143. Même dans ce cas où l'on élimine l'arbitraire des limites de classes, on n'assure pas mieux la qualité de l'adéquation dans les queues de distributions.

Remarque : un trop grand nombre d'observations rend alors le test trop sensible à un écart qui ne pourrait être dû qu'à l'échantillonnage (pour les valeurs extrêmes surtout).

Généralement on confronte $F(x, \theta_1, \ldots, \theta_k)$ fonction dépendant de k paramètres à l'échantillon $x_1, \ldots, x_n$; on estime $\theta_1, \ldots, \theta_k$ à l'aide des observations, ce faisant on diminue le nombre de degrés de liberté de k, dans ce cas on utilisera un X^2 à n-k-1 degrés de liberté.

On voit tout de suite la limitation de ce test en Hydrologie où l'on a affaire à de petits échantillons : pour avoir un effectif de 4 à 5 valeurs par catégorie (α_{j-1}, α_i) on devra prendre un intervalle assez

large, on diminue ainsi la sensibilité du test, d'autre part les valeurs des queues de distribution qui sont les plus intéressantes car elles permettent de départager les fonctions de répartition confondues dans la zone centrale, sont inutilisables puisque d'effectif 1 ou 2.

3.3.2 - On peut alors essayer un test non paramétrique, le X^2 fait en effet partie des tests paramétriques, étant dépendant des paramètres de $F(x)$; Kolmogoroff a proposé de tester le plus grand écart entre la distribution empirique $H_n(x)$ et la distribution théorique $F(x)$, avec un risque α fixé à l'avance, on a calculé la probabilité :

$$\text{Prob}\left\{ \text{borne supérieure} \left| H_n(x) - F(x) \right| > D_n \right\} = \alpha .$$

Exemples des valeurs de $D_{n,\alpha}$

n	1	2	3	4	5	6	7
$\alpha = .10$	.95	.776	.642	.564	.510	.470	.438
$\alpha = .01$	.995	.929	.828	.733	.669	.618	.577

n	8	9	10	15	20	25	30	35
$\alpha = .10$	.411	.388	.368	.304	.264	.24	.22	$\frac{1.22}{\sqrt{n}}$
$\alpha = .01$	.543	.514	.430	.404	.356	.32	.29	$\frac{1.63}{\sqrt{n}}$

Il existe un nombre important d'autres tests paramétriques et non paramétriques, plus ou moins complexes, notre propos est simplement d'illustrer le mécanisme du test en statistique. Les tests ne sont que des garde-fous, et il serait malhonnête ou naïf d'utiliser seulement le résultat d'un test pour affirmer ou rejeter une hypothèse en hydrologie.

TABLE 5
FRACTILES DE LA LOI DE χ^2

La loi de χ^2 est définie par la probabilité élémentaire

$$f(\chi^2,\nu)\, d(\chi^2) = \frac{1}{2^{\nu/2}\,\Gamma(\nu/2)}\,(\chi^2)^{\frac{\nu}{2}-1}\, e^{-\frac{\chi^2}{2}}\, d(\chi^2) \quad (0 \leq \chi^2 < \infty)$$

Si $u_1, u_2, \ldots, u_n$ sont n variables indépendantes, distribuées suivant la loi normale réduite ($m = 0$, $\sigma = 1$), la somme de leurs carrés, $\sum_i u_i^2$ est distribuée en loi de χ^2 à $\nu = n$ degrés de liberté. Il en est de même pour $\sum_i \frac{(x_i - m)^2}{\sigma^2}$, si les n variables x_i indépendantes suivent la loi normale (m, σ). Si dans l'expression précédente on remplace m par son estimation $\bar{x} = \frac{1}{n}\sum_i x_i$, la quantité $\sum_i \frac{(x_i - \bar{x})^2}{\sigma^2}$ est encore distribuée en loi de χ^2, mais avec $\nu = n - 1$ degrés de liberté.

La table donne les fractiles de la loi de χ^2, c'est-à-dire les valeurs χ^2 telles que $Pr[\chi^2 < \chi^2_p] = P$. On a $Pr[\chi^2 > \chi^2_p] = 1 - P = Q$.

Par exemple, pour $\nu = 5$, on a

$$Pr[\chi^2 < 12,8] = 0,975 \qquad Pr[\chi^2 < 0,83] = 0,025$$
$$Pr[0,83 < \chi^2 < 12,8] = 0,95$$

	0,999	0,995	0,990	0,975	0,95	0,90	0,50	0,10	0,05	0,025	0,010	0,005	0,001	Q = 1 - P
P / ν	0,001	0,005	0,010	0,025	0,05	0,10	0,50	0,90	0,95	0,975	0,990	0,995	0,999	
1	-	-	-	0,001	0,004	0,016	0,455	2,71	3,84	5,02	6,63	7,88	10,8	
2	0,002	0,010	0,020	0,051	0,103	0,211	1,39	4,61	5,99	7,38	9,21	10,6	13,8	
3	0,024	0,072	0,115	0,216	0,352	0,584	2,37	6,25	7,81	9,35	11,3	12,8	16,3	
4	0,091	0,207	0,297	0,484	0,711	1,06	3,36	7,78	9,49	11,1	13,3	14,9	18,5	
5	0,210	0,412	0,554	0,831	1,15	1,61	4,35	9,24	11,1	12,8	15,1	16,7	20,5	
6	0,381	0,676	0,872	1,24	1,64	2,20	5,35	10,6	12,6	14,4	16,8	18,5	22,5	
7	0,598	0,989	1,24	1,69	2,17	2,83	6,35	12,0	14,1	16,0	18,5	20,3	24,3	
8	0,857	1,34	1,65	2,18	2,73	3,49	7,34	13,4	15,5	17,5	20,1	22,0	26,1	
9	1,15	1,73	2,09	2,70	3,33	4,17	8,34	14,7	16,9	19,0	21,7	23,6	27,9	
10	1,48	2,16	2,56	3,25	3,94	4,87	9,34	16,0	18,3	20,5	23,2	25,2	29,6	
11	1,83	2,60	3,05	3,82	4,57	5,58	10,3	17,3	19,7	21,9	24,7	26,8	31,3	
12	2,21	3,07	3,57	4,40	5,23	6,30	11,3	18,5	21,0	23,3	26,2	28,3	32,9	
13	2,62	3,57	4,11	5,01	5,89	7,04	12,3	19,8	22,4	24,7	27,7	29,8	34,5	
14	3,04	4,07	4,66	5,63	6,57	7,79	13,3	21,1	23,7	26,1	29,1	31,3	36,1	
15	3,48	4,60	5,23	6,26	7,26	8,55	14,3	22,3	25,0	27,5	30,6	32,8	37,7	
16	3,94	5,14	5,81	6,91	7,96	9,31	15,3	23,5	26,3	28,8	32,0	34,3	39,3	
17	4,42	5,70	6,41	7,56	8,67	10,1	16,3	24,8	27,6	30,2	33,4	35,7	40,8	
18	4,90	6,26	7,01	8,23	9,39	10,9	17,3	26,0	28,9	31,5	34,8	37,2	42,3	
19	5,41	6,84	7,63	8,91	10,1	11,7	18,3	27,2	30,1	32,9	36,2	38,6	43,8	
20	5,92	7,43	8,26	9,59	10,9	12,4	19,3	28,4	31,4	34,2	37,6	40,0	45,3	
21	6,45	8,03	8,90	10,3	11,6	13,2	20,3	29,6	32,7	35,5	38,9	41,4	46,8	
22	6,98	8,64	9,54	11,0	12,3	14,0	21,3	30,8	33,9	36,8	40,3	42,8	48,3	
23	7,53	9,26	10,2	11,7	13,1	14,8	22,3	32,0	35,2	38,1	41,6	44,2	49,7	
24	8,08	9,89	10,9	12,4	13,8	15,7	23,3	33,2	36,4	39,4	43,0	45,6	51,2	
25	8,65	10,5	11,5	13,1	14,6	16,5	24,3	34,4	37,7	40,6	44,3	46,9	52,6	
26	9,22	11,2	12,2	13,8	15,4	17,3	25,3	35,6	38,9	41,9	45,6	48,3	54,1	
27	9,80	11,8	12,9	14,6	16,2	18,1	26,3	36,7	40,1	43,2	47,0	49,6	55,5	
28	10,4	12,5	13,6	15,3	16,9	18,9	27,3	37,9	41,3	44,5	48,3	51,0	56,9	
29	11,0	13,1	14,3	16,0	17,7	19,8	28,3	39,1	42,6	45,7	49,6	52,3	58,3	
30	11,6	13,8	15,0	16,8	18,5	20,6	29,3	40,3	43,8	47,0	50,9	53,7	59,7	

Lorsque $\nu > 30$ on peut admettre que la quantité $\sqrt{2\chi^2} - \sqrt{2\nu - 1}$ suit approximativement la loi normale réduite.

Exemple :

Calculer la valeur χ^2_p correspondant à $P = 0,90$ lorsque $\nu = 41$. La table 1.3 donne, pour $P = 0,90$, $u = 1,2816$. D'où :

$$\chi^2_{0,90} = \frac{[u_{0,90} + \sqrt{2\nu - 1}]^2}{2} = \frac{1}{2}\left[1,2816 + \sqrt{82 - 1}\right]^2 = \frac{1}{2}(10,2816)^2 = 52,9$$

(Valeur qui coïncide d'ailleurs avec la valeur exacte)

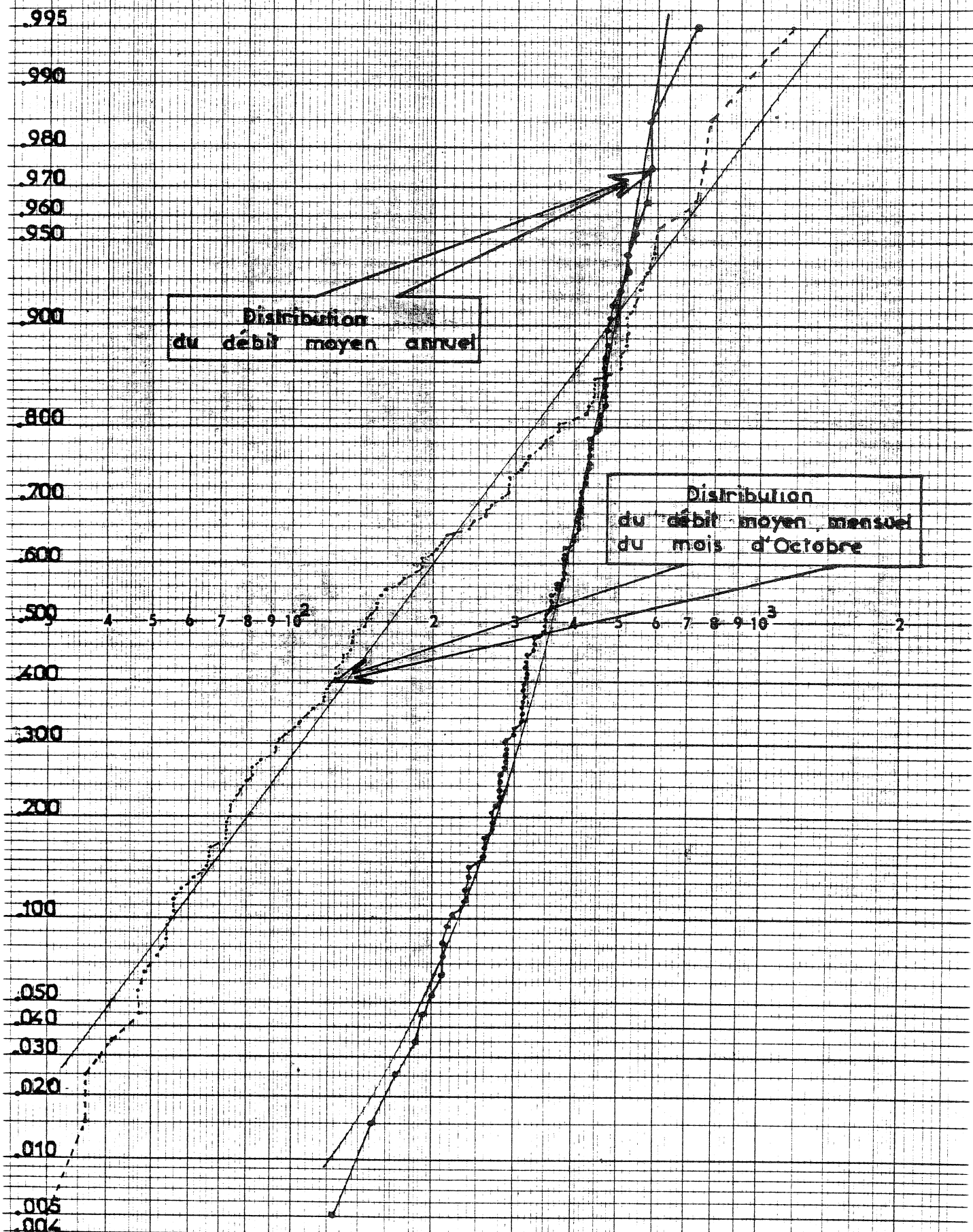
Station de M E S S E I X

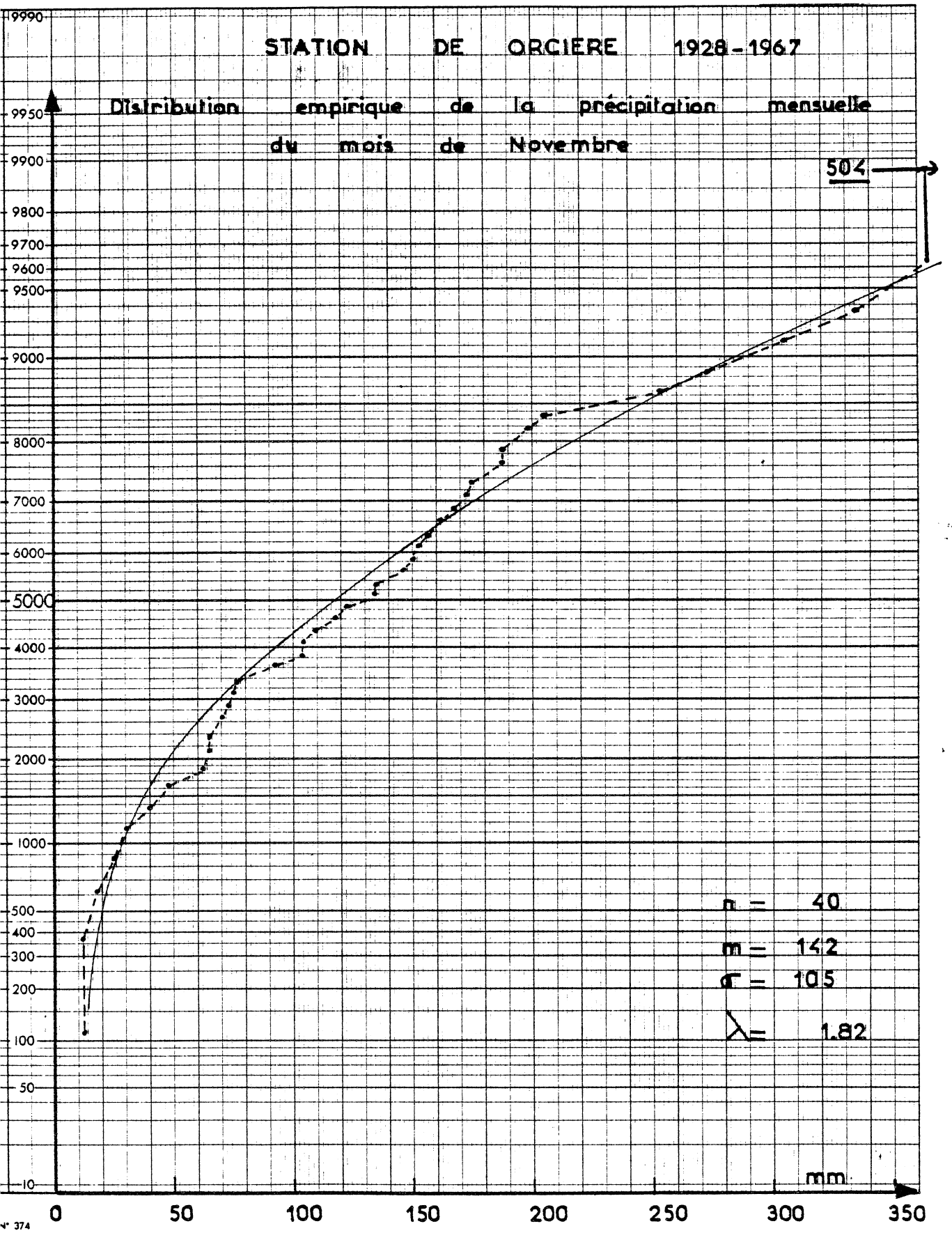
Températures moyennes

1933 - 1967

	15 mai	mai	annuelle
1933	11.5	10.9	8.5
34	13.5	13.2	9.0
35	9.5	9.7	8.1
36	12.5	11.4	8.5
37	12.5	12.7	8.9
38	16.7	9.8	8.8
39	12.0	7.9	8.0
1940	13.0	11.4	7.9
41	11.7	7.8	7.7
42	9.0	12.5	8.3
43	20.0	12.6	9.6
44	10.2	12.2	8.1
45	19.7	14.0	9.7
46	5.0	11.4	8.8
47	13.5	14.3	10.2
48	15.7	12.8	9.8
49	14.7	9.3	9.8
1950	16.0	13.2	8.8
51	5.5	9.1	8.2
52	14.5	13.1	8.7
53	11.0	13.0	8.7
54	10.2	10.4	7.6
55	8.0	10.7	8.5
56	11.0	12.0	6.6
57	14.0	8.7	8.3
58	10.5	13.5	8.3
59	8.7	12.0	9.3
1960	16.3	13.3	8.2
61	18.6	10.3	9.4
62	5.8	10.2	7.6
63	7.0	9.9	7.4
64	13.2	12.3	8.2
65	18.4	10.7	7.7
66	15.0	10.4	8.3
67	11.7	10.4	8.3
m	12.5	11.4	8.5

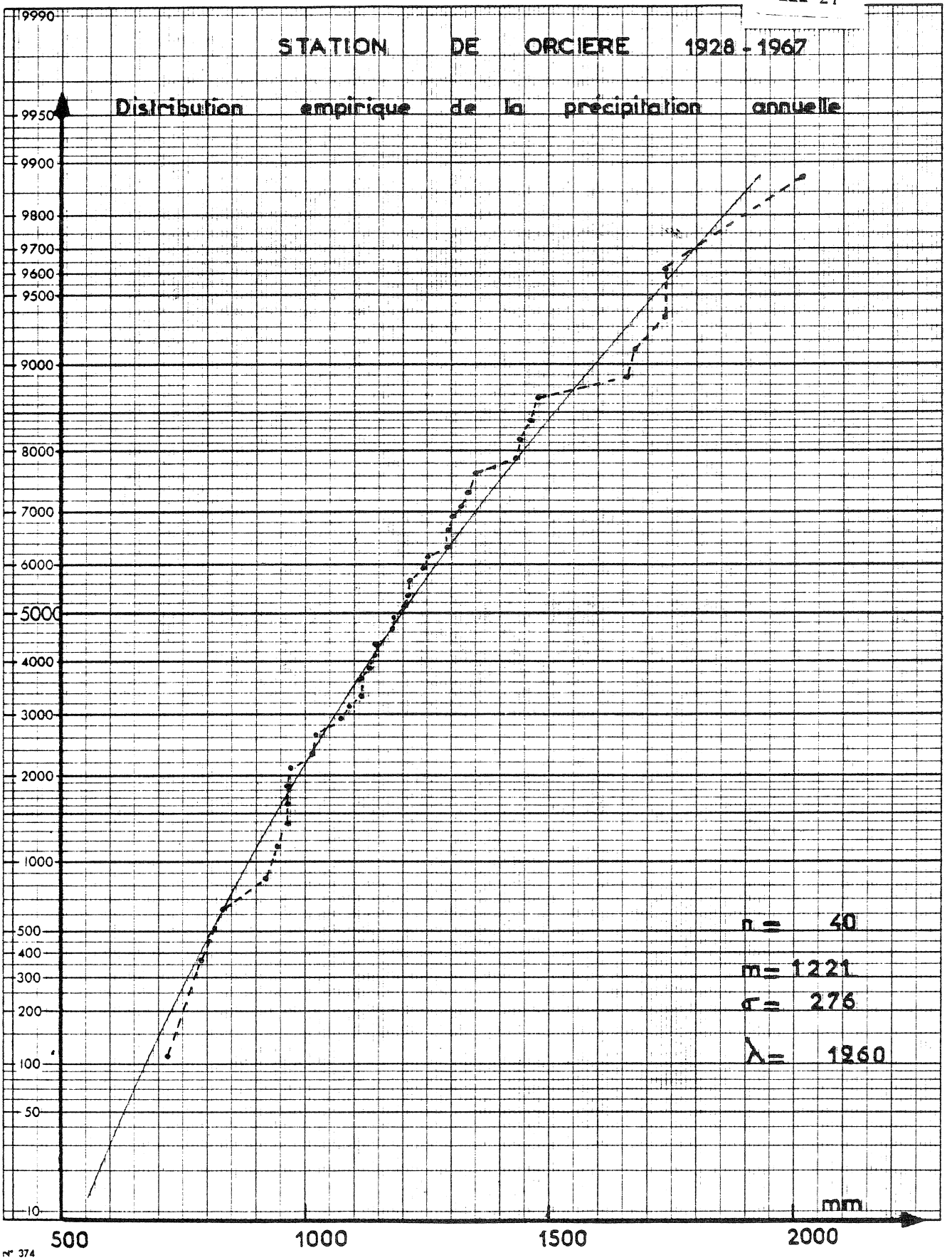
LA LOIRE A BLOIS 1863-1962



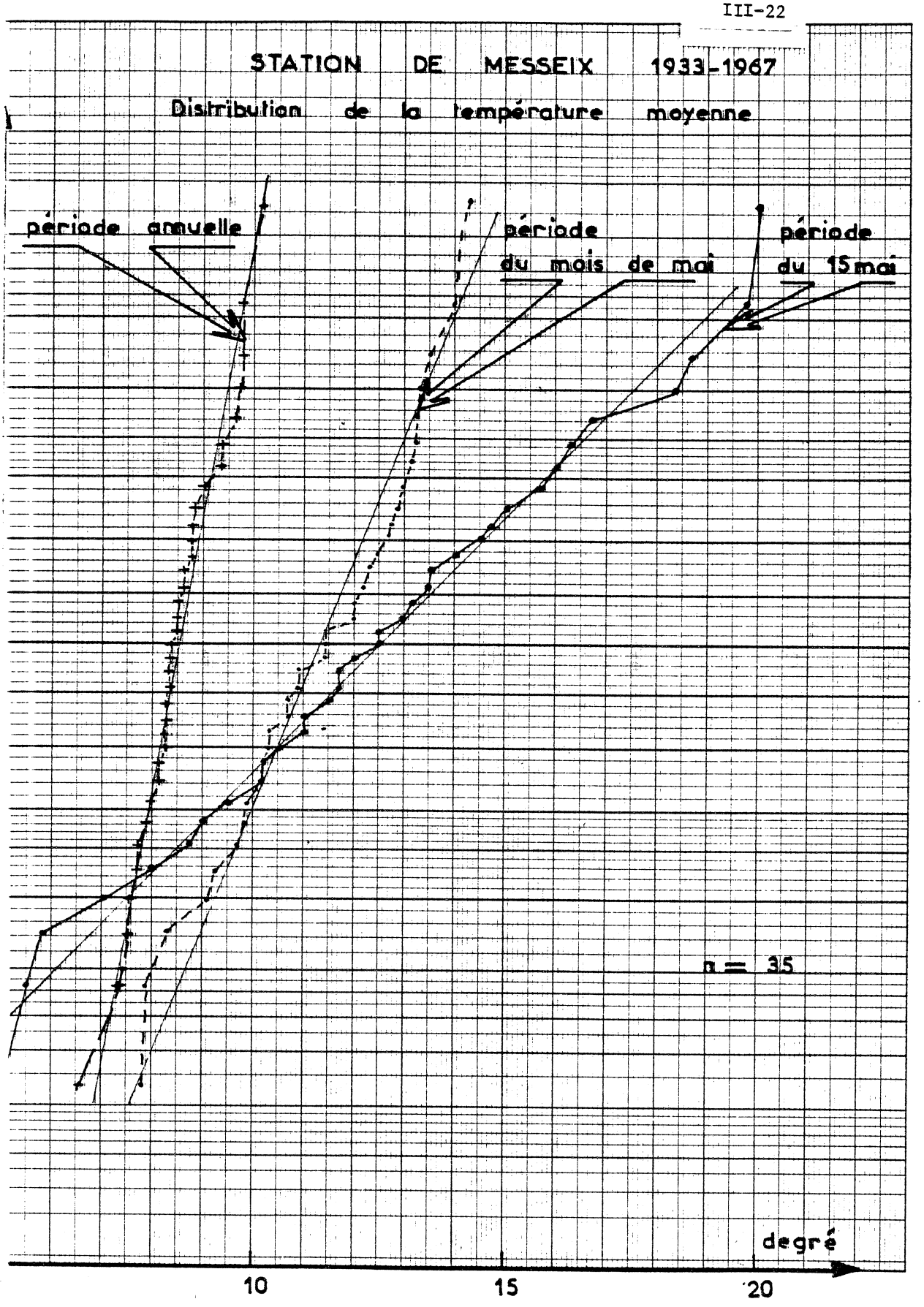

STATION DE ORCIERE 1928-1967
Distribution empirique de la précipitation mensuelle du mois de Novembre
504
n = 40
m = 142
σ = 105
λ = 1.82
mm
9990
9950
9900
9800
9700
9600
9500
9000
8000
7000
6000
5000
4000
3000
2000
1000
500
400
300
200
100
50
10
0
50
100
150
200
250
300
350

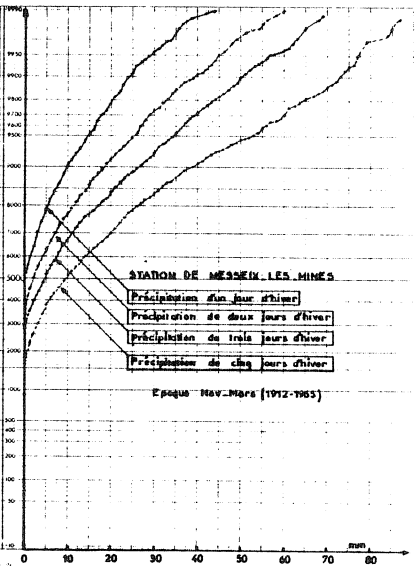
STATION DE ORCIERE 1928 - 1967

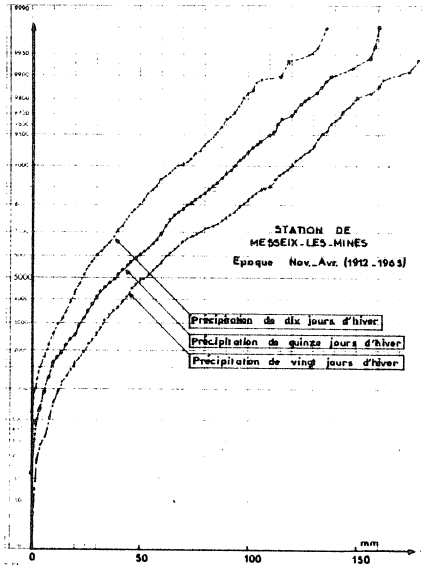
Distribution empirique de la précipitation annuelle

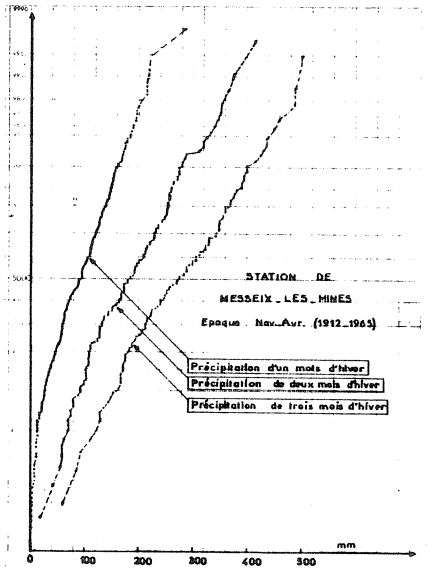


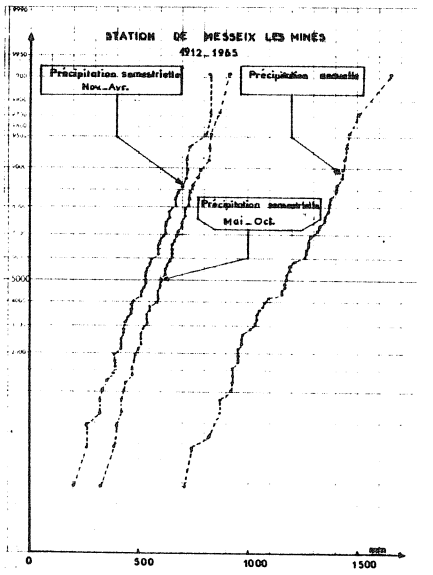
N° 374

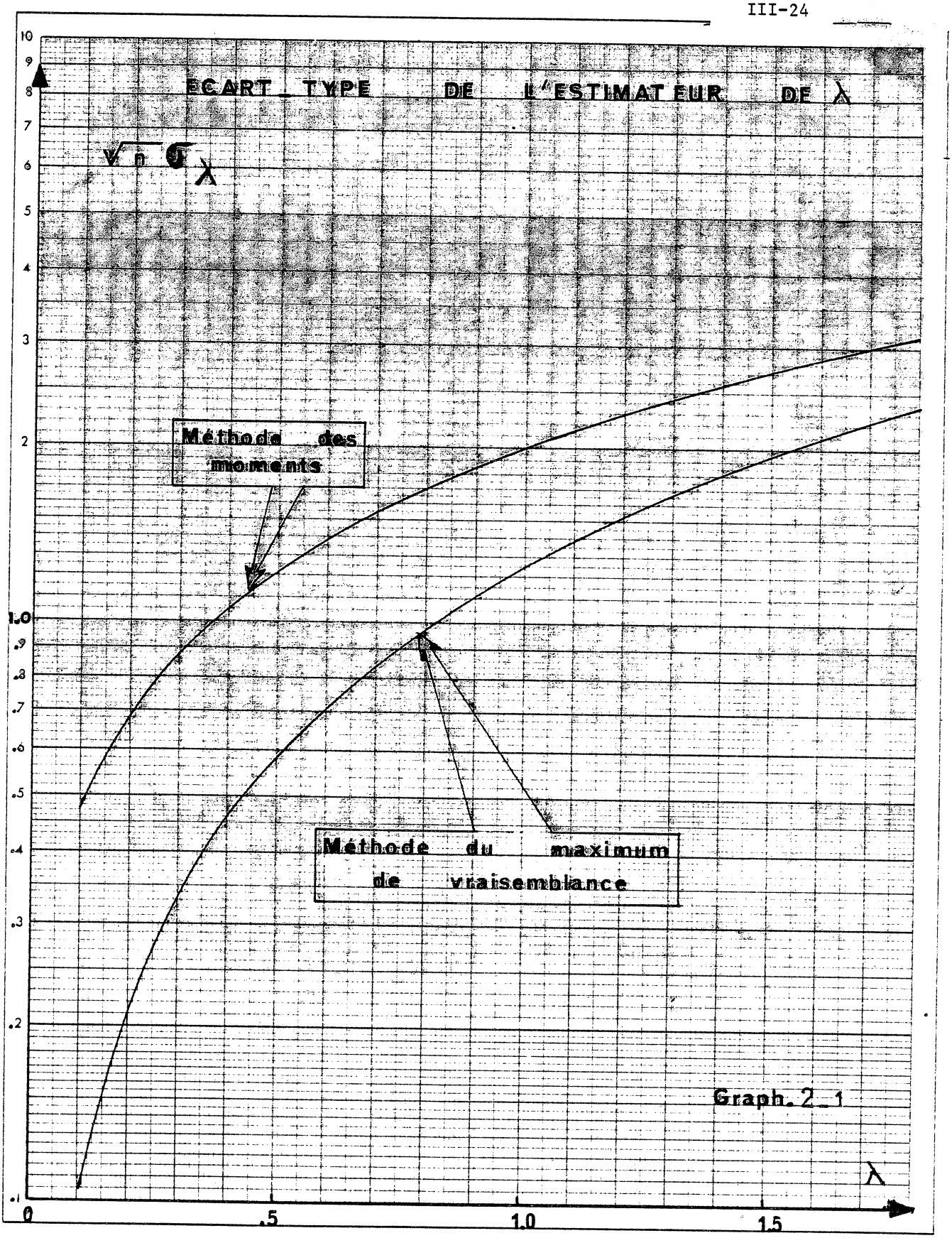

STATION DE MESSEIX 1933-1967
Distribution de la température moyenne
période annuelle
période du mois de mai
période du 15 mai
n = 35
degré
10
15
20


STATION DE MESSEIX - LES - MINES
Précipitation d'un jour d'hiver
Précipitation de deux jours d'hiver
Précipitation de trois jours d'hiver
Précipitation de cinq jours d'hiver
Epoque Nov._Mars (1912-1965)
mm
0
10
20
30
40
50
60
70
80

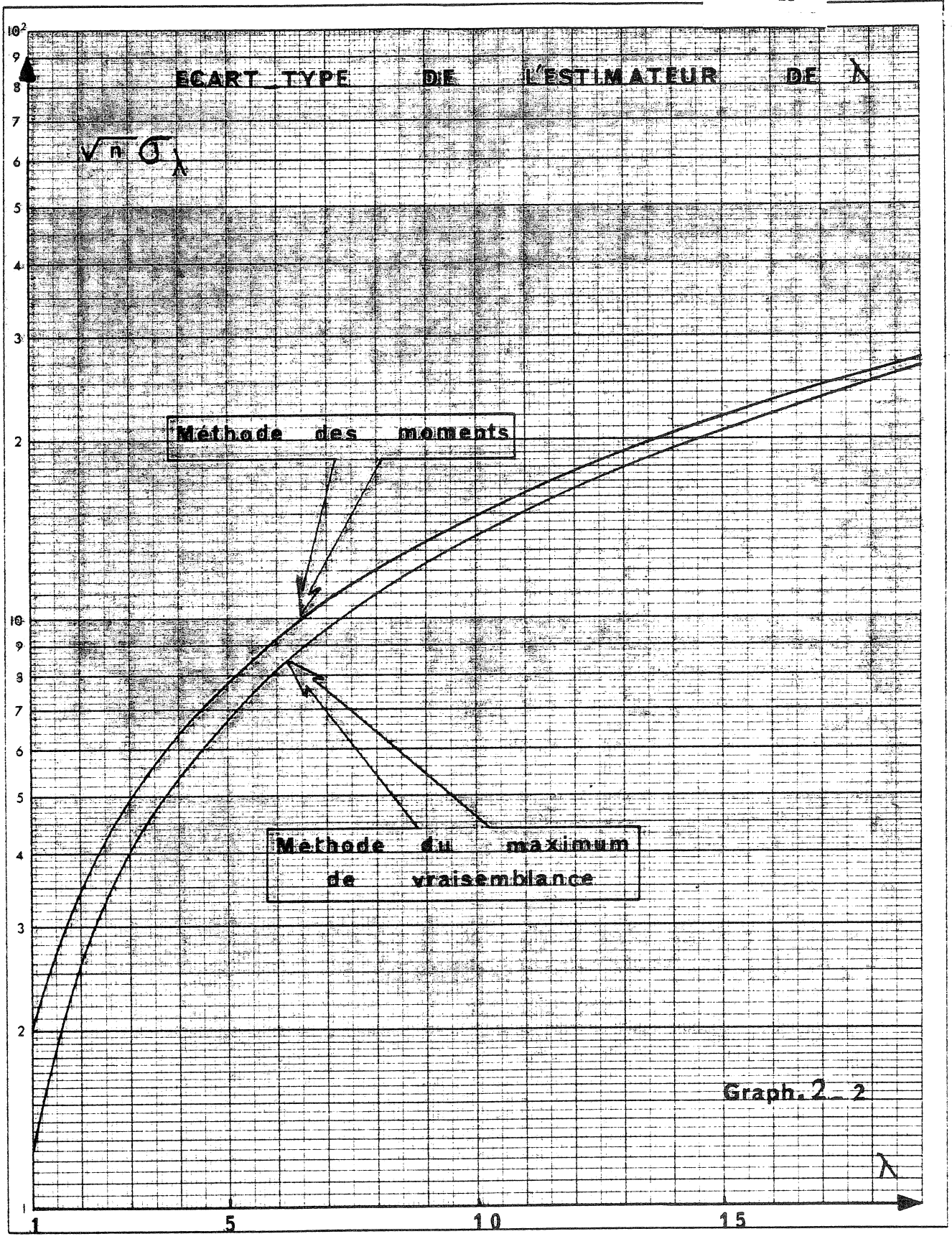

STATION DE
MESSEIX - LES - MINES
Epoque Nov._Avr. (1912 - 1965)
Précipitation de dix jours d'hiver
Précipitation de quinze jours d'hiver
Précipitation de vingt jours d'hiver
mm
0
50
100
150


STATION DE
MESSEIX _ LES _ MINES
Epoque Nov._Avr. (1912_1965)
Précipitation d'un mois d'hiver
Précipitation de deux mois d'hiver
Précipitation de trois mois d'hiver
mm
0
100
200
300
400
500


STATION DE MESSEIX LES MINES
1912_1965
Précipitation semestrielle
Nov._Avr.
Précipitation annuelle
Précipitation semestrielle
Mai_Oct.
mm
0
500
1000
1500


ECART_TYPE DE L'ESTIMATEUR DE λ
√n σλ
Méthode des moments
Méthode du maximum de vraisemblance
λ
0
.5
1.0
1.5
10
9
8
7
6
5
4
3
2
1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1

Graph. 2_1


ECART TYPE DE L'ESTIMATEUR DE λ
√n σλ
Méthode des moments
Méthode du maximum de vraisemblance
λ
1
5
10
15

Graph. 2_2

DEBITS MOYENS DU MOIS D'OCTOBRE

La LOIRE à BLOIS

	m³		m³		m³		m³
1863	545	1888	170	1913	425	1938	194
64	136	89	137	14	149	39	350
65	65	1890	172	15	120	1940	444
66	600	91	317	16	291	41	125
67	371	92	245	17	187	42	81
68	715	93	503	18	141	43	333
69	61	94	66	19	85	44	505
1870	41	95	47	1920	439	45	54
71	105	96	549	21	52	46	54
72	782	97	314	22	147	47	36
73	99	98	80	23	119	48	74
74	155	99	91	24	281	49	30
75	426	1900	508	25	125	1950	49
76	95	01	760	26	57	51	107
77	73	02	201	27	239	52	203
78	192	03	104	28	82	53	131
79	136	04	76	29	120	54	136
1880	265	05	310	1930	441	55	58
81	119	06	36	31	143	56	367
82	501	07	1175	32	289	57	59
83	259	08	73	33	590	58	254
84	122	09	212	34	65	59	73
85	504	1910	291	35	214	1960	562
86	161	11	79	36	136	61	74
87	114	12	151	37	92	62	47

IV - NOTIONS D'ERREUR - AJUSTEMENT D'UNE FONCTION

Nous essaierons dans ce chapitre d'analyser cet ensemble d'incertitudes que le technicien (physicien, ingénieur, ...) baptise "erreurs".

Les quelques réflexions qui suivent, tentent de mieux appréhender la notion d'erreur et remettent en cause des définitions inadéquates, pour répondre à la demande fréquente de ceux qui effectuent la mesure de phénomènes physiques ou qui interprètent (au sens large) ces observations numériques.

Il n'est peut être pas inutile de rappeler que chaque fois que l'on effectue une mesure, on ne trouve qu'une valeur approchée de la grandeur que l'on mesure, et, si l'on répète un grand nombre de fois la mesure de cette grandeur dans les mêmes conditions, on observera une certaine répartition de ces résultats de part et d'autre de la "vraie" valeur.

De plus il est fréquent d'utiliser ces résultats de mesures (x) pour étalonner une relation physique ou statistique de la forme $y = f(x)$. Dans œt ajustement, en général, le modèle dont on cherche à caler les paramètres d'après les observations, n'est lui même qu'une approximation de relations réelles, aussi précis et fidèle soit-il, même s'il est basé sur des considérations théoriques.

On se trouve confronté essentiellement à quatre types d'erreurs :

1°- l'erreur d'adéquation du modèle, c'est-à-dire la distance entre une structure d'hypothèses et la réalité, il serait préférable de dire l'observation de la réalité ; nous incluons dans ce cas l'effet d'échantillonnage, c'est-à-dire domaine de mesures et leur fréquence ;

2°- l'erreur de mesure, purement aléatoire, et qui existe quelles que soient les qualité et précision des appareils de mesure ;

3°- l'erreur absurde (une valeur exceptionnellement fausse pour une cause fortuite);

4°- L'erreur systématique, elle est généralement inhérente à l'appareil de mesure. On peut sans doute inclure dans cette catégorie l'erreur due à la non invariance des conditions de mesure, dans le temps.

Ces deux derniers types d'erreurs sont les plus faciles à détecter.

Une telle classification peut paraître séduisante et simple, mais ce qui fait la difficulté des problèmes d'erreurs est qu'en pratique ces quatre types d'erreurs sont mélangés et qu'il n'est pas toujours aisé d'effectuer une discrimination pour identifier la ou les sources d'erreur et faire la part de leur contribution à l'erreur globale.

Les exemples où de tels problèmes sont rencontrés ne manquent pas :

. ajustement d'une courbe de tarage - contrôle piézométrique - relation pluie-débit ou débit-débit (propagation) - auscultation d'un grand barrage, etc.

4.1 - Revue sommaire de quelques définitions classiques

Dans de nombreux ouvrages (physique ou mathématique appliquée) qui traitent ce problème, on calcule une erreur (absolue ou relative) maximale possible.

Ainsi, soit M une grandeur dont la valeur m est obtenue en mesurant les valeurs x, y, z des grandeurs X, Y, Z et en appliquant la relation :

$$m = f(x, y, z)$$

par exemple,

$$- \; m = x + y - z$$

l'erreur absolue maximale (?) s'évalue comme la somme des erreurs absolues maximales (?) commises sur chacun des termes $\Delta m = \Delta x + \Delta y + \Delta z$

$$- \; m = \frac{xy}{z}$$

l'erreur relative maximale (?) s'évalue comme la somme des erreurs relatives maximales (?)

$$\frac{\Delta m}{m} = \frac{\Delta x}{x} + \frac{\Delta y}{y} + \frac{\Delta z}{z}$$

De façon générale, les deux principales préoccupations sont, sur un exemple simple :

(a) connaissant la limite supérieure des erreurs obsolues faites sur x et y, trouver la limite supérieure de l'erreur faite sur $m = f(x, y)$;

(b) on veut calculer m avec une erreur absolue inférieure à ε, avec quelle précision faut-il connaître x et y ?

Réponse à (a) : k et h étant les erreurs faites sur x et y (sans autre définition), on calcule les dérivées partielles du premier ordre en partant d'une solution approchée :

$$m' = f(x + h, y + k), \text{ alors}$$

$$\Delta f = f(x+h, y+k) - f(x, y) \# h' f_x(x_1, y_1) + k f'_y (x_1, y_2),$$

(x_1 et y_1 valeurs voisines de x, y)

$\left|\Delta f\right| \leq \left|h f'_x\right| + \left|k f'_y\right|$, on calcule alors le second membre par excès.

Réponse à (b) : on veut $\left|\Delta f\right| < \varepsilon$, soit $\left|h f'_x\right| + k \left|f'_y\right| < \varepsilon$, il suffira de prendre $|h| < \frac{\varepsilon}{2\,|f'_x|}$ et $|k| < \frac{\varepsilon}{2\,|f'_y|}$ le second membre étant calculé par défaut.

Ainsi ces calculs d'erreurs ne font explicitement référence qu'à l'erreur de mesure, mais sans évoquer la possibilité d'une référence probabiliste, par l'intermédiaire de la fonction de répartition de l'erreur.

Si, parfois, on aborde l'aspect aléatoire de l'erreur, on fait systématiquement référence à la loi "normale" des erreurs, ce qui est trop restrictif.

Nous développerons dans la suite les différents aspects de l'erreur en étudiant les problèmes d'ajustement d'une fonction ou modèle.

4.2 - Ajustement d'une relation fonctionnelle entre deux variables continues

(approximation au sens mathématique, lorsque l'erreur de mesure est nulle).

Soit $y = g(x)$ cette relation fonctionnelle.

On veut l'approximer par une fonction simple ou "modèle" de la forme :

$$y = f(x) + \varepsilon \qquad \text{pour } x_I \leqslant x \leqslant x_S$$

$f(x, a, b, c, \ldots)$ représente une fonction de x dépendant d'un ou plusieurs paramètres a, b, c, ...; ε représente l'écart entre la liaison réelle et le modèle adopté comme image de cette relation.

Le modèle, défini par une expression analytique, va résulter d'un choix a priori effectué parmi plusieurs fonctions possibles. Il restera alors à calculer les valeurs des paramètres a, b, c, ... en imposant certaines conditions à ε, c'est-à-dire une norme, pour que la distance entre le modèle et la réalité soit la plus petite possible. Citons quelques critères de proximité entre $f(x)$ et $g(x)$:

- minimiser le plus grand écart ε, soit

$$\text{Min}\left\{ \text{Max} \left[g(x) - f(x, a, b, c) \right] \right\} ;$$

- minimiser la somme des valeurs absolues des écarts ε, soit

$$\text{Min} \int_{x_I}^{x_S} |\varepsilon| \, dx = \int_{x_I}^{x_S} \left| g(x) - f(x, a, b, c) \right| \, dx$$

cela revient à minimiser la somme des aires, mesurées en valeur absolue, et situées entre $g(x)$ et $f(x, a, b, c)$;

- minimiser la somme des carrés des projections orthogonales de $g(x)$ sur $f(x, a, b, c)$

- minimiser la somme des carrés des écarts entre $f(x, a, b, c)$ et $g(x)$, soit

$$\text{Min} \int_{x_I}^{x_S} \varepsilon^2 \, dx = \int_{x_I}^{x_S} \left[g(x) - f(x, a, b, c) \right]^2 dx$$

Les deux premiers critères sont plus lourds et plus coûteux à mettre en oeuvre. Le dernier critère est d'un usage plus commode et plus simple, nous l'expliciterons plus en détail, on l'appelle aussi méthode des moindres carrés.

Chercher le minimum de l'expression entre crochets revient à écrire que les dérivées partielles d'ordre 1, par rapport à a, b, c, ... sont nulles et résoudre ce système de p équations à p inconnues si la fonction $f(x, a, b, c, \ldots)$ dépend de p paramètres.

En supposant les conditions de continuité, dérivabilité, ... requises, on peut dériver sous le signe somme et écrire :

$$\int_{x_I}^{x_S} \frac{\delta}{\delta a} \left[g(x) - f(x, a, b, c, \ldots) \right]^2 dx = 0$$

$$\int_{x_I}^{x_S} \frac{\delta}{\delta b} \left[g(x) - f(x, a, b, c, \ldots) \right]^2 dx = 0$$

.....

soit encore

$$-2 \int_{x_I}^{x_S} \left[g(x - f(x, a, b, c, \ldots) \right] \frac{\delta f(x, a, b, c, \ldots)}{\delta a} dx = 0$$

$$-2 \int_{x_I}^{x_S} \left[g(x) - f(x, a, b, c, \ldots) \right] \frac{\delta f(x, a, b, \ldots)}{\delta b} dx = 0$$

.....

Comme exemple d'application nous développerons ces calculs en utilisant le modèle linéaire.

On considère la relation théorique $y = g(x) = e^x$ et l'on va approximer cette fonction par les quatre modèles simples suivants (la liste n'est pas exhaustive) :

$$\left.\begin{aligned} f_1(x) &= a_1\, x + c_1 \\ f_2(x) &= a_2\, x^2 + b_2\, x + c_2 \\ f_3(x) &= a_3\, x^2 + c_3 \\ f_4(x) &= a_4\, x^{\alpha} \end{aligned}\right\} \quad \text{pour } 1 \leqslant x \leqslant 2.$$

4.2.1 - Ajustement de $f_1(x)$

$$\text{Min} \int_1^2 (e^x - a_1\, x - c_1)^2\, dx$$

revient à résoudre :

$$\begin{cases} -2\displaystyle\int_1^2 (e^x - a_1\, x - c_1)\, dx = 0 \\ -2\displaystyle\int_1^2 (e^x - a_1\, x - c_1)\, x\, dx = 0 \end{cases}$$

d'où

$$a_1 = 4.60 \text{ et } c_1 = -2.23$$

$$f_1(x) \simeq 4.60\, x - 2.23$$

4.2.2 - Ajustement de $f_2(x)$

$$\text{Min} \int_1^2 (e^x - a_2\, x^2 - b_2\, x - c_2)^2\, dx$$

revient à résoudre :

$$\begin{cases} -2\int_1^2 (e^x - a_2 x^2 - b_2 x - c_2)\, dx = 0 \\ -2\int_1^2 (e^x - a_2 x^2 - b_2 x - c_2)\, x\, dx = 0 \\ -2\int_1^2 (e^x - a_2 x^2 - b_2 x - c_2)\, x^2\, dx = 0 \end{cases}$$

d'où : $a_2 = 2.458, \quad b_2 = 2.784, \quad c_2 = 3.111$

et $f_2(x) = \underset{\sim}{} 2.46\, x^2 - 2.78\, x + 3.11$

4.2.3 - Ajustement de $f_3(x)$

$$\text{Min} \int_1^2 (e^x - a_3 x^2 - c_3)^2\, dx$$

revient à résoudre :

$$\begin{cases} -2\int_1^2 (e^x - a_3 x^2 - c_3)\, dx = 0 \\ -2\int_1^2 (e^x - a_3 x^2 - c_3)\, x^2\, dx = 0 \end{cases}$$

d'où : $a_3 = 1.537, \quad c_3 = 1.085$

et $f_3(x) = \underset{\sim}{} 1.54\, x^2 + 1.09$

4.2.4 - Ajustement de $f_4(x)$

Dans ce cas on doit effectuer un changement de variable à l'aide de la transformation logarithmique :

$$\log(y) = \log\left[f_4(x)\right] = \alpha \log x + \beta$$

or $\log(y) = \log(e^x) = x$;

Chercher le minimum de $\int_1^2 (x - \alpha \log x - \beta)^2 \, dx$

revient à résoudre :

$$\begin{cases} -2 \int_1^2 (x - \alpha \log x - \beta) \, dx = 0 \\ -2 \int_1^2 (x - \alpha \log x - \beta) \log x \, dx = 0 \end{cases}$$

d'où :

$$\alpha = 1.458, \qquad \beta = .936$$

et

$$f_4(x) \simeq 2.55 \, x^{1.46}$$

4.2.5 - Précision de ces ajustements

Si l'on note s^2 cette précision, il suffira de calculer les quatre valeurs de S^2 :

$$S_j^2 = \int_1^2 \left[e^x - f_j(x) \right]^2 dx \qquad \text{pour } j = 1, 2, 3, 4$$

On voit que le meilleur ajustement est fourni par le modèle $f_2(x)$ dans l'intervalle $1 \leq x \leq 2$.

4.3 - Ajustement d'une relation fonctionnelle sur des couples de valeurs discrètes

La fonction $y = g(x)$ est définie par points, n couples de valeurs (x_i, y_i), avec exactitude; on est donc ramené au problème précédent mais en considérant des valeurs discrètes, l'application des moindres carrés revient à chercher le minimum de :

$$E = \sum_{i=1}^{n} \left[y_i - f(x_i, a, b, c) \right]^2$$

soit à résoudre le système (en se limitant à 3 paramètres) :

$$\begin{cases} \dfrac{\delta E}{\delta a} = 0 \\ \dfrac{\delta E}{\delta b} = 0 \\ \dfrac{\delta E}{\delta c} = 0 \end{cases}$$

et nous sommes ramenés au cas précédent, si l'on prend comme exemple l'ajustement des quatre modèles proposés $f_j(x, a, b, c)$ à la fonction $y = e^x$.

Remarque importante : bien que x_i et y_i soient des valeurs mesurées sans erreur, la répartition des n valeurs de x_i sur le segment $[1, 2]$ va conditionner les ajustements par les moindres carrés.

Prenons comme exemple les dix couples suivants :

$x_1 = 1.05$, $y_1 = 2.8577$; $x_2 = 1.1$, $y_2 = 3.0042$; $x_3 = 1.15$, $y_3 = 3.1582$;
$x_4 = 1.20$, $y_4 = 3.3201$; $x_5 = 1.25$, $y_5 = 3.4903$; $x_6 = 1.30$, $y_6 = 3.6693$;
$x_7 = 1.35$, $y_7 = 3.8574$; $x_8 = 1.40$, $y_8 = 4.0552$; $x_9 = 1.6$, $y_9 = 4.9530$;
$x_{10} = 1.9$, $y_{10} = 6.6859$.

Dans ce cas les paramètres des 4 modèles précédents estimés par les moindres carrés fournissent les ajustements suivants :

$f_1(x) = 4.45\ x - 2$ précision $s_{\varepsilon_1} = .158$
$f_2(x) = 2.18\ x^2 - 1.96\ x + 2.52$ $s_{\varepsilon_2} = .011$
$f_3(x) = 1.52\ x^2 + 1.13$ $s_{\varepsilon_3} = .049$
$f_4(x) = 2.58\ x^{1.42}$ $s_{\log \varepsilon_4} = .025$

Les paramètres de ces ajustements sont à comparer a ceux du § 4.2.

4.4 - Ajustement d'une relation avec erreur de mesure

Nous calculons à présent les paramètres respectifs des quatre modèles précédents, disposant de 20 couples de valeurs (x_i, y_i), x_i étant

une valeur certaine et y_i mesuré avec une erreur dont nous avons imposé la fonction de répartition gaussienne de moyenne nulle et d'écart type .2. Par la méthode de Monte-Carlo, nous avons alors tiré 20 écarts ε_i que l'on a rajouté aux 20 valeurs exactes de Y_i. Pour éviter le problème posé par la répartition des x_i, nous avons adopté une répartition uniforme sur le segment $[1,2]$:

x_i	1.05,	1.10,	1.15,	1.20,	1.25,	1.30,	1.35,	1.40,	1.45,	1.50,
y_i	2.86,	3.10,	3.10,	3.23,	3.58,	3.89,	4.21,	3.68,	4.28,	4.71,

x_i	1.55,	1.60,	1.65,	1.70,	1.75,	1.80,	1.85,	1.90,	1.95,	2.00,
y_i	4.89,	5.22,	5.25,	5.17,	5.54,	5.97,	6.41,	6.42,	7.15,	6.92.

par la méthode des moindres carrés on obtient alors :

$f_1(x) = 4.45\,x - 2$ $\qquad s'_\varepsilon = .23$

$f_2(x) = 1.42\,x^2 + .11\,x + 1.19$ $\qquad s'_\varepsilon = .21$

$f_3(x) = 1.46\,x^2 + 1.27$ $\qquad s'_\varepsilon = .20$

$f_4(x) = 2.6\,x^{1.42}$ $\qquad s_{\log\varepsilon} = .046$

On peut comparer les modèles qui s'ajustent le mieux dans chacun des 3 cas de figure (continu, discret, avec erreur) :

$f_2 = 2.46\,x^2 - 2.78\,x + 3.11$ $\quad f = 2.18\,x^2 - 1.96\,x + 2.52$ $\quad f_3 = 1.46\,x^2 + 1.27$

et en particulier comparer leur valeur lorsqu'on les extrapole à x = 3 et 5 par exemple.

4.5 - Exemple d'application : ajustement d'une courbe de tarage pour le Buech aux Chambons (B.V. 723 km²)

On dispose des résultats de 11 jaugeages, donc 11 couples hauteur-débit (H_i, Q_i) :

H_i	2.255,	2.435,	1.345,	3.060,	0.935,	0.810,	0.830,	0.570,	1.070,	2.640,	1.150
Q_i	42.5,	52.2,	13.6,	96.5,	6.64,	4.76,	5.83,	2.16,	9.1,	62.5,	10.3

l'ajustement d'une relation linéaire donne :

$$Q \# 35\ H - 26 + \varepsilon \quad , \quad s_\varepsilon \# 8\ m^3/s$$

la relation n'est manifestement pas linéaire (systématisme des écarts).

L'ajustement après transformation logarithmique fournit :

$$Q \# 8\ H^{2.2} \quad , \quad \text{l'écart type des logarithmes étant } .066$$

Trois mesures effectuées récemment montrent que malgré le faible nombre de jaugeages utilisés pour ajuster la relation précédente, et, probablement du fait de la stabilité de la section au droit de la station, la courbe de tarage ajustée est correcte :

$$\begin{cases} H = 5.35 & \text{et} & Q = 335 & \text{observé} \\ \quad " & \text{et} & \hat{Q} = 320 & \text{calculé} \end{cases}$$

$$\begin{cases} H = 3.75 & \text{et} & Q = 130 & \text{observé} \\ \quad " & \text{et} & \hat{Q} = 146 & \text{calculé} \end{cases}$$

$$\begin{cases} H = .69 & \text{et} & Q = 3.6 & \text{observé} \\ \quad " & \text{et} & \hat{Q} = 3.5 & \text{calculé} \end{cases}$$

Remarque

Quelques exemples de fonctions utilisées comme "modèles" de représentation de phénomènes physiques observés par l'intermédiaire de mesures. La technique d'ajustement étant celle des moindres carrés on considèrera des combinaisons linéaires de fonctions simples, ou de variables x transformées (log, puissance $\frac{p}{q}$, exponentielle)

- fonction polynomiale :

$$\begin{cases} y = \sum_{i=0}^{K} a_i\ x^i \\ \text{ou } y = \sum_{i=0}^{K} a_i\ x^i + \sum_{j=1}^{m} b_j\ \left(\frac{1}{x}\right)^i \end{cases}$$

- fonction sinusoïdale : $y_i = \sum_{K=0}^{m} c_K \cos\left(\frac{2K\pi i}{T} + \varphi_K\right)$ pour i = 1 à T

4.6 - Comparaison de la méthode des moindres carrés et de la méthode des moindres distances

L'exemple suivant illustre cette comparaison pour l'ajustement d'une relation linéaire à une série de n = 23 couples (x_i, y_i) :

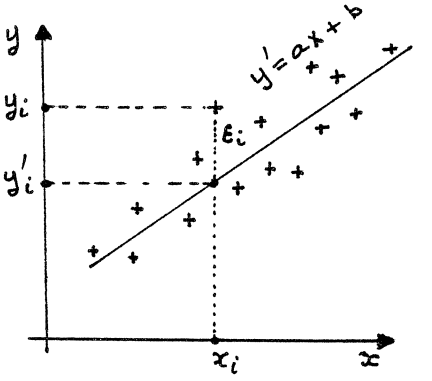
x 96, 219, 149, 126, 184, 76, 219, 170, 127, 184, 166, 136,
y 388, 609, 606, 535, 431, 376, 655, 463, 411, 634, 559, 572.

x 121, 246, 73, 173, 119, 269, 146, 176, 249, 201, 345,
y 613, 601, 449, 423, 544, 794, 584, 717, 631, 769, 737.

4.6.1 - Ajustement de la relation y' = ax + b par les moindres carrés

On minimise $\sum_{i=1}^{n} \varepsilon_i^2 = \Sigma (y_i - y'_i)^2 = \Sigma (y_i - axi - b)^2$

ce qui revient à chercher la solution de :



① $-2 \Sigma (y_i - axi - b)\, x_i = 0$

② $-2 \Sigma (y_i - axi - b) = 0$

Si l'on pose $y_i - \overline{y} = Y_i$
et $x_i - \overline{x} = X_i$ d'après ②, la relation ① s'écrit $\Sigma (Y_i - aX_i)\, X_i = 0$
soit : $a = \dfrac{\Sigma Y_i X_i}{\Sigma X_i^2}$

La précision de l'ajustement est définie par :

$$s_\varepsilon^2 = \frac{1}{n} \Sigma \varepsilon_i^2 = \frac{1}{n} \Sigma (Y_i - aX_i)^2 = s_y^2 - a^2 s_x^2$$

Calculer a, b, ainsi que l'écart type de l'écart ε sachant que :

$\Sigma x_i = 3\,970$; $\Sigma y_i = 13\,101$; $\Sigma x_i y_i = 2\,382\,503$;

$\Sigma x_i^2 = 778\,440$; $\Sigma y_i^2 = 7\,787\,667$

4.6.2 - <u>On peut ajuster la relation $\hat{y} = cx + d$ par les moindres distances</u>

On veut donc minimiser Σq_i^2 ,

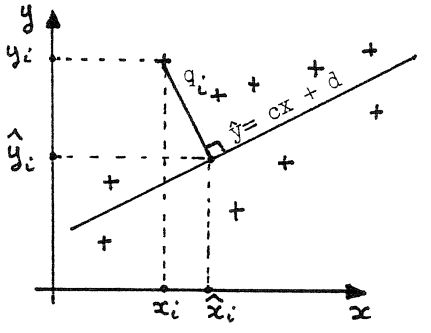
soit $\Sigma q_i^2 = \sum_{i=1}^{n} (x_i - \hat{x}_i)^2 + \Sigma (y_i - \hat{y}_i)^2 = \text{minimum}$

On établira la relation suivante entre $(\hat{x}_i, \hat{y}_i)$ et (x_i, y_i) :

$$\begin{cases} \hat{y}_i = c\hat{x}_i + d \\ \hat{x}_i = (cy_i + x_i - cd)\,\dfrac{1}{c^2 + 1} \end{cases}$$

l'expression précédente devient alors, après transformation :

$$\Sigma q_i^2 = \sum_{i=1} \frac{(y_i - cx_i - d)^2}{c^2 + 1}$$



on cherche alors la solution du système :

③ $$\frac{\delta \Sigma q_i^2}{\delta d} = -2 \Sigma \frac{(y_i - cx_i - d)}{c^2 + 1} = 0$$

④ $$\frac{\delta \Sigma q_i^2}{\delta c} = \frac{-2}{(a^2 + 1)^2} \Sigma (a^2 + 1)(y_i - cx_i - d)\, x_i + c\Sigma (y_i - cx_i - d)$$

d'après ③ la droite passe par le barycentre du nuage de points, ainsi d'ailleurs que la droite des moindres carrés, et si l'on pose :

$$y_i - \overline{y} = Y_i \quad \text{et} \quad x_i - \overline{x} = X_i$$

la relation ④ devient :

⑤ $c(\Sigma Y_i^2 - \Sigma X_i^2) + (1 - c^2)\,\Sigma X_i Y_i = 0$

c est solution d'une équation du 2ème degré :

$$\begin{cases} c = \dfrac{\alpha \pm (\alpha^2 + 4)^{1/2}}{2} \\ \text{avec } \alpha = \dfrac{\Sigma Y_i^2 - \Sigma X_i^2}{\Sigma X_i Y_i} \end{cases}$$

La précision de l'ajustement s'obtient d'après :

$$s_d^2 = \frac{\Sigma d_i^2}{n} = \frac{\Sigma (y_i - cx_i - d)^2}{n(1 + c^2)} = \frac{\Sigma (Y_i - cX_i)^2}{n(1 + c^2)}$$

en développant cette expression et en utilisant la relation ⑤ , on obtient :

$$s_d^2 = \frac{c^2 \Sigma X_i^2 - \Sigma Y_i^2}{(c^2 - 1)\, n}$$

Calculer les valeurs de c, d, et s_d sur l'exemple.

Tracer le graphe des couples (x_i, y_i) ainsi que les 2 droites ajustées $y' = ax + b$ et $\hat{y} = cx + d$

Cas particulier :

Effectuer les mêmes calculs que précédemment sur les couples de valeurs $(\frac{X_i}{s_x} \quad \frac{Y_i}{s_y})$, tracer le graphe de ces couples de valeurs ainsi que les droites ajustées à ces données d'après les moindres carrés et les moindres distances.

4.7 - Cas d'une relation (modèle) non linéaire

Lorsque la fonction $y = f(x, a, b, c, d, \ldots)$ que l'on veut ajuster d'après n couples de valeurs mesurées (x_i, y_i), n'est pas linéaire ou linéarisable par une transformation simple, on peut encore utiliser la méthode des moindres carrés.

Si l'on connaît une valeur approchée des coefficients a, b, c, d que l'on note a', b', c', d' :

$$a = a' + h, \quad b = b' + k, \quad c = c' + l, \quad d = d' + m,$$

en développant en série de Taylor, et en se limitant au premier terme, on peut écrire :

$$\textcircled{1}\quad f(x,a,b,c,d) \approx f(x,a',b',c',d') + \left(\frac{\delta f}{\delta a}\right)' h + \left(\frac{\delta f}{\delta b}\right)' k + \left(\frac{\delta f}{\delta c}\right)' l + \left(\frac{\delta f}{\delta d}\right)' m$$

si on note $\Delta = f(x,a,b,c,d) - f(x,a',b',c',d') = y - f(x,a',b',c',d')$

et $\left(\frac{\delta f}{\delta}\right)'$ la valeur de la dérivée partielle du premier ordre de f dans laquelle on remplace chaque paramètre par sa valeur approchée,

on calculera les coefficients h, k, l, m par la méthode des moindres carrés soit :

$$\textcircled{2}\quad \text{Minimum de } \sum_{i=1}^{n} \left[\Delta_i - (h\, Z_{i1} + k\, Z_{i2} + l\, Z_{i3} + m\, Z_{i4})\right]^2$$

en notant Z_{i1} la valeur de $\left(\frac{\delta f}{\delta a}\right)'$ pour $x = x_i$ et etc.

Exemple :

Ajuster $y = a\, e^{-bx} + c\, e^{-dx}$

connaissant n couples de valeurs (y_i, x_i) avec i = 1 à n.

L'expression de la fonction, connaissant une valeur approchée des paramètres est :

$$y' = a'\, e^{-b'x} + c'\, e^{-d'x}$$

Calcul des dérivées partielles :

$$\left(\frac{\delta y}{\delta a}\right)' = e^{-b'x} \quad ; \quad \left(\frac{\delta y}{\delta b}\right)' = -a'x'\, e^{-b'x} \quad ; \quad \left(\frac{\delta y}{\delta c}\right)' = e^{-d'x} \quad ; \quad \left(\frac{dy}{\delta d}\right) = -c'x' e^{-d'x}$$

on obtient alors les valeurs des corrections h, k, l, m en cherchant le minimum de :

$$\sum_{i=1}^{n} \left[(y_i - a' e^{-b'x_i} - c' e^{-d'x_i}) - h e^{-b'x_i} + k a'x' e^{-b'x_i} - l e^{-d'x_i} + m c' x' e^{-d'x_i} \right]^2$$

les valeurs que l'on obtient ainsi permettent de calculer une nouvelle valeur approchée de chaque paramètre :

$$a'' = a' + h, \quad b'' = b' + k, \quad c'' = c' + l, \quad d'' = d' + m,$$

puis l'on réitère le processus de calcul en (1) en recherchant des accroissements h_1, k_1, l_1, m_1 : $a'' + h_1$, $b'' + k_1$, $c'' + l_1$, $d' + m_1$.

Remarque : il serait préférable, avant d'appliquer la méthode des moindres carrés, d'orthogonaliser les variables :

$$Z_{1i} = e^{-b'x_i}, \quad Z_{2i} = a'x'e^{-b'x_i}, \quad Z_{3i} = e^{-d'x_i}, \quad Z_{4i} = c'x'e^{-d'x_i},$$

qui sont fortement corrélées, par une analyse en composantes principales en ne conservant que les 2 premières composantes pour les calculs (cf. chapitre VI).

V - LES LIAISONS STOCHASTIQUES

5.1 - La corrélation simple

La notion de corrélation est assez intuitive. On dispose de n couples d'observations numériques caractérisant deux phénomènes, par exemple les précipitations cumulées d'hiver (NOV-MARS) $P_{11}^{3} = X$ sur le bassin de la Romanche et les écoulements de printemps-été (AVR-AOU) $E_{4}^{8} = Y$ au Chambon. Pour étudier le degré et le type d'association entre ces 2 séries d'évènements (X_i, Y_i), l'idée la plus simple consiste à porter sur graphique à axes cartésiens les n points représentatifs de chaque année.

On observe une image de points ayant la forme d'une ellipse plus ou moins aplatie. Il s'agit alors de condenser cette information en quelque chose de plus maniable, dans le cas présent une relation linéaire entre X et Y : $Y = aX + b + \varepsilon$, on veut en effet calculer Y en fonction de X. On estimera les coefficients a et b d'après l'échantillon des n couples d'observations, selon les conditions que l'on impose à ε qui est l'écart (positif, négatif ou nul) entre la relation calculée $Y' = aX + b$ et le phénomène observé Y.

5.1.1 - Calcul des coefficients de régression a et b

Notons X_i, Y_i les précipitation et écoulement relatifs à la $i^{ème}$ année d'observation $(i = 1, \ldots, n)$ et ε_i l'écart entre la valeur observée Y_i et la valeur Y'_i calculée par la relation linéaire. On imagine bien que la droite, image de cette relation, passera par le centre de gravité ou barycentre du nuage de points, défini par $\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$ et $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$, mais il faut fixer une condition aux écarts ε_i :

on a choisi de minimiser $\sum_{i=1}^{n} \varepsilon_i^2$, soit encore $E = \sum_{i=1}^{n} (Y_i - aX_i - b) = \text{minimum}$.
On retrouve ici le principe des moindres carrés.

Cette condition se traduit par :

$$\left\{ \begin{array}{l} \dfrac{\delta E}{\delta b} = 0 \\ \\ \dfrac{\delta E}{\delta a} = 0 \end{array} \right\}$$

Les dérivées partielles du premier ordre de la fonction E, par rapport aux inconnues a et b, sont nulles

$$\text{soit} \begin{cases} -2 \Sigma (Y_i - a X_i - b) = 0 & \textcircled{1} \\ -2 \Sigma (Y_i - a X_i - b) X_i = 0 & \textcircled{2} \end{cases}$$

① et ② signifient respectivement :

$\sum_{i=1}^{n} \varepsilon_i = 0$, la moyenne des écarts est nulle et la droite passe par le barycentre ;

$\sum_{i=1}^{n} \varepsilon_i X_i = 0$, les vecteurs ε et X (dans l'espace à n dimensions) sont orthogonaux et il y a indépendance linéaire entre X et ε.

Au lieu de résoudre le système de 2 équations à 2 inconnues ① et ②, une solution élégante consiste à faire un changement d'origine, en centrant les variables, soit $y_i = Y_i - \overline{Y}$ et $x_i = X_i - \overline{X}$.

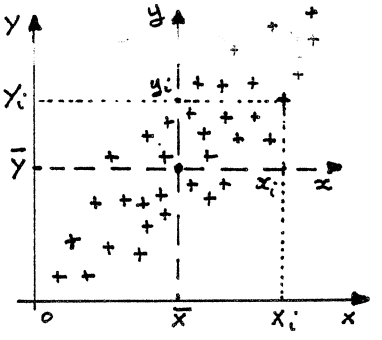
La condition des moindres carrés conduit à résoudre alors :

$$\Sigma (y_i - a x_i)^2 \text{ minimum, soit } \sum_{i=1}^{n} (y_i - a x_i) x_i = 0 ;$$

on obtient alors : $a = \dfrac{\Sigma x_i y_i}{\Sigma x_i^2}$

5.1.2 - Définition du coefficient de corrélation linéaire

Considérons la représentation du nuage de points (X_i, Y_i), et partageons le plan en 4 quadrants à l'aide des parallèles aux axes d'abscisse $\overline{X}$ et d'ordonnée $\overline{Y}$. Tout point M (de coordonnées X_i, Y_i) est défini par les 2 écarts $x_i = X_i - \overline{X}$ et $y_i = Y_i - \overline{Y}$.



Dans le quadrant I , le produit $x_i y_i > 0$
" " II , " $x_i y_i < 0$
" " III, " $x_i y_i > 0$
" " IV , " $x_i y_i < 0$

La quantité $\frac{1}{n} \sum_{i=1}^{n} x_i y_i$ caractérise l'association entre X et Y. Si cette quantité est positive, la plupart des points sont dans les quadrants I et III, si elle est négative les points sont au contraire plus nombreux dans les quadrants II et IV, et lorsque les points sont répartis indifféremment dans les 4 quadrants, la quantité est voisine de 0. On appelle covariance entre X et Y la valeur de $\frac{1}{n} \Sigma x_i y_i$.

Pour rendre cette mesure sans dimension et lui affecter un intervalle de variation borné, on la norme par le produit des écarts types de X et Y, et l'on obtient alors l'expression du coefficient de corrélation linéaire r entre X et Y :

$$r = \frac{\frac{1}{n} \Sigma x_i y_i}{\sqrt{\frac{1}{n} \Sigma x_i^2} \sqrt{\frac{1}{n} \Sigma y_i^2}} = \frac{1}{n} \frac{\Sigma x_i y_i}{S_x S_y}$$

$-1 \leqslant r \leqslant 1$, il y a liaison fonctionnelle entre X et Y si $r = \pm 1$.

On voit immédiatement que le coefficient angulaire de la relation linéaire entre X et Y peut se calculer d'après :

$$a = \frac{\Sigma x_i y_i}{\Sigma x_i^2} = r \cdot \frac{S_y}{S_x} = r \cdot \frac{S_Y}{S_X}$$

5.1.3 - Analyse des variances

Par définition nous avons écrit $y_i = y'_i + \varepsilon_i$ avec $y'_i = a x_i$; en effectuant la somme des n carrés on obtient :

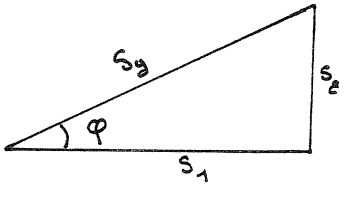
$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} y_i'^2 + \sum_{i=1}^{n} \varepsilon_i^2 + 2 \sum_{i=1}^{n} y'_i \varepsilon_i$$

or $\Sigma y'_i \varepsilon_i = a \Sigma \varepsilon_i x_i = 0$ d'après la condition ②

d'où : $\Sigma y_i^2 = \Sigma y_i'^2 + \Sigma \varepsilon_i^2$ ③

soit $\frac{1}{n} \Sigma y_i^2 = \frac{1}{n} \Sigma y_i'^2 + \frac{1}{n} \Sigma \varepsilon_i^2$

$$S_y^2 = S_{y'}^2 + S_\varepsilon^2 = S_1^2 + S_2^2$$



C'est le théorème de Pythagore, il y a orthogonalité entre les vecteurs S_1 et S_2 (dans l'espace à n dimensions).

S_y^2 est la variance de la distribution libre des y ou Y,

S_1^2 est la variance due à la dépendance linéaire de y en x ou Y en X ou la variance de Y expliquée par X,

S_2^2 est la variance résiduelle ou non expliquée par la relation linéaire entre X et Y.

Divisons l'expression ③ par Σy_i^2, on obtient :

$$\frac{\Sigma y_i'^2}{\Sigma y_i^2} = 1 - \frac{\Sigma \varepsilon_i^2}{\Sigma y_i^2}$$

or $\Sigma y_i'^2 = a^2 \Sigma x_i^2$, d'ou $\frac{a^2 \Sigma x_i}{\Sigma y_i^2} = 1 - \frac{\Sigma \varepsilon_i^2}{\Sigma y_i^2}$

soit $r^2 = 1 - \frac{\Sigma \varepsilon_i^2}{\Sigma y_i^2} = 1 - \frac{S_\varepsilon^2}{S_y^2} = \frac{S_1^2}{S_y^2}$ ④

Le carré du coefficient de corrélation représente la réduction relative de dispersion des y ou Y, obtenue en utilisant la relation linéaire entre X et Y.

Exemple de la Romanche au Chambon, $r^2 \# .84$ signifie que 84 % de la dispersion des écoulements de printemps-été $Y = E_4^8$ est imputable à la variabilité des précipitations d'hiver $X = P_{11}^3$, dont E_4^8 dépend linéairement.

On obtient ainsi l'écart type du résidu ε :

$$S_\varepsilon = S_2 = \sqrt{1 - r^2}\ S_Y$$

La relation ④ fournit également une interprétation géométrique du coefficient de corrélation avec la figure ci-dessus :

$$r = \frac{S_1}{S_Y} = \cos \varphi$$

Remarque :

Nous avons raisonné jusqu'ici sur des valeurs observées, le raisonnement serait parfait si l'on disposait d'un nombre n très grand ou infini (de l'ensemble de tous les couples X_i, Y_i) mais dans la réalité on ne dispose que d'un ou plusieurs échantillons limités en taille de valeurs observables ($20 \leq n \leq 40$), par conséquent les paramètres de la relation $Y' = a X + b$ que l'on aura calculé sur chacun de ces échantillons sont des estimations des vraies valeurs (inconnues) de la relation $Y' = \alpha X + \beta$. A et B sont des variables aléatoires d'espérance mathématique α et β respectivement dont on obtient une ou plusieurs réalisations a et b. On doit tenir compte du nombre de degrés de liberté dont on dispose pour les calculs, dans la relation d'analyse de la variance :

$$\Sigma y_i^2 = \Sigma y'^2_i + \Sigma \varepsilon_i^2 \quad \text{ou} \quad \Sigma(Y_i - \overline{Y})^2 = \Sigma(Y'_i - \overline{Y})^2 + \Sigma (Y_i - Y'_i)^2$$

- Pour la variance due à la liaison linéaire S_1^2 il suffit d'un seul Y'_i pour déterminer les (n-1) autres, $\overline{Y}' = \overline{Y}$ étant fixée, d'où $\nu_1 =$ 1 degré de liberté.

- Pour la variance résiduelle S_2^2, la détermination de Y'_i pris comme origine des différences $Y_i - Y'_i$ nécessite deux relations, une pour déterminer $\overline{Y}$ ou b et une pour déterminer a, d'où $\nu_2 =$ (n-2) degrés de liberté.

Résumons ces résultats dans le tableau suivant :

Source de variation	Somme des carrés	Degrés de liberté	Variance
liaison linéaire	$\Sigma(Y'_i - \overline{Y})^2 = r^2 \Sigma(Y - \overline{Y})^2$	1	$\frac{r^2 \Sigma (Y_i - \overline{Y})^2}{1}$
résiduelle	$\Sigma(Y_i - Y'_i)^2 = (1-r^2) \Sigma(Y - \overline{Y})^2$	n - 2	$\frac{(1-r^2) \Sigma(Y_i-\overline{Y})^2}{n-2}$
totale	$\Sigma(Y_i - \overline{Y})^2$	n - 1	$\frac{1}{n-1} \Sigma(Y_i - \overline{Y})^2$

(La propriété d'additivité ne subsiste plus pour les variances).

5.1.4 - Tests sur les coefficients de régression

On démontre que les estimations a, b des vrais coefficients inconnus α et β, et obtenues par la méthode des moindres carrés sont fournies par des estimateurs A et B sans biais :

$$E(A) = \alpha \text{ et } E(B) = \beta$$

On cherche donc à comparer l'écart entre a et α, puis b et β ; on va calculer la variance des estimateurs A et B (dont a et b sont une réalisation calculée sur l'échantillon des n valeurs).

On démontre que :

$$- V(A) = S_\varepsilon^2 \cdot \frac{1}{\Sigma (X_i - \overline{X})^2} = \frac{1}{n-2} \frac{\Sigma (Y_i - Y_i')^2}{\Sigma (X_i - \overline{X})^2}$$

si l'échantillon est important : $V(A) \neq \frac{n S_\varepsilon^2}{S_X^2}$

$$- V(B) = S_\varepsilon^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\Sigma (X - \overline{X})^2} \right]$$

si n est grand : $V(B) \neq \frac{S_\varepsilon^2}{n}$

La variable aléatoire :

$t_A = \frac{A - \alpha}{S_\varepsilon} \sqrt{\Sigma (X_i - \overline{X})^2}$ suit une loi de Student à n-2 degrés de liberté

$t_B = \frac{B - \beta}{S_\varepsilon} \sqrt{n}$ suit une loi de Student à (n-2) degrés de liberté

Pratiquement, lorsqu'on a obtenu une valeur a de A calculée sur un échantillon de taille n, on teste si cette valeur est significativement différente de 0, c'est-à-dire si l'intervalle à p % centré sur a, soit $a \pm t_p \frac{S_\varepsilon}{\sqrt{\Sigma x_i^2}}$ contient ou non la valeur 0. (Si n est grand $\frac{A}{\sqrt{V(A)}}$ est sensiblement gaussienne)

Remarque :

On peut utiliser le test de Fisher Snedecor pour $\nu_1 = 1$ degré de liberté et $\nu_2 = n-2$ degrés de liberté et qui teste si la quantité $F = \frac{S_1^2}{S_2^2}$ est signi-

ficativement plus grand que 1, c'est-à-dire si la variance due à la liaison linéaire est significativement plus grande que la variance résiduelle.

5.1.5 - Test sur le coefficient de corrélation linéaire

En général le coefficient r est calculé sur un échantillon de n couples d'observations, ce coefficient représente donc une estimation de la vraie valeur ρ inconnue, et r a une distribution d'échantillonnage autour de cette valeur. La fonction de répartition des valeurs de r est dissymétrique lorsque $\rho \neq 0$ et n'est pas d'un usage commode. On utilise alors une approximation en calculant la transformée de Fisher $z = \frac{1}{2} \log \frac{1+r}{1-r}$. Cette nouvelle variable z a un intervalle de variation de $-\infty$ à $+\infty$ et suit sensiblement une loi normale de moyenne ρ et d'écart type $\frac{1}{\sqrt{n-3}}$, indépendant de ρ, on teste si l'écart entre z et $\frac{1}{2} \log \frac{1+\rho}{1-\rho}$ est significativement différent de zéro. Dans la pratique. on teste si r est significativement différent de 0, c'est-à-dire si $z = \frac{1}{2} \log \frac{1+r}{1-r}$ est à l'extérieur de l'intervalle $\pm \frac{2}{\sqrt{n-3}}$ au seuil 5 % par exemple.

5.1.6 - Fonction de répartition gaussienne du couple (X, Y)

Si on considère des valeurs continues de X et Y et si l'on fait l'hypothèse que le couple (X, Y) est distribué selon une loi normale à 2 dimensions, sa densité de répartition s'écrit :

$$⑤ \quad Z = f(X,Y) = \frac{1}{2\pi\, \sigma_X\, \sigma_Y \sqrt{1-\rho^2}}\; e^{-\frac{1}{2(1-\rho^2)} \left(\frac{Y-m_Y}{\sigma_Y}\right)^2 - 2\rho \frac{(Y-m_Y)\,(X-m_X)}{\sigma_X\ \sigma_Y} + \left(\frac{X-m_X}{\sigma_X}\right)^2}$$

définie par cinq paramètres : $m_Y = E(Y)$, $m_X = E(X)$, $\sigma_Y^2 = E(Y-m_Y)^2$, $\sigma_X^2 = E(X-m_X)^2$, $\rho = E(Y-m_Y)\,(X-m_X)$ et que l'on peut écrire sous la forme :

$$⑥ \quad Z = \frac{1}{\sigma_Y \sqrt{2\pi(1-\rho^2)}}\; e^{-\frac{1}{2(1-\rho^2)} \left(\frac{Y-m_Y}{\sigma_Y}\right) - r\left(\frac{X-m_X}{\sigma_X}\right)^2} \cdot \frac{1}{\sigma_X \sqrt{2\pi}}\; e^{-\frac{1}{2}\left(\frac{X-m_X}{\sigma_X}\right)^2}$$

L'expression ⑥ représente une surface de Gauss (chapeau de gendarme) dans l'espace à 3 dimensions Z, X, Y.

Pour X constant, c'est-à-dire en effectuant des coupes de cette surface par des plans verticaux d'abscisse X, on obtient dans ces plans des courbes de Gauss (d'après l'expression ⑥) qui se projettent sur le plan horizontal (X, Y) selon une droite d'équation :

$$Y_X = m_Y + \rho\ (X - m_X)\ \frac{\sigma_Y}{\sigma_X}$$

on reconnaît l'équation de la droite de régression de Y en fonction de X obtenue précédemment. L'écart type lié de cette corrélation est $\sigma_L = \sqrt{1-\rho^2}\ \sigma_Y$, à comparer avec S_ε lorsque $\rho = r$ coefficient de corrélation linéaire entre X et Y. Les sections de la surface ⑤ par des plans horizontaux sont des ellipses homothétiques $e(X, Y) = d^2$ dont la forme dépend de ρ. La droite $Y = m_Y + \rho\,(X - m_X)\ \frac{\sigma_Y}{\sigma_X}$ n'est autre que le diamètre conjugué de la direction OY par rapport à ces ellipses. On obtiendrait de même le diamètre conjugué de la direction OX sour la forme $X = m_X + \rho\,(Y - m_Y)\ \frac{\sigma_X}{\sigma_Y}$.

On dit que ces ellipses sont équiprobables, d'égale densité des variables X et Y : on a une densité plus ou moins élevée de points à l'intérieur de ces ellipses suivant les valeurs de d.

Pour $\rho = 0$, on obtient des cercles, la surface est de révolution, et, pour $\rho = 1$ on obtient une droite : la relation linéaire entre X et Y est fonctionnelle.

5.1.7 - Applications de la corrélation simple

5.1.7.1 - la prévision :

D'après les 29 couples de valeurs observées $(E_4^8,\ P_{11}^3)$ on peut calculer :

- l'équation de régression $E_4^8 = .85\ P_{11}^3 + 341 + \varepsilon$,
- le coefficient de corrélation linéaire $r \neq .92$,
- l'écart type lié $S_\varepsilon = S_L = 67$

- après avoir étudié la distribution empirique des ε_i, gaussienne dans ce cas, on peut calculer chaque année au 1er avril une prévision d'écoulement, c'est-à-dire un intervalle qui a 8 chances sur 10 de contenir l'écoulement de l'été d'après les pluies de l'hiver passé :

$$E_4^8 = .85\ P_{11}^3 + 341 \pm 1.28\ (67)$$

5.1.7.2 - test d'homogénéité :

On veut comparer deux séries X et Y pour lesquelles on a n réalisations simultanées (exemple de la Romanche). On trace la ligne d'écarts cumulés $\sum_{i=1}^{k} (Y_i - Y'_i)$ (pour k = 1 à n) en fonction de k ou en fonction de $\sum_{i=1}^{k} (Y_i + Y'_i)$. Cela permet de déceler les séquences des valeurs systématiquement au-dessus ou en dessous de la droite de régression, donc des modifications de conditions de mesures.

Ainsi, pour la Romanche au Chambon, il semble y avoir une hétérogénéité de faible importance en 1947-49. Bien entendu plus r sera voisin de 1, plus on mettra en évidence les faibles hétérogénéités et par conséquent celles qui sont importantes (cf. méthode de contrôle mise au point par M. BOIS à l'I.N.P.G.).

5.1.7.3 - reconstitution de données manquantes :

Ayant calculé, par exemple, la corrélation entre les précipitations annuelles observées à Villard-de-Lans et Engins sur la période 1920-1967 E = 1.033 V - 18, on peut reconstituer la précipitation E à Engins en 1968 (le pluviomètre étant hors service) affectée d'une plage d'incertitude proportionnelle à l'écart type lié de la corrélation (σ_L = 106) connaissant la précipitation V à Villard-de-Lans V = 1470.

5.2 - La corrélation double

Il s'agit d'une extension de la notion de corrélation simple. On dispose de n observations, simultanément sur 3 variables Y, X, Z (Y_i, X_i, Z_i). X et Z sont les variables explicatives de Y, c'est-à-dire que l'on cherche à établir une relation multilinéaire de la forme :

$$Y = ax + bz + c + \varepsilon$$

Le résidu ε représente l'écart entre l'équation ajustée $Y' = ax + bz + c$ et la valeur observée Y.

En d'autres termes, dans l'espace à trois dimensions (Y, X, Z) on va chercher à faire passer un plan qui s'ajuste au mieux au nuage des n points observations (Y_i, X_i, Z_i). On utilisera pour cela la méthode des moindres carrés qui consiste à minimiser la somme des carrés des écarts entre valeurs ajustées par le plan et valeurs observées $(\varepsilon_i = Y_i - Y'_i)$ parallèlement à la direction OY, on déterminera ainsi les coefficients de régression a, b et c.

5.2.1 - Calcul des coefficients de régression

Satisfaire à la condition :

$$\Sigma (Y_i - a X_i - b Z_i - c)^2 = \Sigma \varepsilon_i^2 = \text{minimum}$$

revient à résoudre le système d'équations suivant :

$$\begin{cases} \dfrac{\delta \Sigma \varepsilon_i^2}{\delta a} = 0 \\ \dfrac{\delta \Sigma \varepsilon_i^2}{\delta b} = 0 \\ \dfrac{\delta \Sigma \varepsilon_i^2}{\delta c} = 0 \end{cases}$$

On doit donc résoudre le système :

$$2 \Sigma X_i (Y_i - a X_i - b Z_i - c) = 0 \qquad (1)$$

$$2 \Sigma Z_i (Y_i - a X_i - b Z_i - c) = 0 \qquad (2)$$

$$2 \Sigma (Y_i - a X_i - b Z_i - c) = 0 \qquad (3)$$

On remarquera que les deux premières relations établissent l'orthogonalité, donc l'indépendance linéaire au sens artistique, du résidu et de chacune des variables explicatives puisque :

$$\Sigma X_i \varepsilon_i = 0 \quad \text{et} \quad \Sigma Z_i \varepsilon_i = 0$$

De plus, la relation ③ après division par n (nombre d'observations) devient :

$$c = \overline{Y} - a\overline{X} - b\overline{Z} \quad , \quad \text{(le plan passe par le barycentre du nuage)}$$

$\overline{Y}$, $\overline{X}$, $\overline{Z}$ étant les moyennes arithmétiques respectivement des Y_i, X_i, Z_i, on obtient facilement c connaissant a et b.

Pour la commodité des calculs, nous travaillerons en variables centrées réduites, soit :

$$y_i = Y_i - \overline{Y} \quad , \qquad x_i = X_i - \overline{X} \quad , \qquad z_i = Z_i - \overline{Z} \quad ;$$

la condition des moindres carrés :

$$\Sigma (y_i - a x_i - b z_i)^2 = \text{minimum}$$

consiste à résoudre le système à deux équations :

$$\Sigma x_i (y_i - a x_i - b z_i) = 0$$
$$\Sigma z_i (y_i - a x_i - b z_i) = 0$$

soit :
$$a = \frac{\Sigma y_i x_i \ \Sigma z_i^2 - \Sigma y_i z_i \ \Sigma x_i z_i}{\Sigma x_i^2 \ \Sigma z_i^2 - (\Sigma x_i z_i)^2} \qquad ④$$

$$b = \frac{\Sigma y_i z_i \ \Sigma x_i^2 - \Sigma y_i x_i \ \Sigma x_i z_i}{\Sigma x_i^2 \ \Sigma z_i^2 - (\Sigma x_i z_i)^2} \qquad ⑤$$

Notons :

- r_1 le coefficient de corrélation <u>totale</u> ou simple entre y et x

$$r_1^2 = \frac{(\Sigma x_i\ y_i)^2}{(\Sigma x_i^2)\ (\Sigma y_i^2)}$$

- r_2 le coefficient de corrélation totale entre y et z

$$r_2^2 = \frac{(\Sigma z_i\ y_i)^2}{(\Sigma z_i^2)\ (\Sigma y_i^2)}$$

- ρ le coefficient de corrélation totale entre x et z

$$\rho^2 = \frac{(\Sigma x_i\ z_i)^2}{(\Sigma x_i^2)\ (\Sigma z_i^2)}$$

On peut écrire l'expression ④ ainsi (en supprimant l'indice i pour simplifier l'écriture) :

$$a = \frac{\dfrac{\Sigma xy\ \ \Sigma z^2}{\Sigma x^2\ \ \Sigma y^2\ \ \Sigma z^2} - \dfrac{\Sigma yz\ \ \Sigma xz}{\Sigma x^2\ \ \Sigma y^2\ \ \Sigma z^2}}{\dfrac{\Sigma x^2\ \ \Sigma z^2}{\Sigma x^2\ \ \Sigma y^2\ \ \Sigma z^2} - \dfrac{(\Sigma xz)^2}{\Sigma x^2\ \ \Sigma y^2\ \ \Sigma z^2}}$$

sachant que :

$$n\,S_y = \Sigma y_i^2$$
$$n\,S_x = \Sigma x_i^2$$
$$n\,S_z = \Sigma z_i^2$$

on obtient :

$$a = \frac{\dfrac{r_1}{n\,S_x\,S_y} - \dfrac{r_2\,\rho}{n\,S_x\,S_y}}{\dfrac{1}{n\,S_y^2} - \dfrac{\rho^2}{n\,S_y^2}}$$

$$\boxed{a = \frac{r_1 - r_2\rho}{(1 - \rho^2)}\ \frac{S_y}{S_x}}$$

on trouverait de même :

$$b = \frac{r_2 - r_1 \rho}{1 - \rho^2} \frac{S_y}{S_x}$$

Les coefficients de régression s'obtiennent directement à l'aide des coefficients de corrélation totale r_1, r_2, ainsi que des écarts types S_x, S_z, S_y.

5.2.2 - Calcul du coefficient de corrélation multiple

La qualité de l'ajustement effectué à l'aide de la relation multilinéaire précédente , c'est-à-dire la proximité entre valeurs calculées Y'_i et valeurs observées Y_i, peut être caractérisée par un coefficient sans dimension ainsi défini (coefficient de détermination multiple) :

$$R^2 = 1 - \frac{\Sigma \varepsilon_i^2}{\Sigma Y_i^2} \qquad (6)$$

Lorsque la relation entre Y et Z est fonctionnelle, la dispersion est nulle ($\Sigma \varepsilon_i^2 = 0$) le coefficient $R^2 = 1$; dans le cas contraire lorsqu'il n'y a aucune relation linéaire entre Y et le couple (Z, X) la dispersion naturelle des y_i n'est pas réduite et $\Sigma \varepsilon_i^2 = \Sigma y_i^2$ soit $R^2 = 0$.

La corrélation multiple entre la variable dépendante Y et les variables explicatives sera d'autant plus élevée que R^2 sera voisin de 1. R est également le coefficient de corrélation multiple entre valeurs calculées et valeurs observées :

$$R = \frac{\Sigma y_i \, y'_i}{\sqrt{\Sigma y_i^2} \sqrt{\Sigma y_i'^2}}$$

On peut établir facilement à l'aide de (6) la relation suivante entre le carré du coefficient de corrélation multiple et les coefficients de corrélation totale :

$$R^2 = \frac{r_1^2 + r_2^2 - 2\rho r_1 r_2}{1 - \rho^2}$$

On remarque en particulier que lorsqu'il n'y a pas de corrélation entre X et Z ($\rho = 0$) : $R^2 = r_1^2 + r_2^2$

5.2.3 - Calcul de la variance liée ou résiduelle

Notons S_ε l'écart type des écarts résiduels, d'après ⑥ on peut écrire :

$$R^2 = 1 - \frac{S_\varepsilon^2}{S_y^2}$$

soit : $$S_\varepsilon^2 = (1 - R^2)\, S_y^2$$

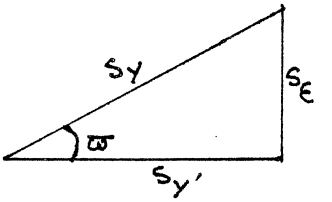
La variance résiduelle ou variance liée représente la part de variance totale de Y inexpliquée par la relation multilinéaire entre Y, X et Z : elle sera d'autant plus faible que R^2 sera voisin de 1.

$$y_i^2 = \Sigma (y_i' + \varepsilon_i)^2 = \Sigma y_i'^2 + 2 \Sigma y_i' \varepsilon_i + \Sigma \varepsilon_i^2$$

or $\Sigma y_i' \varepsilon_i = \Sigma (a x_i + b z_i) \varepsilon_i = 0$ d'après ① et ②

donc $\Sigma y_i^2 = \Sigma y_i'^2 + \Sigma \varepsilon_i^2$

soit S $S_y^2 = S_{y'}^2 + S_\varepsilon^2$



La variance totale se décompose en variance expliquée par la relation linéaire et variance non expliquée ou résiduelle.

Pour estimer les variances réelles, il faut tenir compte du nombre de degrés de liberté, ainsi :

- pour la variance liée à la relation linéaire $\frac{1}{\nu_1} \Sigma (Y_i' - \overline{Y})^2$ on dispose de 2 degrés de liberté, $\overline{Y}$ étant donné il suffit de 2 points pour définir le plan, $\nu_1 = 2$;

- pour la variance résiduelle $S_\varepsilon^2 = \frac{1}{\nu_2} \Sigma (Y_i - Y_i')^2$ on dispose de $\nu_2 = n-3$ degrés de liberté (3 relations pour déterminer les 3 coefficients);

- pour la variance globale S_y^2 on dispose de n-1 degrés de liberté (une relation pour déterminer $\overline{Y}$).

$$\begin{cases} S_{Y'}^2 = \frac{R^2}{2} \Sigma (Y_i' - \overline{Y})^2 \\ S_\varepsilon^2 = \frac{1 - R^2}{n-3} \Sigma (Y_i - \overline{Y})^2 \\ S_Y^2 = \frac{1}{n-1} \Sigma (Y_i - \overline{Y})^2 \end{cases}$$

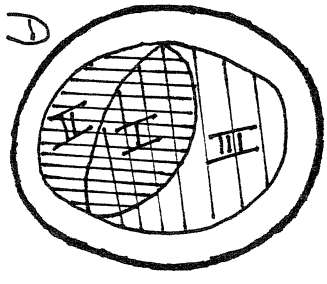
5.2.4 - Calcul des coefficients de corrélation partielle

Les coefficients de corrélation totale r_1 et r_2 ne sont pas de bons témoins de la liaison réelle qui existe respectivement entre Y et X, Y et Z.

On a donc défini un coefficient de corrélation partielle qui mesure l'intensité de la relation entre Y et X, par exemple, lorsqu'on élimine l'influence de Z sur chacune de ces variables, on le notera $r_{yx,\ z}$; de même le coefficient de corrélation partielle entre Y et Z sera note $r_{yz,\ x}$.

Essayons tout d'abord d'illustrer cette notion de corrélation partielle à l'aide d'une représentation graphique.

Schématisons l'information totale concernant le phénomène Y par l'aire D que l'on prend égale à 1. La part Y expliquée par les phénomènes X et Z est définie par toute l'aire hachurée soit R^2. La part inexpliquée, zone en blanc, est donc mesurée par $(1 - R^2)$.
La corrélation totale entre Y et X est définie par l'aire hachurée horizontalement soit r_1^2 (II); la corrélation totale entre Y et Z est définie par l'aire hachurée verticalement (III) soit r_2^2.



Les corrélations peuvent alors être définies par :

$$r^2_{yx,\ z} = \frac{R^2 - r_2^{\ 2}}{1 - r_2^{\ 2}}$$ coefficient proportionnel aux aires I + II

$$r^2_{yz,\ x} = \frac{R^2 - r_1^{\ 2}}{1 - r_1^{\ 2}}$$ coefficient proportionnel aux aires I + III.

L'aire I caractérise la dépendance entre X et Z (elle est nulle dans le cas d'indépendance).

On remarquera que lorsqu'il y a indépendance entre X et Z, $R^2 = r_1^{\ 2} + r_2^{\ 2}$ les corrélations partielles sont différentes des corrélations totales, en effet :

$$r^2_{yx,\ z} = \frac{r_1^{\ 2}}{1 - r_2^{\ 2}}$$

$$r^2_{yz,\ x} = \frac{r_2^{\ 2}}{1 - r_1^{\ 2}}$$

Le coefficient de corrélation partielle a été défini comme coefficient de corrélation entre la variable dépendante et l'une des variables explicatives, Y et X par exemple, lorsqu'on enlève à chacune d'elles la part due à l'influence de la troisième, Z.

Il faudra donc calculer les écarts résiduels ξ et δ à l'aide des relations linéaires ajustées par les moindres carrés :

$$\begin{cases} Y = \alpha_1\ Z + \beta_1 + \xi_1 \\ X = c_1\ Z + d_1 + \delta_1 \end{cases}$$

Le coefficient de corrélation partielle entre Y et X sera défini ainsi :

$$r_{yx,\ z} = \frac{\Sigma\ \xi_1^i\ \delta_1^i}{\sqrt{\Sigma\ \xi_1^{i\,2}}\ \sqrt{\Sigma\ \delta_1^{i\,2}}}$$

De même on obtiendra le coefficient de corrélation partielle entre Y et Z en calculant :

$$\begin{cases} Y = \alpha_2 X + \beta_2 + \xi_2 \\ Z = c_2 X + d_2 + \delta_2 \end{cases}$$

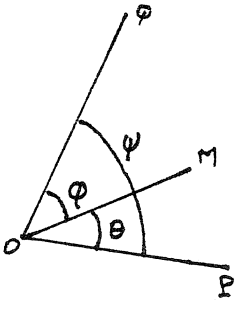
En développant ces relations, on obtient les expressions suivantes, en fonction des coefficients de corrélation totale :

$$r_{yx,\,z} = \frac{r_1 - r_2 \rho}{\sqrt{1 - r_2^2}\ \sqrt{1 - \rho^2}}$$

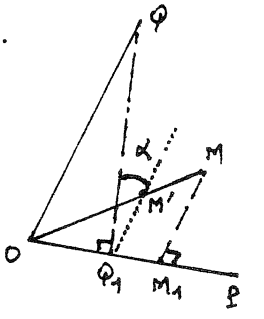
$$r_{yz,\,x} = \frac{r_2 - r_1 \rho}{\sqrt{1 - r_1^2}\ \sqrt{1 - \rho^2}}$$

5.2.5 - Signification géométrique des coefficients de corrélation simple, partielle et multiple

Dans l'espace à n dimensions, on définit les vecteurs OM, OP, OQ, d'origine O (centre de gravité des observations, ou point le coordonnées $\overline{X}$, $\overline{Z}$, $\overline{Y}$) et dont les extrémités M, P, Q ont pour coordonnées respectivement $(x_1, x_2, \ldots, x_n)$, $(z_1, \ldots, z_n)$, $(y_1, \ldots, y_n)$.

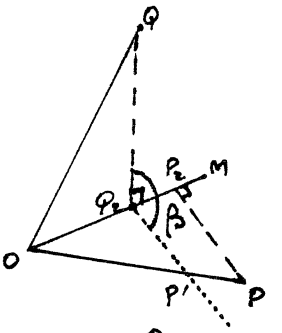


Le cosinus de l'angle des vecteurs OM et OQ n'est autre que le coefficient de corrélation simple ou totale entre x et y : cos $\varphi = r_1$, de même le cosinus de l'angle des vecteurs OP et OQ est le coefficient de corrélation totale entre z et y : cos $\psi = r_2$, enfin le cosinus de l'angle des vecteurs OP et OQ est le coefficient de corrélation totale entre x et z. (on rappelle que $OM^2 = n\,S_x^2$, $OP^2 = n\,S_z^2$, $OQ^2 = n\,S_y^2$).

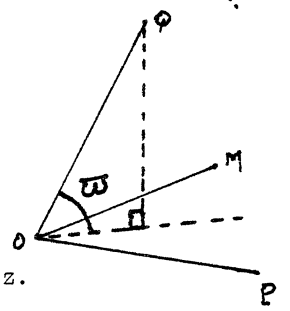


Le coefficient de corrélation partielle entre y et x n'est autre que le cosinus de l'angle des projections des vecteurs OQ et OM sur le plan orthogonal à OP soit : $r_{yx,z} = \cos \alpha$ (Q_1 M' est // à M_1M)

De même le coefficient de corrélation partielle entre y et z est le cosinus de l'angle des projections des vecteurs OQ et OP sur le plan orthogonal à OM soit : $r_{yz,x} = \cos\beta$ (Q_2P' est // à $P\,P_2$).



Le coefficient de corrélation multiple est le cosinus de l'angle entre la projection orthogonale de OQ sur le plan MOP : $R = \cos\varpi$; on voit immédiatement que si OQ est orthogonal à MOP, il n'y a pas de corrélation multilinéaire entre y et x, z, et que si OQ est dans le plan MOP, il y a liaison linéaire fonctionnelle entre y et x, z.



5.2.6 - Test sur les coefficients de la régression double

On démontre que l'espérance mathématique des coefficients de régression a, b, c calculés sur un échantillon de taille n, par la méthode des moindres carrés, est :

$$E(a) = A_o\ ,\quad E(b) = B_o\ ,\quad E(c) = C_o \qquad \text{si } A_o,\ B_o,\ C_o$$

sont les valeurs théoriques inconnues.

Les coefficients estimés d'après n observations ont une dispersion d'échantillonnage et on démontre que leur écart type respectif est :

$$\text{pour c} \qquad S_c = \frac{S_\varepsilon}{n} = \sqrt{1 - R^2}\ \frac{S_Y}{\sqrt{n-3}}$$

$$\text{pour a} \qquad S_a = \sqrt{\frac{1 - R^2}{1 - \rho^2}} \cdot \frac{S_Y}{S_X} \cdot \frac{1}{\sqrt{n-3}}$$

$$\text{pour b} \qquad S_b = \sqrt{\frac{1 - R^2}{1 - \rho^2}} \cdot \frac{S_Y}{S_Z} \cdot \frac{1}{\sqrt{n-3}}$$

généralement on teste si les variables de Student $t_c = \frac{c}{S_c}$, $t_a = \frac{a}{S_a}$, $t_b = \frac{b}{S_b}$ sont significativement différentes de 0 au seuil α % (α = 5 % par exemple) avec n-3 degrés de liberté.

Pour tester si les différents coefficients de corrélation partielle sont significativement différents de 0, on peut appliquer la transformée de Fisher, mais avec un écart type égal à $\frac{1}{\sqrt{n-4}}$

Le test sur le coefficient de détermination multiple R^2 revient à tester si le rapport de Fisher Snedecor $F = \frac{S_1^2}{S_2^2}$ avec $\nu_1 = 2$ degrés de liberté et $\nu_2 = n-3$ degrés de liberté est significativement différent de 1 au seuil α %. C'est un test sur les variances, variance expliquée par la relation multilinéaire comparée à la variance résiduelle, en se reportant au paragraphe 5.2.3 on calcule F d'après la relation :

$$F = \frac{(n-3)R^2}{2\ (1-R^2)}$$

Remarque

L'analyse des variances permet d'obtenir une estimation R'^2 sans biais du coefficient de détermination multiple R^2, en effet :

$$R'^2 = \frac{S_Y^2 - S_E^2}{S_Y^2} = 1 - \frac{n-1}{n-3}\ (1 - R^2)$$

$$R'^2 = \frac{(n-1)\ R^2 - 2}{n-3}$$

Exemples de corrélation simple, double, triple entre les écoulements du Drac au Sautet E_4^7 (variable principale) et les précipitations P_{10}^3 et écoulements E_{10}^3 d'hiver, les précipitations d'été P_4^7.

PREVISION d'APPORTS de la ROMANCHE au CHAMBON

E_4^8 P_{11}^3

	1	2
M	743.3	475.6
S	166.2	180.5

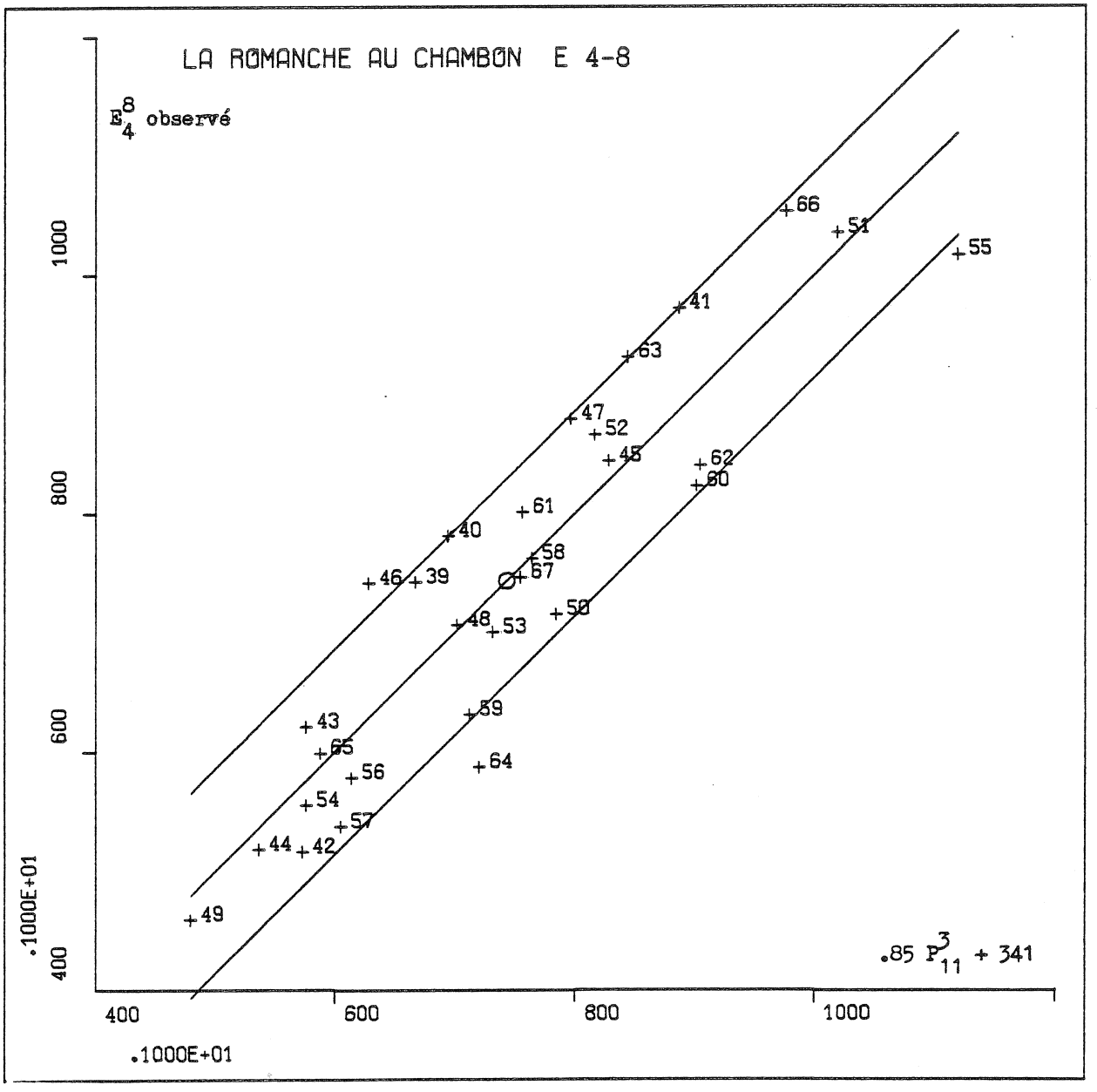
	RP	B	T	BR
1				
2	.9180	.8454	12.0	.92

A= 341.2295

R2= .8369 R= .9148 SL= 67.1254

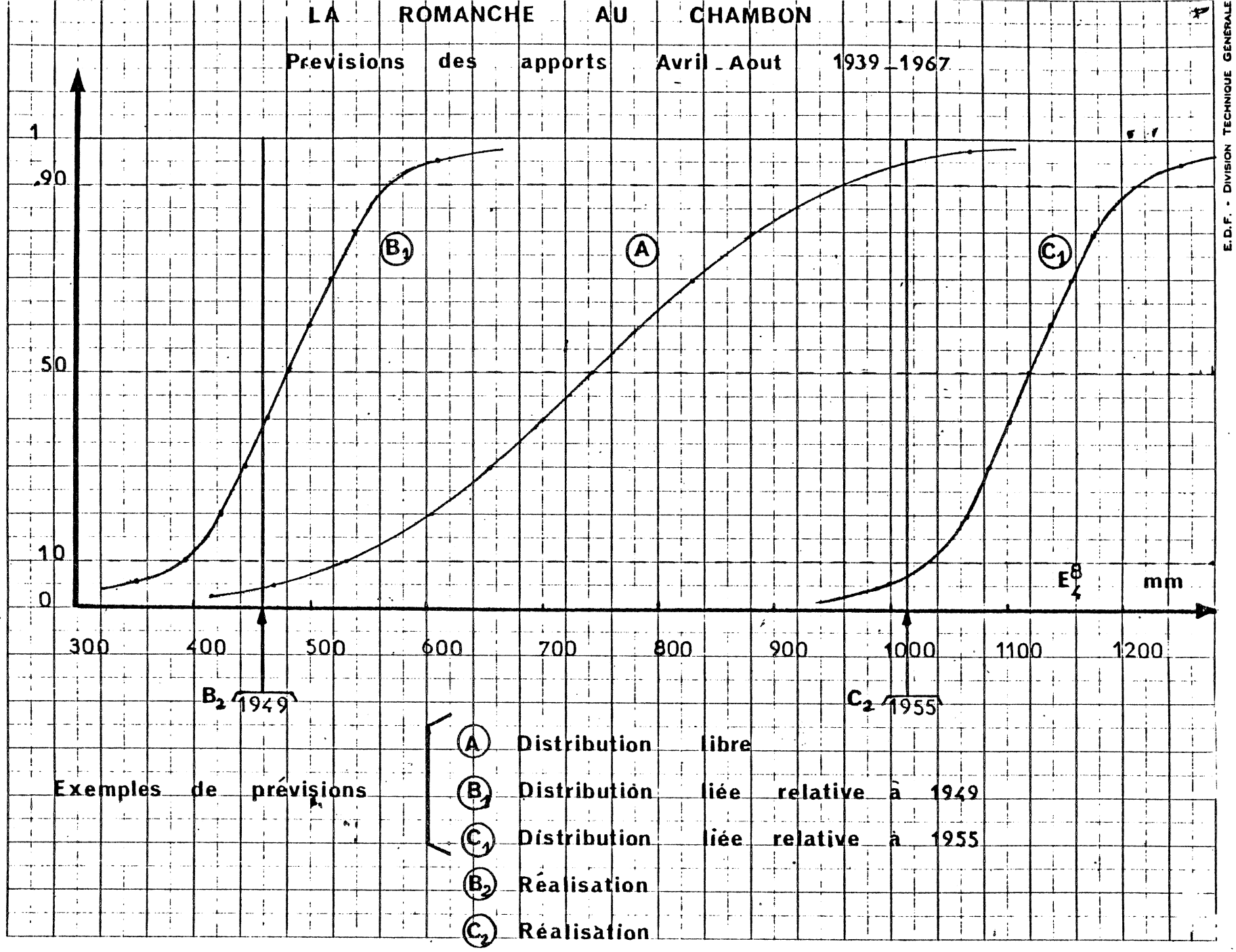
CLE 1	CLE 2	CLE 3
667.5467	742.0000	74.4533
694.5989	781.0000	86.4011
886.5004	972.0000	85.4996
572.8640	516.0000	-56.8640
576.2455	621.0000	44.7545
536.5126	518.0000	-18.5126
827.3238	844.0000	16.6762
628.6592	741.0000	112.3408
796.0446	879.0000	82.9554
702.2073	706.0000	3.7927
479.8721	459.0000	-20.8721
784.2093	715.0000	-69.2093
1019.2253	1035.0000	15.7747
816.3338	866.0000	49.6662
731.7957	700.0000	-31.7957
576.2455	555.0000	-21.2455
1118.9803	1017.0000	-101.9803
614.2877	578.0000	-36.2877
604.9885	537.0000	-67.9885
763.9202	762.0000	-1.9202
712.3519	631.0000	-81.3519
900.8719	823.0000	-77.8719
756.3117	801.0000	44.6883
904.2534	841.0000	-63.2534
844.2314	931.0000	86.7686
720.8057	587.0000	-133.8057
588.0809	599.0000	10.9191
976.1109	1053.0000	76.8891
754.6210	746.0000	-8.6210

MX	MY	SX	SY
743.3104	743.3103	152.5842	166.2133
BX	**BY**	**R2**	**R**
.8427	1.0000	.8369	.9148
A0	**LAMBDA**	**CONF 80**	**SYX**
-.0002	.4039	85.9205	67.1254


LA ROMANCHE AU CHAMBON E 4-8
E_4^8 observé
.85 P_{11}^3 + 341
.1000E+01
.1000E+01
400
600
800
1000
400
600
800
1000
66
51
55
41
63
47
52
45
62
60
61
40
58
46
39
67
48
53
50
59
43
65
56
64
54
57
44
42
49

LA ROMANCHE AU CHAMBON

Previsions des apports Avril - Aout 1939 - 1967



Exemples de prévisions

A Distribution libre

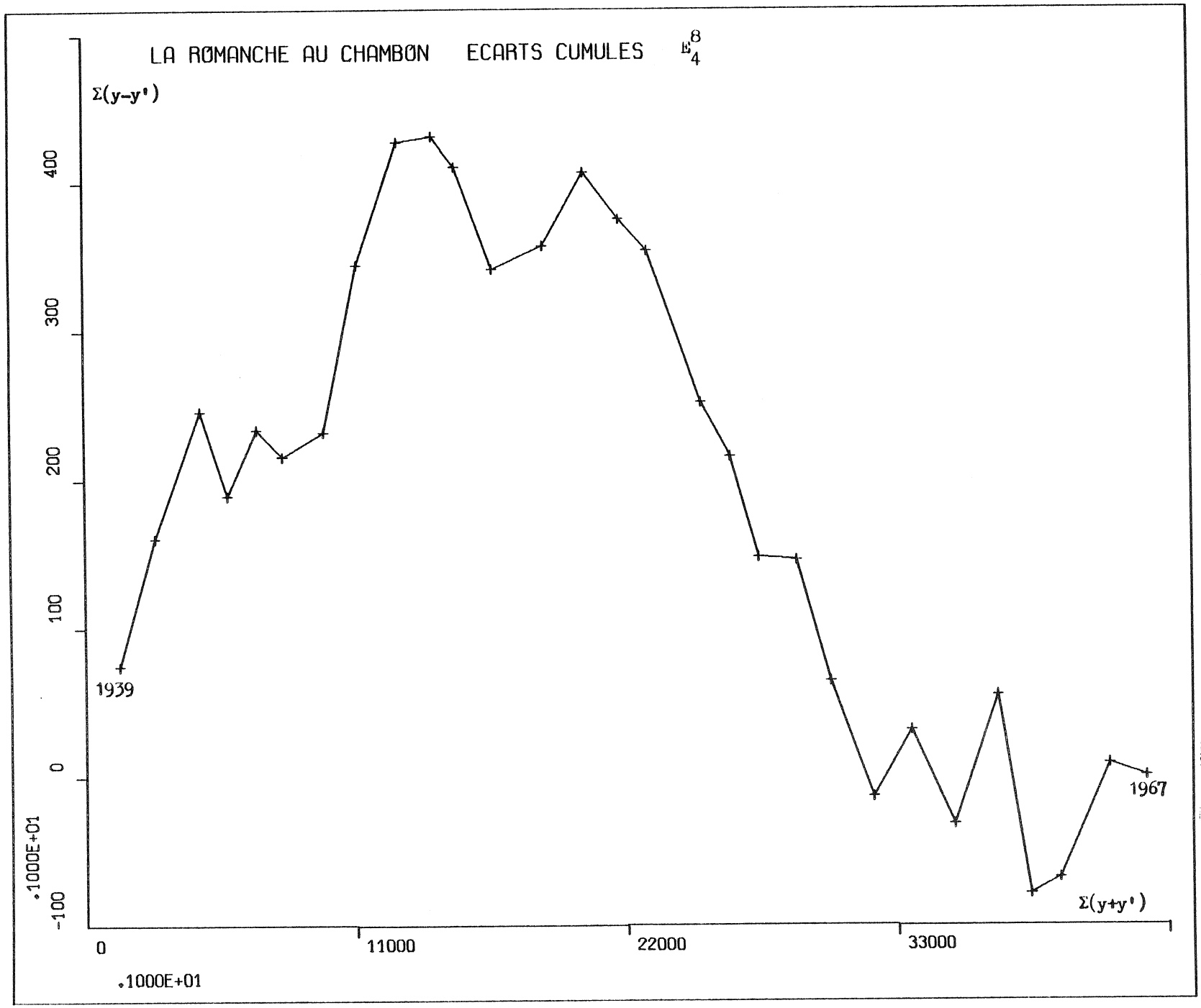
B_1 Distribution liée relative à 1949

C_1 Distribution liée relative à 1955

B_2 Réalisation

C_2 Réalisation

E.D.F. - Division Technique Générale


LA ROMANCHE AU CHAMBON ECARTS CUMULES E_4^8
Σ(y-y')
400
300
200
100
0
-100
.1000E+01
1939
1967
Σ(y+y')
0
11000
22000
33000
.1000E+01

V 23

Le DRAC au SAUTET

P_{10}^{3} : précipitations cumulées du 1er octobre au 31 mars (mm)

E_{10}^{3} : écoulement calculé en mm du 1er octobre au 31 mars

E_{4}^{7} : écoulement calculé en mm du 1er avril au 31 juillet

P_{4}^{7} : précipitations cumulées du 1er avril au 31 juillet

		P_{10}^{3}	E_{10}^{3}	E_{4}^{7}	P_{4}^{7}
1	1941-42	376	231	263	200
2	43	610	417	400	301
3	44	384	260	270	243
4	45	660	465	395	232
5	46	529	283	514	350
6	47	639	300	541	277
7	48	531	325	538	497
8	49	324	207	251	255
9	50	611	291	370	231
10	51	949	457	864	432
11	52	819	517	551	307
12	53	604	428	358	243
13	54	415	293	369	316
14	55	967	511	729	334
15	56	517	281	475	416
16	57	367	215	363	320
17	58	565	247	536	378
18	59	610	416	439	305
19	60	866	431	651	296
20	61	847	749	504	351
21	62	838	436	582	230
22	63	650	193	695	381
23	64	744	516	413	224
24	65	385	185	352	275
25	1965-66	748	506	633	263
26	67	618	368	434	239
27	68	604	280	518	338
28	69	630	384	583	359
29	70	732	348	743	317
30	71	724	322	695	440
31	72	469	175	453	338
32	73	476	285	463	483
33	74	474	229	394	258
34	75	451	232	512	355
35	76	359	231	241	164
36	77	989	597	852	571

DEFINITION DES SYMBOLES FIGURANT DANS LA SORTIE EN PROGRAMME REMULOB

M — moyenne arithmétique des n observations pour chaque variable (1, 2, 3, ...)

S — écart type des n observations pour -d°-

demi-matrice de corrélation totale entre tous les couples de variables

R_p — coefficient de corrélation partielle entre la variable principale et chacune des variables explicatives

RT — coefficient de corrélation totale entre chaque variable explicative et la variable principale

B — coefficient de régression partielle

T — t de student, soit $\frac{B}{S_B}$

beta — soit $B \cdot \frac{S_x}{S_y}$, S_y étant l'écart type de la variable principale et S_x l'écart type de chaque variable explicative

A — terme constant de la relation multilinéaire

R^2 — coefficient de détermination multiple

R — coefficient de corrélation multiple

lambda — $\lambda = \sqrt{1 - R^2}$

S_L — écart type du résidu ou écart type lié $S_L = \lambda S_y$

F — test de Fisher Snedecor : $F = \frac{(n-K)\ R^2}{(K-1)\ (1-R^2)}$, (K étant le nombre total de variables utilisées)

Remarque : lorsque R^2 n'est pas spécifié coefficient brut, il s'agit de l'estimation sans biais $R'^2 = \frac{(n-1)\ R^2 - (K-1)}{n-K}$

REMULOR 0 (MAI 1972) LE 26/11/74 A 11/25/59

4 VARIABLES
24 OBSERVATIONS

DRAC au SAUTET (1 018 km²)
1941-42 à 1964-65
(observations pluies et débits en mm)

	P_{10}^{3}	E_{10}^{3}	E_{4}^{7}	P_{4}^{7}
	1	2	3	4
M	617.0	360.6	476.0	308.1
S	191.1	137.4	153.9	74.5
I	324.0	185.0	251.0	200.0
H	967.0	749.0	864.0	497.0

$$F = \frac{s_1^2}{s_2^2} = \frac{(n-K)\,R^2}{(K-1)\,(1-R^2)} \qquad \nu_1 = K-1 \qquad \nu_2 = n-K$$

$$R'^2 = \frac{(n-1)\,R^2 - (K-1)}{n-K}$$

R

1	1.0000			
2	0.7785	1.0000		
3	0.7956	0.3476	1.0000	
4	0.1657	0.0017	0.5879	1.0000

NOMBRE DE VARIABLES UTILISEES : 2 ; SEUILS D'ELIMINATION : GLOBALE = 0.100 ; PROGRESSIVE = 0.500
ORDRE DE CES VARIABLES
3 1

	RP	RT	B	T	BETA
3					
1	0.7956	0.7956	0.6406	6.2	0.80

A = 80.7516

COEFF.BRUTS R2 = 0.6330 R = 0.7956 LAMBDA = 0.6058 SL = 93.2382 1.28*SL = 119.3449 F = 37.0390
COEFF.CORRIGES R2 = 0.6163 R = 0.7850 LAMBDA = 0.6195 SL = 95.3337 1.28*SL = 122.0272

NOMBRE DE VARIABLES UTILISEES : 3 ; SEUILS D'ELIMINATION : GLOBALE = 0.100 ; PROGRESSIVE = 0.500
ORDRE DE CES VARIABLES
3 1 2

	RP	RT	B	T	BETA
3					
1	0.8920	0.7956	1.0729	9.0	1.33
2	-0.7146	0.3476	-0.7728	-4.7	-0.69

A = 92.6735

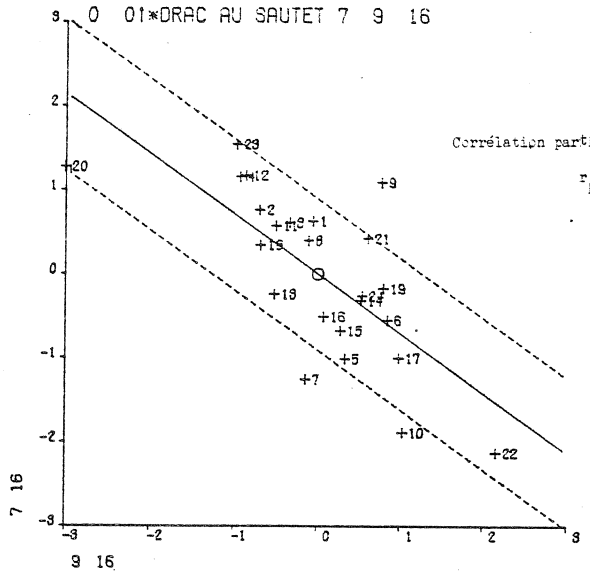
COEFF.BRUTS R2 = 0.8204 R = 0.9057 LAMBDA = 0.4238 SL = 65.2255 1.28*SL = 83.4887 F = 47.2560
COEFF.CORRIGES R2 = 0.8033 R = 0.8963 LAMBDA = 0.4435 SL = 68.2609 1.28*SL = 87.3740

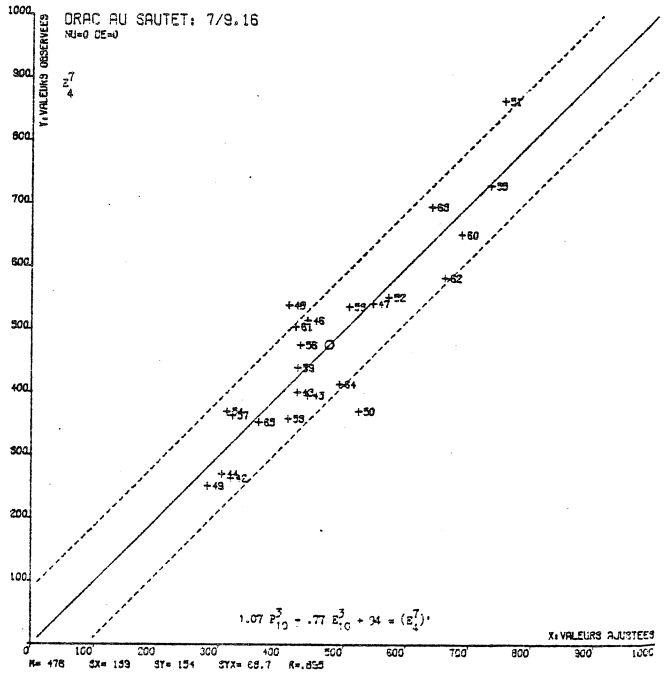
NOMBRE DE VARIABLES UTILISEES : 4 ; SEUILS D'ELIMINATION : GLOBALE = 0.100 ; PROGRESSIVE = 0.500
ORDRE DE CES VARIABLES
3 1 2 4

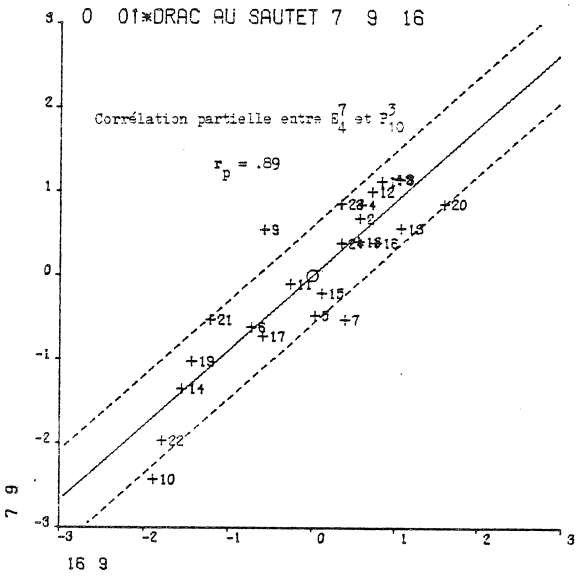
	RP	RT	B	T	BETA
3					
1	0.9674	0.7956	0.9401	17.1	1.17
2	-0.8820	0.3476	-0.6296	-8.4	-0.56
4	0.9003	0.5879	0.8173	9.3	0.40

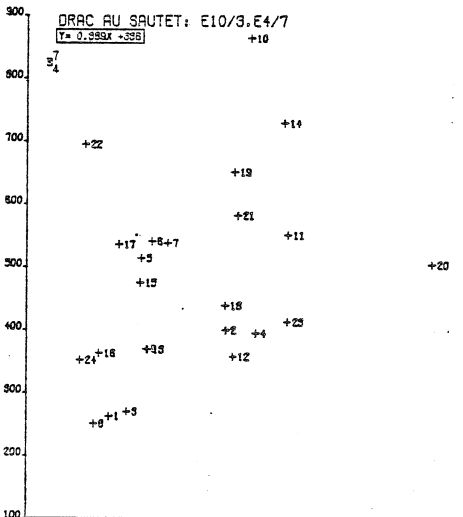
A = -128.7665

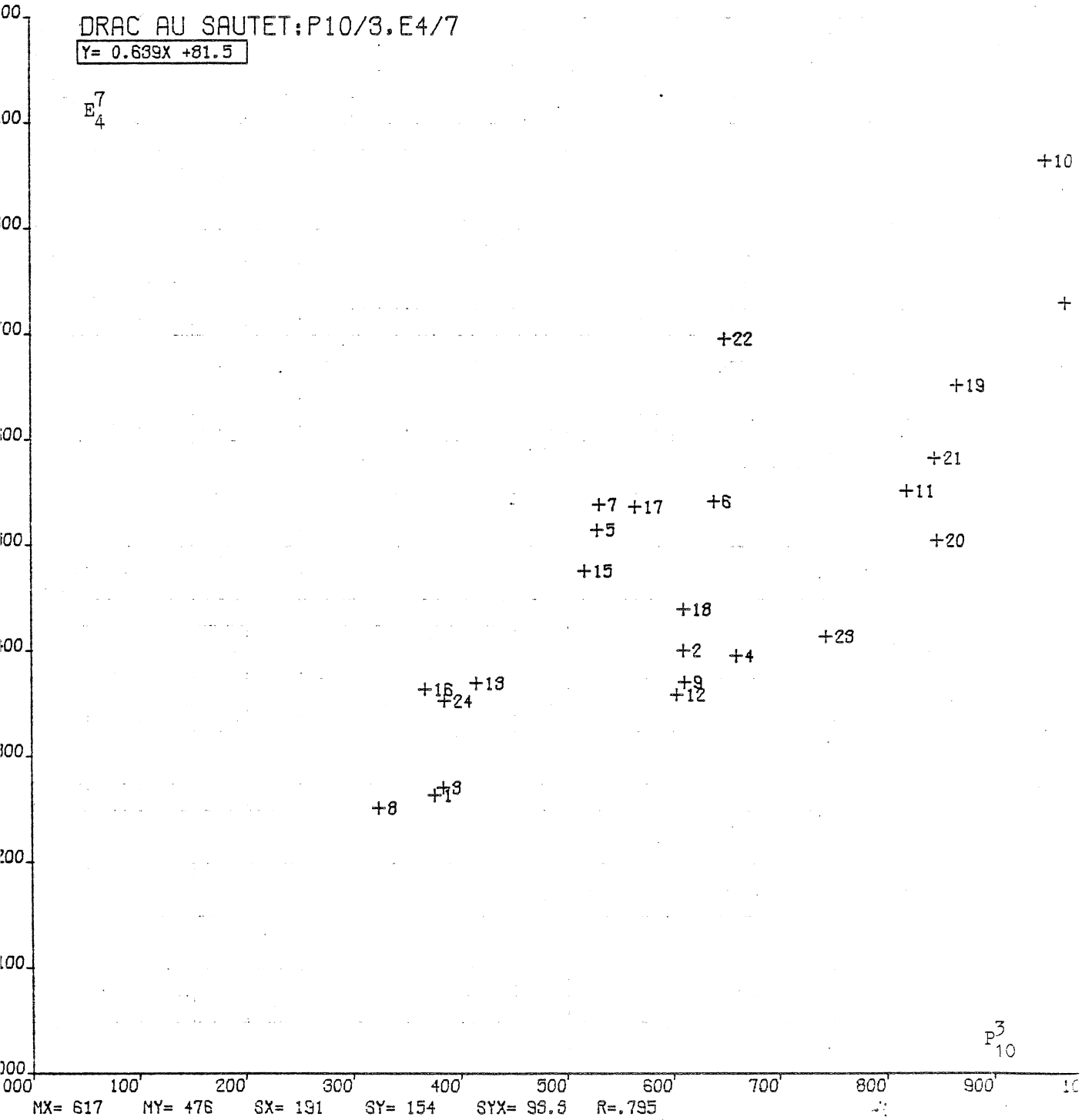
COEFF.BRUTS R2 = 0.9660 R = 0.9828 LAMBDA = 0.1944 SL = 28.3241 1.28*SL = 36.2316 F = 189.3254
COEFF.CORRIGES R2 = 0.9609 R = 0.9802 LAMBDA = 0.1678 SL = 30.4385 1.28*SL = 38.9613


0 01*DRAC AU SAUTET 7 9 16
Corrélation partielle entre E_4^7 et E_{10}^3
$r_p = -.71$
7 16
9 16


DRAC AU SAUTET: 7/9.16
NU=0 DE=0
Y:VALEURS OBSERVEES
X:VALEURS AJUSTEES
1.07 P_{10}^3 = .77 E_{10}^3 + 34 = $(E_4^7)'$
N= 476 SX= 159 SY= 154 SYX= 68.7 R=.855

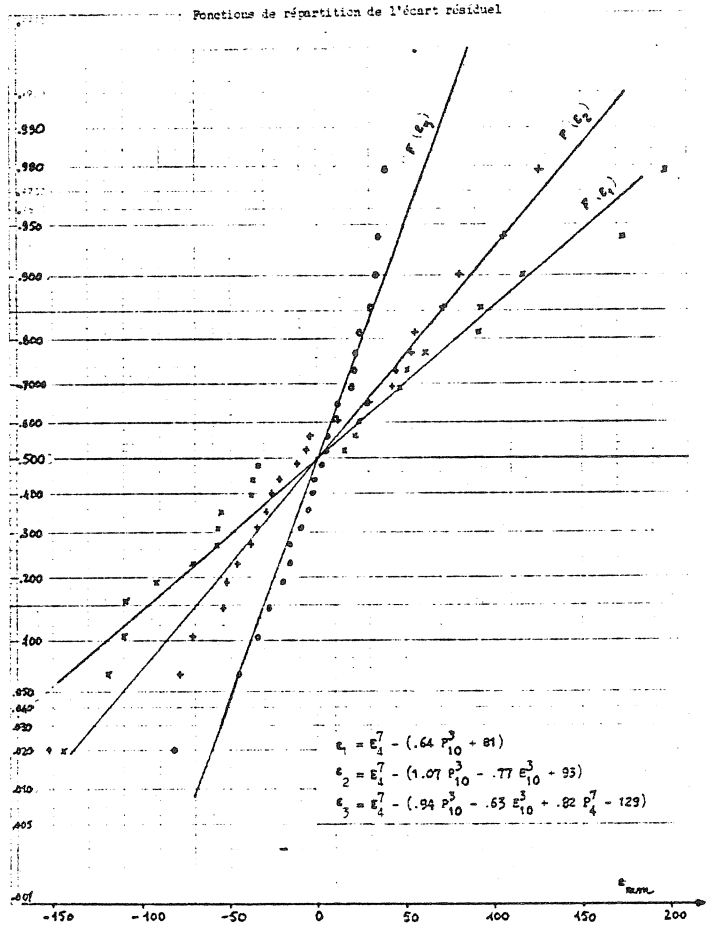

0 01*DRAC AU SAUTET 7 9 16
Corrélation partielle entre E_4^7 et E_{10}^3
r_p = .89
7 9
16 9


DRAC AU SAUTET: E10/3.E4/7
Y= 0.989X +336


DRAC AU SAUTET:P10/3,E4/7
Y= 0.639X +81.5
E 7 4
P 3 10
+10
+22
+19
+21
+11
+7
+17
+6
+5
+20
+15
+18
+23
+2
+4
+16
+13
+24
+9
+12
+3
+1
+8
000
100
200
300
400
500
600
700
800
900
MX= 617
MY= 476
SX= 191
SY= 154
SYX= 99.8
R=.795

Le DRAC au SAUTET

Fonctions de répartition de l'écart résiduel



Le DRAC au SAUTET

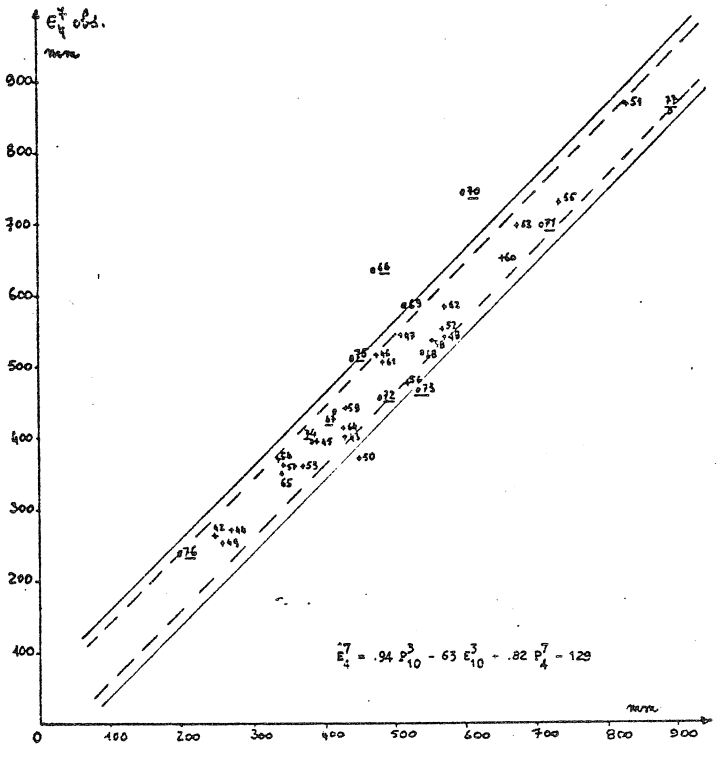


TABLE 4
FRACTILES t_p de la LOI de STUDENT

	0,50	0,40	0,30	0,20	0,10	0,05	0,025	0,010	0,005	0,001	0,0005	Q = 1 - P
ν \ P	0,50	0,60	0,70	0,80	0,90	0,95	0,975	0,990	0,995	0,999	0,9995	
1	0,000	0,325	0,727	1,376	3,078	6,314	12,71	31,82	63,66	318,3	636,6	
2	0,000	0,289	0,617	1,061	1,886	2,920	4,303	6,965	9,925	22,33	31,60	
3	0,000	0,277	0,584	0,978	1,638	2,353	3,182	4,541	5,841	10,22	12,94	
4	0,000	0,271	0,569	0,941	1,533	2,132	2,776	3,747	4,604	7,173	8,610	
5	0,000	0,267	0,559	0,920	1,476	2,015	2,571	3,365	4,032	5,893	6,859	
6	0,000	0,265	0,553	0,906	1,440	1,943	2,447	3,143	3,707	5,208	5,959	
7	0,000	0,263	0,549	0,896	1,415	1,895	2,365	2,998	3,499	4,785	5,405	
8	0,000	0,262	0,546	0,889	1,397	1,860	2,306	2,896	3,355	4,501	5,041	
9	0,000	0,261	0,543	0,883	1,383	1,833	2,262	2,821	3,250	4,297	4,781	
10	0,000	0,260	0,542	0,879	1,372	1,812	2,228	2,764	3,169	4,144	4,587	
11	0,000	0,260	0,540	0,876	1,363	1,796	2,201	2,718	3,106	4,025	4,437	
12	0,000	0,259	0,539	0,873	1,356	1,782	2,179	2,681	3,055	3,930	4,318	
13	0,000	0,259	0,538	0,870	1,350	1,771	2,160	2,650	3,012	3,852	4,221	
14	0,000	0,258	0,537	0,868	1,345	1,761	2,145	2,624	2,977	3,787	4,140	
15	0,000	0,258	0,536	0,866	1,341	1,753	2,131	2,602	2,947	3,733	4,073	
16	0,000	0,258	0,535	0,865	1,337	1,746	2,120	2,583	2,921	3,686	4,015	
17	0,000	0,257	0,534	0,863	1,333	1,740	2,110	2,567	2,898	3,646	3,965	
18	0,000	0,257	0,534	0,862	1,330	1,734	2,101	2,552	2,878	3,611	3,922	
19	0,000	0,257	0,533	0,861	1,328	1,729	2,093	2,539	2,861	3,579	3,883	
20	0,000	0,257	0,533	0,860	1,325	1,725	2,086	2,528	2,845	3,552	3,850	
21	0,000	0,257	0,532	0,859	1,323	1,721	2,080	2,518	2,831	3,527	3,819	
22	0,000	0,256	0,532	0,858	1,321	1,717	2,074	2,508	2,819	3,505	3,792	
23	0,000	0,256	0,532	0,858	1,319	1,714	2,069	2,500	2,807	3,485	3,767	
24	0,000	0,256	0,531	0,857	1,318	1,711	2,064	2,492	2,797	3,467	3,745	
25	0,000	0,256	0,531	0,856	1,316	1,708	2,060	2,485	2,787	3,450	3,725	
26	0,000	0,256	0,531	0,856	1,315	1,706	2,056	2,479	2,779	3,435	3,707	
27	0,000	0,256	0,531	0,855	1,314	1,703	2,052	2,473	2,771	3,421	3,690	
28	0,000	0,256	0,530	0,855	1,313	1,701	2,048	2,467	2,763	3,408	3,674	
29	0,000	0,256	0,530	0,854	1,311	1,699	2,045	2,462	2,756	3,396	3,659	
30	0,000	0,256	0,530	0,854	1,310	1,697	2,042	2,457	2,750	3,385	3,646	
40	0,000	0,255	0,529	0,851	1,303	1,684	2,021	2,423	2,704	3,307	3,551	
60	0,000	0,254	0,527	0,848	1,296	1,671	2,000	2,390	2,660	2,232	3,460	
80	0,000	0,254	0,527	0,846	1,292	1,664	1,990	2,374	2,639	3,195	3,415	
100	0,000	0,254	0,526	0,845	1,290	1,660	1,984	2,365	2,626	3,174	3,389	
200	0,000	0,254	0,525	0,843	1,286	1,653	1,972	2,345	2,601	3,131	3,339	
∞	0,000	0,253	0,524	0,842	1,282	1,645	1,960	2,326	2,576	3,090	3,291	

TABLE 4
FRACTILES DE LA LOI DE t
(Loi de Student)

La loi de t (loi de Student) est définie par la probabilité élémentaire

$$f(t,\nu) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} dt$$

Le paramètre ν est le nombre de degrés de liberté (d.d.l.).

La table donne les fractiles de la loi de t, c'est-à-dire les valeurs t_p telles que $Pr[t < t_p] = P$ pour $P \geq 0,50$ $(t \geq 0)$. Pour les valeurs de $P < 0,50$ $(t < 0)$, on a $t_p = - t_{1-p}$.

Lorsque $\nu \to \infty$, la loi de Student tend vers la loi normale réduite : la ligne $\nu = \infty$ donne les valeurs que l'on trouve dans la table 1.3 pour les mêmes valeurs de P. On peut considérer l'approximation normale comme valable pour $\nu > 100$, et même pratiquement pour $\nu > 60$.

Par exemple

pour $\nu = 9$ $Pr[t < 2,262] = 0,975$ $Pr[t < - 2,262] = 0,025$
$Pr[- 2,262 < t < 2,262] = 0,95$

pour $\nu = 140$ $Pr[- 1,96 < t < 1;96] \# 0,95$

5.3 - La corrélation multiple

Nous généralisons à présent toutes les propriétés développées dans le cas de la corrélation double, en considérant le traitement de l'information hydrométéorologique destiné essentiellement à mettre au point des "modèles" de prévision.

Et plutôt que d'exposer des méthodes de calcul (abondamment développées dans la littérature mathématico-scolaire) dont le formalisme pour certains, ou l'apparente simplicité pour d'autres, détourne trop souvent l'attention au détriment de l'intérêt et des difficultés de l'application pratique, nous essaierons de montrer, sur des cas concrets, quelles sont les préoccupations du "tailleur" de prévision. En particulier les choix qu'il est amené à faire lorsqu'il utilise le calcul automatique pour mettre au point ses schémas de prévision.

En préambule, on ne peut qu'insister une nouvelle fois sur l'indispensable qualité des données utilisées dans ces calculs. On peut établir un modèle très sophistiqué, si les données d'entrée sont mauvaises, la sortie rendra caduque toute application. Il n'est pas exagéré d'affirmer que dans bien des cas, la moitié du travail consiste à contrôler et critiquer les séries de mesures (précipitations, débits, température) en s'appuyant sur des corrélations spatiales et temporelles.

Une erreur courante, mais grave, consiste à croire que le fait d'effectuer un traitement des données par ordinateur suffit à leur délivrer le "label de qualité" : on ne peut se faire une opinion qu'en associant constamment les résultats numériques et graphiques.

5.3.1 - Généralités et définition des variables utilisées

L'écoulement entre deux instants t_1 et t_2 (l'intervalle de temps entre t_1 et t_2 pourra être de 24 h - 48 h - 7 jours - 1 mois - 5 mois - 1 an, etc.) dépend de deux ensembles de facteurs :

- ceux qui caractérisent l'état antérieur ou initial du bassin versant avant l'instant t_1, sa mémoire ou inertie (I)

- ceux qui caractérisent les conditions météorologiques (P) régnant pendant la période prévisionnelle (entre les instants t_1 et t_2).

Il ne peut être question de rechercher de liaison fonctionnelle entre l'écoulement et les autres facteurs, pour deux raisons :

. on dispose seulement de quelques longues séries de mesures ponctuelles de débits, précipitations et températures de l'air, piézométres (pour le niveau des nappes souterraines) observées en continu (diagrammes - cassettes) ou discontinu (hauteurs, compteurs d'accumulation) ;

. la complexité, l'interférence, la variabilité spatiale et temporelle des facteurs conditionnant l'écoulement, rendent illusoire et arbitraire une formulation mathématique directe en vraie grandeur des termes du bilan d'écoulement.

La seule solution rationnelle et objective consiste à traiter ces influences en valeur relative, et à les représenter par des indices ou témoins, dont on recherche, d'après les séries d'observations, la corrélation avec l'écoulement à prévoir.

La structure de la relation la plus simple qu'on puisse imaginer est un modèle multilinéaire, modèle auquel se ramène toute relation si compliquée soit-elle, si on néglige, dans une première étape des recherches, les termes de second ordre :

$$E_{t_1}^{t_2} = I_t^{t_1} + P_{t_1}^{t_2} + \varepsilon$$

On notera pour simplifier :

- Y la variable à expliquer (ou variable andogène), définie également comme prévisande ,

- X_j les variables explicatives (ou variables exogènes), définies comme préviseurs.

5.3.2 - Calcul des coefficients de régression multiple

On cherchera à définir un invariant, dans le cas présent une relation linéaire définie par p coefficients a_j $(0 \leqslant j \leqslant p)$, de telle sorte que :

$$\textcircled{1} \qquad \sum_{i=1}^{n} \left[Y_i - \sum_{j=1}^{p} a_j X_j - a_o \right]^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \text{Minimum}$$

c'est la condition des moindres carrés qui va permettre de calculer les coefficients a_j.

En notation matricielle, si on note :

- le vecteur colonne des coefficients : a (p,1)
- la matrice des données centrées $(x_i = X_i - \overline{X})$: x (n,p)
- le vecteur colonne des données y_i centrées $(y_i = Y_i - \overline{Y})$: y (n,1)
- le vecteur colonne écarts résiduels ε_i : ε (n,1)

la relation multilinéaire s'écrit alors : $y = x\,a + \varepsilon$

Chercher le minimum de $E(a) = (y - xa)'\,(y - xa)$, notant par ' l'opération matrice transposée, conduit à résoudre :

$$a = (x'\,x)^{-1} \,.\, x'\,y$$

En notation classique, résoudre ① revient à résoudre le système de p + 1 relations linéaires à p + 1 inconnues $(a_o, a_1, \ldots, a_p)$ suivant :

$$\frac{\delta \sum_{i=1}^{n} \varepsilon_i^2}{\delta\, a_j} = 0 \qquad \text{pour } j = 0, 1, \ldots, p$$

ce qui revient à résoudre :

$$\begin{cases} \sum_{i=1}^{n} X_{ji}\ \varepsilon_i = 0 & \text{pour } j = 1 \text{ à } p \qquad \textcircled{2} \\ \sum \varepsilon_i = 0 & \text{pour } j = 0 \qquad \textcircled{3} \end{cases}$$

- la relation ③ signifie que la moyenne arithmétique des écarts résiduels est nulle et que l'hyperplan $Y' = \sum_{j=1}^{p} a_j X_j + a_o$ passe par le centre de gravité des observations $(\overline{Y}, \overline{X}_1, \overline{X}_2, \ldots, \overline{X}_p)$;

- la relation ② signifie l'indépendance <u>linéaire</u>, au sens statistique, entre les vecteurs X_j et ε (ils sont en effet orthogonaux).

Si l'on effectue les calculs sur des variables centrées ($x_{ji} = X_{ji} - \overline{X}_j$), le système ② et ③ se simplifie, on doit alors résoudre :

$$\sum_{i=1}^{n} x_{ji}\, \varepsilon_i = 0 \qquad \text{pour } j = 1 \text{ à } p \qquad ④$$

soit encore :

$$\begin{cases} a_1 \Sigma\, x_{1i}^2 + a_2 \Sigma\, x_{1i}\, x_{2i} + \ldots + a_p \Sigma\, x_{1i}\, x_{pi} = \Sigma\, x_{1i}\, y_i \\ \ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots \\ a_1 \Sigma\, x_{pi}\, x_{1i} + \ldots\ldots\ldots\ldots\ldots + a_p \Sigma\, x_{pi}^2 = \Sigma\, x_{pi}\, y_i \end{cases}$$

en résolvant les p systèmes suivants :

$$(1) \quad \begin{cases} c_{11} \Sigma\, x_{1i}^2 + \ldots + c_{1p} \Sigma\, x_{1i}\, x_{pi} = 1 \\ c_{11} \Sigma\, x_{2i}\, x_{1i} + \ldots + c_{1p} \Sigma\, x_{2i}\, x_{pi} = 0 \\ \ldots\ldots\ldots\ldots\ldots \\ c_{11} \Sigma\, x_{pi}\, x_{1i} + \ldots + c_{1p} \Sigma\, x_{pi}^2 = 0 \end{cases}$$

$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$

$$(p) \quad \begin{cases} c_{p1} \Sigma\, x_{1i}^2 + \ldots + c_{pp} \Sigma\, x_{1i}\, x_{pi} = 0 \\ \ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots \\ c_{p1} \Sigma\, x_{pi}\, x_{1i} + \ldots + c_{pp} \Sigma\, x_{pi}^2 = 1 \end{cases}$$

on montre que l'on obtient :

$$\begin{cases} a_1 = \Sigma\, (c_{11}\, x_{1i} + c_{12}\, x_{2i} + \ldots + c_{1p}\, x_{pi})\, y_i \\ \vdots\vdots\vdots \\ a_p = \Sigma\, (c_{p1}\, x_{1i} + \ldots\ldots\ldots\ldots + c_{pp}\, x_{pi})\, y_i \end{cases}$$

On peut d'ailleurs obtenir ces coefficients à partir de la matrice des coefficients de corrélation totale, entre tous les couples possibles (X_j, X_k) et (Y, X_j), il suffira d'inverser cette matrice et de calculer le rapport de certains de ses éléments.

Considérons le déterminant à p lignes et colonnes, des coefficients de corrélation simple :

n° ligne / n° colonne	0	1	2	...	p
0	1	r_{X_1Y}	r_{X_2Y}	...	r_{X_pY}
1	r_{YX_1}	1	$r_{X_2X_1}$		
2	r_{YX_2}				
⋮					
p	r_{YX_p}				1

$$\Longrightarrow A = \begin{bmatrix} 1 & r_{X_1Y} & \cdots & r_{X_pY} \\ r_{YX_1} & 1 & & \\ & & & \\ r_{YX_p} & & & 1 \end{bmatrix}$$

On note A_{Kj} le déterminant obtenu en supprimant la $K^{ième}$ ligne et la $j^{ième}$ colonne $(0 \leqslant K, j \leqslant p)$, 0 désigne la première ligne ou colonne des coefficients de corrélation simple entre chaque variable explicative et la variable principale. On trouve ainsi que :

$$a_j = a_{oj} = -(-1)^j \frac{A_{oj}}{A_{oo}} \frac{S_Y}{S_{X_j}} \quad \text{(on tiend compte de la règle des signes)}$$

5.3.3 - Calcul des coefficients de corrélation partielle

Le coefficient de corrélation partielle entre la variable principale et l'une des variables explicatives, $X_j = X_1$ par exemple, est le coefficient de corrélation entre les variables Y et X_1 désinfluencées chacune des autres variables explicatives $X_2, \ldots, X_p$.

Soit : $\xi_1 = Y - (\alpha_2 X_2 + \alpha_3 X_3 + \ldots + \alpha_p X_p + \alpha_o)$

$\theta_1 = X_1 - (\beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p + \beta_o)$

les coefficients α et β étant obtenus par moindres carrés

le coefficient de corrélation partielle entre Y et X_1 :

$$r_{Y\,X_1,\,X_2\,X_3\,X_4\,\ldots\,X_p} = \frac{\sum_{i=1}^{n} \xi_{1i}\ \theta_{1i}}{\sqrt{\Sigma\ \xi_{1i}^2}\ \sqrt{\Sigma\ \theta_{1i}^2}}$$

on peut l'obtenir par récurrence :

$$r_{Y\,X_1,\,X_2} = \frac{r_{Y\,X_1} - r_{Y\,X_2}\ r_{X_1\,X_2}}{\sqrt{1 - r_{Y}^2\,X_2}\ \sqrt{1 - r_{X_1}^2\,X_2}}$$

$$r_{Y\,X_1,\,X_2\,X_3} = \frac{r_{Y\,X_1,\,X_2} - r_{Y\,X_3,\,X_2}\ r_{X_1\,X_3,\,X_2}}{\sqrt{1 - r_{Y\,X_3}^{\ 2},\,X_2}\ \sqrt{1 - r_{X_1}^{\ 2}\,X_3,\,X_2}}$$

mais c'est une méthode extrêmement lourde.

On peut calculer ces coefficients de corrélation partielle par les déterminants, en effet :

$$r_{Y\,X_1,\,X_2\,X_3\,\ldots\,X_p} = + \frac{A_{o1}}{\sqrt{A_{oo}}\ \sqrt{A_{11}}}$$

plus généralement :

$$r_{Y\,X_j,\,X_2\,\ldots\,X_p} = - (-1)^j \frac{A_{oj}}{\sqrt{A_{oo}}\ \sqrt{A_{jj}}}$$

Remarque : les coefficients de corrélation partielle que l'on calcule sur un échantillon sont différents de ceux de la population totale, on obtient une estimation sans biais $\hat{r}$ d'après :

$$\hat{r}_{Y}^{\ 2}\,X_j,\,X_1\,\ldots\,X_p = \frac{(n-p)\ r_{Y}^{\ 2}\,X_j,\,X_1\,\ldots\,X_p - 1}{n - p - 1}$$

5.3.4 - Calcul du coefficient de corrélation multiple R

On obtient une estimation du coefficient de détermination multiple (R^2) en calculant la corrélation entre valeurs ajustées par la relation multilinéaire $Y'_i = \sum_{j=1}^{p} a_j x_{ji} + a_o$ et valeurs observées Y_i

$$R^2 = \frac{\left[\Sigma (Y_i - \overline{Y})(Y'_i - \overline{Y})\right]^2}{\Sigma (Y_i - \overline{Y})^2 \; \Sigma (Y'_i - \overline{Y})^2}$$

ou encore :
$$R^2 = \frac{\Sigma (Y'_i - \overline{Y})^2}{\Sigma (Y_i - \overline{Y})^2}$$

On peut également calculer ce coefficient d'après les coefficients de corrélation partielle :

$$R^2 = 1 - (1 - r_{Y^2 X_1})(1 - r_{Y^2 X_2, X_1}) \ldots (1 - r_{Y^2 X_p, X_1 X_2 \ldots X_{p-1}})$$

on obtient, en utilisant les déterminants :

$$R^2 = 1 - \frac{A}{A_{oo}}$$

5.3.5 - Analyse des variances

On définit la variance résiduelle S^2_ε ou variance de Y liée à X_1, ..., X_p soit S^2_L par

$$S^2_\varepsilon = S^2_L = \frac{\Sigma (Y - Y')^2}{n}$$

on montre aisément que la variance totale est la somme des variances dues à la liaison multilinéaire et résiduelle

$$S^2_Y = S^2_1 + S^2_\varepsilon$$

Mais il s'agit d'échantillon et il faut tenir compte du nombre de degrés de liberté :

On démontre que la densité de répartition s'écrit (lorsque $\rho = 0$) :

$$f_n(r) = \frac{1}{\sqrt{\pi}} \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n-2}{2})} (1-r^2)^{(\frac{n-4}{2})}$$

La variable transformée $t = \sqrt{n-2} \frac{r}{\sqrt{1-r^2}}$ suit une loi de Student à (n-2) degrés de liberté.

Ainsi si t_p est la valeur correspondant à la probabilité p %, nous aurons p chances sur 100 d'obtenir $|t| > t_p$ ou encore : $|r| > \frac{t_p}{\sqrt{t_p^2 + n-2}}$

Test appliqué à un coefficient de corrélation partielle

Supposons la variable dépendante définie par K variables explicatives. Soit θ un coefficient de corrélation partielle dans laquelle on a éliminé l'influence de K-1 variables.

La densité de répartition θ s'écrit :

$$g_n(\theta) = \frac{1}{\sqrt{\pi}} \frac{\Gamma(\frac{n-k}{2})}{\Gamma(\frac{n-k-1}{2})} (1-\theta^2)^{(\frac{n-k-3}{2})}$$

Il s'ensuit que la variable transformée t : $\sqrt{n-k-1} \cdot \frac{\theta}{\sqrt{1-\theta^2}}$ suit une loi de Student à (n-k-1) degrés de liberté.

Donc si t_p est la valeur correspondant à la probabilité p %, nous aurons p chances sur 100 d'obtenir $|t| > t_p$ ou $\theta > \frac{t_p}{\sqrt{t_p^2 + n-k-1}}$

Remarques :

1°- Il existe également un test dû à Fisher et utilisant la variable transformée $z = \frac{1}{2} \text{Log} \frac{1+r}{1-r}$ (r étant un coefficient de corrélation totale ou partielle). Fisher a montré que cette variable suit approximativement une loi normale de moyenne nulle et d'écart type $\frac{1}{\sqrt{n-(2+p)}}$ (p = 1 pour une corrélation simple).

2°- Nombre de variables à introduire dans une régression multiple : la somme des carrés de résidus $\Sigma (y-y')^2$; (y' = valeur ajustée par équation de régression, y = valeur observée) est d'autant plus petite que le nombre K des variables explicatives est plus élevé. On peut se demander si cette amélioration est justifiée par l'introduction de nouvelles variables.

On remarque que R_K^2 étant le coefficient de corrélation multiple obtenu en utilisant K variables explicatives, R_{K+1}^2 étant celui obtenu en introduisant une variable explicative supplémentaire, les variances résiduelles sont respectivement :

$$S_{2_K}^2 = \frac{1 - R_K^2}{(n-k-1)} \Sigma (y- \overline{y})^2$$

$$S_{2_{K+1}}^2 = \frac{1-R_{K+1}^2}{(n-k-2)} \Sigma (y - \overline{y})^2$$

Pour que l'influence de cette nouvelle variable soit significative, il faut au moins que :

$$\frac{1-R_{K+1}^2}{(n-k-2)} < \frac{1-R_K^2}{(n-k-1)}$$

soit : $$R_{K+1}^2 > 1 - \frac{(n-k-2)\ (1-R_K^2)}{(n-k-1)}$$

Exemple d'application : contrôle d'homogénéité d'une série

Explication de l'écoulement annuel (année hydrologique) de la Cère à St-Etienne-Cantalès en fonction des précipitations et températures mensuelles, il ne s'agit pas ici d'un calcul destiné à la prévision mais plutôt de contrôler l'homogénéité de la série des débits.

Précipitations -

Pour caractériser les précipitations reçues par le bassin de la Cère, on dispose d'une seule longue série d'observations à Marmanhac, série qui est en bon accord avec celles de Messeix-les-Mines et Argentat (B.V. Dordogne). Il se trouve que la pluie à Marmanhac est proche de la pluviométrie

moyenne sur le bassin de la Cère, mais ce n'est pas là une condition nécessaire pour qu'il soit un bon témoin - on a jugé utile de prendre un témoin plus étoffé que Marmanhac seul en ajoutant la moyenne Vic-sur-Cère + Saint-Etienne-Cantalès + Marmanhac de 1948 à 1966 : sur les années récentes, la corrélation entre les valeurs annuelles des deux témoins atteint 0.99 avec un coefficient de régression voisin de 1.

Températures -

Il s'agit de la série des températures observées à Messeix qui est en étroite corrélation avec celle de Saint-Flour. La série a également été contrôlée par celle du Mont Dore et de Clermont-les-Landais.

On peut considérer que cette station est un bon témoin de la température moyenne de l'air sur le plateau du Massif Central.

Débits -

Jusqu'en 1945, les débits mesurés à l'usine de Montvert (B.V. 764 km²), contrôlés avec la série de la Maronne aux Estourocs, ont été affectés d'un coefficient 0.9 (rapport des B.V.), ensuite on a utilisé la série des débits mesurés à l'usine de St-Etienne-Cantalès.

On contrôle habituellement la cohérence entre débits et précipitations en mettant en corrélation E^{9}_{10}, écoulements d'octobre à septembre avec P^{9}_{10}, ou mieux $P^{4}_{10} + .5\ P^{9}_{5}$. Dans le cas de la Cère, on a utilisé une pondération plus raffinée en comparant l'écoulement annuel à une combinaison linéaire, obtenue par corrélation multiple, comprenant :

- les précipitations des 12 mois de l'année hydrologique,
- un index de l'état initial : le logarithme des 7 derniers jours du mois de septembre précédant l'année hydrologique,
- les températures de mai et septembre (mesurées à Messeix)

- Tableau I, figure 2 -

On peut alors s'assurer d'une part, de la représentativité individuelle des pluviomètres, et également de l'hétérogénéité éventuelle d'une série par la méthode des lignes d'écarts cumulés, qui consiste à prendre

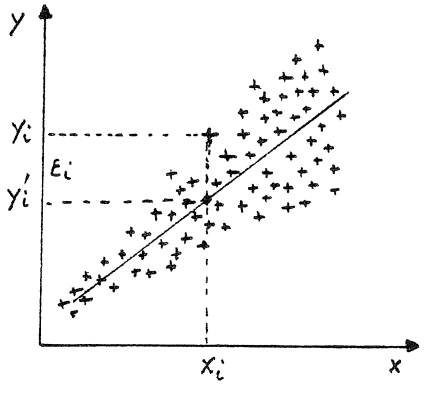
comme nouveaux axes les bissectrices des axes précédents gradués en $\Sigma (E_o + E_c)$ et $\Sigma (E_o - E_c)$, ce qui a pour effet d'amplifier considérablement tout écart systématique. Ainsi, sur la figure 3, apparaît un écart systématique d'environ + 2 % en moyenne jusqu'en 1945-46, date de la mise en exploitation de l'usine de Cantalès, et de - 1,5 % ensuite, écart que nous n'avons pas jugé critique pour la suite des calculs, la corrélation entre écoulement calculé et l'écoulement observé étant très étroite. La solution la plus rationnelle serait de n'utiliser que la série 1947-1965.

5.4 - Conditions d'application de la technique des corrélations multiples

Les réflexions et commentaires qui suivent nous ont été suggérés par la pratique fréquente de ces techniques; une réponse précise, des démonstrations complètes pourront être trouvées dans la bibliographie (non exhaustive), pour certains problèmes.

5.4.1 - L'homoscédasticité

C'est l'invariance de l'écart type des résidus ε_i, quelle que soit la valeur de X. Cette propriété est fondamentale puisqu'elle intervient directement dans le calcul des coefficients de régression. En effet si la dispersion des écarts est plus grande dans les fortes valeurs de x, il est évident que ces données conditionneront les valeurs des coefficients a_j obtenus en résolvant le système :



$$\sum_{i=1}^{n} X_{ji}\ \varepsilon_i = 0 \qquad \text{pour } j = 1 \text{ à } p$$

On peut évidemment effectuer un tel calcul sans aucune précaution, si l'on a des raisons de donner un poids privilégié aux fortes valeurs.

Une procédure, pour calculer les coefficients sans transformation de variable, consiste à utiliser la méthode des moindres carrés pondérés :

$$\text{on minimise} \quad \sum_{i=1}^{n} \varpi_i^2 \left[Y_i - a X_i - b \right]^2 ,$$

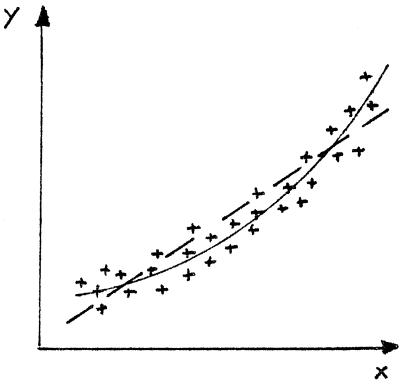
ϖ_i est une fonction de X : $\varpi_i = \varpi(X_i)$; généralement ϖ_i est inversement proportionnel à la variance des Y_{ik} (k = 1 à m) pour une valeur de X_i donnée (ou pour une classe sur X) $\varpi_i^2 \neq \frac{1}{\sigma Y_i^2}$.

C'est d'ailleurs le principe utilisé dans la méthode des Probits [1] .

Plus simplement, et surtout si l'on dispose de peu d'observations, on effectue une transformation élémentaire sur la variable principale.

5.4.2 - La linéarité

Cette condition est importante mais non essentielle, car on peut généralement effectuer une transformation simple, en logarithme ou puissance fractionnaire, sur x pour retrouver une relation quasi linéaire



$$Y = \sum_{j=1}^{p} a_j Z_j + a_o + \varepsilon$$

avec $$Z_j = \begin{cases} X_j & \text{ou} \\ \alpha X_j^p + \beta X_j^{p-1} + \ldots & \text{ou} \\ \log X_j \end{cases}$$

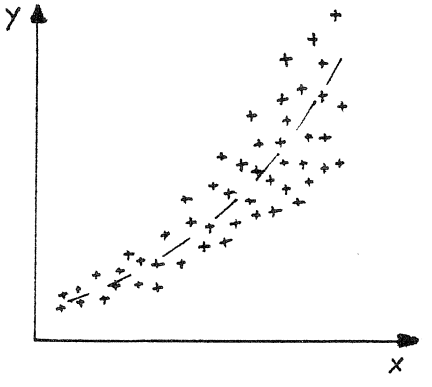
On voit ainsi qu'il s'agit d'un problème d'ajustement, puis de corrélation; généralement on obtient directement les coefficients de pondération des transformées de X_j en traitant simultanément les variables explicatives transformées, ou non, par les moindres carrés.

Les graphiques de corrélation partielle entre la variable principale Y et chaque variable explicative X_j peuvent suggérer le type de transformation, on peut songer à procéder ainsi de façon itérative jusqu'à ce que

le nuage de points soit sensiblement elliptique, c'est sans doute illusoire car on ne sait pas à quel moment l'optimum est atteint et on ne doit pas oublier que l'on traite souvent un échantillon de 30 à 50 observations.

Pour toutes ces raisons, la transformation doit être simple et toute extrapolation, en dehors du domaine observé sur lequel ont été effectués des calculs de corrélation, est à proscrire. Dans ces conditions la forme mathématique exacte de la transformation n'est pas critique.

Lorsque la relation entre Y et X n'est pas linéaire, mais est homoscédastique (2), il convient d'effectuer une transformation sur X. Si la relation en X et Y est non linéaire et hétéroscédastique, on effectue une transformation sur Y et éventuellement sur X



$$\begin{cases} \log Y = \Sigma a_j X_j + a_o + \varepsilon \\ \log Y = \Sigma a_j \operatorname{Log} X_j + a_o + \varepsilon \end{cases}$$

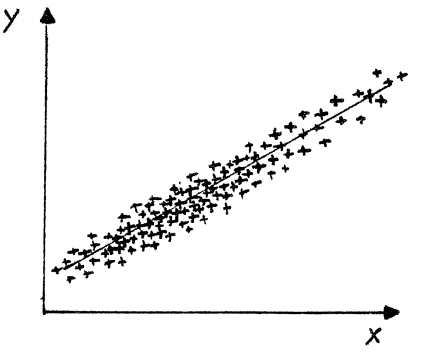
Un moyen pratique de tester la linéarité dans le cas de la corrélation simple, et pour de petits échantillons, est de découper l'intervalle de variation de X en trois classes et de tester si dans chacune d'elles la moyenne des écarts résiduels ε_i est significativement différente de 0.

Si l'on dispose d'un échantillon important de données, on peut effectuer une segmentation, c'est-à-dire découper l'intervalle de variation de x en classes, à l'intéroeur desquelles on ajustera des segments de droites par les moindres carrés, en lissant éventuellement les paramètres de ces droites. On substitue à une relation curviligne, une représentation "polygonale".

5.4.3 - La dissymétrie

La symétrie de la distribution des valeurs de ε_i (la condition d'homoscédasticité étant respectée) n'est pas fondamentale pour l'application des moindres carrés. Cette propriété est surtout utile lorsqu'on teste la signification statistique ces coefficients a_j (test de Student) puisque cela suppose la normalité des variables et résidus. En fait, ces tests sont robustes et restent encore applicables avec des variables dont la distribution n'est pas trop dissymétrique. La droite de régression qui passe par les moyennes des Y, liée à X, n'est plus confondue avec le lieu des médianes de la distribution liée.

Toutefois, si la variable explicative et la variable principale ont une distribution dissymétrique, et s'il y a homoscédasticité, cela a une répercussion sur l'estimation de la pente a_1.



On peut s'affranchir de l'incidence des fonctions de répartition en calculant le coefficient de rang.

Il s'agit d'un index destiné à évaluer le degré d'association entre deux séries de valeurs, par exemple les précipitations du mois de juin observées en deux stations au cours des 40 dernières années (voir exemple).

Le principe consiste à remplacer des variables continues par leur rang; en effet on dispose de deux chroniques de mesures X_i et Y_i $(1 \leqslant i \leqslant n)$, n étant le nombre d'années d'observations, on classe ces valeurs dans l'ordre croissant $(\hat{x}_1, \ldots \hat{x}_k, \ldots \hat{x}_n)$ et $(\hat{y}_1, \ldots \hat{y}_n)$. On constate alors que les observations effectuées la même année n'occupent pas forcément le même rang.

Le coefficient de corrélation de rang ou de Spearman se calcule ainsi :

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} di^2}{n\,(n^2-1)}$$

di = la différence de rang qu'occupent X_i et Y_i dans le classement.

Lorsque $\Sigma d_i^2 = 0$, il y a concordance parfaite et $\rho = 1$; il y a discordance parfaite lorsque $\rho = -1$.

Propriétés de ce coefficient - Comparaison avec le coefficient de corrélation linéaire r

- On peut comparer deux séries dont la distribution est très dissymétrique, alors que le calcul de r suppose la normalité des distributions.

- Il est indépendant de toute transformation monotone effectuée sur les variables.

- Il est rapide à calculer lorsque $n < 50$.

- Les coefficients ρ et r sont en étroite corrélation, cette liaison étant bien sûr fonction de n. Toutefois le coefficient r est toujours plus élevé que ρ, de 2 à 15 % suivant les valeurs de n et r; lorsque les variables sont normales : $r = 2 \sin \left(\frac{\pi \rho}{6}\right)$.

- Il est non paramétrique (ne dépend ni de la moyenne ni de l'écart type).

- On peut tester ce coefficient à l'aide du test de Student $t = \rho \sqrt{\frac{n-2}{1-\rho^2}}$ avec n-2 degrés de liberté.

- Il n'y a d'ailleurs pas d'inconvénient à adopter pour ρ l'écart type de r : $\sigma_r = \frac{1-r^2}{\sqrt{n-1}}$

Traitement multidimensionnel des coefficients de rang

On peut effectuer une analyse multidimensionnelle de la matrice des coefficients de Spearman calculés sur tous les couples de variables $[X_i, X_j]$ avec $1 \leqslant i, j \leqslant P$.

Les moyennes, ainsi que les variances des séries de rang, sont égales respectivement à :

On dispose d'un tableau d'observations sur p+1 variables :

Y	X_1	X_2	X_3	X_4	X_p
y_1	x_{11}	x_{21}	x_{31}	x_{41}	x_{p1}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
y_n	x_{1n}	x_{2n}			x_{pn}

Dans le cas général $Y = \Sigma\, a_j\, X_j + a_o + \varepsilon$; si l'on impose la condition $a_o = 0$ on devra résoudre le système linéaire suivant :

$$\sum_{i=1}^{n} \left(Y_i - \Sigma\, a_j\, X_{ji}\right) X_{ji} = 0, \qquad \text{pour } j = 1 \text{ à } p$$

1/- Cas où $j = 1$: $\qquad Y = a\,X + \varepsilon$

$$\Sigma \left(Y_i - a\, X_i\right) X_i = 0$$

conduit à la relation :

$$a = \frac{\sum_{i=1}^{n} X_i\, Y_i}{\sum_{i=1}^{n} X_i^{\,2}}$$

Le programme de corrélation multiple calcule habituellement :

- les moyennes arithmétiques $\overline{X}$, $\overline{Y}$
- les écarts types S_x, S_y
- le coefficient de corrélation $r_{xy} = r$

il est facile d'obtenir alors :

$$a = \frac{r\, S_x\, S_y + \overline{X}\,\overline{Y}}{S_x^{\,2} + \overline{X}^2}$$

2/ - Cas où $j = 2$: $\qquad Y = b\,X + cZ + \varepsilon$

soit à résoudre :

$$\begin{cases} \sum_{i=1}^{n} \left(Y_i - b\, X_i - c\, Z_i\right) X_i = 0 \\ \sum_{i=1}^{n} \left(Y_i - b\, X_i - c\, Z_i\right) Z_i = 0 \end{cases}$$

on obtient :

$$c = \frac{\Sigma Y_i Z_i - b \Sigma X_i Z_i}{\Sigma Z_i^2}$$

et :

$$b = \frac{(\Sigma X_i Z_i)(\Sigma Y_i Z_i) - (\Sigma Z_i^2)(\Sigma X_i Y_i)}{(\Sigma X_i Z_i)^2 - (\Sigma X_i^2)(\Sigma Z_i^2)}$$

Autre exemple :

On impose aux valeurs calculées de satisfaire une contrainte $\sum_{j=1}^{P} a_j X_{ji} + a_o \geq 0$ quel que soit i ; on peut alors utiliser une méthode itérative pour estimer les coefficients tout en respectant la condition des moindres carrés; on utilise alors des techniques de programmation linéaire [7] .

Ces méthodes sont généralement lourdes et nécessitent des temps de calcul importants sur ordinateur, d'où un prix de revient élevé. Il pourra parfois être plus astucieux et aussi performant, d'effectuer des approximations ou transformations simples sur les variables.

5.4.9 - Les corrélations factices

Cela peut être le cas lorsqu'on calcule une corrélation entre des rapports de variables, par exemple entre $Y_1 = \frac{X_1}{X_3}$ et $Y_2 = \frac{X_2}{X_3}$, X_1, X_2, X_3 étant des variables pour lesquelles on dispose de n observations; si, en particulier, leur variabilité est faible et égale (coefficients de variation $\frac{\sigma}{m}$ égaux) on trouve une corrélation de 0.50 entre Y_1 et Y_2 bien que X_1 et X_2 soient indépendants [8].

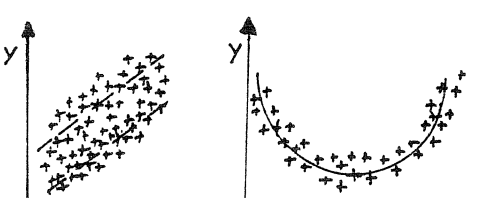
De même, bien que X et Y ne soient pas corrélées, le coefficient de corrélation entre X et (X + Y) est égal à $\left[1 + \left(\frac{\sigma_X}{\sigma_Y}\right)^2\right]^{-\frac{1}{2}}$

5.4.10 - Les pièges de la corrélation

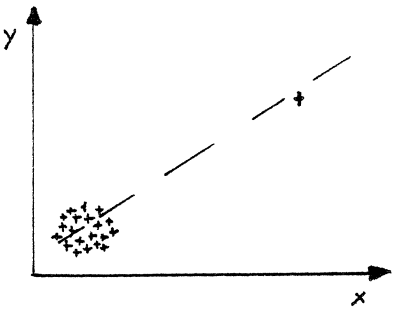
a) - coefficient de corrélation nul :

. la liaison est du type parabolique,

. il y a hétérogénéité dans la série, les échantillons partiels étant for-



b) - le coefficient de corrélation est voisin de 1 ; du fait d'une densité très dissymétrique, la ou les valeurs élevées conditionnent les valeurs des coefficients a_j.



Exemples : Oued Nahal Ayalon

5.5 - Choix des variables explicatives

Pour déterminer la "meilleure" corrélation multiple entre un prévisande (ou variable principale) et p préviseurs (ou variables explicatives) choisis parmi m variables, il existe deux méthodes : la régression ascendante ou progressive et la régression descendante ou régressive.

La méthode ascendante englobe différents procédés (Stepwise, Stagewise ...), sous le vocable régression filtrante (screening regression) très en faveur outre Atlantique, le principe consistant à établir des équations de régression successives dans lesquelles le nombre des préviseurs croît de 1 à p. Une bonne partie de ces procédés comporte un biais, par exemple supposons que l'on ait retenu 2 variables explicatives ($Y = a X_1 + b X_3 + c + \varepsilon$), le principe consiste à orthogonaliser chacun des m-2 préviseurs restant avec X_1 et X_3 ($X_K = \alpha_K X_1 + \beta_K X_3 + \gamma_K + \xi_K$) on retient comme nouvelle variable celle qui correspond à la plus forte corrélation entre Y et ξ_K, on n'utilise pas réellement les propriétés d'orthogonalité de la corrélation partielle.

La méthode descendante consiste à mettre en corrélation les m préviseurs et la variable principale Y, puis à éliminer les variables non significatives en testant le T de Student ou le coefficient de corrélation partielle. En fait, la méthode que nous utilisons comporte 2 seuils de signification (ou 2 valeurs de T, $T_1 < T_2$). Après la première corrélation globale on élimine toutes les variables dont le T est inférieur à T_1, puis on calcule de nouvelles équations en éliminant une par une les variables dont le T de student $T_1 \leqslant T \leqslant T_2$, la procédure s'arrête lorsque les p variables ont des $T > T_2$.

Mais il serait vain de se fier entièrement à une sélection automatique. On tient compte de la continuité dans le temps des équations de régression, et de la cohérence spatiale des variables sélectionnées par comparaison avec des bassins voisins.

La stabilité et la signification des coefficients de régression partielle sont d'autant mieux assurées que le coefficient de corrélation multiple est voisin de 1. Ce phénomène est bien illustré par le tableau 3 (Cère à Cantalès) où l'on compare une corrélation multiple avec ou sans termes prévisionnels.

A ce propos, lorsqu'on introduit les précipitations de la période prévisionnelle, on a intérêt à utiliser un découpage fin (mois) car l'effet de la pluie varie avec sa répartition dans le temps. Mais il y a une limite car, pourquoi ne pas rechercher une liaison de la forme :

$$E = BI + \int_{t_1}^{t_2} C(t)\, P(t)\, dt$$

Comme mise en garde, nous proposons le tableau suivant, extrait de l'article de Fisher "On the Influence of Rainfall ...", et montrant les dangers d'augmenter abusivement le nombre des variables explicatives relativement au nombre d'observations.

Avec un échantillon de 13 observations, tel que le coefficient de corrélation multiple réel entre une variable dépendante et K variables explicatives soit nul, Fisher a calculé la probabilité d'obtenir un R dépassant .5, .7, .9 :

K	R = .5	.7	.9
4	.633	.200	.0055
6	.897	.519	.0505
8	.984	.825	.2424

Un dernier écueil à éviter est l'utilisation de variables explicatives très corrélées, cas de colinéarité. Ainsi, dans le cas d'une corrélation double, on imagine aisément que le plan de régression risque de pivoter autour d'une droite, au gré des échantillons (cf. interprétation géométrique).

Utilisation de l'équation de prévision

Une façon commode d'utiliser la relation :

$$\sqrt[2]{E^6} = I + P + 2.40 \quad (\varepsilon \text{ est négligeable devant } P)$$

est de la représenter graphiquement, en portant en abscisse l'index de l'état initial I, et en ordonnée l'écoulement en valeur naturelle E_4^6, on trace différentes intersections à P constant; plus exactement ces courbes sont graduées en probabilité de ne pas dépasser P. On a en effet une estimation bien étayée et robuste de la distribution des précipitations par la fonction gamma incomplète.

La figure 7 restitue bien la non linéarité de la relation entre E_4^6 et P, et surtout met en évidence l'influence dissymétrique des termes prévisionnels sur l'écoulement de printemps.

Connaissant l'état initial I du bassin versant, par exemple I = 6.6 au 1er avril 1967, la prévision est lue en traçant la demi-droite d'abscisse 6.6.

Conventionnellement on définit la prévision par l'intervalle (fourchette) entre les quantiles 10 % et 90 %. Ainsi, sur un grand nombre de prévisions, on a 80 % des réalisations à l'intérieur de la fourchette.

Enfin dernière remarque, en utilisant cette méthode de prévision on admet implicitement que la variance du résidu est négligeable devant la variance de P :

$$\sigma_p^2 = 8.2$$

$$\sigma_\varepsilon^2 = .33$$

Il paraît plus justifié d'associer ce résidu à l'imparfaite description de l'état initial plutôt qu'aux termes prévisionnels, comme en témoignent les coefficients de corrélation partielle entre chacun des 2 facteurs et l'écoulement.

Dans l'exemple de la Cère, la largeur de la fourchette est due à l'influence importante des termes prévisionnels. Pour les bassins à régime nival, ou l'état initial, caractérisé par le stock neigeux, est prépondérant, la fourchette est plus réduite, exemple du Drac (figure 8).

Un moyen simple et évocateur, permettant d'apprécier l'information apportée par une prévision, est de tracer sur un graphique arithmétique gradué en écoulement et en fréquence cumulée, la distribution liée aux conditions initiales d'une année particulière (figure 14).

L'information fournie par la prévision est alors caractérisée par la "pente" de la courbe en S (figure 13), porportionnelle à l'écart type lié S_L. Ainsi la gamme possible des prévisions se situe entre $S_L = S_1$ - pas d'influence de l'état initial, les courbes graduées en probabilité de P sont alors parallèles à l'axe des I sur les figures 7 et 8 - et $S_L = 0$ - prédiction, les courbes P sont confondues en une seule.

Conclusion

Le choix de l'équation de régression multiple finale doit être mûrement réfléchi. On ne retiendra pas nécessairement celle qui conduit à la variance résiduelle minimale, cet optimum relatif est d'ailleurs obtenu sur un échantillon limité, parfois il vaut mieux utiliser une relation robuste avec un plus petit nombre de facteurs dont l'influence est solidement établie. Dans l'application en prévision, il se peut que certains écarts soient dus à une variable latente (à distribution très dissymétrique en général) et dont on n'a pu discriminer l'influence dans le bruit ε, sur l'échantillon d'ajustement.

L'usage des corrélations multiples est le seul critère numérique objectif qui permette de faire de la prévision opérationnelle car une prévision est définie par une graduation en probabilité de l'écart réalisation-calcul.

IV-C

LA CERE A CANTALES — TABLEAU I

0
REMULOB 3 DECEMBRE 1967

NB. VAR. ET OBS.
22 32

	1	2	3	4	5	6	7	8	9
M	149.0	121.7	103.8	104.1	120.9	102.4	85.3	111.2	112.8
S	80.1	82.3	68.3	56.1	47.4	46.0	51.2	62.8	52.6

	10	11	12	13	14	15	16	17	18
M	112.4	147.8	163.0	861.1	478.6	411.8	423.2	156.7	11.35
S	67.0	77.6	90.6	269.6	145.2	115.7	171.2	46.9	1.75

	19	20	21	22
M	14.89	16.62	16.03	13.79
S	1.34	1.82	1.38	1.63

NOMBRE DE VARIABLES UTILISEES:
16
ORDRE DE CES 16 VARIABLES
13 10 11 12 1 2 3 4 5 6 7 8 9 17 18 22

	RP	B	T	BR
13				
10	.7665	1.1305	4.8	.28
11	.8668	1.3830	7.0	.40
12	.8524	.9740	6.5	.33
1	.8459	1.0691	6.3	.32
2	.8212	.9984	5.8	.30
3	.8715	1.5021	7.1	.38
4	.6420	.8386	3.3	.17
5	.6228	.9725	3.2	.17
6	.3308	.5300	1.4	.09
7	.4883	.7353	2.2	.14
8	.4235	.3835	1.9	.09
9	.3986	.5317	1.7	.10
17	.7092	1.1778	4.0	.20
18	-.2543	-7.5647	-1.1	-.05
22	.5345	23.1633	2.5	.14
	A=	-908.5255		

R2= .9546 R= .9771 SL= 57.4307

Ecoulement : OCT-SEP
Précipitations : OCTOBRE
" : NOVEMBRE
" : DECEMBRE
" : JANVIER
" : FEVRIER
" : MARS
" : AVRIL
" : MAI
" : JUIN
" : JUILLET
" : AOUT
" : SEPTEMBRE
Logarithme des 7 derniers jours SEP
Température moyenne de MAI
" " de SEPTEMBRE

V 57

TABLEAU 2

LA CERE à CANTALES

CLE 1	CLE 2	CLE 3	
1089.5063	1046.0000	-43.5063	1934 - 1935
1320.7110	1363.0000	42.2890	- 1936
1056.2307	1078.0000	21.7693	- 1937
572.2322	565.0000	-7.2322	- 1938
736.4678	721.0000	-15.4678	- 1939
1008.5935	1106.0000	97.4065	- 1940
1233.7810	1245.0000	11.2190	- 1941
582.6935	577.0000	-5.6935	- 1942
508.2782	474.0000	-34.2782	- 1943
592.8892	620.0000	27.1108	- 1944
1010.7747	1127.0000	116.2253	- 1945
492.7150	549.0000	56.2850	- 1946
717.5858	702.0000	-15.5858	- 1947
695.7414	714.0000	18.2586	- 1948
193.8538	215.0000	21.1462	- 1949
709.8221	668.0000	-41.8221	- 1950
1312.7488	1326.0000	13.2512	- 1951
761.2952	729.0000	-32.2952	- 1952
1006.3353	929.0000	-77.3353	- 1953
761.9606	771.0000	9.0394	- 1954
990.1461	994.0000	3.8539	- 1955
720.1107	724.0000	3.8893	- 1956
686.2484	708.0000	21.7516	- 1957
803.3620	744.0000	-59.3620	- 1958
802.4520	814.0000	11.5480	- 1959
1117.0414	1132.0000	14.9586	- 1960
1099.8717	1044.0000	-55.8717	- 1961
976.5103	970.0000	-6.5103	- 1962
1015.6545	993.0000	-22.6545	- 1963
807.5704	779.0000	-28.5704	- 1964
851.8547	840.0000	-11.8547	- 1965
1319.9618	1288.0000	-31.9618	1965 - 1966

MX	MY	SX	SY
861.0938	861.0937	266.4685	269.6439
BX	**BY**	**R2**	**R**
.9766	1.0000	.9546	.9771
A0	**LAMBDA**	**CONF 80**	**SYX**
-.0005	.2130	73.5107	57.4303

IV-C

REMULOB 3 DECEMBRE 1967

TABLEAU 3

NB. VAR. ET OBS.
8 30

LA CERE A CANTALES

	1	2	3	4	5	6	7	8
M	13.23	216.5	494.2	102.0	99.3	122.0	103.3	114.2
S	3.84	31.3	305.3	58.9	54.0	48.6	47.5	17.8

R

1	1.0000							
2	.7012	1.0000						
3	.2433	.5294	1.0000					
4	.6258	.7920	.7243	1.0000				
5	.7369	.3715	-.1177	.0914	1.0000			
6	.7079	.2218	-.0050	.2797	.4698	1.0000		
7	.5057	.3368	.2378	.4542	.1524	.2051	1.0000	
8	-.2217	-.0612	.0869	-.0801	.0537	-.1105	-.3501	1.0000

NOMBRE DE VARIABLES UTILISEES:
4
ORDRE DE CES 4 VARIABLES
1 2 3 4

	RP	B	T	BR	
1					$\sqrt{\text{Ecoulement d'AVRIL à JUIN}}^2$
2	.4267	.0621	2.4	.50	Log des 7 derniers jours MARS
3	-.3847	-.0050	-2.1	-.39	Produit température . écoulement MARS
4	.3625	.0334	2.0	.51	Ecoulement MARS
	A=	-1.1620			

R2= .5296 R= .7277 SL= 2.6366

NOMBRE DE VARIABLES UTILISEES:
8
ORDRE DE CES 8 VARIABLES
1 2 3 4 5 6 7 8

	RP	B	T	BR	
1					$\sqrt{\text{Ecoulement d'AVRIL à JUIN}}^2$
2	.5112	.0211	2.8	.17	Log des 7 derniers jours MARS
3	-.1680	-.0005	-.8	-.04	Produit écoulement . température MARS
4	.6677	.0206	4.2	.32	Ecoulement de MARS
5	.9192	.0341	10.9	.48	Précipitation d'AVRIL
6	.8639	.0250	8.0	.32	" de MAI
7	.5943	.0106	3.5	.13	" de JUIN
8	-.6193	-.0275	-3.7	-.13	Température moyenne de MAI
	A=	2.4038			

R2= .9725 R= .9861 SL= .6380

IV-C

TABLEAU 4

REMULOB 1 DECEMBRE 1967

NB. VAR. ET OBS.
3 30

LA CERE A CANTALES

	1	2	3
M	13.23	6.42	4.39
S	3.84	1.68	2.86

R

1	1.0000		
2	.7053	1.0000	
3	.9006	.3582	1.0000

NOMBRE DE VARIABLES UTILISEES:
3
ORDRE DE CES 3 VARIABLES
1 2 3

	RP	B	T	BR
1				
2	.9430	1.0047	14.7	.44
3	.9790	.9982	24.9	.74
	A=	2.3985		

$\sqrt[2]{}$ Ecoulement d'AVRIL à JUIN
Index de l'état initial
Termes prévisionnels

R2= .9775 R= .9887 SL= .5761

CLE 1 CLE 2 CLE 7 CLE 8

EXPOSANT CHOISI POUR Y NATURELLE
1 2
EXPOSANT CHOISI POUR Z AJUSTEE
1 2
DONNEES NON TRANSFORMEES : TAPER 1 SINON 2
2

I

MX	MY	SX	SY
13.2317	13.2317	3.8038	3.8442
BX	BY	R2	R
.9791	1.0000	.9775	.9887
A0	LAMBDA	CONF 80	SYX
.0000	.1499	.7374	.5761

4

MX	MY	SX	SY
189.0636	189.3623	103.9837	104.0697
BX	BY	R2	R
.9875	.9892	.9751	.9875
A0	LAMBDA	CONF 80	SYX
2.3440	.1576	20.9990	16.4055

TABLEAU 5

LA CERE A CANTALES

	X'_1	X_1	$(X'_1)^2$	$(X_1)^2$	$X_1 - X'_1$
1936	13.9403	14.7600	194.3328	217.8576	.8197
7	18.5777	18.7300	345.1294	350.8129	.1523
8	8.1291	7.8700	66.0829	61.9369	-.2591
9	12.9774	12.9600	168.4119	167.9616	-.0174
40	18.1122	17.4600	328.0508	304.8516	-.6522
1	16.7726	18.0000	281.3204	324.0000	1.2274
2	13.4265	13.1500	180.2719	172.9225	-.2765
3	8.5262	8.8300	72.6968	77.9689	.3038
4	7.2196	7.6200	52.1222	58.0644	.4004
5	8.5827	8.3700	73.6634	70.0569	-.2127
6	12.6546	12.1200	160.1396	146.8944	-.5346
7	12.0032	12.5300	144.0762	157.0009	.5268
8	13.1625	14.4900	173.2509	209.9601	1.3275
9	8.3353	7.6200	69.4780	58.0644	-.7153
50	12.9550	13.0400	167.8322	170.0416	.0850
1	20.8697	20.4200	435.5425	416.9764	-.4497
2	11.4315	11.5300	130.6798	132.9409	.0985
3	8.8983	9.0600	79.1798	82.0836	.1617
4	14.3014	14.8300	204.5303	219.9289	.5286
5	8.2001	8.6600	67.2418	74.9956	.4599
6	11.9370	11.4000	142.4917	129.9600	-.5370
7	11.7012	11.1400	136.9170	124.0996	-.5612
8	14.9236	14.7000	222.7135	216.0900	-.2236
9	17.0393	16.6700	290.3366	277.8889	-.3693
60	9.5944	8.7700	92.0530	76.9129	-.8244
1	12.3494	12.0400	152.5078	144.9616	-.3094
2	16.0659	16.4300	258.1137	269.9449	.3641
3	18.4717	18.8900	341.2023	356.8321	.4183
4	18.6096	18.2200	346.3189	331.9684	-.3896
5	17.1820	16.6400	295.2196	276.8896	-.5420

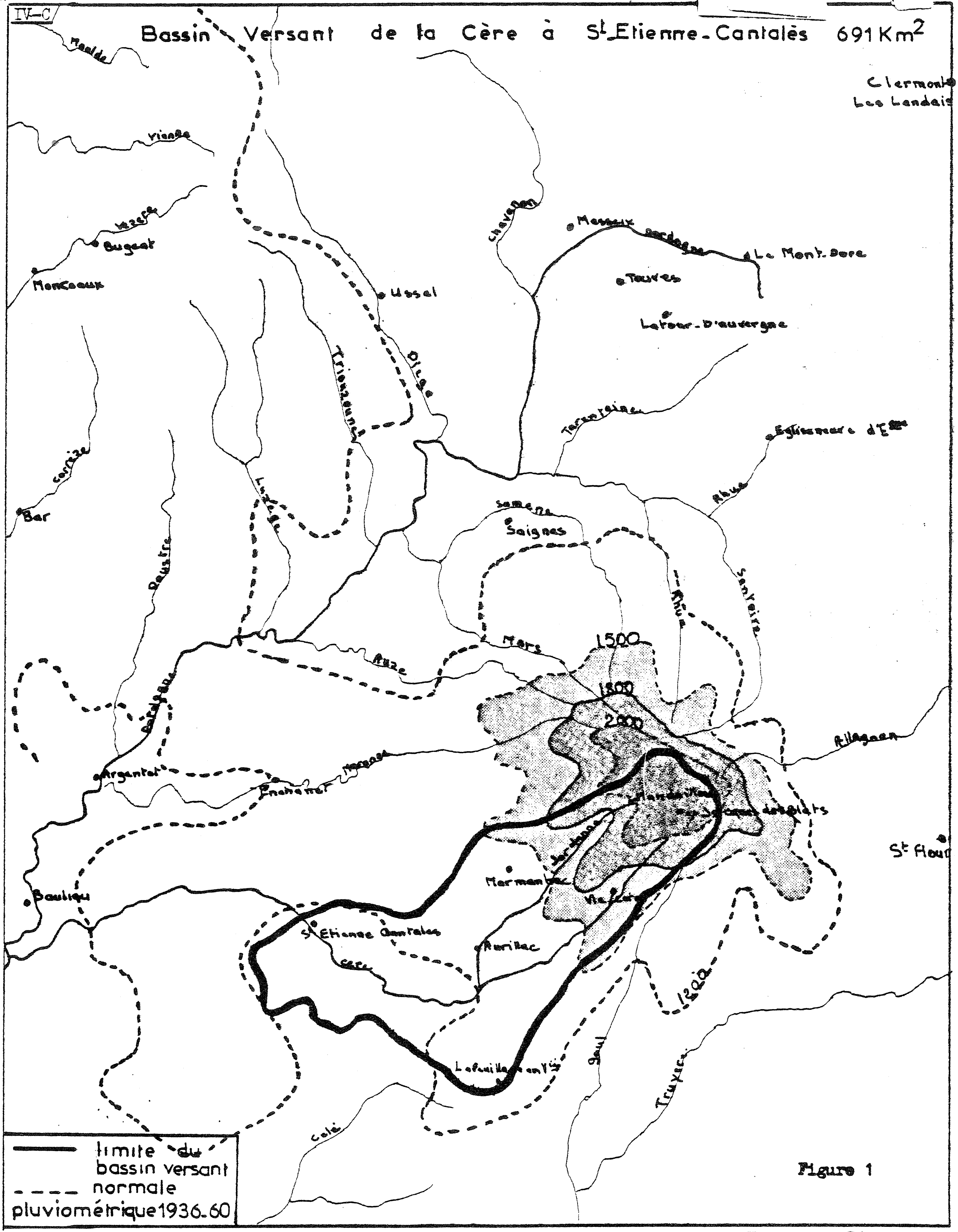

IV-C
Bassin Versant de la Cère à St Etienne-Cantalès 691 Km²
Clermont
Vienne
Ussel
Diège
Triouzoune
Bugeat
Moncoaux
Chavanon
Dordogne
Le Mont-Dore
Tauves
Latour-D'auvergne
Tarentaine
Rhue
Saignes
Mars
Auze
Santoire
1500
1800
2000
Allagnon
Argentat
Maronne
St Flour
Beaulieu
Marmanhac
Vic-s-Cère
St Etienne Cantalès
Aurillac
Cère
Goul
Truyère
Célé
limite du bassin versant
normale pluviométrique 1936-60

Figure 1

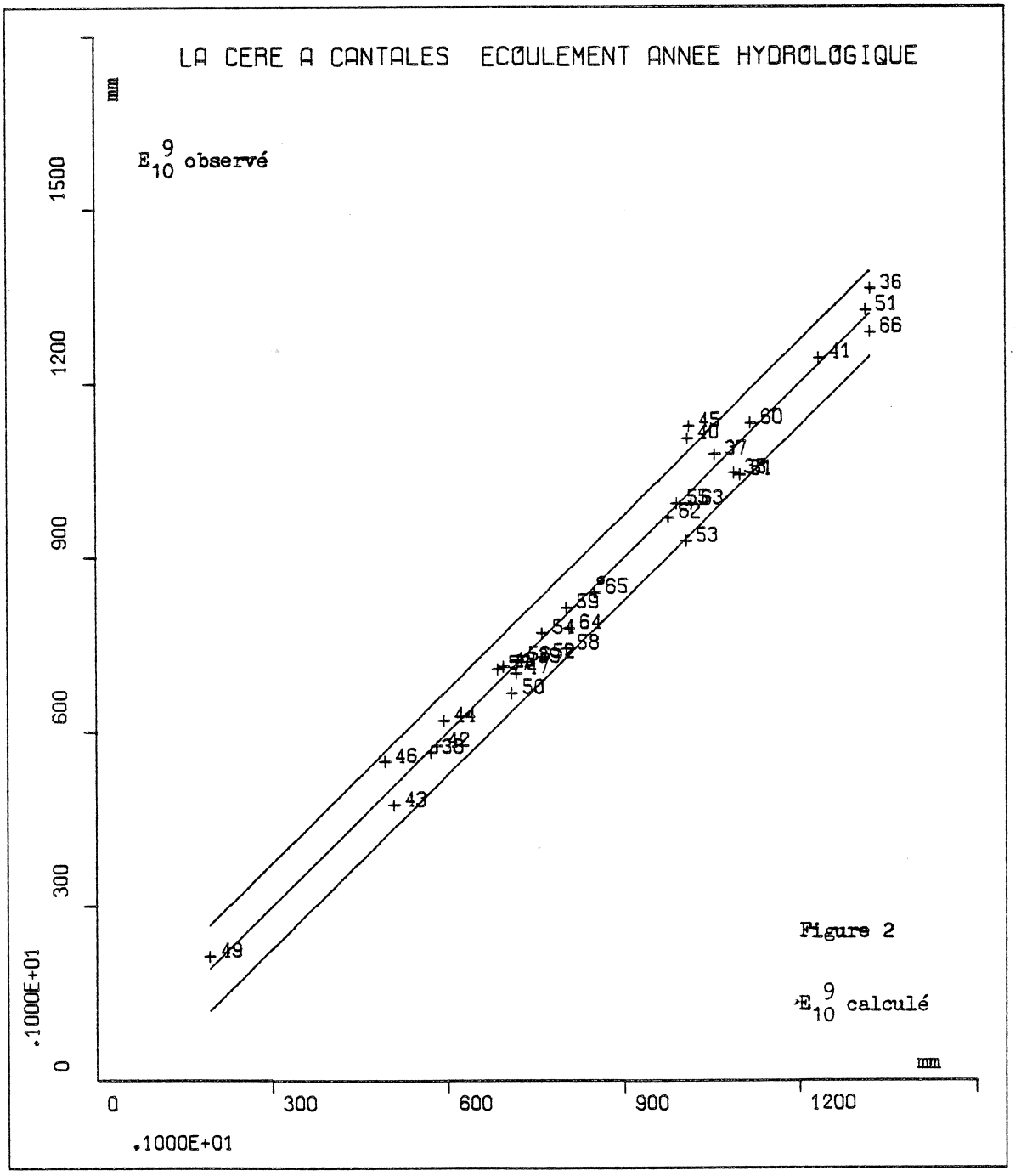

LA CERE A CANTALES ECOULEMENT ANNEE HYDROLOGIQUE
mm
E_{10}^{9} observé
1500
1200
900
600
300
0
.1000E+01
0
300
600
900
1200
mm
.1000E+01
E_{10}^{9} calculé

Figure 2

V 63

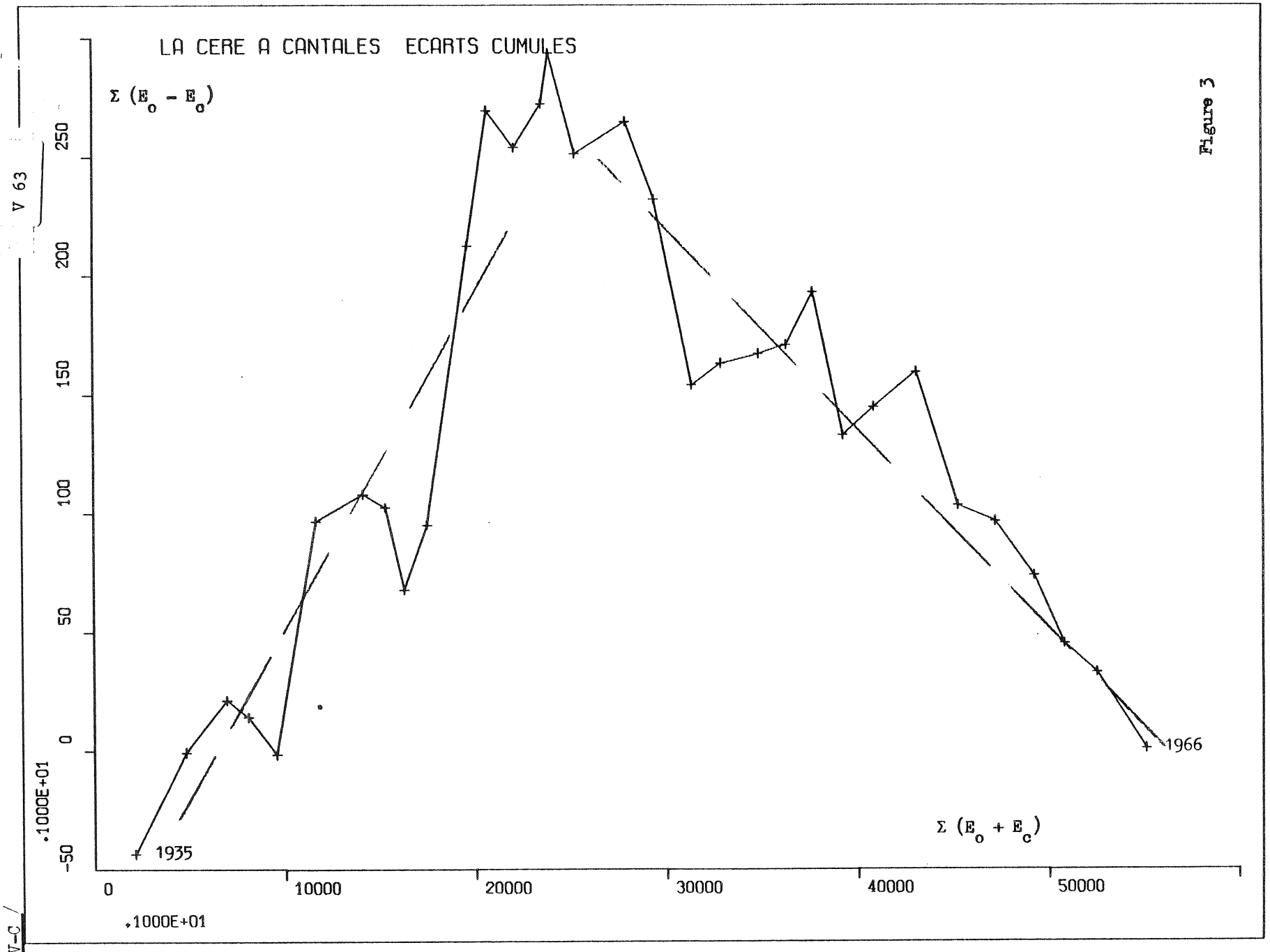


Figure 3

IV-C

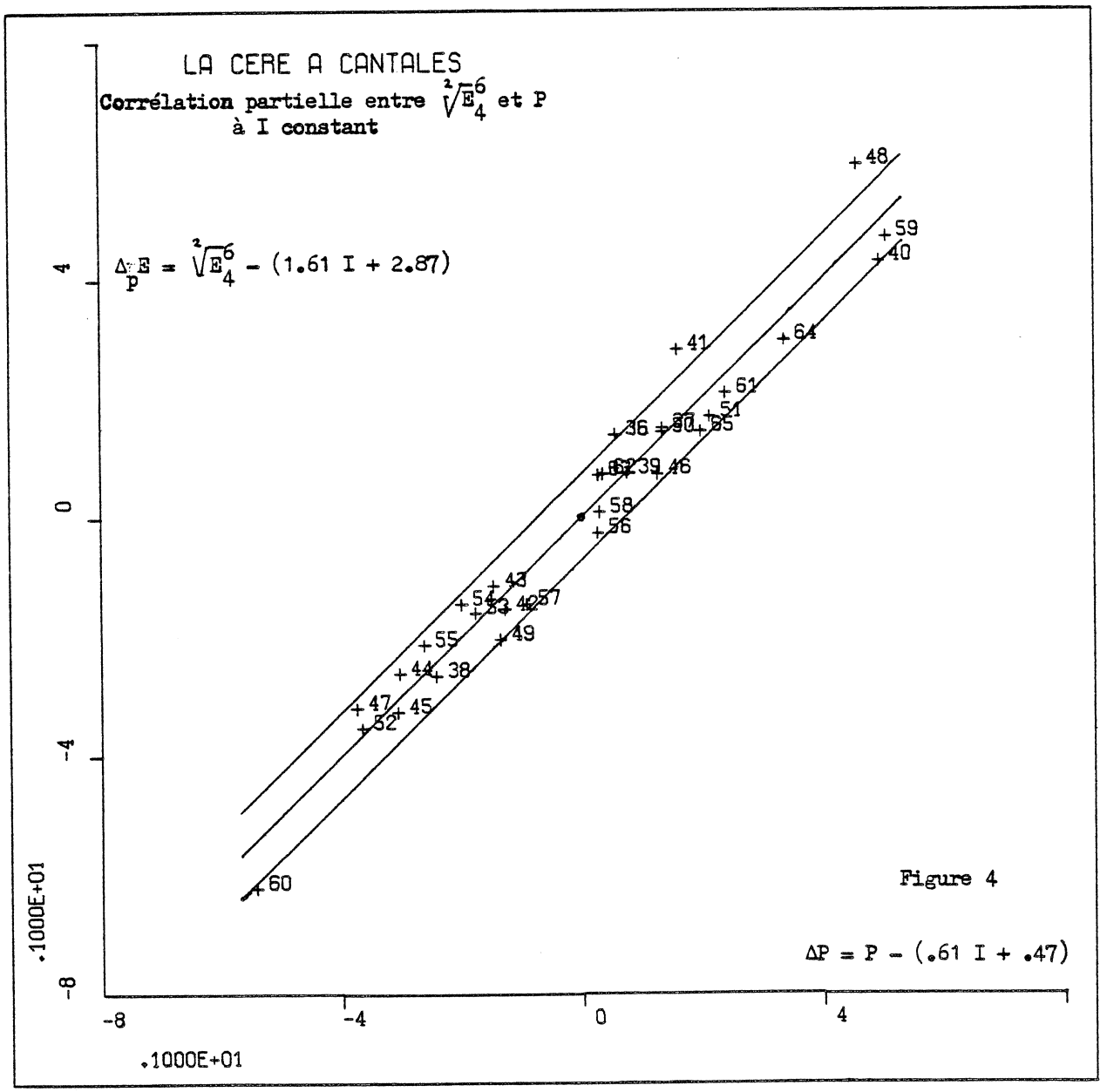

LA CERE A CANTALES
Corrélation partielle entre $\sqrt[2]{E_4^6}$ et P
à I constant
$\Delta_p E = \sqrt[2]{E_4^6} - (1.61\ I + 2.87)$
$\Delta P = P - (.61\ I + .47)$
.1000E+01
-8
-4
0
4
.1000E+01
-8
-4
0
4

Figure 4

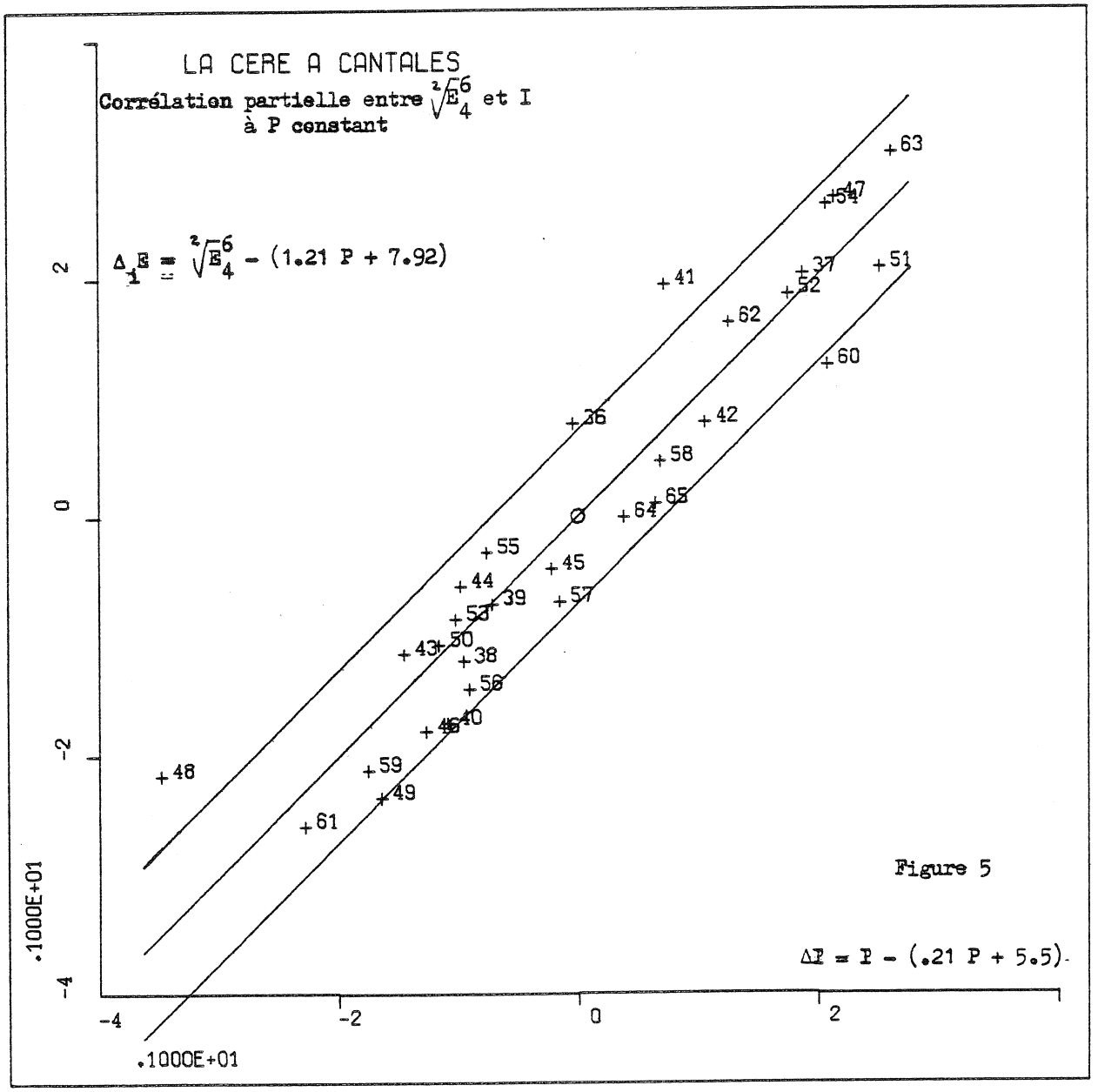

LA CERE A CANTALES
Corrélation partielle entre $\sqrt[2]{E_4^6}$ et I
à P constant
$\Delta_1 E = \sqrt[2]{E_4^6} - (1.21\ P + 7.92)$
$\Delta I = I - (.21\ P + 5.5)$
.1000E+01
-4
-2
0
2
-4
-2
0
2
.1000E+01

Figure 5

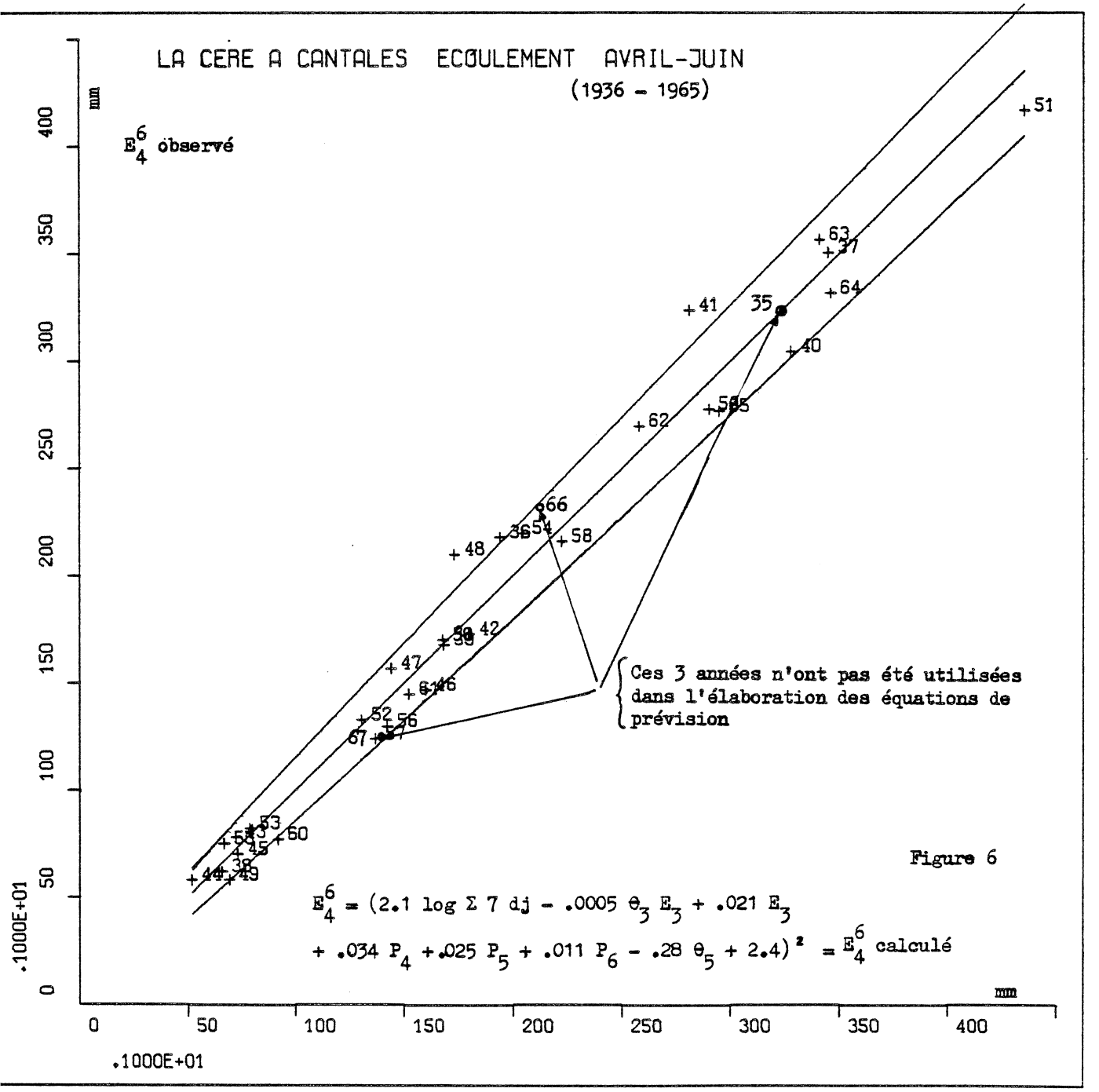

LA CERE A CANTALES ECOULEMENT AVRIL-JUIN
(1936 - 1965)
mm
E_4^6 observé
Ces 3 années n'ont pas été utilisées dans l'élaboration des équations de prévision
$E_4^6 = (2.1 \log \Sigma 7\ dj - .0005\ \theta_3\ E_3 + .021\ E_3 + .034\ P_4 + .025\ P_5 + .011\ P_6 - .28\ \theta_5 + 2.4)^2 = E_4^6$ calculé
mm
0
50
100
150
200
250
300
350
400
.1000E+01

Figure 6

V '67

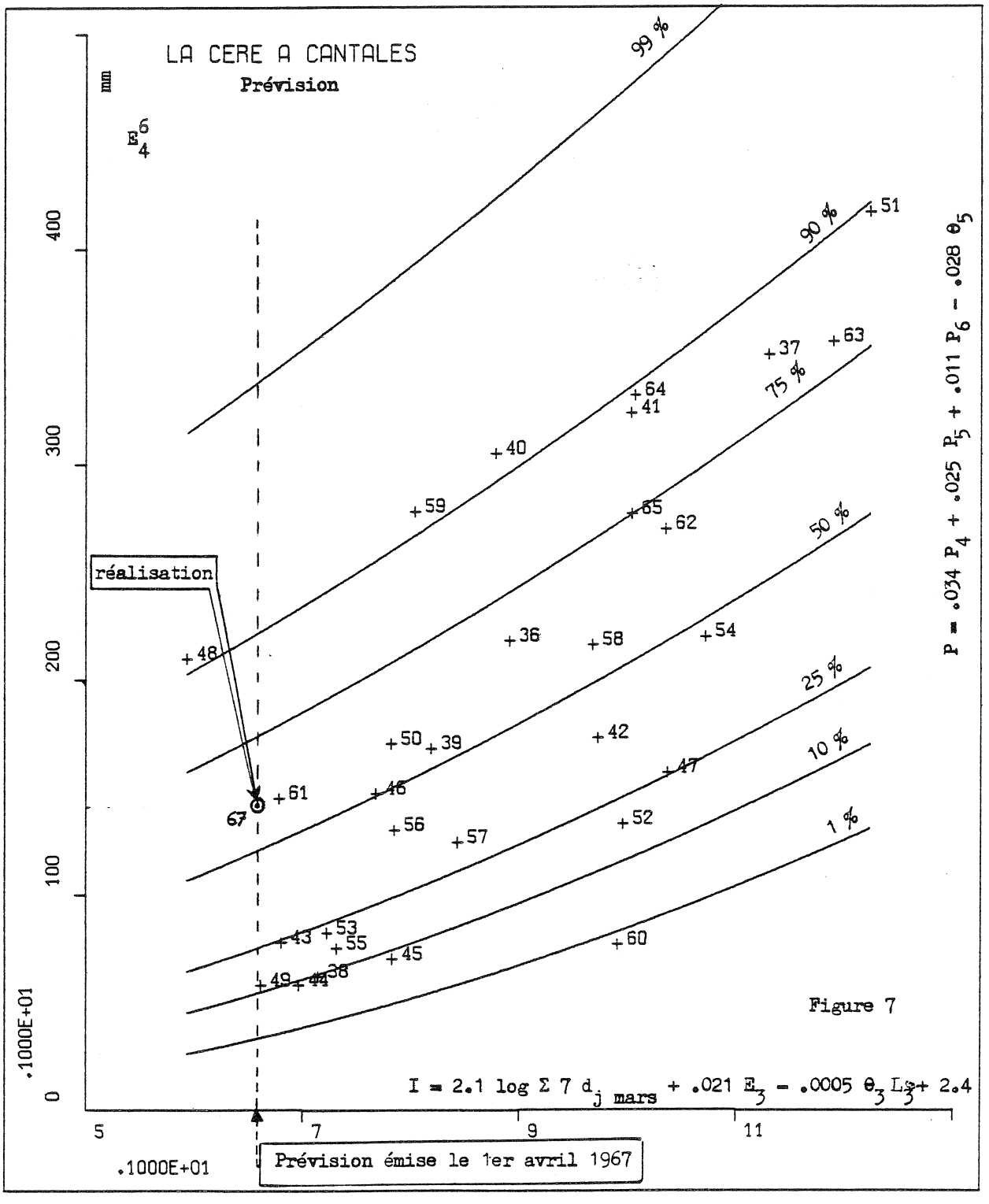

LA CERE A CANTALES
Prévision
mm
E_4^6
99 %
90 %
75 %
50 %
25 %
10 %
1 %
réalisation
+48
+59
+40
+64
+41
+51
+37
+63
+65
+62
+36
+58
+54
+50
+39
+42
+61
67
+46
+47
+56
+57
+52
+43
+53
+55
+45
+49
+44
+38
+60
400
300
200
100
0
.1000E+01
5
7
9
11
.1000E+01
Prévision émise le 1er avril 1967
$P = .034\ P_4 + .025\ P_5 + .011\ P_6 - .028\ \theta_5$
$I = 2.1 \log \Sigma 7\ d_j \text{ mars} + .021\ E_3 - .0005\ \theta_3 L_3 + 2.4$

Figure 7

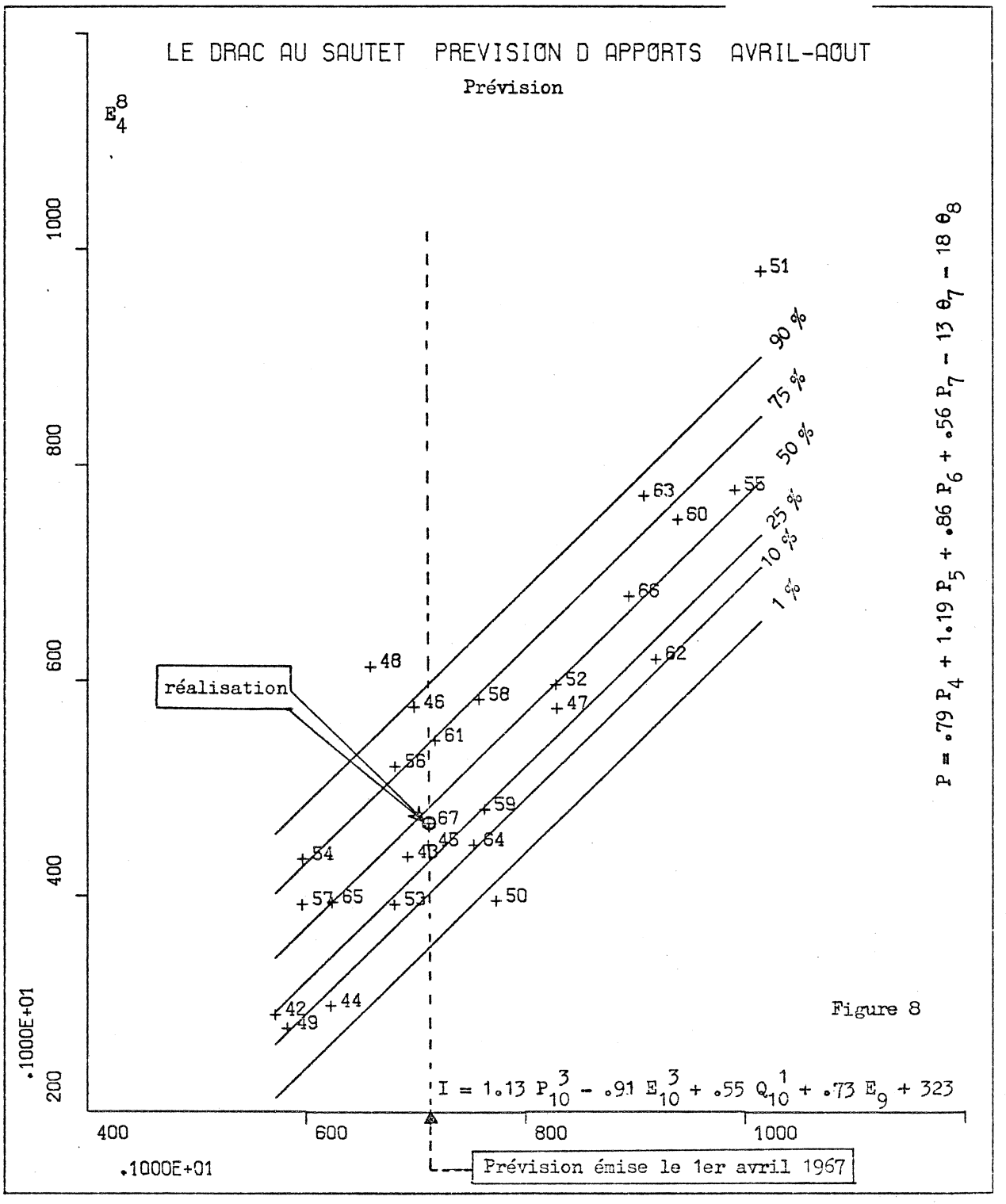

LE DRAC AU SAUTET PREVISION D APPORTS AVRIL-AOUT
Prévision
E_4^8
P = .79 P_4 + 1.19 P_5 + .86 P_6 + .56 P_7 - 13 θ_7 - 18 θ_8
90 %
75 %
50 %
25 %
10 %
1 %
réalisation
I = 1.13 P_{10}^3 - .91 E_{10}^3 + .55 Q_{10}^1 + .73 E_9 + 323
Prévision émise le 1er avril 1967
.1000E+01
200
400
600
800
1000
400
600
800
1000
.1000E+01
+51
+63
+60
+55
+66
+48
+46
+58
+52
+47
+62
+61
+56
+67
+59
+45
+64
+54
+43
+57
+65
+53
+50
+42
+44
+49

Figure 8

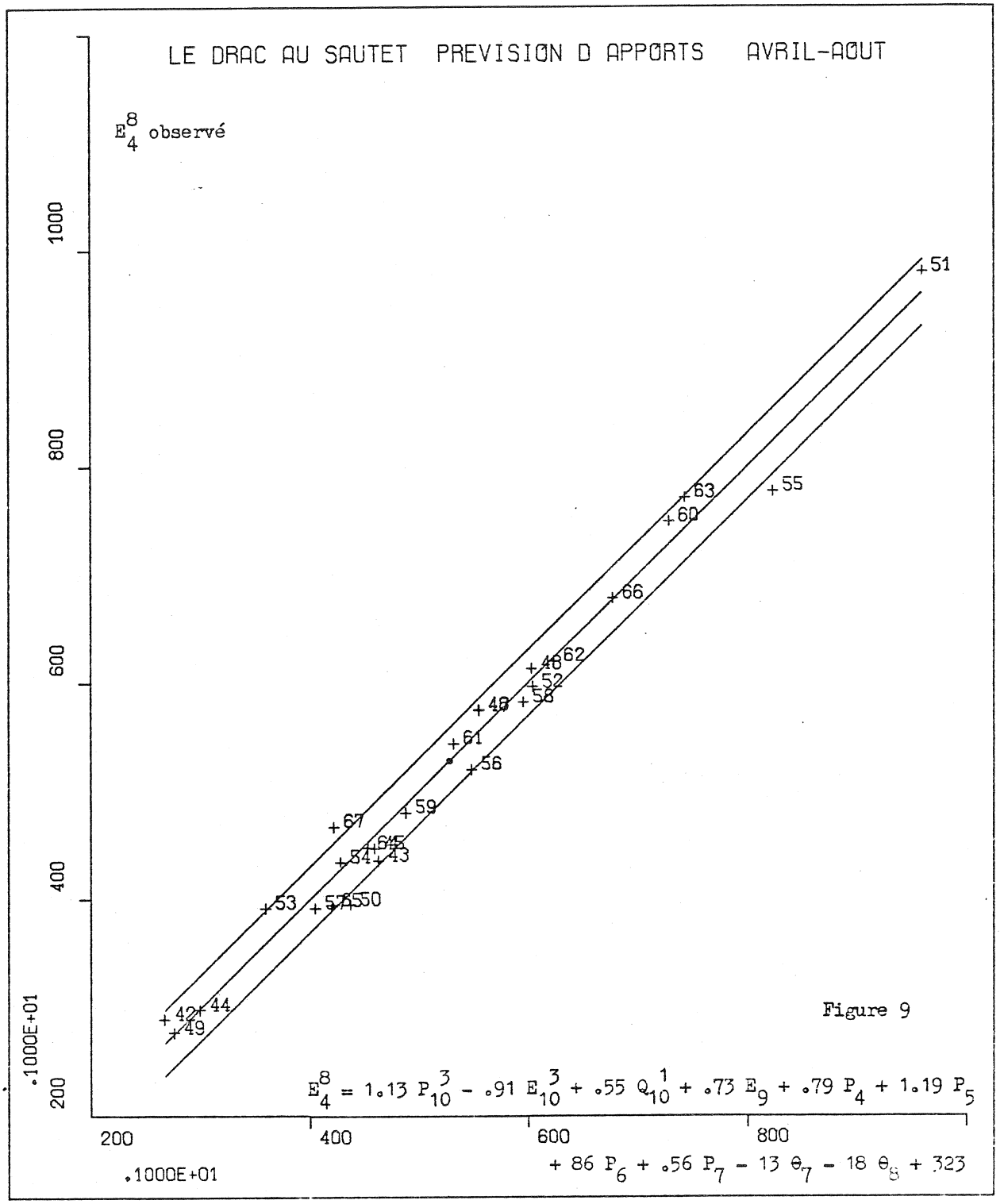
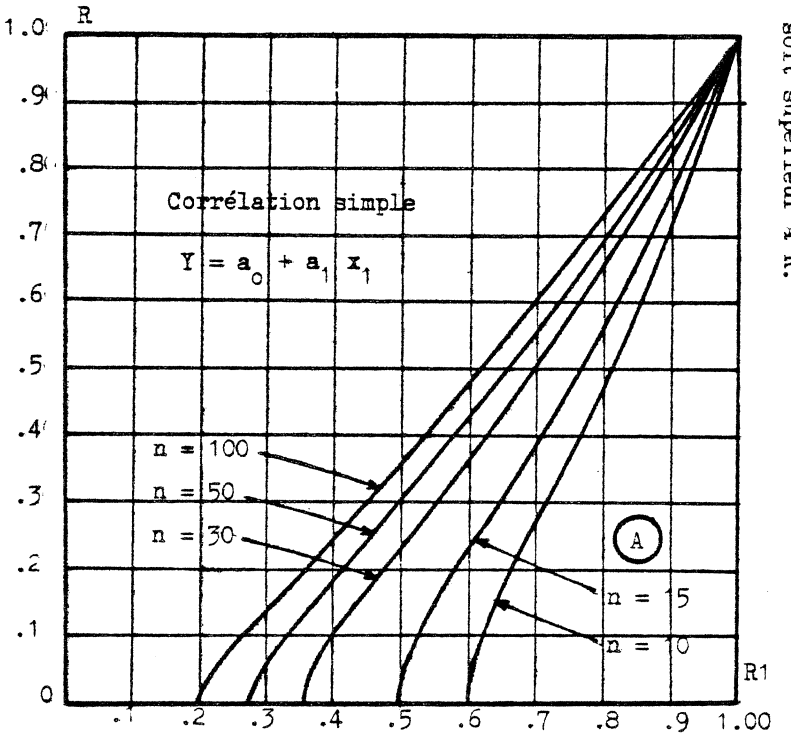

LE DRAC AU SAUTET PREVISION D APPORTS AVRIL-AOUT
E_4^8 observé
1000
800
600
400
200
.1000E+01
200
400
600
800
.1000E+01
51
63
60
55
66
46
62
52
58
48
61
56
59
67
53
54
43
42
44
49
$E_4^8 = 1.13\ P_{10}^3 - .91\ E_{10}^3 + .55\ Q_{10}^1 + .73\ E_9 + .79\ P_4 + 1.19\ P_5$
$+ 86\ P_6 + .56\ P_7 - 13\ \theta_7 - 18\ \theta_8 + 323$

Figure 9

SIGNIFICATION D'UN COEFFICIENT DE CORRELATION MULTIPLE

ABAQUES A, B, C

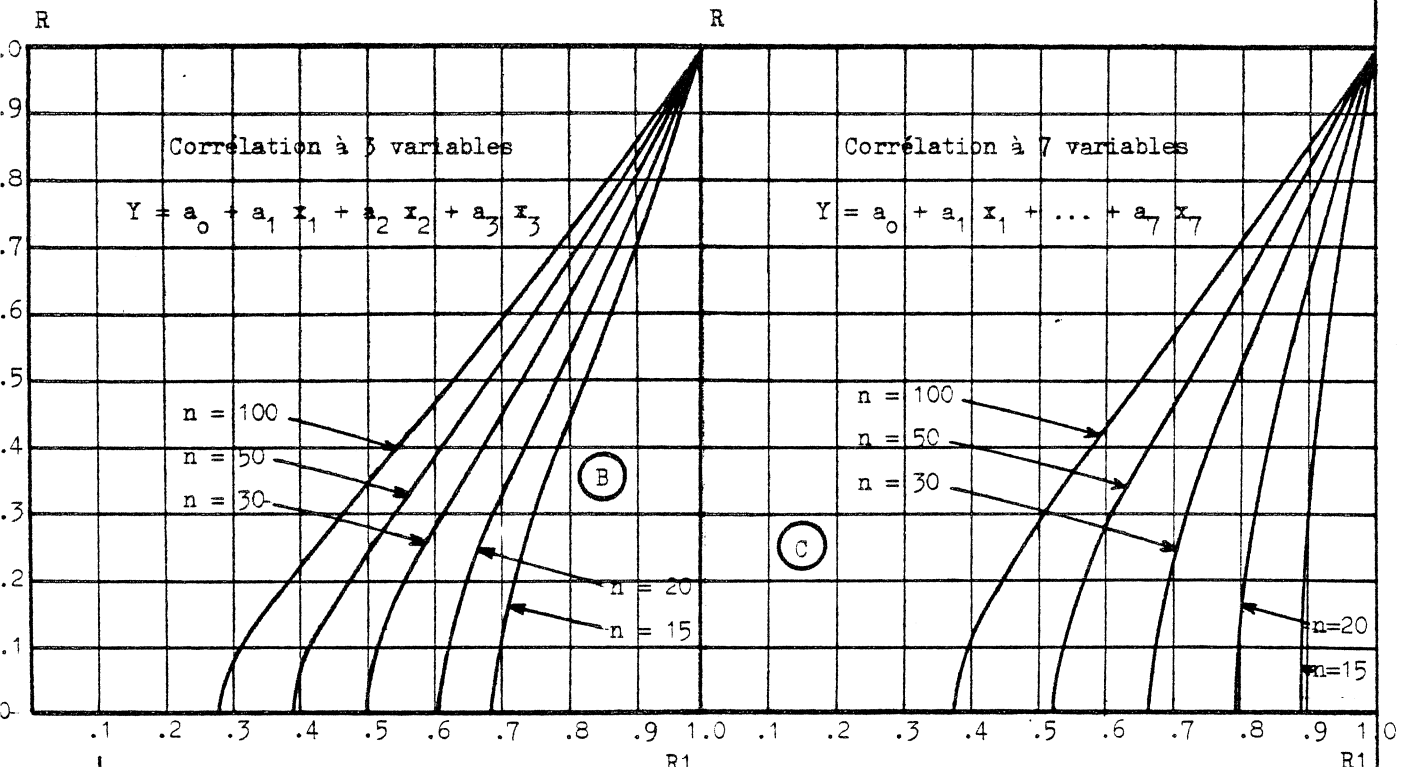


R_1 : Coefficient de corrélation multiple obtenu d'après un échantillon de n observations dans 5 % des cas.

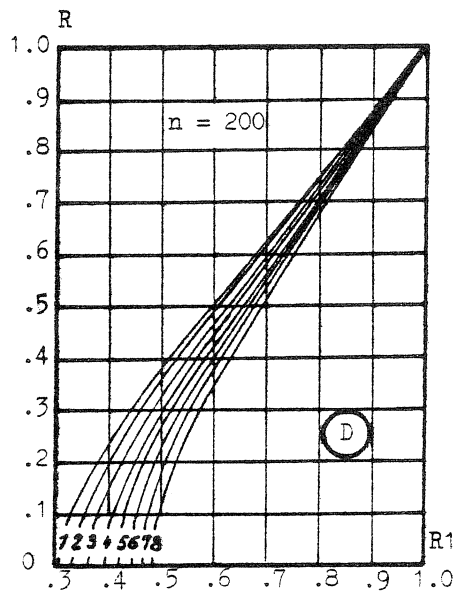
R : Coefficient de corrélation multiple théorique

Pour un n donné, et avec un coefficient calculé R_1, on a 95 chances sur 100 que le vrai coefficient soit supérieur à R.

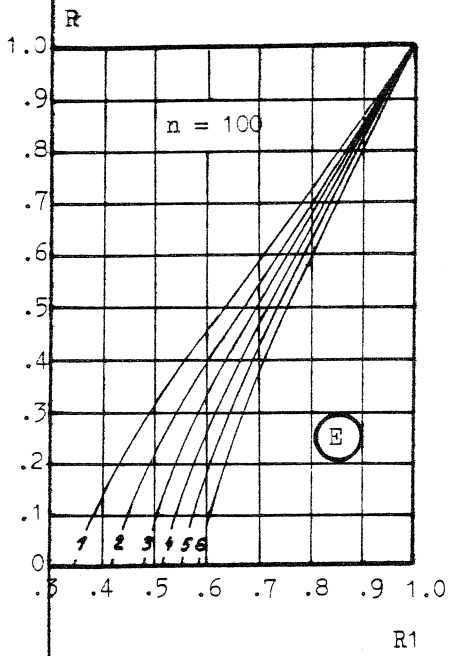
(1) - La taille de l'échantillon est variable, le nombre des variables explicatives est fixé.

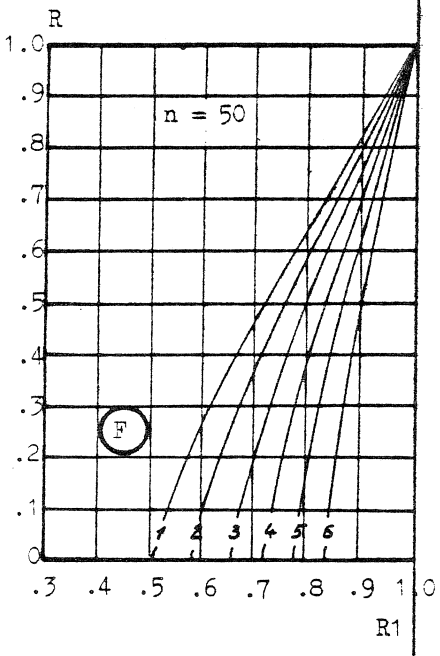


ABAQUES D, E, F



(2) - La taille de l'échantillon est fixée, le nombre de variables explicatives varie.





L'exemple de l'oued Nahal Ayalon illustre les points suivants :

- hétéroscédasticité (figure 1)
  non linéarité (figure 1)

- choix de la transformation à appliquer à E (figures 2 et 3)

- incidence de ces transformations sur les corrélations partielles (figures 6, 7 et 8) et sur la corrélation multiple (figures 9, 10 et 11)

- robustesse de la transformation en racine carrée (figure 12)

- régression avec contrainte, $\sum_{j=1}^{p} a_i X_{ij} + a_o > 0$, simplicité et économie en temps de calcul obtenues par une transformation simple sur E, par rapport à un algorithme sophistiqué et onéreux à mettre en oeuvre, pour un résultat presque trivial (il consiste à ajouter un terme positif à la constante de l'équation de régression).

Calcul des coefficients de dissymétrie (β_1) et d'aplatissement (β_2) pour les variables utilisées dans les corrélations multiples :

	P	I	S	Log_E	$E^{1/12}$	$E^{1/6}$	$E^{1/3}$	$E^{5/12}$	$E^{4/9}$	$E^{1/2}$	$E^{7/12}$	$E^{3/4}$	E
$\sqrt{\beta_1}$	-.12	-.10	.06	-.57	-2.29	-1.65	-.24	.22	.37	.60	.91	1.40	1.97
β_1	.01	.01	0	.33	5.22	2.73	.06	.05	.13	.36	.84	1.96	3.88
β_2	2.05	1.47	2.03	2.56	6.67	5.28	2.89	2.81	2.87	3.05	4.45	4.45	6.12

(rappelons que pour une variable normale $\beta_1 = 0$ et $\beta_2 = 3$)

Référence: Thèse de Milu Rosenberg,Hydrologie d'Israel,Grenoble 1970

Exemple : Apports annuels de l'oued NAHAL AYALON
(années hydrologiques 1938-39 à 1959-60)

Année	n°	P	I	S	
1938-39	1	795	.404	6	66.0
40	2	530	.491	8	15.8
41	3	566	.370	7	13.3
42	4	587	.488	9	13.6
43	5	848	.300	10	33.0
44	6	480	.486	8	4.5
45	7	803	.336	12	21.0
46	8	595	.377	7	12.0
47	9	357	.456	6	.5
48	10	571	.364	7	4.7
49	11	816	.305	12	23.0
50	12	663	.479	7	30.0
51	13	310	.362	4	0
52	14	807	.496	6	84.0
53	15	610	.405	10	5.6
54	16	620	.354	9	9.1
55	17	432	.486	4	13.0
56	18	830	.364	10	37.5
57	19	615	.309	10	2.7
58	20	505	.472	5	14.0
59	21	448	.468	9	2.3
1959-60	22	270	.411	4	0

P = précipitation annuelle (septembre-août)

I = indice de concentration des pluies (Roche), $I = \left[\frac{1}{12 \times 11} \sum_{i=1}^{12} \left(\frac{P_i - P}{P} \right)^2 \right]^{\frac{1}{2}}$

avec P_i = pluie mensuelle du $i^{ème}$ mois et $P = \frac{1}{12} \sum P_i$

S = nombre de séquences pluvieuses

Tableau des coefficients de corrélation partielle

$(E)^q = a_o + a_1 P + a_2 I + a_3 S$

q =	1/12 .083	2/12 .166	4/12 .333	5/12 .417	4/9 .444	6/12 .500	7/12 .583	9/12 .750	12/12 1.000
P =	.7874	.898	.978	.986	.987	.987	.983	.971	.942
I =	.629	.728	.860	.882	.881	.868	.819	.712	.563
S =	.255	-.007	-.733	-.870	-.887	-.900	-.896	-.871	-.813
R =	.843	.914	.9780	.9855	.9861	.9854	.9807	.967	.936

Tableau des écarts réduits

$$\frac{(E_o)^q - (E_c)^q}{{}^o(E_o)^q}$$

(1946-47)	-.314	+.102	-.105	-.096	-.083	-.062	-.013	.072	.174
(1950-51)	-.822	-.651	-.202	-.068	-.022	+.049	.130	.246	.343
(1959-60)	.097	-.019	.029	.072	.090	.123	.172	.270	.280
(1960-61)	-.972	-.755	-.201	-.031	.022	.106	.207	.334	.435

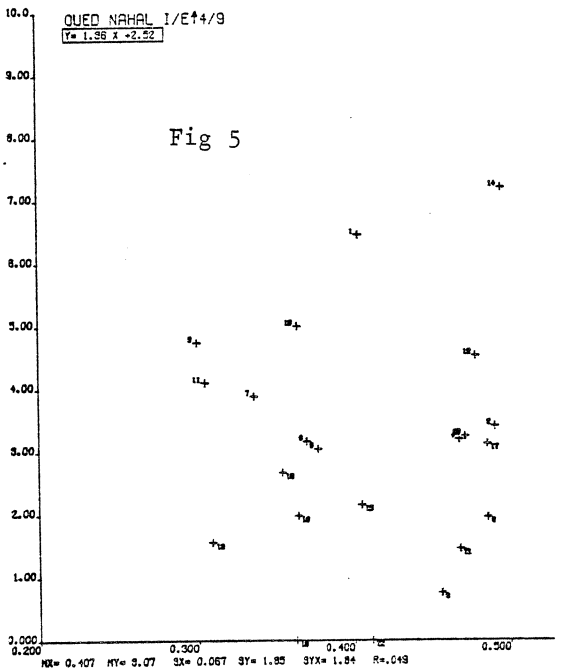


Fig 5

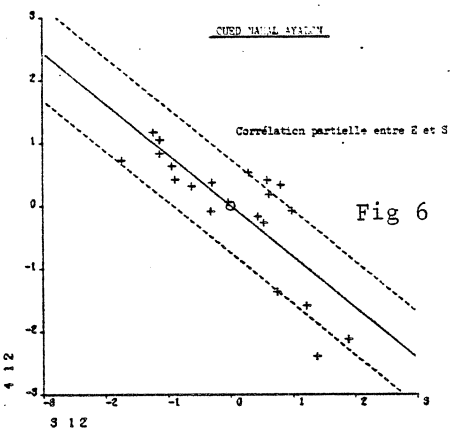
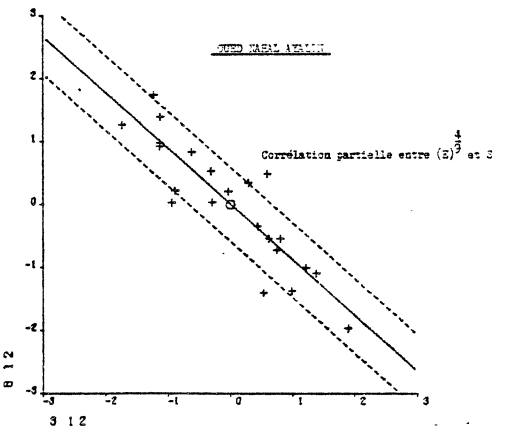


Fig 6



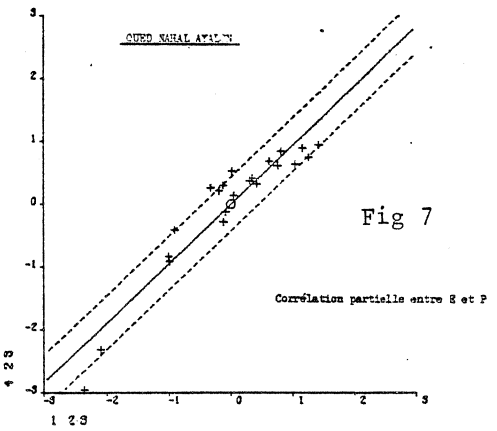
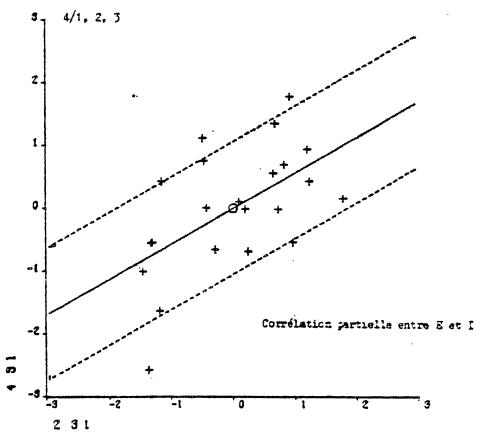


Fig 7



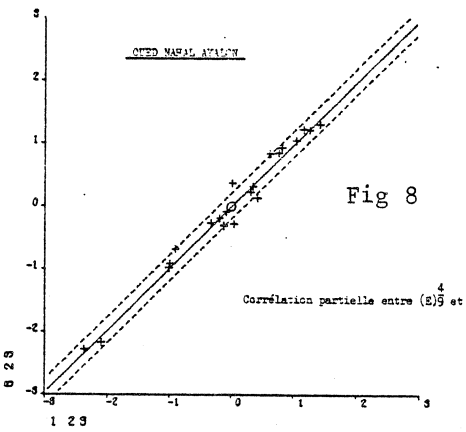
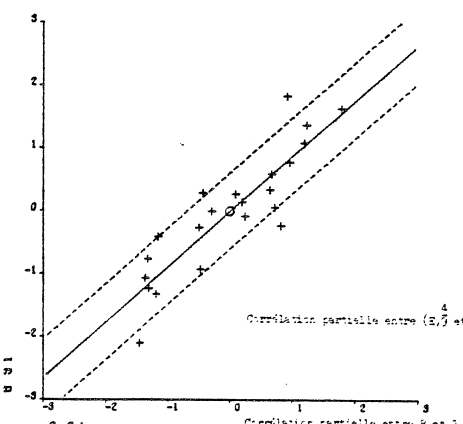


Fig 8



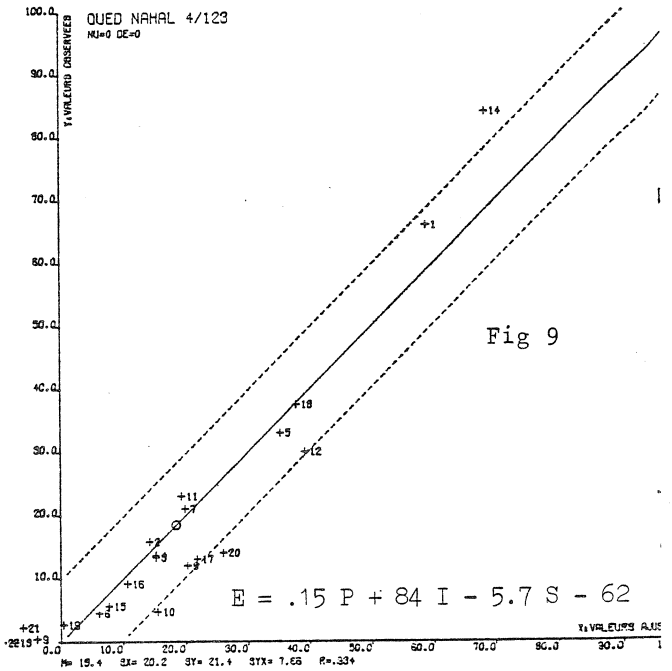


Fig 9

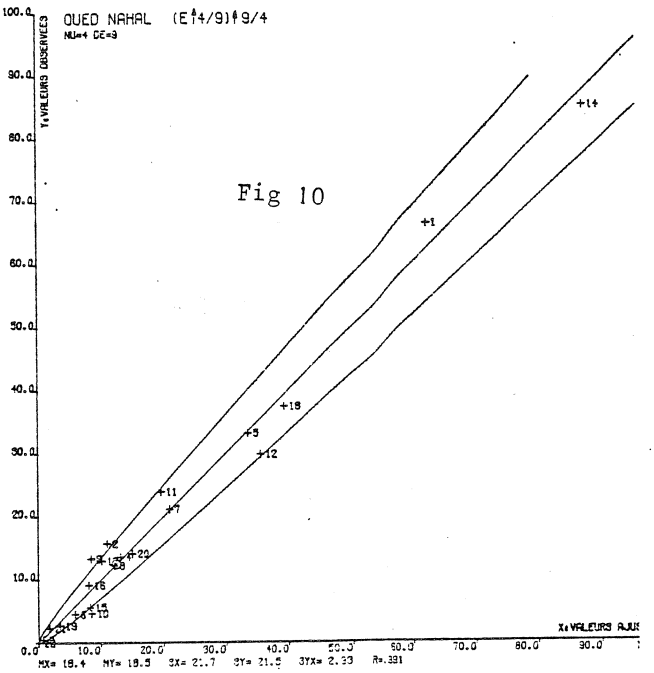


Fig 10

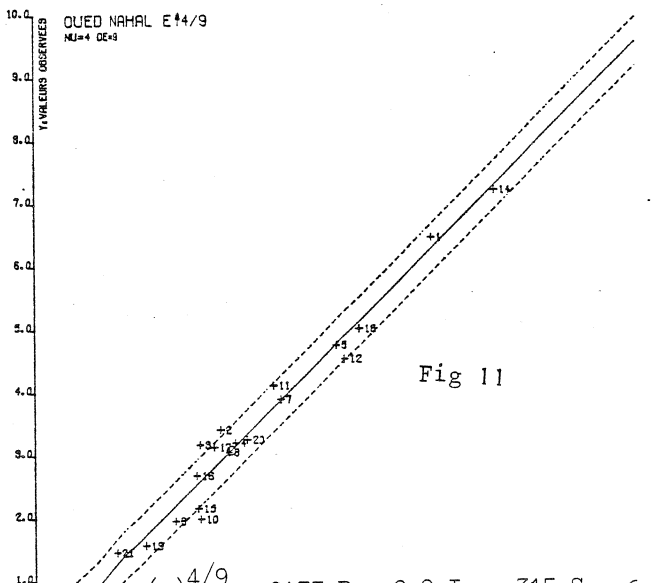


Fig 11

Cette technique d'analyse n'est cependant pas un "fourre-tout" et nous avons essayé de montrer sur des exemples le rôle de l'utilisateur, qui cherche à maintenir en permanence un équilibre entre les rigueurs théoriques de la statistique mathématique et les réalités physiques et pratiques, attitude raisonnée qui n'est pas susceptible d'un traitement automatique.

5.6 - L'autocorrélation ou corrélation en chaîne

Lors du premier exposé, nous avons vu comment utiliser des séries statistiques de débits moyens journaliers pour décrire la répartition du débit selon l'époque de l'année en un point d'une rivière (Loire, Drac, Romanche). Prenons l'exemple de la Cère à St-Etienne-Cantalès pour laquelle on disposait de 30 ans d'observations journalières. A partir de la distribution empirique des débits du 5, 15 et 25 de chacun des mois de l'année, on a graphiqué les valeurs des quantiles 10 %, 50 %, 75 %, 90 % puis un lissage a été effectué pour obtenir des "courbes quantiliques" régulières tout au long de l'année.

Ce faisant on a admis implicitement qu'il n'y avait pas de tendance à long terme, c'est-à-dire qu'à une date donnée, dans dix ans par exemple, on aurait autant de chances de ne pas dépasser tel débit. Cette tendance à long terme aurait pu être due à une modification de climat par exemple (on sait qu'un tel phénomène n'est pas perceptible à l'échelle du siècle mais plutôt du millénaire).

Comme exemple typique de phénomène ayant une tendance, on peut rappeler la "loi" du doublement tous les dix ans de la consommation d'électricité en France, il ne s'agit en fait que d'une relation moyenne.

En étudiant le graphique du régime des débits de la Cère à St-Etienne-Cantalès, on constate que les courbes quantiliques ont une variation saisonnière, d'une part leur position sur le graphique "date-débit" évolue suivant les mois et que, d'autre part, la situation relative des courbes est également variable suivant l'époque.

Il y a cependant des périodes telles que décembre et janvier, février et mars, juillet et août par exemple, où ces courbes sont relativement stables, c'est-à-dire que la distribution empirique du débit journalier est sensiblement identique quelle que soit la date, à l'intérieur de l'une de ces périodes : le phénomène est stationnaire. Cela ne signifie pas que les débits journaliers d'une saison durant une année particulière, 1966 par exemple, se maintiendront au niveau d'une même courbe quantilique; car si le débit d'un jour peut alors être considéré comme indépendant de la saison, il sera soumis à deux influences principales, d'une part son inertie qui intègre le comportement du bassin versant la veille et les jours antérieurs, et d'autre part les évènements météorologiques qui surviennent le jour même ou la veille (mais dont les effets sont différes de n h), cette corrélation étant supposée stable pendant la saison étudiée.

Donc si l'on veut disposer d'une information exhaustive, il faut compléter le graphique descriptif du régime des débits par les fonctions d'autocorrélation propres à chaque saison.

Avant d'illustrer ces généralités par un exemple concret, nous essaierons de leur donner un contenu plus formel.

5.6.1 - Analyse des séries chronologiques -

Pour différentes raisons : séries disponibles, lissages d'erreurs de mesures, on étudie généralement le débit moyen journalier, qui est une moyenne de valeurs instantanées; ce faisant, on remplace un phénomène continu par un phénomène discret; c'est aussi le cas des températures par exemple.

Le tableau de ces valeurs, en fonction du temps, constitue une chronique, dont on veut étudier les propriétés statistiques, car le débit à l'exutoire d'un bassin est la résultante de différentes causes dont les effets sont complexes et interférents et lorsqu'on veut le prévoir, on ne peut envisager une approche déterministe.

Il serait plus exact d'étudier la fonction aléatoire Q(t) dont les valeurs sont des variables aléatoires qui constituent une suite infinie dépendant du paramètre continu t, mais en fait on ne dispose que d'une réalisation discrétisée de cette fonction, la chronique des débits journaliers Qt du 1er janvier 1936 au 31 décembre 1965 pour la Cère par exemple; grâce au théorème ergodique on verre que ceci n'est pas grave.

On définit la stationnarité au sens large du phénomène, par la constance des paramètres suivants quel que soit t :

le moment d'ordre 1 $M(t) = E\left[Q(t)\right] = m$

la covariance $C(t,\tau) = E\left[Q(t) - M(t)\right]\left[Q(\tau) - M(\tau)\right] = C(t-\tau)$
$t < \tau < t + Kt$

La covariance ne dépend que de l'intervalle de temps $(t-\tau)$ mais pas de t.

Ces propriétés générales, issues du calcul des probabilités, s'appliquent également dans la pratique où l'on ne dispose que d'une chronique, et non d'une infinité de chroniques, grâce au théorème ergodique.

Ce théorème assure que l'espérance mathématique de Q(t) et Q(t) . Q(τ) obtenue comme moyenne sur un ensemble de chroniques peut être remplacée par la moyenne temporelle des mêmes quantités sur une seule chronique.

Ayant discrétisé, la schématisation la plus générale que l'on puisse donner d'une série chronologique consiste à écrire :

$$Q_t = T_t + S_t + U_t \qquad (1)$$

Il existe également un modèle multiplicatif $Q_t = T_t \cdot S_t \cdot U_t$ que l'on rend additif par transformation logarithmique.

5.6.2 - Quelques commentaires sur le modèle général (1)

La tendance T_t se détermine en considérant les valeurs moyennes annuelles de Q_t soit $\overline{Q}_n$, et, en recherchant la relation entre $\overline{Q}_n$ et le temps n en années $(1 \leqslant n \leqslant N)$ d'après N années d'observations.

La composante saisonnière, l'influence de la tendance étant éliminée de Q_t, s'obtient en considérant les N séries de 365 valeurs journalières (si l'intervalle unitaire est la journée) ou de 52 valeurs hebdomadaires (si l'intervalle unitaire est la semaine).

La variable U_t, les effets de tendance et saisonniers éliminés, se présente sous forme d'une liaison en chaîne simple ou multiple, par exemple :

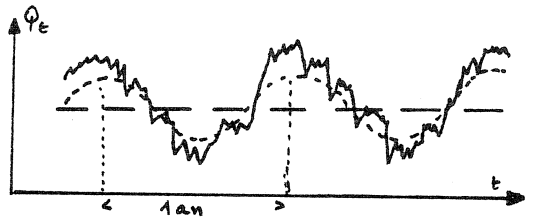
$$U_t = \sum_{i=1}^{K} \alpha_i U_{t-i} + \sum_{e=1}^{L} \beta_e V_{t-e} + \varepsilon_t$$

(V_{t-e} étant une variable explicative et ε_t l'aléa résiduel).

Nous donnons ci-après quelques exemples de tendances à long terme.

- Tendance nulle :

$$Q_t = S_t + U_t \quad ,$$



- Tendance linéaire :

$$Q_t = (\alpha t + \beta) + S_t + U_t \quad ,$$

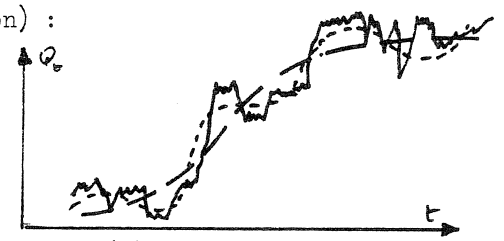
- Tendance exponentielle :

$$Q_t = \alpha e^{\beta t} + S_t + U_t$$

$$Q_t = \gamma t^{p} + S_t + U_t \quad ,$$

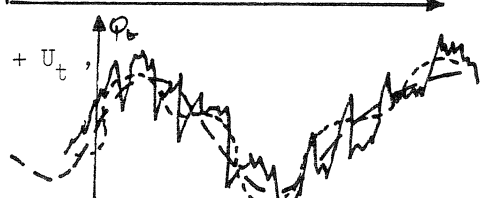
- Tendance logistique (avec saturation) :

$$Q_t = \frac{1}{1 + \alpha e^{-\beta t + \gamma}} + S_t + U_t$$



- Tendance périodique :

$$Q_t = a_o + \sum_{K=1}^{N} a_K \cos\left(\frac{2\pi K t}{T} + \varphi_K\right) + S_t + U_t \quad ,$$



Cette panoplie de modèles de tendances n'est pas limitative. Ce qui fait la difficulté d'identification du type ou de l'existence des tendances est que l'écart type de U_t est souvent supérieur au gradient $(T_N - T_o)$. (on utilise généralement les moindres carrés pour caler les modèles).

Nous avons admis que la tendance T_t n'est pas discernable à l'échelle du siècle pour les débits : $T_t \not\equiv 0$. Encore que ce ne soit pas le cas semble-t-il pour les températures de l'air.

On peut d'ailleurs proposer un modèle pour T_t qui soit en lignes brisées (voir travaux de Mandelbrot et Meija) et qui, contrairement aux modèles précédents, en particulier celui qui relève de l'analyse harmonique et décrit des "cycles", suppose que la tendance est constituée d'une suite de segments d'égale durée dont les ordonnées des intersections sont aléatoires. Ce modèle séduisant car il rendrait compte de séquences d'années sèches ou humides, sans périodicité, reste délicat à ajuster sur les séries observées.

Pour les débits, en supposant $T_t \not\equiv 0$ (non identifiable), on devra donc éliminer l'effet saisonnier pour obtenir les paramètres α_i et β_e de U_t.

L'opération consiste à ajuster une courbe sinusoïdale, aux moyennes des 365 débits journaliers (1er janvier au 31 décembre) calculées sur n années :

ou
$$\begin{cases} S_{t_j} = \sum_{K=1}^{2} \gamma_K \cos\left(\frac{2K\pi j}{365} + \varphi_K\right) + \overline{Q} & (2) \\ S_{t_j} = \sum_{K=0}^{2 \text{ ou } 3} \left(a_K \cos\frac{2K\pi j}{365} + b_K \sin\frac{2K\pi j}{365}\right) & (3) \end{cases}$$

On peut d'ailleurs effectuer la même opération pour les écarts types des 365 jours et étudier Q_t sous la forme d'une variable centrée ou centrée-réduite désaisonnalisée :

$$U_t = Q_{tj} - m_{tj} \quad \text{ou} \quad Y_t = \frac{Q_{tj} - m_{tj}}{\sigma_{tj}}$$

Or, cette opération ne suffit pas à assurer la stationnarité des débits ainsi transformés, car il faut aussi vérifier que la covariance $\Sigma Y_t Y_{t+k}$ est constante durant les 365 jours, et que la loi de probabilité du résidu ε de l'équation d'autorégression est constante durant toute l'année : les applications montrent que cette hypothèse d'invariance de la structure interne des chroniques de débits au cours de l'année n'est pas toujours vérifiée : supposons que l'on dispose de N années ($1 \leqslant i \leqslant N$) d'observations journalières. On devrait étudier l'évolution saisonnière (avec j constant, pour $1 \leqslant j \leqslant 365$) des coefficients d'autocorrélation d'ordre K ($1 \leqslant K \leqslant 10$) :

$$r_{j,K} = \left[\sum_{i=1}^{N} (Y_{i,j}) (Y_{i,j+K}) \right] \frac{1}{N}$$

ceci est possible si N est suffisamment grand ($N > 100$). Dans la pratique N est souvent inférieur à 50, et la dispersion d'échantillonnage des $r_{j,K}$ ne permet guère d'identifier leur composante saisonnière. Mais dans ce cas, le calcul généralement effectué d'après :

$$r_K = \left[\sum_{i=1}^{N} \sum_{j=1}^{365-K} (Y_{i,j}) (Y_{i,j+K}) \right] \frac{1}{N(365-K)} ,$$

qui suppose implicitement que $r_{j,K} = r_K$, et ne dépend plus de la date dans l'année, est généralement incorrect.

Pour ces raisons, il est préférable, le plus souvent, de ne pas effectuer d'analyse harmonique (dont l'appareil mathématique peut aussi faire illusion) pour éliminer la composante saisonnière, mais de limiter l'analyse aux périodes de l'année (2 à 5 mois) pour lesquelles on s'est assuré de la stationnarité à l'aide des courbes quantiliques du graphique descriptif du régime des débits, et du corrélogramme calculé par bimestre ou trimestre.

Par exemple, si l'on considère le 1er trimestre (1er janvier-31 mars), il sera plus exact de calculer :

$$r_K = \left[\sum_{i=1}^{N} \sum_{j=1}^{90-K} (Y_{i,j}) (Y_{i,j+K}) \right] \frac{1}{N(90-K)}$$

Nous donnons en exemple le cas de la Loire à Blois où l'on a ajusté à titre d'exemple, pour les débits moyens décadaires, la sinusoïde ② par les moindres carrés (sur les données naturelles) et la sinusoïde ③ par l'analyse harmonique (aux données en logarithmes népériens) :

- ajustement des 36 moyennes
- ajustement des 36 écarts types
- ajustement des 36 coefficients d'autocorrélation d'ordre 1. Il aurait fallu compléter par les coefficients d'autocorrélation d'ordre $K > 1$ et les paramètres des fonctions de répartition décadaires des écarts résiduels ε_{ij}, car :

$$Y_{tj} = \frac{Q_{tj} - m_j}{\sigma_j} \qquad ④$$

$$Y_{tj} = \sum_{i=1}^{K} \alpha_{ij} Y_{tj-i} + \varepsilon_{tj} \qquad ⑤$$

la relation ⑤ se simplifie si l'on admet que la liaison en chaîne est d'ordre 1.

Remarque : l'ajustement de la composante saisonnière par les moindres carrés a l'avantage, par rapport à l'analyse harmonique, de fournir la précision de l'écart entre le modèle et les normales calculées sur les observations.

5.6.3 - Calcul pratique -

Considérons le cas de la Cère à St-Etienne-Cantalès; on veut établir la relation qui existe entre le débit d'un jour quelconque de la période 1er février - 31 mars et les débits des jours précédents. Pour tenir compte de la non linéarité de la relation et pour rendre homoscédastique l'écart entre la relation calculée et le débit observé, on effectue une transformation en logarithme :

$X_j = \log 10\ Q_j$ (le 10 élimine les décimales pour faciliter la perforation des données) ;

on cherche à établir une relation de la forme :

$$X_j = \alpha_1 X_{j-1} + \alpha_2 X_{j-2} + \alpha_3 X_{j-3} + \ldots + \varepsilon_j$$

quel que soit j entre le 1er février et le 31 mars. Pour cela on calcule, si n est le nombre de débits des 2 mois pendant N années (n = 59 N) :

$$C_{KK} = \frac{1}{n-KN} \sum_{i=1}^{N} \sum_{j=1}^{59-K} (X_{i,j})(X_{i,j+K}) - \frac{1}{(n-KN)^2} \left(\sum_{i=1}^{N} \sum_{j=1}^{59-K} X_{i,j} \right) \left(\sum_{i=1}^{N} \sum_{j=1}^{59} X_{i,j+K} \right)$$

$$C_{oK} = \frac{1}{n-KN} \sum_{i=1}^{N} \sum_{j=1}^{59-K} (X_{i,j})^2 - \frac{1}{(n-KN)^2} \left(\sum_{i=1}^{N} \sum_{j=1}^{59-K} X_{i,j} \right)^2$$

$$r_K = \frac{C_{KK}}{C_{oK}}$$

Le graphique r_K en fonction de K est appelé corrélogramme; il fournit une première indication sur la structure de la relation interne des débits.

On démontre que l'on peut obtenir les coefficients $\alpha_1 \ldots \alpha_k$ en résolvant :

$$\begin{bmatrix} 1 & r_1 & \ldots & r_{k-1} \\ r_1 & 1 & \ldots & r_{k-2} \\ r_{k-1} & & \ldots & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{bmatrix} = \begin{bmatrix} r_1 \\ \vdots \\ r_k \end{bmatrix}$$

Il est équivalent d'utiliser la méthode des moindres carrés en cherchant à minimiser le résidu ε.

Le point important est de savoir combien de coefficients on retiendra, et quel sera le niveau de signification, c'est-à-dire à partir de quel moment peut-on considérer que tel coefficient n'est pas significativement différent de 0.

Pour cela on effectue les calculs progressivement, on part d'une matrice d'ordre 3, puis 4, ..., etc; on teste soit la variance résiduelle, soit les coefficients de corrélation partielle entre X_j et X_{j-p}, les effets des X_{j-1}, X_{j-2}, ..., étant éliminés.

Dans le cas des débits, généralement 2 ou 3 coefficients suffisent. On voit que le fait d'augmenter l'ordre de la matrice de corrélation laisse les premiers coefficients très stables, alors que les autres se mettent à osciller autour de 0.

Lorsque la liaison entre valeurs successives est une chaîne de Markov simple, soit $r_K = r_1^K$, il est facile de voir que :

$$\begin{cases} \dfrac{X_j - \overline{X}}{\sigma} = r_1 \left(\dfrac{X_{j-1} - \overline{X}}{\sigma}\right) + \varepsilon_j \\ \text{avec } \sigma_\varepsilon = \sqrt{1 - r_1^2} \end{cases}$$

toute l'information passée est contenue dans la valeur précédente X_{j-1}. Les coefficients de corrélation partielle entre X_j et X_{j-K} $(1 < K \leqslant \infty)$ sont nuls.

Dans le cas d'une chaîne d'ordre 2, toute l'information passée est contenue dans les 2 derniers états (X_{j-1}, X_{j-2}) :

$$\begin{cases} \dfrac{X_j - \overline{X}}{\sigma} = \dfrac{r_1 (1 - r_2^2)}{1 - r_1^2} \dfrac{X_{j-1} - \overline{X}}{\sigma} + \dfrac{r_2 (1 - r_1^2)}{1 - r_1^2} \dfrac{X_{j-2} - \overline{X}}{\sigma} + \varepsilon_j \\ \sigma_\varepsilon = \sqrt{\dfrac{1 + r_2 (r_2 - 2 r_1^2)}{1 - r_1^2}} \end{cases}$$

Ces relations supposent évidemment la stationnarité des moyenne - écart type - fonction de répartition de ε_j pendant la saison.

On calcule alors les résidus $\varepsilon_j = X_j - \alpha_1 X_{j-1} - \alpha_2 X_{j-2} \cdots$ et on établit leur distribution empirique. Il est alors aisé d'effectuer une prévision :

- sur le graphique X_j, ($\alpha_1 X_{j-1} + \alpha_2 X_{j-2}$, ...) il suffit de tracer les droites à p %, parallèles entre elles, ε_p étant le quantile correspondant à la probabilité p %.

On peut vérifier que la distribution des résidus par classe de ($\alpha_1 X_{j-1} + \alpha_2 X_{j-2} + \ldots$) est constante quelle que soit la classe.

Cette analyse permet de faire deux opérations :

- la prévision par inertie, que l'on pourrait améliorer en décomposant ε en effet de la pluie et un autre terme aléatoire. En réalité, il y a un effet d'autocorrélation pour les résidus supérieurs à un seuil, mais compte tenu de leur petit nombre, cet effet est noyé dans l'ensemble des résidus sans autocorrélation : $\varepsilon_j = \alpha R_{j-1} - \beta R_{j-2} + \varepsilon'_j$, pour $R_j > R_o$ (seuil), ε'_j étant réellement un bruit blanc (la pluie peut être traitée en $\sqrt{}$: $R_j = \sqrt{P_j}$) ;

- la simulation : on peut, à l'aide de nombres au hasard, générer des séries fictives de débits connaissant les coefficients saisonniers α_1, α_2, ..., α, β, les lois de probabilité de R_j et ε'_j, par saison.

Remarque : il est assez dangereux d'appliquer cette technique sur les valeurs journalières d'une saison (2 ou 3 mois) dont on ne dispose que de 1, 2, 3 ou 4 années d'observations; en effet, on peut montrer que les coefficients d'autocorrélation calculés sont alors affectés d'un biais systématique important. Ce biais tend à disparaître pour des échantillons de 1 000 à 2 000 valeurs.

Exemple pour un schéma Markovien simple, la mémoire à l'instant t est entièrement définie par la mémoire à l'instant t-1 : $X_t = \rho X_{t-1} + \sqrt{1 - \rho^2} \cdot \varepsilon_t$; avec un coefficient d'autocorrélation théorique ρ et une taille d'échantillon n le coefficient d'autocorrélation empirique r calculé sur l'échantillon est :

	ρ =	0	0.5	0.95	0.99	1
u = 25	r =	0	0.44	0.84	0.87	1
u = 50	r =	0	0.47	0.93	0.93	1

5.6.4 - Autocorrélation multiple

Lorsque l'on recherche une corrélation multiple entre une variable principale temporelle (débit journalier d'un grand fleuve) et des variables explicatives temporelles (débits témoins d'affluents amont et éventuellement précipitations de bassins amont), il faut non seulement calculer les coefficients d'autocorrélation propres à chaque station mais les coefficients de corrélation sériale croisés, avec déphasage.

Ce que l'on cherche à obtenir c'est une relation (stationnaire pendant une saison) de la forme :

$$Y_j = \sum_{K=1}^{m} a_k Y_{j-k} + \sum_{l=p}^{q} b_l Z_{j-l} + \sum_{i=s}^{t} c_i U_{j-i} + a_o + \varepsilon_j$$

Z et U étant les débits témoins amont (en logarithme éventuellement).

Il est conseillé, pour trouver les déphasages :

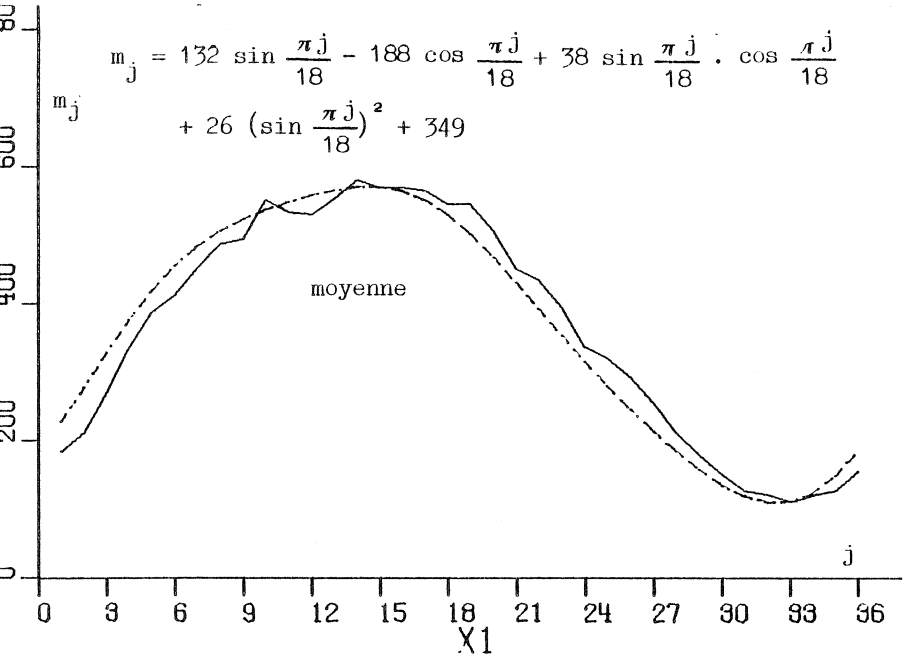
p à q pour Z et s à t pour U ,

de calculer les équations d'autorégression propres à chaque station :

$$\begin{cases} Y_j = \sum_{k=1}^{K} \alpha_K Y_{j-K} + \alpha_o + \xi_j \\ Z_j = \sum_{l=1}^{L} \beta_l Z_{j-l} + \beta_o + \theta_j \\ U_j = \sum_{i=1}^{I} \gamma_i U_{j-i} + \gamma_o + \eta_j \end{cases}$$

on calcule ensuite les coefficients d'autocorrélation croisés (avec décalage) entre les chroniques de ε_j d'une part et θ_j, η_j d'autre part.

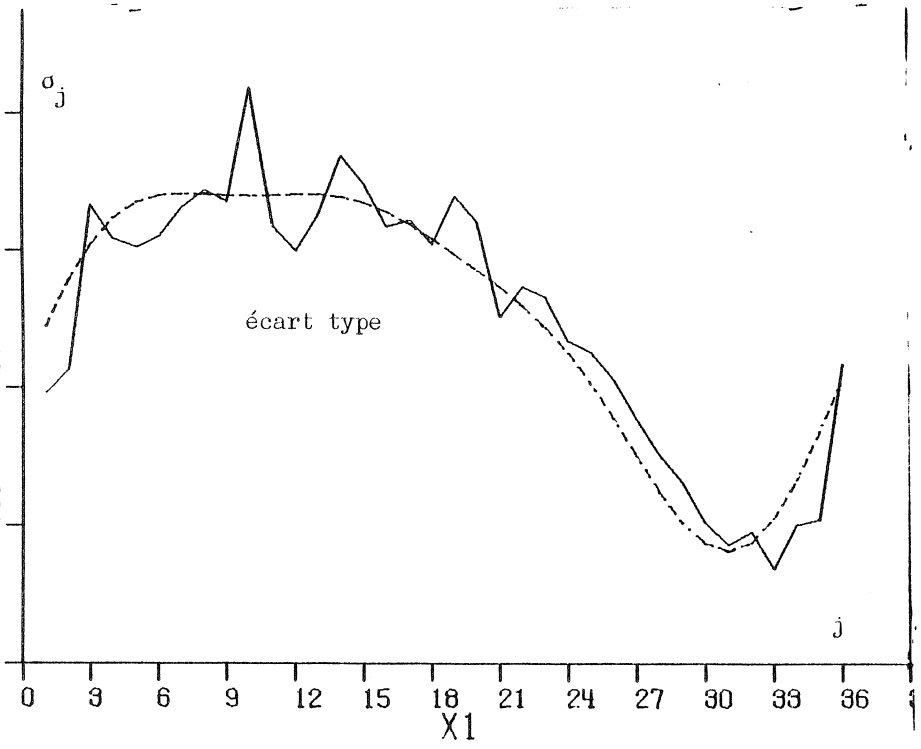
$$m_j = 132 \sin \frac{\pi j}{18} - 188 \cos \frac{\pi j}{18} + 38 \sin \frac{\pi j}{18} \cdot \cos \frac{\pi j}{18} + 26 \left(\sin \frac{\pi j}{18}\right)^2 + 349$$

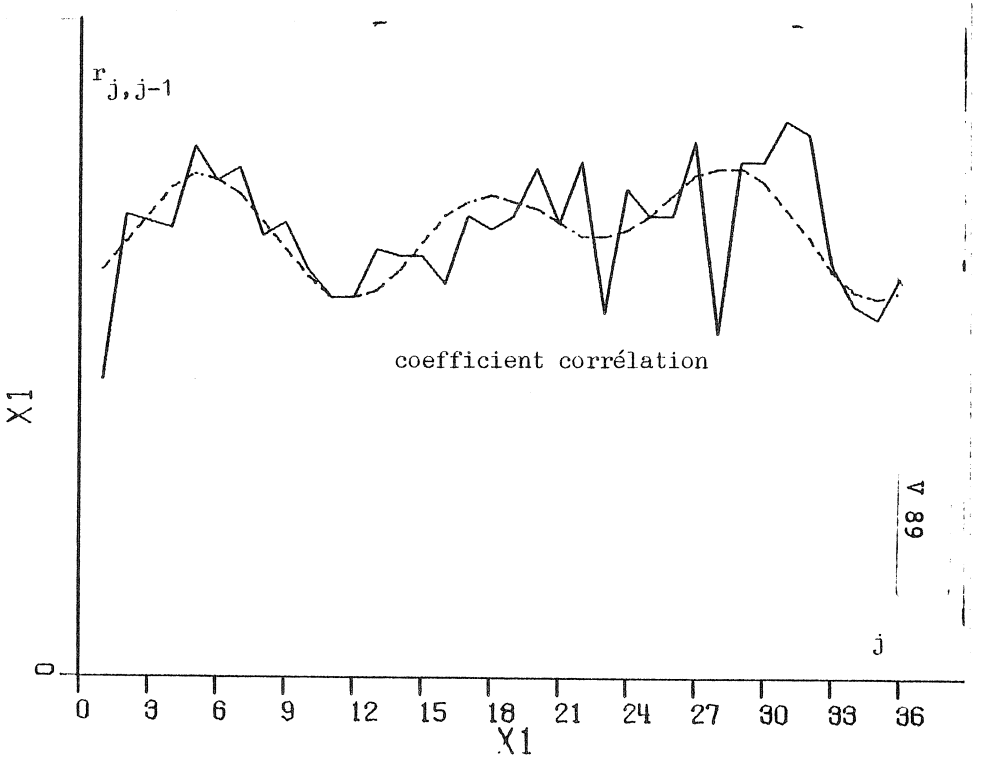


La LOIRE à BLOIS

Débits moyens décadaires
(1-10) octobre à (21-30) septembre
1863-1972

ajustement par $y_j = \sum_{k=1}^{2} \gamma_k \cos \left(\frac{k\pi j}{18} + \varphi_k\right) + \bar{y}$
(moindres carrés)

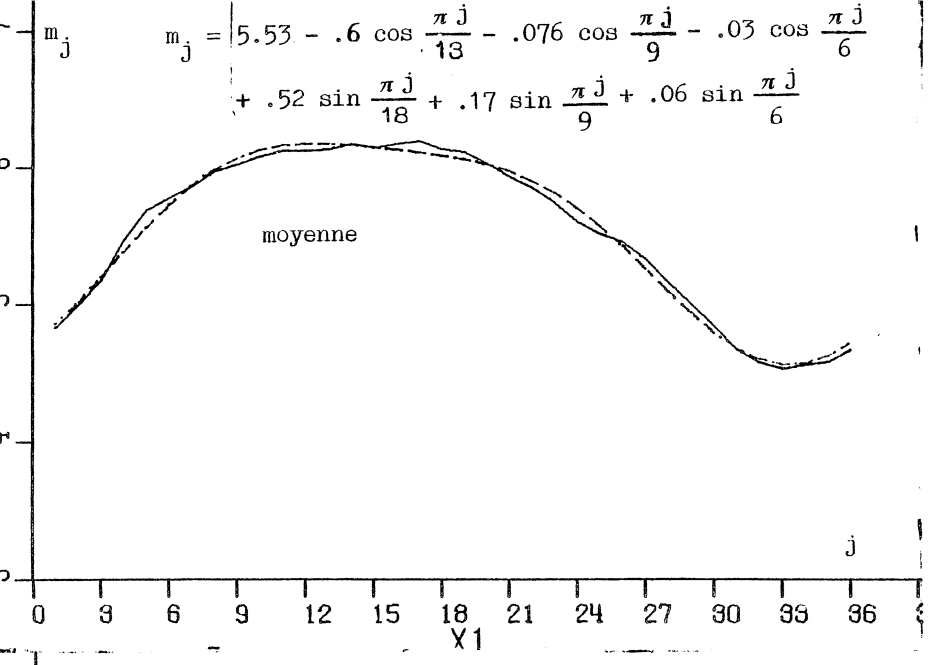


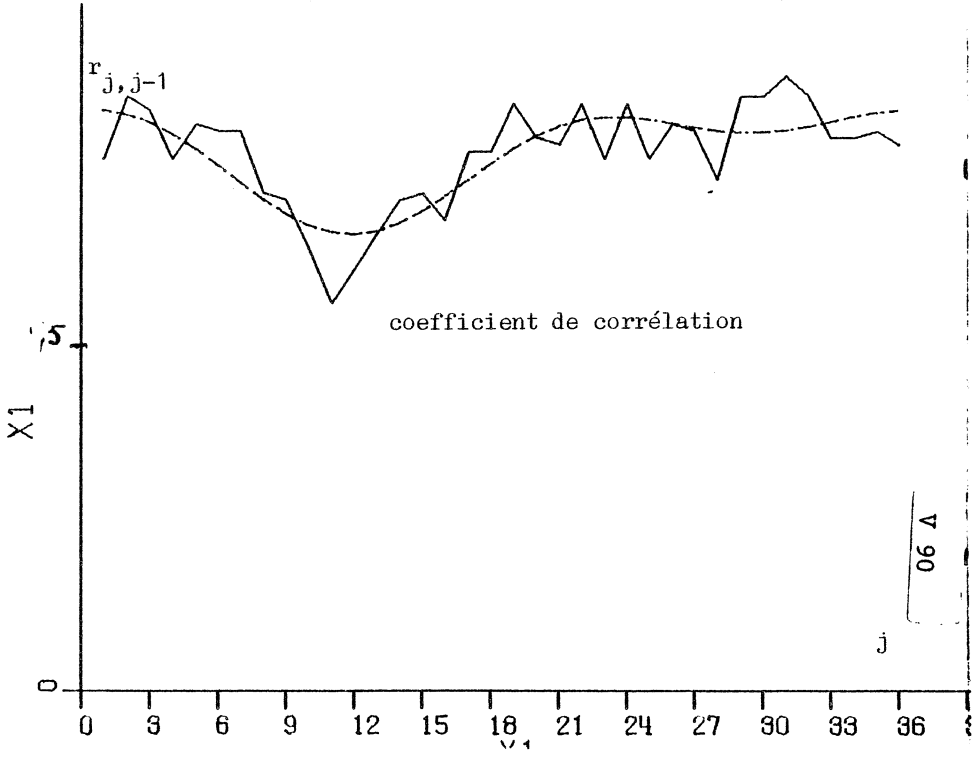


La LOIRE à BLOIS

Logarithmes népériens des
débits moyens décadaires
(1-10) octobre à (21-30) septembre
1863-1972

ajustement par $y_j = \sum_{k=0}^{3} \left(a_k \cos \frac{k\pi j}{18} + b_k \sin \frac{k\pi j}{18}\right)$





VI-D

LA CERE à SAINT-ETIENNE-CANTALES

Autocorrélation des débits journaliers (en logarithme)
durant la période 1er FEVRIER - 31 MARS (1936-1965)

Coefficient de corrélation entre les débits des jours j et j+k	
r_j, j+1	.941
r_j, j+2	.866
r_j, j+3	.796
r_j, j+4	.730
r_j, j+5	.673
r_j, j+6	.622
r_j, j+7	.582
r_j, j+8	.548
r_j, j+9	.515
r_j, j+10	.488
r_j, j+11	.464
r_j, j+12	.441
r_j, j+13	.420
r_j, j+14	.397
r_j, j+15	.377

```
COSTE:GAUROS TERMINEE
AE6036452 FP012223 DP007444
GENERATION TERMINEE
CLE 2:LECTURE DE LA MATRICE TRIANGULAIRE COMPLETE
SINON LECTURE DE LA PREMIERE LIGNE UNIQUEMENT

CLE 1:SUR RUBAN
SINON AU CLAVIER

NB D OBSERVATIONS ET ORDRE MAXIMUM
1800 9
1 .9413 .8664 .7960 .7301 .6728 .6216 .5824 .5478
          N=                                          coefficients d'autocorrelation partielle   coefficients d'autoregression
              3          I=  1
                             J=  1            -1.0000      -1.0000
                             J=  2              .7460       1.1036
                             J=  3             -.1724       -.1724
 R2= .8889
          N=  4          I=  1
                             J=  1            -1.0000      -1.0000
                             J=  2              .7421       1.1069
                             J=  3             -.1299       -.1937
                             J=  4              .0193        .0193
 R2= .8890
 S = .9996           SEUIL= .9994VARIABLE NON SIGNIFICATIVE
          N=  5          I=  1
                             J=  1            -1.0000      -1.0000
                             J=  2              .7421       1.1071
                             J=  3             -.1304       -.1961
                             J=  4              .0220        .0329
                             J=  5             -.0122       -.0122
 R2= .8890
 S = .9999           SEUIL= .9994VARIABLE NON SIGNIFICATIVE
          N=  6          I=  1
                             J=  1            -1.0000      -1.0000
                             J=  2              .7424       1.1075
                             J=  3             -.1311       -.1972
                             J=  4              .0261        .0392
                             J=  5             -.0324       -.0483
                             J=  6              .0326        .0326
 R2= .8891
 S = .9989           SEUIL= .9994VARIABLE SIGNIFICATIVE
          N=  7          I=  1
                             J=  1            -1.0000      -1.0000
                             J=  2              .7422       1.1075
                             J=  3             -.1309       -.1970
                             J=  4              .0260        .0391
                             J=  5             -.0317       -.0477
                             J=  6              .0196        .0293
                             J=  7              .0030        .0030
 R2= .8891
 S = 1.0000           SEUIL= .9994VARIABLE NON SIGNIFICATIVE
          N=  8          I=  1
                             J=  1            -1.0000      -1.0000
                             J=  2              .7430       1.1072
                             J=  3             -.1326       -.1991
                             J=  4              .0283        .0425
                             J=  5             -.0336       -.0504
                             J=  6              .0288        .0432
                             J=  7             -.0505       -.0753
                             J=  8              .0707        .0707
 R2= .8897
 S = .9950           SEUIL= .9994VARIABLE SIGNIFICATIVE
          N=  9          I=  1
```

VI D

LA CERE à St-ETIENNE-CANTALES

FEV-MARS (1936-1965)

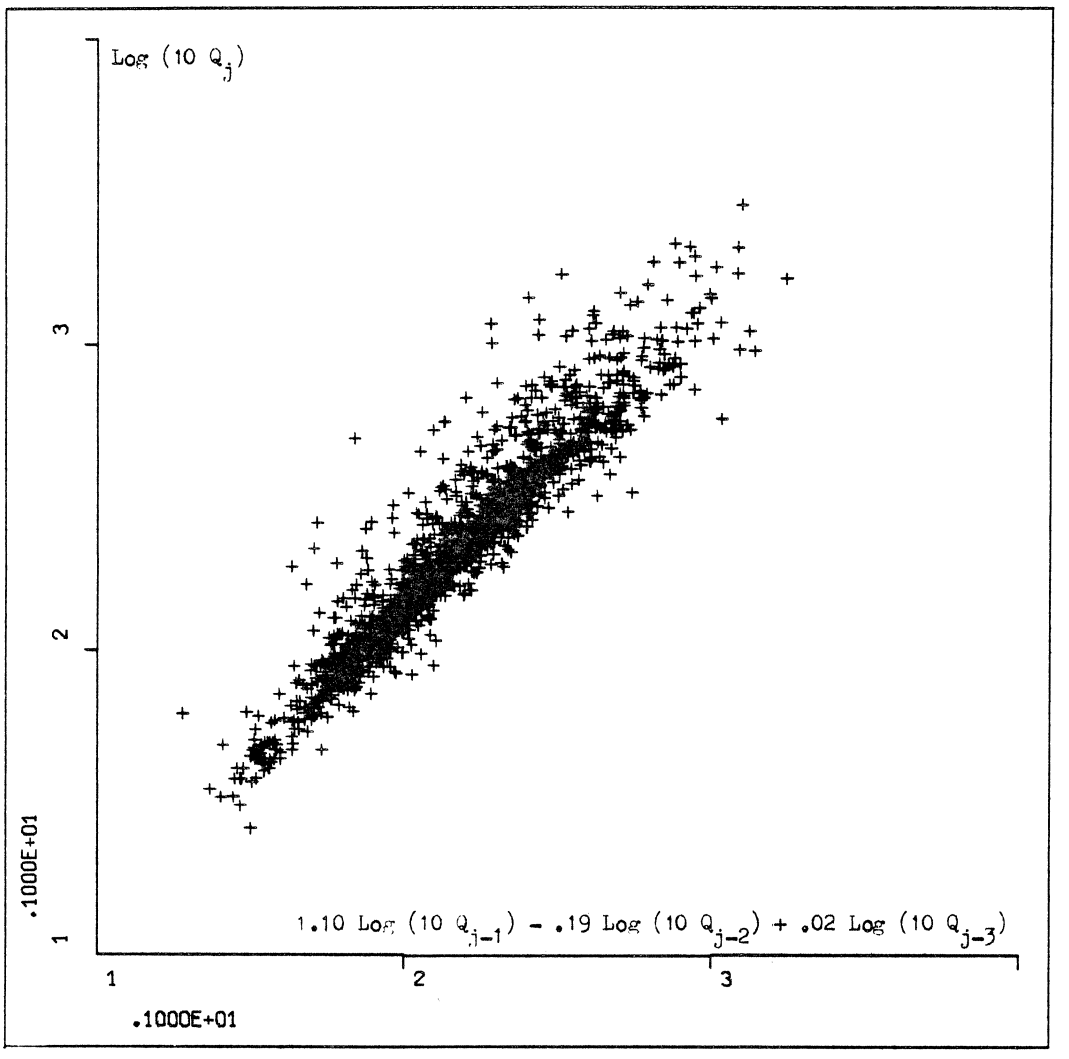
Fréquence de $\varepsilon_j = \text{Log } 10\ Q_j - (1.1 \text{ Log } 10\ Q_{j-1} - .19 \text{ Log } 10\ Q_{j-2} + .02 \text{ Log } 10\ Q_{j-3})$

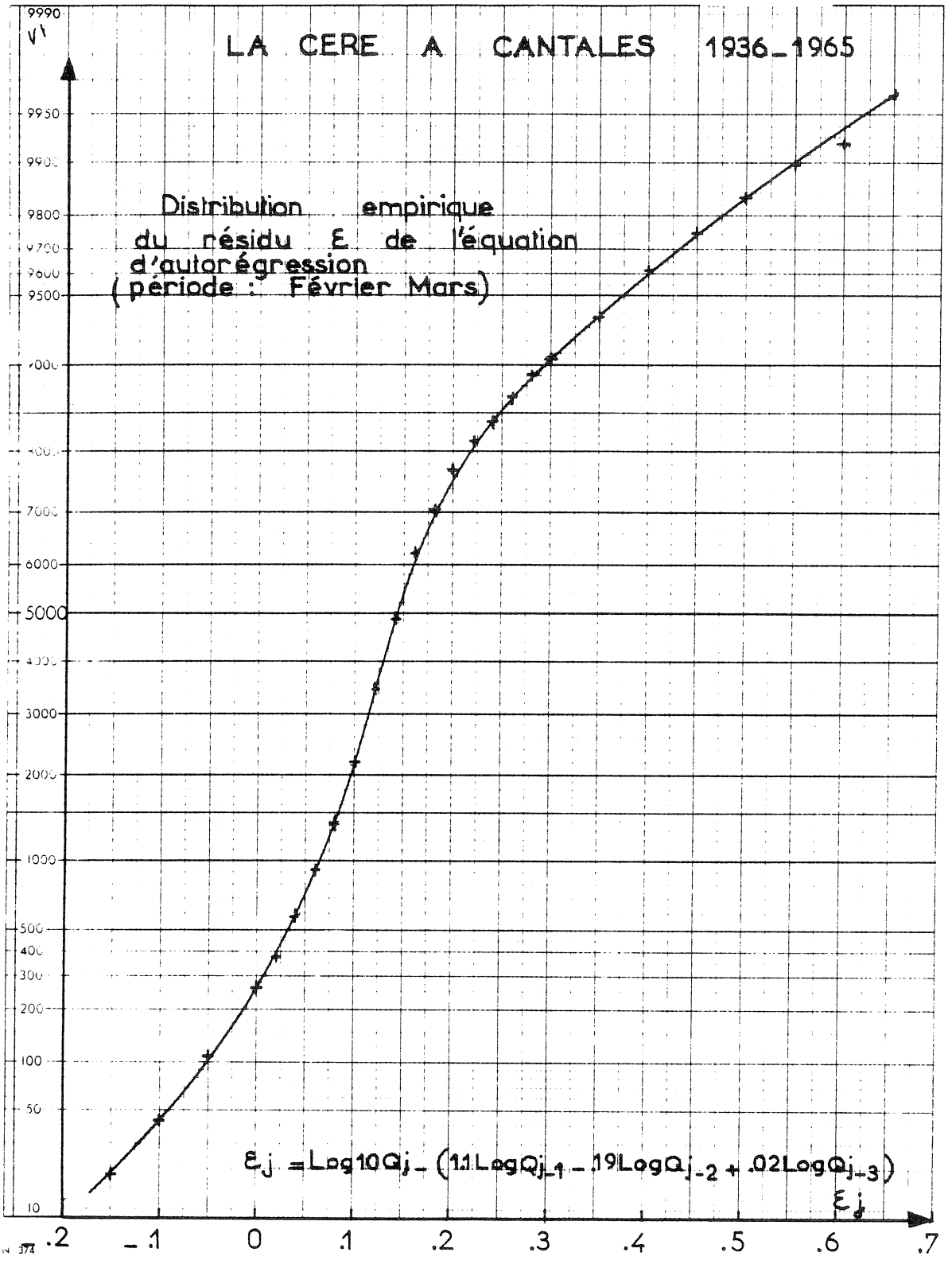
MOYENNE .162 ECART TYPE .108

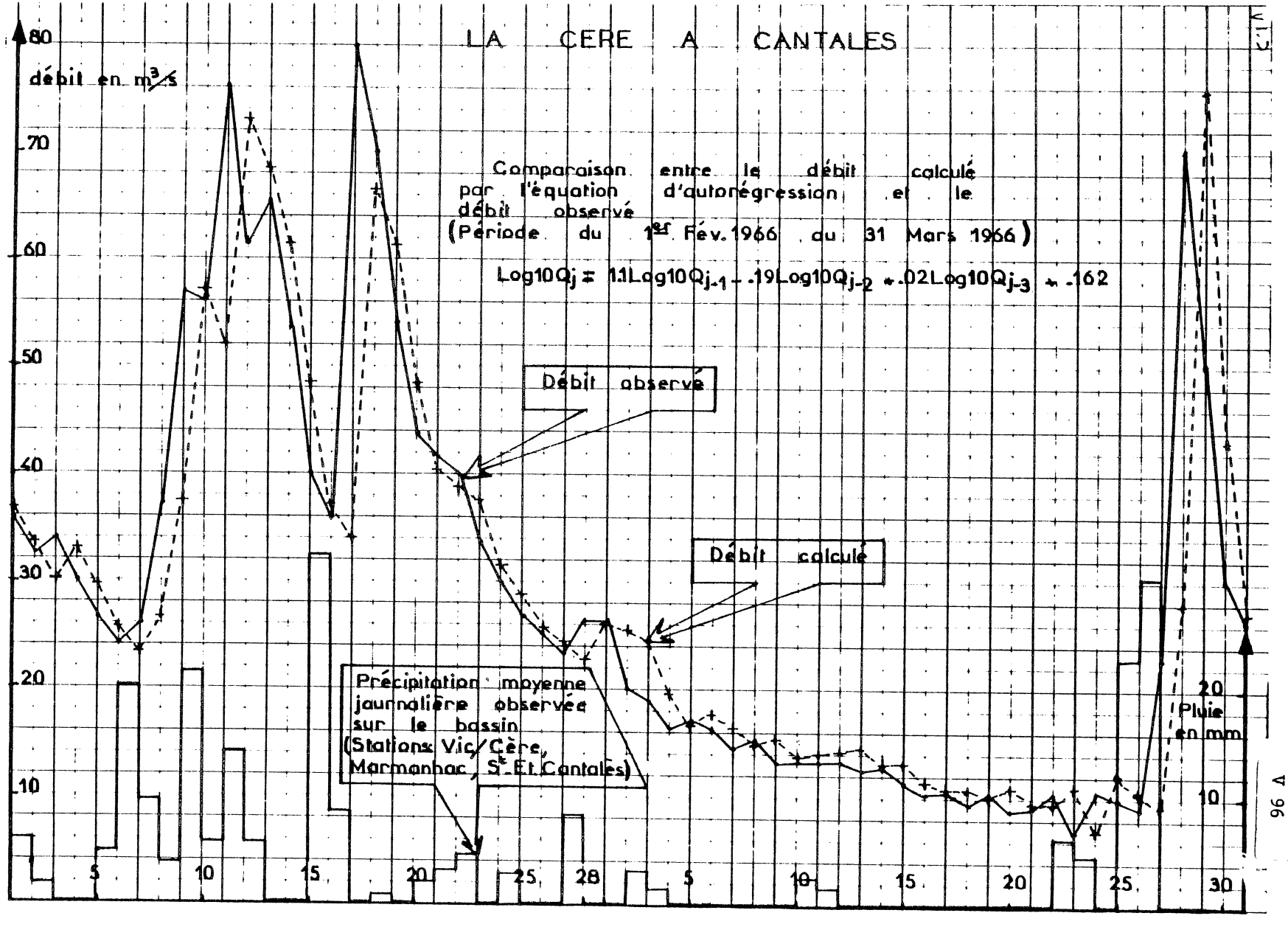
NO	LIM.SUP	N	FREQUENCE	F.CUMULEE
2	-.25	1	.06	.06
3	-.20	1	.06	.12
4	-.15	1	.06	.18
5	-.10	4	.24	.42
6	-.05	11	.65	1.07
7	.00	27	1.61	2.68
8	.02	19	1.13	3.81
9	.04	36	2.14	5.95
10	.06	54	3.21	9.17
11	.08	81	4.82	13.99
12	.10	134	7.98	21.96
13	.12	211	12.56	34.52
14	.14	243	14.46	48.99
15	.16	219	13.04	62.02
16	.18	147	8.75	70.77
17	.20	108	6.43	77.20
18	.22	68	4.05	81.25
19	.24	42	2.50	83.75
20	.26	53	3.15	86.90
21	.28	38	2.26	89.17
22	.30	25	1.49	90.65
23	.35	45	2.68	93.33
24	.40	50	2.98	96.31
25	.45	22	1.31	97.62
26	.50	13	.77	98.39
27	.55	11	.65	99.05
28	.60	3	.18	99.23
29	.65	7	.42	99.64
31	.80	5	.30	99.94
32	.90	1	.06	100.00

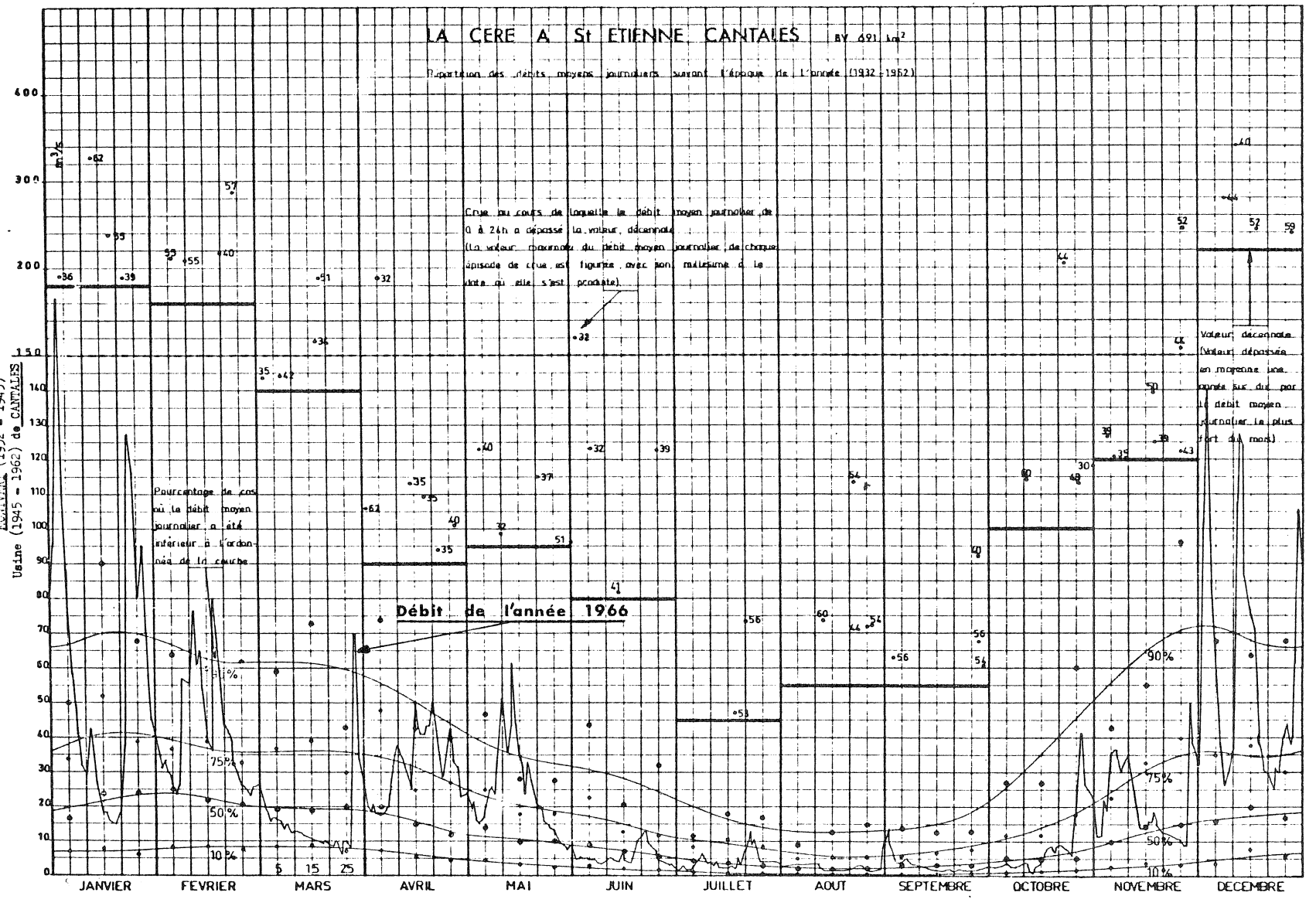
LA CERE à SAINT-ETIENNE-CANTALES

Prévision du débit moyen journalier par autocorrélation
en février et mars (1936-1965) n = 1680




LA CERE A CANTALES 1936_1965
Distribution empirique
du résidu ε de l'équation
d'autorégression
(période : Février Mars)
V'
9990
9950
9900
9800
9700
9600
9500
7000
6000
5000
3000
2000
1000
500
400
300
200
100
50
10
$\varepsilon_j = \text{Log}10Q_j - (1.1\text{Log}Q_{j-1} - .19\text{Log}Q_{j-2} + .02\text{Log}Q_{j-3})$
ε_j
-.2
-.1
0
.1
.2
.3
.4
.5
.6
.7


LA CERE A CANTALES
débit en m³/s
Comparaison entre le débit calculé par l'équation d'autorégression et le débit observé
(Période du 1er Fév. 1966 au 31 Mars 1966)
Log10Qj = 1.1Log10Qj-1 - .19Log10Qj-2 + .02Log10Qj-3 + .162
Débit observé
Débit calculé
Précipitation moyenne journalière observée sur le bassin (Stations: Vic/Cère, Marmanhac, St-Et.Cantalès)
Pluie en mm
80
70
60
50
40
30
20
10
5
10
15
20
25
28
5
10
15
20
25
30



Débits reconstitués à partir de la Station Usine de MONTVERT (1932 - 1945)
Usine (1945 - 1962) de CANTALES

Origine des données :

- E. D. F. - Division Technique Générale -

V 97

1 W.D. ASHTON - The logit transformation - Griffin's statistical monograph and courses - E^d by A. Stuart

D.J. FINNEY - Probit analysis - Cambridge, University Press

2 G. UDNY YULE and M.G. KENDALL - An introduction to the theory of statistics - E^d Griffin

3 A.E. HOERL and R.W. KENNARD - Ridge regression : biased estimation for non orthogonal problems - Technometrics, vol. 12, n° 1, février 1970

4 N.C. MATALAS and M.A. BENSON - Effect of interstation correlation on regression analysis - Journal of Geophysical Research, vol. 66, n° 10, october 1961

5 J. JOHNSTON - Econometric Methods - E^d M.G. Graw-Hill Book Company

6 R.K. BHUIYA and V. YEVJEVICH - Effects of truncation on dependance in hydrological time serie - Hydrology paper, Colorado States University, nov. 1968
Rapport interme D.T.G. : variables aléatoires censurées (janvier 1969)

7 M. ROSENBERG - Une méthode générale pour résoudre des problèmes de corrélation multiple en hydrologie lorsqu'il y a des contraintes - C.R. Académie des Sciences Paris, t 269, série D, novembre 1969 et également publication dans le bulletin de l'I.A.S.H., XV, 3-9/1970

8 M.A. BENSON - Spurious correlation in Hydraulics and Hydrology. Journal of the Hydraulics Division, vol. 91, n° HY 4, juillet 1965

9 O.V. BATYREVA - Calculs concernant la signification du coefficient de corrélation multiple et choix du nombre optimal de prévi-seurs. Meteorologija i Gidrologija, n° 3 Moskwa, Moskovskoe Otdelemie Gidrometeorizdata, mars 1969

10 B. MANDELBROT - Une classe de processus stochastiques homothétiques à soi; application à la loi climatologique, compte rendu Académie des Sciences Paris. 260 - 3274, 1965.

11 B. MANDELBROT and J.R. WALLIS - Noah Joseph and operational Hydrology, W.R.R., vol. 4, n° 5, 1968

12 B. MANDELBROT and J.R. WALLIS - Some long run properties of Geophysical records, W.R.R., Vol. 5, n° 2, 1969

13 I. RODRIGUEZ ITURBE, J.M. MEJIA, D.R. DAWDY -
Streamflow simulation
1/ A new look at Markovian Models, Fractionnal Gaussian noise and Crossing theory
2/ The broken line process as a potential model for Hydrologic Simulation, W.R.R., Vol. 8, n° 4, 1972

VI - L'ANALYSE EN COMPOSANTES PRINCIPALES (A.C.P.) - APPLICATIONS

6.1 - DEFINITIONS

6.1.1 - Définition algébrique

On considère p variables X $\left[X_1, X_2, \ldots, X_j, \ldots, X_p\right]$ pour lesquelles on dispose de n observations (mesures) simultanées.

$$\begin{array}{ll} \text{1ère observation} & \left[\begin{array}{cccc} x_{11} & x_{21} & \ldots\ldots\ldots & x_{p1} \\ x_{12} & x_{22} & \ldots\ldots\ldots & x_{p2} \\ \ldots & \ldots & \ldots\ldots\ldots & \ldots \\ x_{1n} & x_{2n} & & x_{pn} \end{array}\right] \\ \text{2ème observation} & \\ \ldots\ldots\ldots\ldots & \\ n^{\text{ième}}\text{observation} & \end{array}$$

Notons $\left[X_{pn}\right]$ cette matrice de données.

On calcule alors la moyenne de chacune de ces p variables, leur écart type ainsi que les coefficients de corrélation totale entre tous les couples possibles de variables :

moyennes $M_1, \ldots, M_p$ avec : $M_j = \frac{1}{n} \sum_{i=1}^{n} x_{ji}$

écarts types $\sigma_1, \ldots, \sigma_p$ avec : $\sigma_j = \left[\frac{1}{n-1} \sum_{i=1}^{n} (x_{ji} - M_j)^2\right]^{\frac{1}{2}}$

coefficient de corrélation r_{jk} entre les variables X_j et X_k

$$r_{jk} = \frac{\sum_{i=1}^{n} (x_{ji} - M_j)(x_{ki} - M_k)}{\left[\sum_{i=1}^{n} (x_{ji} - M_j)^2 \sum_{i=1}^{n} (x_{ki} - M_k)^2\right]^{\frac{1}{2}}}$$

Calculer les composantes principales de ces p variables X, définies par n observations, revient à déterminer p relations linéaires de la forme :

$$C_1 = \alpha_{11} X_1 + \alpha_{12} X_2 + \ldots + \alpha_{1j} X_j + \ldots + \alpha_{1p} X_p + \alpha_{10}$$

$$C_2 = \alpha_{21} X_1 + \ldots\ldots\ldots\ldots + \alpha_{2p} X_p + \alpha_{20}$$

$$C_p = \alpha_{p1} X_1 + \ldots\ldots\ldots\ldots + \alpha_{pp} X_p + \alpha_{p0}$$

Ces nouvelles variables ayant la propriété essentielle d'être orthogonales, c'est-à-dire que les coefficients de corrélation entre tous les couples de C.P. sont nuls, la nouvelle base définie par la matrice $\left[\alpha_{e,j}\right]$ étant orthogonale.

En pratique, nous effectuons les calculs de composantes principales sur des variables centrées réduites, procédure qui sera justifiée dans la suite. On obtient alors :

$$Z_1 = a_{11} \frac{X_1 - M_1}{\sigma_1} + a_{12} \frac{X_2 - M_2}{\sigma_2} + \ldots + a_{1p} \frac{X_p - M_p}{\sigma_p}$$

$$Z_2 = a_{21} \frac{X_1 - M_1}{\sigma_1} + \ldots\ldots\ldots\ldots + a_{2p} \frac{X_p - M_p}{\sigma_p}$$

$$\ldots \quad \ldots\ldots\ldots \quad \ldots\ldots\ldots\ldots\ldots \quad \ldots\ldots\ldots$$

$$Z_p = a_{p1} \frac{X_1 - M_1}{\sigma_1} + \ldots\ldots\ldots\ldots + a_{pp} \frac{X_p - M_p}{\sigma_p}$$

On calcule généralement q composantes $(q < p)$

$$Z_1 = \sum_{j=1}^{p} a_{1j} \frac{X_j - M_j}{\sigma_j} \quad \text{avec } 1 \leqslant l \leqslant q$$

A une matrice de données $\left[X_{pn}\right]$ correspond alors une nouvelle matrice de données $\left[Z_{qn}\right]$ de dimensions réduites en colonnes.

$$\text{avec } \left[Z_{qn}\right] = \begin{bmatrix} z_{11} & z_{21} \ldots\ldots z_{q1} \\ z_{12} & z_{22} \ldots\ldots z_{q2} \\ \ldots & \ldots \ldots\ldots \ldots \\ z_{1n} & z_{2n} \ldots\ldots z_{qn} \end{bmatrix}$$

Les éléments de ce tableau sont obtenus en appliquant à la matrice X_{pn} les q premières relations linéaires précédentes :

$$z_{li} = \sum_{j=1}^{p} a_{lj} \left(\frac{x_{ji} - M_j}{\sigma_j}\right)$$

Les coefficients a_{lj} (pour j = 1 à p) sont appelés cosinus directeurs de la C.P. d'ordre l; ils sont orthonormés :

$$\sum_{j=1}^{p} a_{lj} \cdot a_{mj} = \begin{cases} 1 \text{ si } m = l \\ 0 \text{ si } m \neq l \end{cases}$$

Ce sont les coordonnées des vecteurs propres de la matrice des coefficients de corrélation r_{jk}; on les calcule en diagonalisant la matrice :

$$[R] = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1p} \\ & 1 & r_{23} & \cdots & r_{2p} \\ & & 1 & \cdots & \\ & & & \cdots & \\ & & & & 1 \end{bmatrix}$$

Les valeurs propres λ_l de cette matrice ne sont autres que les variances des nouvelles variables Z_l, c'est-à-dire :

$$\lambda_l = \frac{1}{n} \sum_{i=1}^{n} z_{li}^2 \quad ; \quad \text{(la moyenne de } Z_l \text{ est nulle par construction)}$$

N.B. - Dans certains textes, les C.P. sont également appelées fonctions orthogonales empiriques (ou naturelles).

- On peut toujours diagonaliser une matrice de coefficients de corrélation, même lorsque le nombre des variables est supérieur au nombre d'observations ($p \geqslant n$), mais dans ce cas il y aura p - (n-1) valeurs propres strictement nulles : $\lambda_p = \lambda_{p-1} = \lambda_{p-2} = \ldots \lambda_{p-n-1} = 0$; ce qui n'est pas gênant pour des calculs ultérieurs car on n'utilise pas

plus des 3 à 5 premières composantes principales lorsque $30 < n < 100$. Pour des échantillons si peu importants, il est en effet illusoire et risqué d'attribuer une signification aux C.P. d'ordre supérieur à 5 eu égard à la dispersion d'échantillonnage de la matrice des coefficients de corrélation totale ainsi qu'au "bruit" engendré par les erreurs de mesure sur les observations.

On trouve également des valeurs propres nulles, lorsqu'il existe une relation fonctionnelle linéaire entre 2 ou plusieurs couples (X_j, X_K).

Dans le processus de calcul, ces vecteurs propres sont déterminés dans l'ordre des valeurs propres λ_l décroissantes. Une propriété remarquable des valeurs propres ainsi obtenues est que leur somme égale la dimension de la matrice R :

$$\sum_{l=1}^{P} \lambda_l = p \qquad \text{(trace)}$$

Cette relation permet de connaître la contribution, en variance, de chacune des composantes principales à la variance totale du système à p dimensions.

En particulier, un des critères permettant de choisir le nombre de composantes "utiles" (nous verrons qu'il n'est pas suffisant et peut être parfois sujet à caution) consiste à éliminer les composantes Z_m d'ordre élevé $(q < m \leq p)$ de telle sorte que la perte d'information soit faible par rapport à la variance totale :

$$\frac{1}{p} \sum_{m=q+1}^{p} \lambda_m = \varepsilon \qquad \text{avec } \varepsilon = 1\ \%,\ 5\ \% \text{ ou } 10\ \%$$

Il est équivalent de conserver un nombre minimum de C.P. qui maximise le coefficient de corrélation multiple entre chaque variable et ces q premières C.P. soit :

$$q \text{ minimum et } R_j^2 = \sum_{l=1}^{q} a_{lj}^2 \lambda_l \text{ aussi voisin de 1 que possible}$$

Le coefficient de corrélation entre chaque variable initiale et la composante principale d'ordre k est en effet égal à :

$$\rho_{jk} = \frac{\sum_{i=1}^{n} (x_{ji} - m_j)\ z_{ki}}{\left[\sum_{i=1}^{n} (x_{ji} - m_j)^2 \sum_{i=1}^{n} z_{ki}^2\right]^{\frac{1}{2}}} = a_{jk}\sqrt{\lambda_k}$$

et d'après la propriété d'orthogonalité des C.P., le coefficient de corrélation multiple entre la variable X_j et les q premières C.P. est obtenu par :

$$R_j^2 = \sum_{l=1}^{q} \rho_{jl}^2 = \sum_{l=1}^{q} a_{jl}^2\ \lambda_l$$

6.1.2 - Définition géométrique

On peut considérer que dans l'espace à p dimensionsqui constitue le système de référence initial pour le phénomène étudié, chaque observation ou ensemble de p valeurs x_{ji} est représentée par un point, par conséquent la matrice d'observations X_{pn} est représentée par n points (les p axes ne constituent pas une base). Dans l'espace à n dimensions, les p vecteurs X_1, ... X_p de coordonnées respectives $x_{11},..,x_{1n}-x_{21},..,x_{2n}-..x_{p1},..,x_{pn}-$)ne sont pas orthogonaux.

Calculer la première C.P. revient à chercher l'axe tel que la somme des carrés des distances des n points à cet axe soit minimale, ou encore que la somme des carrés des projections des points sur cet axe soit maximale.

Puis on détermine le second axe, après projection des n points dans un hyperplan orthogonal au premier axe, tel que la somme des carrés des distances des points-observations à celui-ci soit minimale. Ce processus de calcul se réitère p fois. Si l'on calcule et conserve p composantes principales, cela revient à effectuer une rotation du système d'axes origine. Lorsqu'on ne conserve que les q premières C.P., on considère alors un nuage de n points qui sont les projections des originaux dans un sous-espace de dimensions q. Dans l'espace à n dimensions les vecteurs $Z_1,...Z_p$ (de coordonnées $z_{11},...$ z_{1n} - ... - $z_{p1},...,$ z_{pn} -) sont alors orthogonaux.

6.2 - CONDITIONS D'UTILISATION DE L'A.C.P.

La plupart des auteurs affirment dans les ouvrages ou articles se rapportant à cette technique, que le domaine d'utilisation de l' A.C.P. est

très large - sans inconvénient - sans danger - ne nécessite pas au préalable d'hypothèses restrictives et, de plus, que l'interprétation des résultats n'est pas "induite" comme c'est le cas pour d'autres procédés d'analyse factorielle.

S'il est vrai que les conditions de normalité ne sont pas impératives, ni même nécessaires pour pratiquer l'A.C.P., en particulier que la fonction de répartition empirique des séries x_{ji} (pour un j donné et $1 \leqslant i \leqslant n$) soit gaussienne, on ne doit pas oublier que la matière première de l'A.C.P. est la matrice des coefficients de corrélation totale entre tous les couples (X_j, X_k). C'est elle, en effet, qui après diagonalisation permet de calculer les valeurs et vecteurs propres.

Or, la valeur du coefficient de corrélation linéaire r_{jk}, entre X_j et X_k, dépend essentiellement des couples d'observations x_{ji}, x_{ki} et en particulier de la configuration de ce nuage de points dans le plan (X_j, X_k).

En particulier, si l'on applique une transformation monotone aux p variables X_j, du type $\log_e X_j$ ou $(X_j)^{\frac{p}{q}}$ on modifiera les coefficients de corrélation r_{jk} et, de ce fait, la structure des composantes principales.

Autre remarque importante : il n'est pas équivalent de calculer les C.P. sur des variables centrées réduites, donc avec une matrice de corrélation, ou sur des variables centrées donc avec une matrice de covariances, lorsque les variables X_j ont des variabilités très différentes, lorsque par exemple $\frac{\sigma_j}{\sigma_k} = 10$, car alors les premières composantes risquent d'être définies uniquement par les variables X_j à fortes variances, les variables du type X_k ayant une faible pondération, alors qu'au contraire celles-ci conditionnent presque totalement les composantes correspondant aux plus faibles valeurs propres λ_m, qui sont celles que l'on élimine généralement.

Il est donc indispensable d'être conscient de tous ces aspects et de leurs conséquences lorsqu'on effectue une A.C.P.; cette mise en garde ne diminue en rien le potentiel de ce puissant outil d'analyse.

6.3 - APPLICATIONS DE L'A.C.P.

Cette technique peut être utilisée pour traiter des variables caracrisant :

. un phénomène spatial (champ de température de l'air, précipitations, de pressions atmosphériques, débits, etc, dans une région ou pays);

. un phénomène temporel (chronique des débits, précipitations, températures, pressions journalières - hebdomadaires - décadaires en un lieu).

On peut distinguer deux aspects principaux de l'A.C.P. :

. analyse descriptive (structure d'une matrice d'observation),

. analyse opérationnelle (optimisation d'un réseau de mesures, critique des données, prévision, simulation ,...).

6.3.1 - Analyse descriptive

6.3.1.1. - Interprétation dans l'espace des variables

Si l'on calcule les vecteurs propres de la matrice des coefficients de corrélation, les relations entre C.P. et variables initiales sont de la forme :

$$Z_l = \sum_{j=1}^{p} a_{lj} \left(\frac{X_j - M_j}{\sigma_j}\right)$$

ou, en variables réduites :

$$Y_l = \frac{Z_l}{\sqrt{\lambda_l}}$$

On interprète généralement la signification des coefficients de pondération a_{lj} pour l = 1, 2, 3, 4, 5.

Lorsque l'on traite des variables X_j homogènes, représentant un même phénomène spatial, la suite des coefficients a_{1j} est relativement uniforme; dans certains cas $a_{1j} \neq \left(\frac{1}{p}\right)^{1/2}$ quel que soit j. La première C.P. représente alors une moyenne de variables centrées réduites, c'est un facteur de taille.

L'interprétation, recherche d'une typologie ou d'une proximité entre stations, s'effectue principalement sur les couples de coefficients a_{2j} et a_{3j} et moins fréquemment sur les couples a_{4j}, a_{5j}.

Deux modes de représentation sont utilisés pour visualiser les résultats numériques et orienter l'analyse descriptive.

6.3.1.1.1 - Représentation graphique en coordonnées cartésiennes des couples de valeurs de cosinus ou des coefficients de corrélation (les axes étant gradués soit en valeurs de a_{ej} soit l_{ej}) - Ellipse de proximité

On porte sur le graphique les p points (a_{21}, a_{31}), (a_{22}, a_{32}) ... (a_{2p}, a_{3p}). On voit alors apparaître des "nuages" de points : une typologie sera d'autant mieux définie que les nuages de points sont bien distincts les uns des autres; de plus la redondance, c'est-à-dire le fait que plusieurs stations représentent la même information, sera d'autant plus importante que chaque nuage sera bien concentré, forte compacité interne des "nuages".

On graphique généralement les couples de coefficients de corrélation entre les C.P. et chacune des p variables soit (ρ_{2j}, ρ_{3j}) et éventuellement (ρ_{4j}, ρ_{5j}), ce qui permet de tracer les ellipses de proximité des variables (cf. représentation des rayons de bicyclette et ellipses de proximité).

Ellipse de proximité : exemple de l'ALLIER supérieur (pluies d'octobre et novembre en Ardèche).

n° variables	ρ_{2j} = Z	ρ_{3j} = T	i
32	-.106	-.039	1
33	.063	.143	2
34	.322	.279	3
35	.342	.247	4
36	.473	.197	8
37	.476	.225	6
38	.395	.167	7
39	.098	.101	8

on suppose que les variables Z et T ont une densité de répartition gaussienne à 2 dimensions :

$$f(Z,T) \quad \frac{1}{2\pi s_Z s_T \sqrt{1-r^2}} \; e^{-\frac{E^2}{2}}$$

avec $E^2 = \frac{1}{1-r^2}\left[\left(\frac{Z-\bar{Z}}{s_Z}\right)^2 - 2r\left(\frac{Z-\bar{Z}}{s_Z}\right)\left(\frac{T-\bar{T}}{s_T}\right) + \left(\frac{T-\bar{T}}{s_T}\right)^2\right]$

les moyennes : $\overline{Z} = \frac{1}{8} \sum_{i=1}^{8} Z_i$; $\overline{T} = \frac{1}{8} \sum_{i=1}^{8} T_i$

les variables : $s_Z^2 = \frac{1}{7} \Sigma (Z_i - \overline{Z})^2$; $s_T^2 = \frac{1}{7} \Sigma (T_i - \overline{T})^2$

le coefficient de corrélation : $r = \frac{\Sigma (Z_i - \overline{Z})(T_i - \overline{T})}{\sqrt{\Sigma (Z_i - \overline{Z})^2 \times \Sigma (T_i - \overline{T})^2}}$

On effectue un changement d'axes en prenant le barycentre comme origine :

$$① \quad \begin{cases} u = (Z - \overline{Z}) \cos \theta + (T - \overline{T}) \sin \theta \\ v = (Z - \overline{Z}) \sin \theta + (T - \overline{T}) \cos \theta \end{cases}$$

θ est l'angle de l'axe U avec l'axe Z.

Pour que les nouvelles variables soient orthogonales, en exprimant E^2 en fonction de u et v d'après les relations ① , il faut éliminer le terme rectangle uv, donc écrire que son coefficient est **nul**, soit :

$$② \quad \operatorname{tg} 2\theta = \frac{2r\, s_Z\, s_T}{s_Z^2 - s_T^2}$$

(lorsque $s_Z = s_T$: $\theta = \frac{\pi}{4}$)

Remarque : le jacobien de la transformation ① est égal à l'unité

$\Rightarrow f(Z,T)\, dZ\, dT = g(u, v)\, du\, dv$

Pour l'exemple étudié :

$$\begin{cases} \overline{Z} = .258 \quad , \quad s_Z = .214 \\ \overline{T} = .165 \quad , \quad s_T = .101 \\ r = .821 \end{cases}$$

$\operatorname{tg} 2\theta \# 1$ d'où $\theta = 22°.5$

- l'axe U est défini par l'équation : $T - \overline{T} = (Z - \overline{Z}) \operatorname{tg} \theta$ ou $T = \overline{T} + .414\ (Z - \overline{Z})$
- l'axe V " " " : $Z - \overline{Z} = -(T - \overline{T}) \operatorname{tg} \theta$ ou $Z = \overline{Z} - .414\ (T - \overline{T})$
- les écarts types des nouvelles variables sont définis par les deux relations suivantes, aisées à obtenir d'après ② :

$$\begin{cases} s_U^2 \, s_V^2 = (1 - r^2) \, s_Z^2 \, s_T^2 = P \\ s_u^2 \, s_v^2 = s_Z^2 + s_T^2 = S \end{cases}$$

cette dernière propriété résulte de l'invariance de la distance des points observations au barycentre du nuage dans le changement d'axes.
s_u^2 et s_v^2 sont racines de l'équation :

$$X^2 - SX + P = 0$$

soit $s_u = .231$, $s_v = .054$

La somme des carrés de 2 variables gaussiennes centrées réduites suit une loi du X^2 à 2 degrés de liberté, soit :

$$X^2 = \frac{u^2}{s_u^2} + \frac{v^2}{s_v^2}$$

on peut donc construire les ellipses de concentration à p % :

$$\int_0^{X_p^2} \frac{1}{2} \, e^{-\frac{X^2}{2}} \, dX^2 = p$$

pour p = .50 $X_p^2 = 1.386$
= .80 = 3.219
= .90 = 4.605
= .99 = 9.210

Le tracé des ellipses de proximité met en évidence l'importance de la concentration des stations appartenant à un même bassin, fait apparaître (par la proximité ou l'intersection des ellipses associées) la redondance d'un réseau de mesures couvrant plusieurs bassins :

- la redondance est d'autant plus forte pour un bassin que les axes de l'ellipse sont de faibles dimensions et égaux; une ellipse allongée signifiant qu'une seule station ne suffit pas à représenter l'information ;

- le chevauchement d'ellipses signifie redondance de l'information spatiale mesurée sur différents bassins ;

- lorsqu'une grande ellipse englobe plusieurs nuages distincts, il est préférable de refaire un tracé d'ellipses en tenant compte de ces sous-groupes ;

- les intersections d'ellipses mettent en évidence la continuité spatiale entre bassins ;

- dans l'analyse en C.P. des séries d'un réseau de mesures, il vaut mieux n'utiliser qu'un nombre réduit de stations pour définir les quelques ellipses de proximité, les stations non utilisées se distribuant au voisinage des nuages avec lesquels elles ont des affinités. (On peut faire le même raisonnement à 3, 4, 5 dimensions, on considère alors des hyperellipsoïdes de proximité).

6.3.1.1.2 - Représentation cartographique de chaque série a_{1j} (j=1 à p)

Sur une carte où sont représentées les stations de mesures, on porte la valeur des cosinus directeurs affectant chaque variable correspondante, pour une même C.P. On trace alors les lignes d'égaux coefficients ou lignes d'isocosinus directeurs, qui sont les "lignes de force" du phénomène étudié. Le graphique peut suggérer une interprétation physique de la C.P. étudiée, en s'appuyant sur la géographie, le climat, etc.

6.3.1.1.3 - Représentation graphique interprétée d'une série de coefficients a_{1j} pour j = 1 à p

On peut mettre en évidence la signification physique d'une C.P. en représentant sur graphique cartésien les couples de valeurs (a_{1j}, D_j) pour j = 1 à p, D_j caractérisant le paramètre physique que l'on identifie à la $1^{ième}$ composante. Ainsi, dans le cas d'une A.C.P. sur des températures mensuelles au cours d'une saison, et relevées dans des stations de plaine et de montagne de la moitié Sud de la France, on constate qu'il y a une corrélation étroite et non linéaire entre l'altitude et les cosinus directeurs de la 2ème C.P.

Cette démarche peut également suggérer une interprétation typologique.

6.3.1.2 - Interprétation dans l'espace des observations

Ce que l'on va analyser à présent, c'est la situation des couples de valeurs (y_{1i}, y_{ki}) pour i = 1 à n observations, les indices 1 et k prenant les valeurs 1,2 - 3,4 - 5,6 ... ou 2,3 - 4,5 ...

En représentant, par exemple, sur graphique cartésien tous les couples (y_{1i}, y_{2i}) valeurs des 1ère et 2ème C.P., on examinera :

- si certains points sont voisins, donc s'il y a ressemblance de certains individus i et si cette ressemblance se maintient pour y_{3i} et y_{4i}, etc, la proximité sera d'autant meilleure et l'analyse facilité; ceci permet de cartographier l'isohyète moyenne d'épisodes analogues en répartition spatiale ;

- si les nuages de points apparaissent plus ou moins séparés (familles d'analogues caractérisant des états particuliers distincts du phénomène) plus ou moins agrégés (redondance des mesures ou fréquence d'apparition du même état).

On peut définir la signification physique d'une C.P. (Y_1) en mettant en corrélation les n valeurs y_{1i} avec les n observations d'un phénomène physique externe.

On pourra également associer tout point du plan (y_{1i}, y_{ki}), pour i = 1 à n, à des valeurs (dichtomiques comme 0 et 1 ou continues) caractérisant les états d'un autre phénomène lié aux valeurs des composantes, et matérialiser la ou les frontières entre nuages de points de même cote. Nous verrons dans la suite de cette note comment quantifier cette description à l'aide de l'analyse discriminante.

6.3.2 - Formalisation et exploitation numérique des résultats de l'A.C.P.

6.3.2.1 - Régression orhtogonale

On veut établir une relation multilinéaire entre une variable principale V et des variables explicatives $X_1, \ldots, X_j, \ldots, X_p$. On dispose pour cela de n observations sur ces (p+1) variables :

$$\begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} \qquad \begin{bmatrix} x_{11} & \cdots & x_{p1} \\ \vdots & & \vdots \\ x_{1n} & & x_{pn} \end{bmatrix}$$

En utilisant la méthode des moindres carrés, la détermination des coefficients "b" de cette relation ($V = b_o + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p + \varepsilon$) consiste à minimiser $\sum_{i=1}^{n} \varepsilon_i^2$ avec :

$$\varepsilon_i = v_i - (b_o + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_p x_{pi})$$

Cette opération est toujours possible et parfaitement rigoureuse quelle que soit l'importance des corrélations entre variables explicatives X_j. Toutefois lorsqu'on a affaire à de petits échantillons ($15 \leqslant n \leqslant 50$), du fait des fluctuations d'échantillonnage, des coefficients de corrélation r_{jk} entre plusieurs couples de variables $[X_j, X_k]$ voisins de 1 (colinéarité) peuvent rendre l'estimation de certains coefficients de régression instable. Une procédure, proposée pour remédier à cet inconvénient, consiste à orthogonaliser le système $(X_1, \ldots, X_p)$ puis à ne retenir que les q premières C.P. qui expliquent entre 90 et 98 % de la variance totale; s'il y a des colinéarités, le rang de la matrice de corrélation est inférieur à p, ce qui revient à éliminer les dernières C.P.

Cette procédure a son intérêt et est recommandée lorsqu'on traite des variables X_j homogènes (en ne mélangeant pas des observations de débits, de pluies et de températures). Toutefois lorsque les variables X_j sont trop hétérogènes, en particulier en variabilité, cette méthode est sujette à caution car, en éliminant les composantes ayant les plus faibles valeurs propres, on peut se priver d'une information importante pour la prévision, précisément située dans ces éléments.

6.3.2.2. - Rationalisation d'un réseau de mesure

C'est un problème à la mode : disposant d'un réseau de stations de mesures trop dense et redondant, dont l'entretien est onéreux, on se propose de "l'optimiser".

Il faudrait préciser ce que l'on entend par "optimiser" : par rapport à quel type d'information et selon quel critère. Il est toujours difficile de préjuger des besoins à venir, et même dans une situation actuelle il est presque impossible de s'appuyer sur des considérations exhaustives pour définir rigoureusement l'optimisation.

C'est pourquoi nous ne traiterons que d'un aspect limité et partiel de ce problème : rationalisation relativement à la mesure actuelle et passée du phénomène étudié. Un autre aspect important est celui de l'unité de temps de la mesure; veut-on rationaliser par rapport à une information moyenne annuelle - mensuelle - hebdomadaire - journalière - horaire ... ?

Il faut se garder d'éliminer toute redondance, car elle peut être utile pour atténuer l'effet d'erreurs de mesures aléatoires; lorsqu'on exploite les résultats de mesures d'un ensemble de stations, la redondance permet un lissage des erreurs accidentelles.

Une forte redondance entre stations est bien mise en évidence par les analyses graphiques 6.3.1.1.1., toutefois il est souhaitable de la compléter par l'analyse cartographique 6.3.1.1.2.

Le choix des stations témoins doit satisfaire aux conditions suivantes :

- bien corrélées avec la moyenne générale (bonnes corrélations avec la première composante) ;

- représentatives de groupements homogènes caractéristiques de la typologie; bonne corrélation avec les 2ème et 3ème C.P., éventuellement les 4ème et 5ème C.P. ;

- perdre le minimum d'information, c'est-à-dire que le coefficient de corrélation multiple entre chaque station et les q premières composantes

doit être aussi voisin de 1 que possible et, en particulier, le coefficient de corrélation multiple de la station témoin d'un même groupement avec les q C.P. doit être le plus important parmi les stations de ce groupe ;

- conserver un minimum de redondance en retenant 2 ou 3 témoins par groupe.

Une procédure numérique analogue à la régression descendante consisterait à :

(a) - calculer les q premières C.P. qui représentent par exemple 95 % de la variance totale,

(b) - calculer les C.P. de toutes les associations possibles de (q-1) stations parmi q,

(c) - retenir le groupement des (q-1) stations dont les premières C.P. sont les mieux corrélées avec les C.P. de (a),

(d) - calculer les C.P. de toutes les combinaisons de (q-2) stations prises dans le sous-ensemble (c) de (q-1) stations,

(e) - retenir le groupement des (q-2) stations dont les premières C.P. sont les mieux corrélées avec les C.P. de (a),

(f) - réitérer l'opération avec toutes les combinaisons (q-3) parmi le groupement retenu en (e).

La sélection s'arrête d'après l'un ou l'autre des critères suivants : on a fixé le nombre des stations du réseau réduit - on se fixe un niveau minimum de corrélation entre les premières composantes des p stations et les composantes de (p-k) stations.

6.3.2.3. - Reconstitution partielle ou totale d'une série d'observations en une station.

6.3.2.3.1 - Reconstitution partielle d'une série

On dispose de n observations sur p stations et de m observations communes ($m < n$) sur une $(p+1)^{ème}$ station :

$$\left\{\begin{array}{l} m \\ n-m \end{array}\right. \left[\begin{array}{ccc} X_1 & : X_p & X_{p+1} \\ x_{11} & : x_{p1} & x_{p+1,1} \\ \vdots & \vdots & \vdots \\ x_{1m} & : x_{pm} & x_{p+1,m} \\ x_{1m+1} & : x_{pm+1} & 0 \\ \vdots & \vdots & \vdots \\ x_{1n} & : x_{pn} & 0 \end{array}\right]$$

On calcule les q premières C.P. des p stations $(X_1, \ldots, X_p)$ d'après le tableau des $p \times n$ observations.

On établit la corrélation multiple entre X_{p+1} et ces q composantes sur la période commune (m observations); puis on applique la relation linéaire ainsi obtenue aux $(n-m) \times p$ observations de la période, on obtient ainsi une estimation des valeurs $\hat{x}_{p+1,\ m+1}$ à $\hat{x}_{p+1,\ n}$, affectées d'une marge d'incertitude calculée.

6.3.2.3.2. - Reconstitution totale d'une série en un lieu

Pour cela, on utilisera les propriétés de la représentation cartographique étudiées au paragraphe 6.3.1.1.1.

Rappelons que si l'on considère un réseau de p stations et que l'on calcule les q premières C.P. :

$$Y_l = \frac{1}{\sqrt{\lambda_l}} \sum_{j=1}^{p} a_{lj} \left(\frac{X_j - M_j}{\sigma_j}\right)$$

réciproquement, on peut exprimer les variables X en fonction de Y, en transposant la matrice des coefficients a_{lj} (l'inverse d'une matrice orthogonale est égale à sa transposée) :

$$\frac{X_j - M_j}{\sigma_j} = \sum_{l=1}^{q} a_{jl} \left(\sqrt{\lambda_l}\, Y_l\right) + \varepsilon_j$$

ε_j étant une variable de moyenne nulle et d'écart type :

$$\sigma_{\varepsilon j} = \left(1 - \sum_{l=1}^{q} a_{jl}^2 \lambda_l\right)^{\frac{1}{2}}$$

On cartographie donc les moyennes M_j, les écarts types Σ_j, les cosinus directeurs relatifs aux q premières C.P. puis l'on trace les lignes d'égales valeurs de chacune de ces grandeurs respectivement; cette opération nécessite une densité importante de stations et également une répartition spatiale bien homogène (équirépartition) de ces dernières. Par interpolation avec les valeurs voisines on calculera M_t, σ_t, a_{1t}, a_{2t}, ..., a_{qt} sur chacune des cartes; on pourra alors calculer les n observations de cette station fictive X_t à l'aide du tableau des nq valeurs des q composantes Y, d'après la relation :

$$X_t = M_t + \sigma_t \sum_{l=1}^{q} a_{tl}\left(\sqrt{\lambda_l}\, Y_l\right) + \sigma_t \pm \varepsilon_t$$

l'écart type du $\sigma_t \pm \varepsilon_t$ étant estimé par :

$$\sigma_t \pm \left[1 - \left(\sum_{l=1}^{q} a_{tl}^2 \lambda_l\right)\right]^{\frac{1}{2}}$$

La légitimité de cette méthode suppose la stationnarité spatiale des moyennes - écarts types - cosinus directeurs, qu'il n'y ait pas de discontinuité entre les points de mesure : l'isotropie des champs de mesures. Il n'est pas recommandé de reconstituer les stations qui se trouvent en bordure de la grille.

6.3.2.4. - Critique des données

La détection d'une ou deux valeurs très erronées dans une série de mesures X_j peut se faire de deux façons :

- on considère le tableau des cosinus directeurs relatifs à chaque composante principale; si dans l'une des composantes les cosinus directeurs sont faibles, sauf celui relatif à X_j (qui peut être très voisin de 1), on identifie la station douteuse. Généralement cette station est peu corrélée avec les autres et occupe une place particulière dans les C.P. L'examen comparatif de cette série avec des séries de stations voisines permet de détecter l'erreur ;

- toutefois on peut détecter l'erreur en calculant les q premières C.P sur l'ensemble des stations puis en reconstituant la série de chaque station par corrélation multiple entre X_j et Y_l ($l = 1$ à q) selon la procédure du paragraphe 6.3.2.3.2. On teste alors l'écart entre valeurs reconstituées et valeurs observées pour chaque station; par exemple si cet écart est supérieur à 2 ou 3 fois l'écart type $\sigma_t \neq \varepsilon_t$, on identifie la ou les observations suspectes.

S'il s'agit d'une erreur extrêmement importante dans une série, elle se répercutera dans la plupart des valeurs correspondantes, reconstituées par l'intermédiaire des composantes (pour cette observation). Le test des écarts détectera alors une anomalie très significative sur un ensemble de stations, alors que celle-ci n'est due qu'à une erreur localisée dans le temps et dans l'espace. Si l'erreur est importante mais non excessive, le test sur les écarts identifiera uniquement l'observation et la station en cause.

La détection d'une erreur systématique ou hétérogénéité dans une série présente plus de difficultés. Sur le plan formel, c'est encore la matrice de corrélation entre tous les couples de variables qui sera le révélateur, à travers les composantes principales.

En effet, la station suspectée d'hétérogénéité aura une moins bonne corrélation avec le reste du réseau. Cela va se traduire par de faibles coefficients de corrélation entre cette série et les 1ère, 2ème, éventuellement 3ème C.P. alors que la série sera fortement corrélée avec une composante d'ordre K. On peut alors reconstituer les p séries à l'aide des q premières composantes, puis tracer des lignes d'écarts cumulés $\Sigma(X_{ji}-X'_{ji})$

en fonction de i pour $1 \leqslant i \leqslant n$; on observera en particulier une "cassure" ou discontinuité, pour la série qui présente une hétérogénéité.

6.3.2.5. - Prévision

6.3.2.5.1 - Dans les prévisions à court terme, moyen terme ou long terme, il s'agit essentiellement de calculer les probabilités d'occurrence de débits ou d'apports, indépendamment du temps, puisque ce sont des valeurs moyennes ou cumulées sur quelques jours (1 à 10 jours).

On utilise la régression orthogonale $Y = f(X_1, \ldots, X_p, X_{p+1}, \ldots, X_p)$ mais en calculant les C.P. d'un ou plusieurs sous-ensembles de variables homogènes $(X_1, \ldots, X_p)$, $(X_m, X_{m+1}, \ldots X_k)$, etc. Car il est évident que la répartition spatiale d'un phénomène tel que la précipitation est mieux caractérisée par les 3 ou 4 premières C.P. d'un réseau de stations pluviométriques que par une simple moyenne. D'ailleurs la corrélation multiple permettra d'attribuer à chaque station des pondérations plus adaptées et plus objectives, par l'intermédiaire des C.P.

Il est recommandé de n'utiliser, dans de tels calculs, que 2 à 4 composantes structurées, un trop grand nombre de variables explicatives d'une part, et l'adjonction de variables trop spécifiques ou éphémères d'autre part, risque d'aboutir à des corrélations factices.

6.3.2.5.2.- Dans le calcul de certaines régressions multiples

Après une analyse en C.P. d'un champ de mesure, qui a pour effet de réduire les dimensions et orthogonaliser les variables, on peut définir sur ces composantes un domaine borné dans lequel les relations entre un phénomène à prévoir et ces variables sont approximables par des régressions multiples (on a linéarisé le domaine d'application). Ce qui n'est pas le cas lorsque l'on considère l'espace complet de variation des paramètres.

Cette technique est utilisée dans la "Reconnaissance dynamique de la forme des situations météorologiques".

Chaque jour à 0 h, l'état de l'atmosphère, sur l'Europe et la bordure atlantique, est caractérisé par 37 niveaux des surfaces 700 mb et 1000 mb (obtenus d'après les mesures de radiosondages). Un fichier ayant été constitué pour les mois d'hiver depuis 1959, soit plus de 3 000 journées , on orthogonalise chacun de ces 2 ensembles de 37 variables pour les réduire à six composantes principales (Y_K, Z_K, pour K = 1 à 6). Ce fichier est complété par les 6 variations ΔZ_K en 24 heures des composantes Z_K, ainsi que par la précipitation en 24 heures R_1 (8 h - 8 h) de 33 groupements pluviométriques. Soit y_{0K} (K = 1 à 6) les valeurs caractérisant la carte d'aujourd'hui, on recherche dans le fichier historique la trentaine de situations analogues qui se trouvent dans la boule de proximité centrée sur y_{0K} : $\sum_{K=1}^{6} (y_{iK} - y_{0K})^2 \leqslant d^2$ et i varie de 1 à 30. Le rayon de cette boule étant indicé à la distance à l'origine. On extrait dans un fichier provisoire les valeurs z_{iK}, Δ_{iK}, R_{i1}, associées à ces analogues et l'on établit l'équation de régression multiple $\sqrt{R_1} = \Sigma \alpha_{1K} Z_K + \Sigma \beta_{1K} \Delta Z_K + C_1$ en éliminant les variables explicatives dont le coefficient de corrélation partiel avec $\sqrt{R_1}$ n'est pas significativement différent de 0. Puis on calcule $\sqrt{R_{01}} = \Sigma \alpha_{1K} z_{0K} + \Sigma \beta_{1K} \Delta z_{0K} + C_1$ affecté d'un intervalle de confiance proportionnel à l'écart type lié de la régression. Les variations de la surface 1000 mb en 24 h sont obtenues en calculant les valeurs de C.P. z_{1K} de la situation du lendemain, prévue par la Météorologie Nationale.

Comme conclusion à ce rapide inventaire, on peut faire les remarques suivantes, à la suite de nombreuses applications et essais d'analyse : lorsque l'on cherche à prévoir l'écoulement pendant un intervalle de temps donné, en plusieurs stations (usines hydroélectriques - stations limnimétriques) d'une même région ou de plusieurs régions, en fonction des précipitations et écoulements passés, il est préférable de mettre individuellement en corrélation les variables à prévoir avec les C.P. des évènements passés, tels que précipitations spatiales ou enneigements, plutôt que d'associer les C.P. de 2 ensembles de mesures, car on caractérise mieux ainsi le comportement spécifique de chaque bassin versant.

Lorsqu'il y a forte homogénéité entre plusieurs points de mesures, et s'il existe plusieurs ensembles de stations bien typés par un ou deux témoins, on pourra effectuer une prévision individuelle sur chacun de ces témoins puis éclater cette prévision sur toutes les stations par l'intermédiaire de composantes principales.

6.3.2.6 - Calcul de lois de probabilité multidimensionnelles

On peut, par l'intermédiaire des C.P., évaluer la probabilité d'occurrence d'un évènement spatial, résultant de la conjonction d'évènements ponctuels, ce qui serait impossible sur les variables initiales du fait des intercorrélations.

6.4 - APPLICATIONS DE L'A.C.P. DANS LE DOMAINE TEMPOREL

On considère à présent des variables temporelles, c'est-à-dire la discrétisation d'une fonction aléatoire : A_1, ... A_t, ..., A_p. Ce sera le débit moyen en 24 heures de chacun des 365 jours de l'année en un lieu, les apports des 52 semaines, les températures décadaires pendant une saison, etc.

L'avantage de cette technique sur l'analyse spectrale est qu'il est plus facile de faire abstraction de la non-stationnarité (au second ordre) du processus, puisqu'elle est prise en compte dans les calculs de l'A.C.P. par l'intermédiaire des moyennes - variances - coefficients de corrélation totale entre couples A_t et $A_{t\pm k}$.

On peut transposer au domaine temporel la plupart des applications de l'A.C.P. qui ont été décrites précédemment dans le domaine spatial :

- analyse descriptive dans l'espace des variables, qui met éventuellement en évidence des relations entre les observations à deux époques distinctes de l'année,

- analyse dans l'espace des observations, qui montre l'analogie de séquences d'évènements pour 2 ou plusieurs années passées,

- critique des données,

- prévision.

La corrélation entre composantes, des apports hebdomadaires du printemps-été pour des rivières à alimentation nivale prédominante et les précipitations, écoulements et températures de l'hiver précédent, permet de prévoir, non seulement l'importance de l'hydrogramme de fusion nivale, mais également sa forme.

On peut proposer une "modèlisation" des premières composantes principales empiriques, pour lisser les dispersions d'échantillonnage des cosinus directeurs.

Notons D_t les variables A_t centrées réduites :

$$Z_K = \Sigma \alpha_{tK} D_t = \sum_{t=1}^{p} (a_K t + b_K) D_t \quad \text{avec} \sum_{t=1}^{p} (a_K t + b_K)(a_{jt} + b_j) = 0 \text{ si } K \neq j, \; = 1 \text{ si } K = j$$

éventuellement $\alpha_t = a_K t + b_K$ pour $1 \leqslant t \leqslant r$
$\alpha_t = c_K t + d_K$ pour $r < t \leqslant q$
$\alpha_t = f_K t + g_K$ pour $q < t \leqslant p$

représentation en lignes brisées.

(un modèle consiste à rendre nul le coefficient de t et prendre le terme constant égal à $\frac{1}{\sqrt{p}}$).

Inversement, il est possible de calculer les variables A_t en fonction des 3 ou 5 premières C.P. :

$$A_t = \sum_{K=1}^{5} \beta_{Kt} Z_K + \beta_{0t} + \varepsilon_t \qquad (1)$$

on peut effectuer un lissage des coefficients β_{Kt}, pour $1 \leqslant t \leqslant p$, qui modulent l'effet de la C.P. Z_K, d'après :

$$\hat{\beta}_{Kt} = \beta_K(t) = \sum_{m=0}^{3} \left(\gamma_{Km} \cos m \frac{2\pi t}{p} + \delta_{Km} \sin \frac{2\pi t}{p} \right)$$

On établit alors une corrélation multiple entre les composantes Z_K et les précipitations et écoulements d'hiver (P_s) :

$$Z_K = \Sigma B_{Ks} P_s + B_K \qquad ②$$

le produit des matrices de coefficients β_{Kt} et B_{Ks} permet d'obtenir :

$$A_t = \Sigma C_{ts} P_s + C_t \qquad ③$$

et donc de calculer la prévision de A_t.

Simulation : connaissant les fonctions de répartition de Z_K et ε_t dans ①, on peut simuler aisément des suites de A_t $(1 \leqslant t \leqslant p)$; en utilisant la relation ③ on peut également effectuer une simulation spatio-temporelle sur plusieurs bassins versants voisins soumis au même régime météorologique, en simulant les composantes principales des précipitations (cohérence spatiale), puis en générant par bassin des chroniques A_t d'après ③, la relation ② assurant la cohérence temporelle propre à chaque bassin.
(Exemple de prévision des apports hebdomadaires d'été de la DURANCE à SERRE-PONCON).

6.5 - QUELQUES REMARQUES PRATIQUES SUR L'A.C.P.

Lorsqu'on calcule les C.P. d'un ensemble de variables représentant un phénomène spatial ou temporel, on a intérêt à calculer les C.P. sur domaine spatial ou temporel plus étendu que le domaine utile, pour éliminer les "effets de bord".

La densité spatiale ou temporelle des variables X a son importance dans le calcul des C.P.; voici 2 exemples extrêmes :

- calculer les C.P. d'un ensemble ultra redondant, la pression atmosphérique mesurée toutes les minutes entre 9 h et 12 h le 10 janvier pendant n années, ou encore les C.P. de la température moyenne de l'air en décembre,

sur 1 km², mesurée par une centaine de thermomètres uniformément répartis sur cette surface (colinéarité + bruit = 1 composante + des composantes bruits) ;

- il y a discontinuité, particularité locale ou instantanée, dont le réseau de mesures ne rend pas compte, le réseau étant trop lâche; anisotropie des champs de mesure.

Le nombre d'observations a également son importance : avec 30 observations, il vaut mieux s'en tenir aux 2 ou 3 premières C.P. alors qu'avec 1000 observations on peut calculer 8 ou 10 C.P. avec garantie de stabilité des cosinus directeurs respectifs.

L'analyse préliminaire de la distribution empirique des observations de chaque série peut suggérer une transformation simple qui pondère l'influence de valeurs extrêmes.

Il faut se garder des habituels pièges relatifs aux coefficients de corrélation simple entre variables : non linéarité - hétéroscédasticité - hétérogénéité - points aberrants, etc.

Enfin, lorsqu'on calcule les C.P. de variables résultant de cumuls progressifs, il y a déformation continue des cosinus directeurs et valeurs propres correspondants.

Lorsqu'on calcule les C.P. sur plusieurs échantillons d'observations effectuées sur le même phénomène, il peut arriver que pour une même composante il y ait inversion de signes ou qu'il y ait inversions dans la suite des composantes.

6.6 - CONCLUSIONS

L'analyse en composantes principales est un outil puissant, efficace et assez objectif pour traiter des tableaux de données, particulièrement les mesures de phénomènes physiques homogènes.

Elle permet une synthèse et une meilleure appréhension de l'information (volume d'observations); elle peut également orienter la recherche d'une typologie, aussi bien dans le domaine des variables que dans celui des observations.

Cette méthode n'est cependant pas universelle, on ne peut en faire une application massive et sans discernement. Ses limites sont essentiellement :

- les corrélations totales entre variables doivent être linéaires ou linéarisables par une transformation simple ;

- on ne traite qu'une partie de l'information puisque les calculs ne font intervenir que les moments de 1er ou 2ème ordre (moyenne, écart type, covariance) ;

- elle est sensible à la dissymétrie de la répartition en fréquence des observations, on peut d'ailleurs s'affranchir de cet inconvénient en traitant la matrice des coefficients de corrélation de rang ;

- on connaît assez mal la distribution d'échantillonnage des composantes principales individuellement (cosinus directeurs et valeurs propres) ;

- de même, on ignore l'incidence réelle des erreurs de mesures faites sur les observations, dans le calcul des composantes principales ;

- le calcul des C.P. sur des ensembles de données hétérogènes (températures et pluies par exemple) est une opération délicate et sujette à caution, cela nécessite une certaine prudence dans l'interprétation et l'utilisation des résultats.

Ces restrictions étant faites, les applications de l'A.C.P. sont nombreuses, à la fois lorsqu'on traite les mesures ponctuelles d'un phénomène spatial ou la chronique des mesures, en un lieu, d'un phénomène temporel.

Dans ce dernier cas, en particulier, cette technique paraît supérieure et plus prometteuse au plan opérationnel que l'analyse spectrale.

o

o o

HIVER 1959-1968

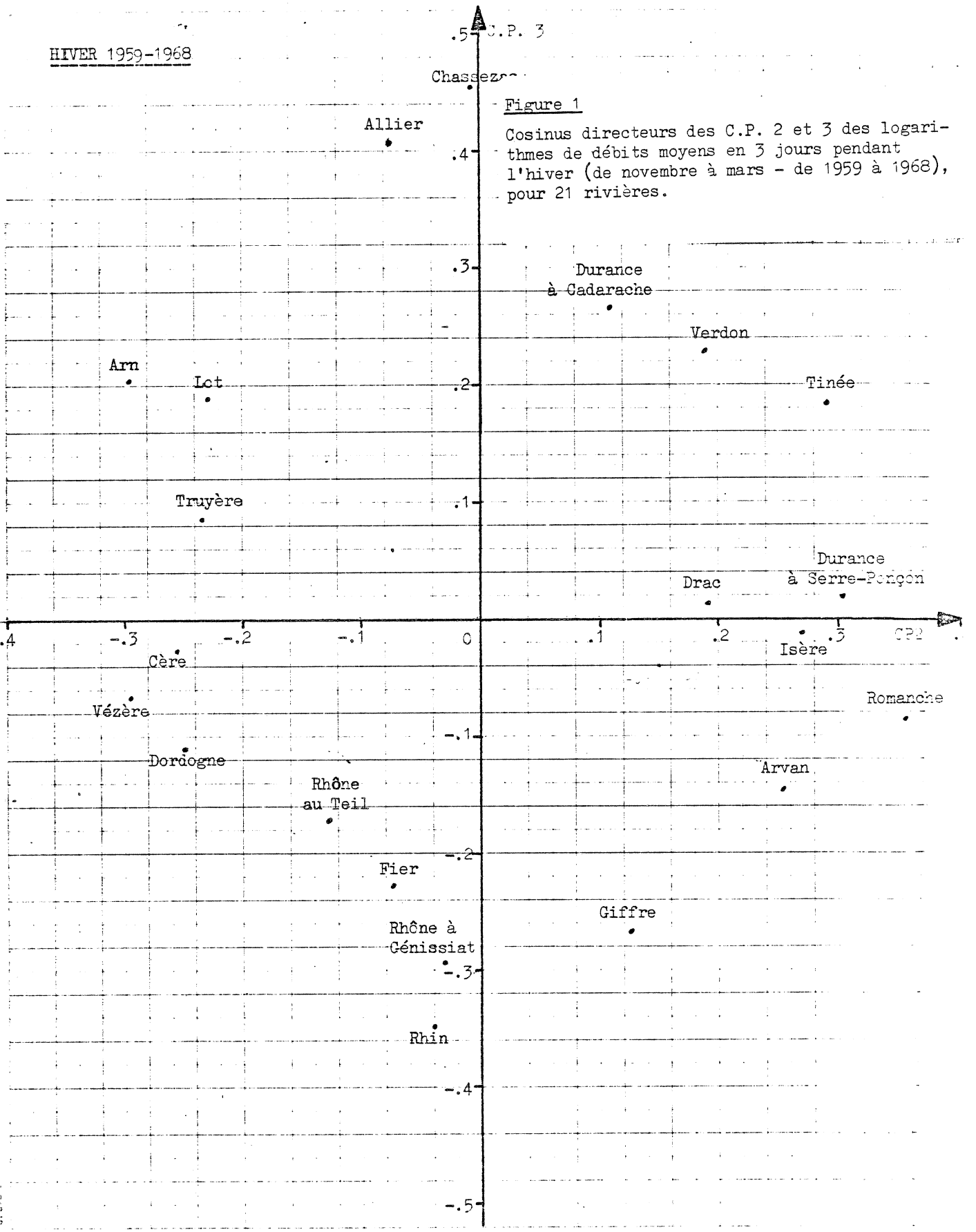


Figure 1

Cosinus directeurs des C.P. 2 et 3 des logarithmes de débits moyens en 3 jours pendant l'hiver (de novembre à mars - de 1959 à 1968), pour 21 rivières.

PREVISION NUMERIQUE DES PRECIPITATIONS

Stations de radiosondage (altitude des surfaces 700 mb et 1 000 mb)

1. Point I
2. Torshaw'n
3. Oslo
4. Jokioinen
5. Point J
6. Stornoway
7. Kastrup
8. Valentia
9. Aughton
10. De Bilt
11. Emden
12. Lindenberg
13. Legionowo
14. Point K
15. Brest
16. Trappes
17. Payerne
18. Munich
19. Prague
20. Budapest
21. La Corogne
22. Bordeaux
23. Nimes
24. Milan
25. Zagreb
26. Belgrade
27. Bucarest
28. Lisbonne
29. Madrid
30. Palma
31. Cagliari
32. Rome
33. Brindisi
34. Funchall
35. Gibraltar
36. Malte
37. Poprad Tatry

o
o o

Niveaux de la surface 700mb en Hiver (dam)

1ère Composante principale

j =	a_{1j}	e_{1j}	e_{1j}^2	ℓ_{1j}^2	M moyenne	σ ecart type
1	0.000753	0.003150	0.000010	0.000010	287	14
2	0.035937	0.150275	0.022583	0.022583	286	13
3	0.110662	0.462745	0.214133	0.214133	289	13
4	0.076632	0.320448	0.102687	0.102687	287	12
5	0.033314	0.139307	0.019406	0.019406	294	14
6	0.081013	0.338766	0.114762	0.114762	290	13
7	0.168655	0.705252	0.497381	0.497381	292	12
8	0.112520	0.470515	0.221384	0.221384	296	13
9	0.152743	0.638715	0.407956	0.407956	295	13
10	0.204469	0.855011	0.731044	0.731044	295	11
11	0.197198	0.824508	0.679978	0.679978	295	11
12	0.207288	0.866802	0.751345	0.751345	295	11
13	0.173789	0.726718	0.528119	0.528119	294	10
14	0.080566	0.336897	0.113500	0.113500	305	12
15	0.172960	0.723252	0.523093	0.523093	293	11
16	0.212285	0.887696	0.788004	0.788004	298	11
17	0.227171	0.949942	0.902390	0.902390	293	9
18	0.231541	0.968217	0.937445	0.937445	298	9
19	0.224208	0.937554	0.879008	0.879008	296	10
20	0.205329	0.858608	0.737208	0.737208	298	9
21	0.135373	0.566080	0.320446	0.320446	303	10
22	0.193421	0.808814	0.654180	0.654180	302	10
23	0.211386	0.883936	0.781343	0.781343	301	9
24	0.225132	0.941419	0.886269	0.886269	300	9
25	0.216910	0.907037	0.822716	0.822716	299	9
26	0.185787	0.776890	0.603558	0.603558	300	8
27	0.144076	0.602469	0.362969	0.362969	299	8
28	0.104363	0.436407	0.190451	0.190451	307	8
29	0.151500	0.633515	0.401342	0.401342	305	8
30	0.172853	0.722805	0.522447	0.522447	304	8
31	0.172039	0.719404	0.517542	0.517542	302	8
32	0.193354	0.808531	0.653723	0.653723	301	9
33	0.159321	0.666221	0.443850	0.443850	301	8
34	0.033992	0.142142	0.020204	0.020204	311	6
35	0.104803	0.438247	0.192060	0.192060	308	7
36	0.122381	0.511751	0.261389	0.261389	305	7
37	0.197133	0.824336	0.679530	0.679530	296	9

$\lambda_1 = 1.74859549$ & 1 2 1.74859549 & 1 $0.472593 = \frac{\lambda_1}{37}$ 1 0.472593 0.354218

2ème Composante principale

	a_{2j}	e_{2j}	e_{2j}^2	$e_{1j}^2 + e_{2j}^2$
1	0.276074	0.696222	0.484725	0.484735
2	0.325957	0.822021	0.675718	0.698301
3	0.266154	0.671204	0.450515	0.644648
4	0.176837	0.445959	0.198880	0.301567
5	0.192530	0.485534	0.235743	0.255149
6	0.331239	0.835340	0.697793	0.812555
7	0.205801	0.519002	0.269343	0.766744
8	0.209057	0.527213	0.277954	0.499338
9	0.243685	0.614541	0.377640	0.785616
10	0.158388	0.399433	0.159546	0.890591
11	0.184202	0.464533	0.215791	0.895768
12	0.113354	0.285862	0.081717	0.833002
13	0.087380	0.220361	0.048559	0.576678
14	0.001295	0.003265	0.000011	0.113510
15	0.076628	0.193246	0.037344	0.560437
16	0.051736	0.130472	0.017023	0.805027
17	-0.051893	-0.130866	0.017126	0.919516
18	-0.013071	-0.032962	0.001086	0.938551
19	0.037540	0.094671	0.008963	0.887971
20	-0.026215	-0.066111	0.004371	0.741578
21	-0.084280	-0.212542	0.045174	0.365620
22	-0.076866	-0.193846	0.037576	0.691756
23	-0.116912	-0.294837	0.086929	0.868272
24	-0.069993	-0.176513	0.031157	0.917426
25	-0.062323	-0.157171	0.024703	0.847419
26	-0.071055	-0.179191	0.032109	0.635668
27	-0.044625	-0.112538	0.012665	0.375634
28	-0.189938	-0.478997	0.229438	0.419889
29	-0.175980	-0.443793	0.196957	0.598298
30	-0.191245	-0.482295	0.232609	0.755056
31	-0.183799	-0.463516	0.214847	0.732389
32	-0.137694	-0.347245	0.120579	0.774302
33	-0.120491	-0.303862	0.092332	0.536182
34	-0.209196	-0.527564	0.278323	0.296528
35	-0.224438	-0.566002	0.320358	0.512419
36	-0.150106	-0.378547	0.143298	0.405187
37	0.011671	0.029433	0.000866	0.680396

$\lambda_2 = 6.359809018$ 0 10 2.384576392 1 2 0.644480 $= \frac{\lambda_1+\lambda_2}{37}$ 2 0.171887 $= \frac{\lambda_2}{37}$ 0.656897

	a_{3j}	e_{3j}	e_{3j}^2	$e_{1j}^2 + e_{2j}^2 + e_{3j}^2$
1	0.112287	0.276017	0.072909	0.557642
2	0.020608	0.049556	0.002456	0.700757
3	-0.100543	-0.241825	0.058479	0.723127
4	-0.165334	-0.397580	0.158070	0.459636
5	0.219624	0.528132	0.278924	0.534073
6	0.097894	0.235406	0.055416	0.867971
7	-0.108153	-0.260076	0.067639	0.834383
8	0.253014	0.608426	0.370182	0.869520
9	0.158835	0.382073	0.145980	0.931596
10	0.045278	0.108879	0.011855	0.902445
11	-0.012444	-0.029925	0.000895	0.896664
12	-0.112882	-0.271449	0.073685	0.906747
13	-0.193509	-0.465332	0.216534	0.793212
14	0.300460	0.722519	0.522033	0.635544
15	0.252557	0.607325	0.369844	0.929281
16	0.137133	0.329765	0.108745	0.913772
17	0.042082	0.101196	0.010241	0.929757
18	-0.047068	-0.113185	0.012811	0.951342
19	-0.098742	-0.237446	0.056381	0.944351
20	-0.176380	-0.424142	0.179897	0.921475
21	0.304055	0.731163	0.534600	0.900220
22	0.206199	0.495848	0.245866	0.937622
23	0.077790	0.187063	0.034992	0.903264
24	-0.034319	-0.082527	0.006811	0.924237
25	-0.129948	-0.312487	0.097648	0.945007
26	-0.195609	-0.470383	0.221260	0.856928
27	-0.216062	-0.519567	0.269950	0.645564
28	0.256191	0.616064	0.379534	0.799424
29	0.227589	0.547286	0.299522	0.897820
30	0.105591	0.253915	0.064473	0.819529
31	-0.045465	-0.109330	0.011953	0.744342
32	-0.106679	-0.256532	0.065809	0.840110
33	-0.175417	-0.421827	0.177938	0.714120
34	0.108015	0.259745	0.067467	0.365995
35	0.190215	0.457411	0.209225	0.721643
36	-0.143489	-0.345048	0.119058	0.524245
37	-0.188075	-0.452260	0.204544	0.884940

3ème Composante principale

$\lambda_3 = 5.782626828$ 0 2 2.962839058 1 0.800767 $= \frac{\lambda_1+\lambda_2+\lambda_3}{37}$ 3 0.156287 $= \frac{\lambda_3}{37}$ 0.946975

Figure 3

Représentation dans le plan des coefficients de corrélation entre les 3 premières composantes des niveaux de la surface 700 mb, mesurés à 0 h pendant 15 hivers et chacune des 37 stations de radiosondage du réseau d'observations.

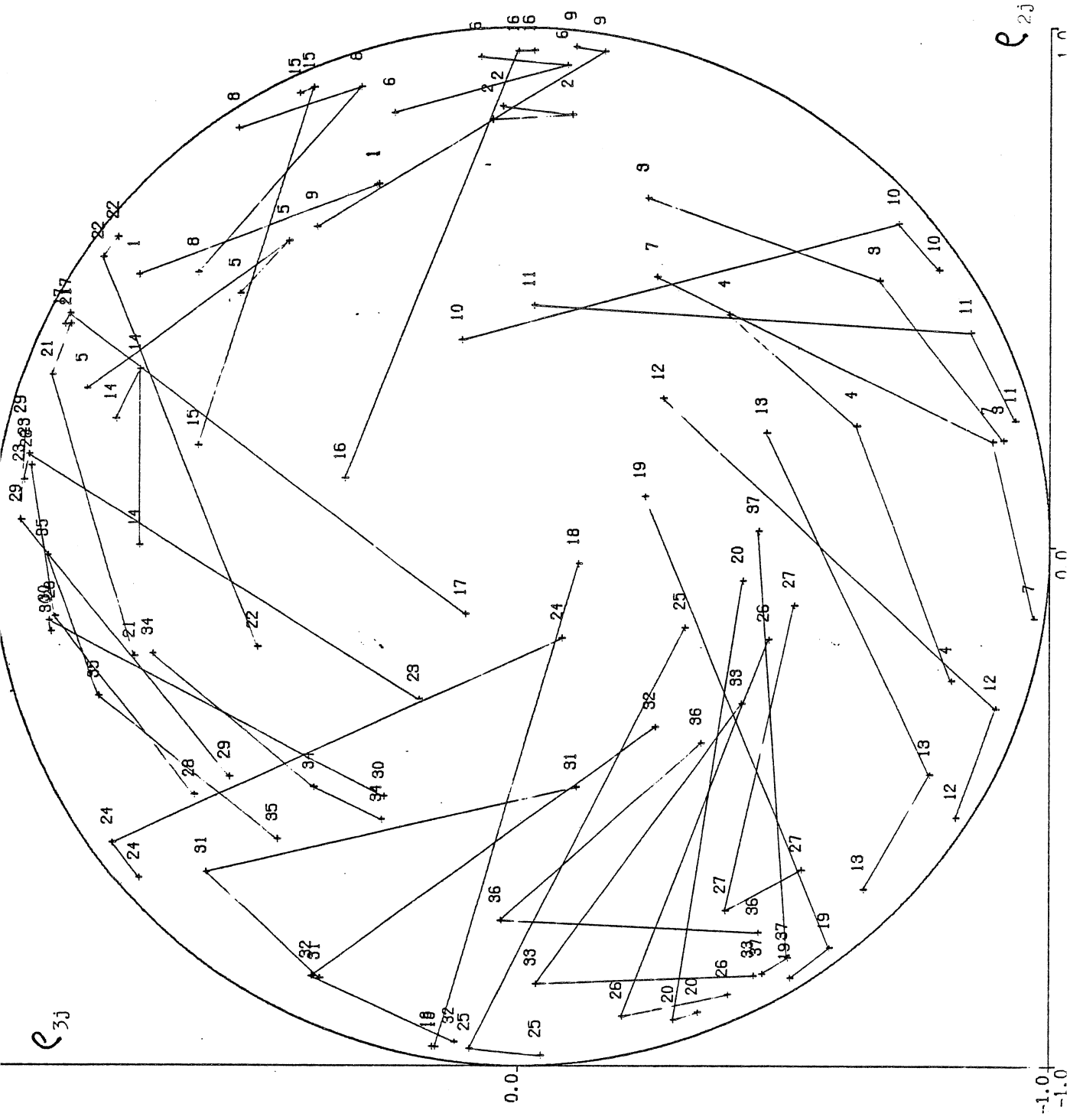


Figure 2

Représentation spatiale des cosinus directeurs et lignes d'isocosinus de la C.P. 2 du champ journalier des géopotentiels 700 mb (altitudes en 37 stations de la surface 700 mb) mesuré en hiver de 1959 à 1973.

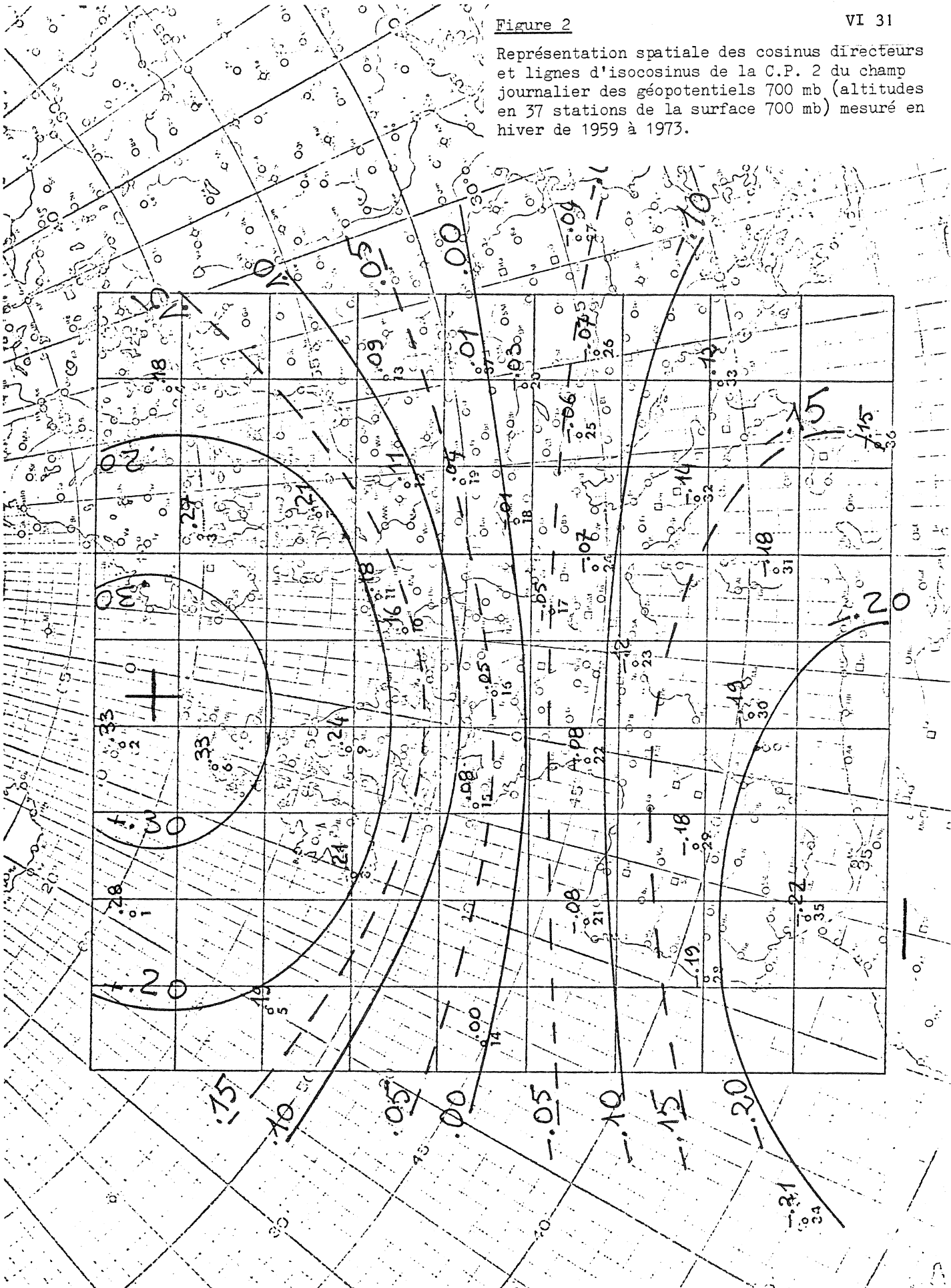


Figure 4

Ellipses de proximité à 80 % établies d'après les coefficients de corrélation entre les 2ème et 3ème composantes principales et les 37 stations de radiosondage.

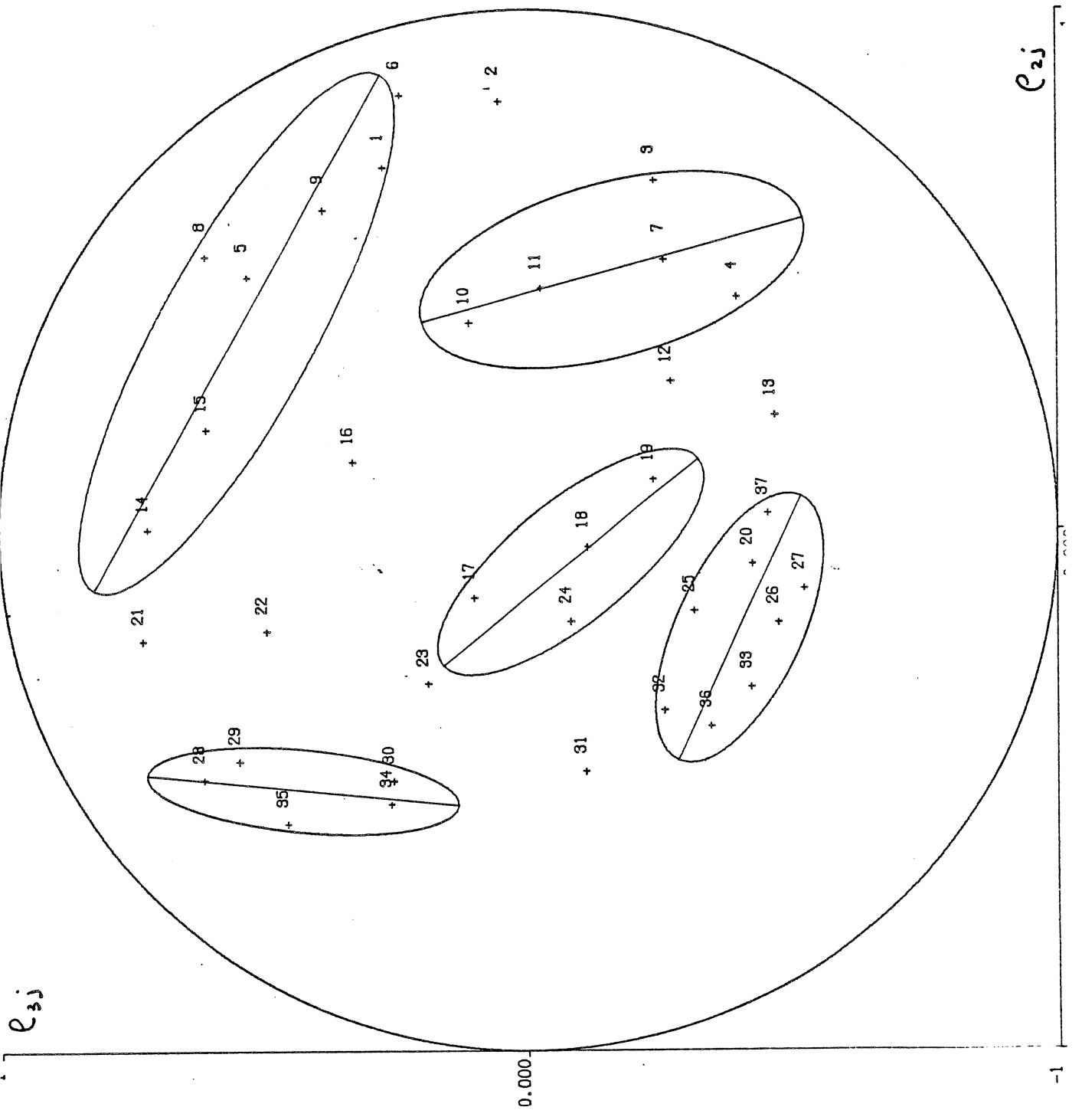
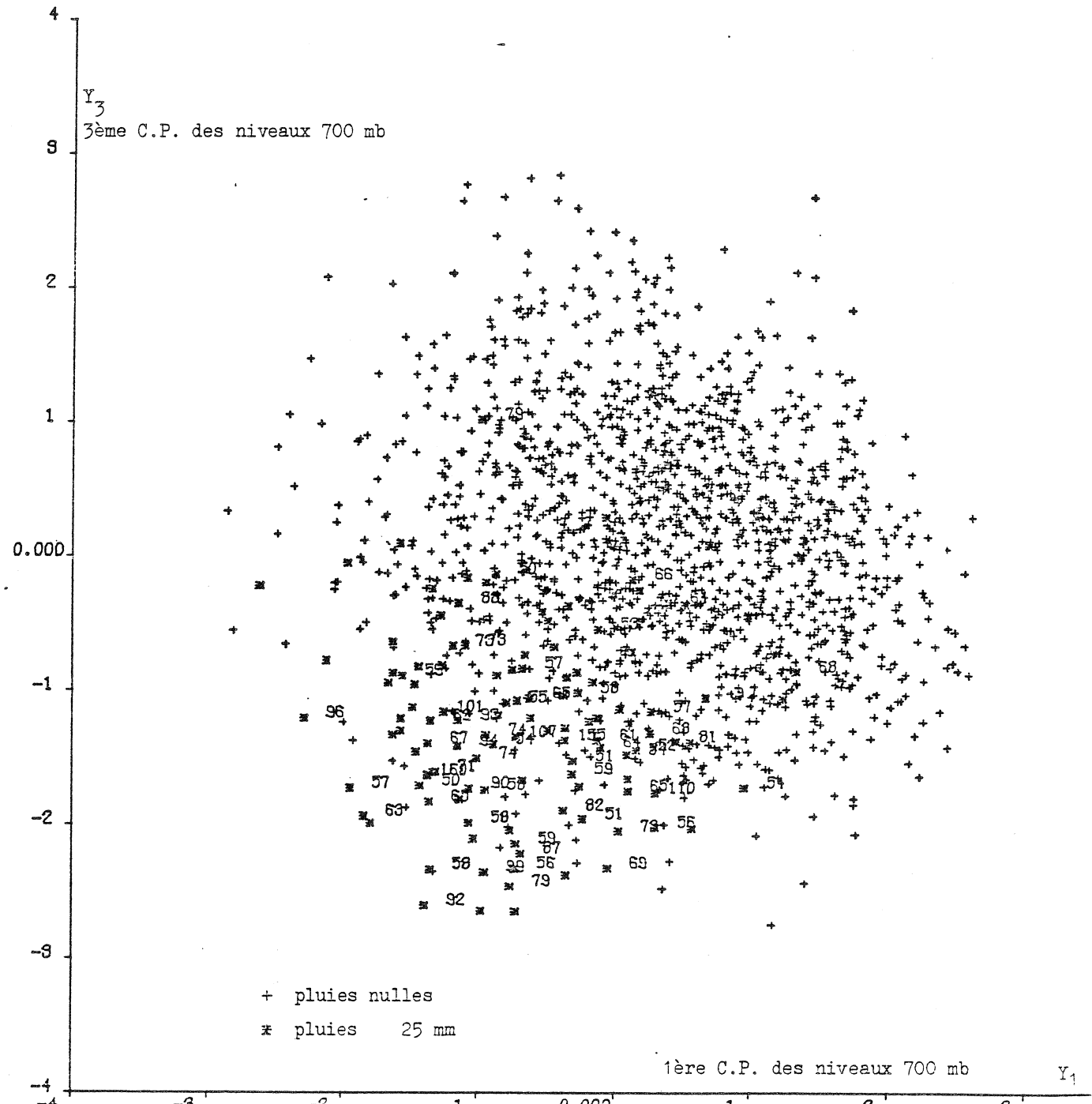


Figure 5

Représentation des valeurs des composantes principales 1 et 3 des niveaux journaliers de la surface 700 mb en hiver de 1959 à 1973 pour les jours sans pluies et les jours avec pluies supérieures à 25 mm sur le bassin du TARN.



P Y R E N E E S

Ecoulements de MAI

	Naguilhes	Lanoux	Izourt	Gnioure	Caillaouas	Bleu
1950	232	180	450	450	391	163
51	228	155	355	337	271	110
52	416	344	391	376	306	125
53	479	370	503	490	387	234
54	323	250	358	334	293	162
55	379	260	288	269	432	351
56	423	325	476	505	380	144
57	154	141	215	197	137	37
58	523	400	567	590	516	337
59	440	340	337	364	318	137
1960	478	370	412	441	518	314
61	431	329	365	386	313	241
62	359	294	313	358	274	160
63	295	271	318	305	208	104
64	464	360	381	415	597	406
65	366	285	451	428	228	139
66	472	353	478	489	377	223
67	383	310	396	404	215	66
68	370	320	423	449	242	95
69	417	359	403	447	372	181
1970	334	238	393	400	197	87
71	447	370	471	459	348	170
72	273	242	322	335	205	78

6 VARIABLES 23 OBSERVATIONS

Ecoulements (mm) de Mai 1950-1972

	NAGUILHES X 1	LANDUX X 2	IZOURT X 3	GNIOURE X 4	CAILLAOUAS X 5	BLEU X 6
M	377.7	298.5	394.2	401.2	327.2	176.7
S	93.3	71.7	78.9	85.3	115.0	97.8

R

1	1.0000					
2	0.9627	1.0000				
3	0.6363	0.5998	1.0000			
4	0.7030	0.6817	0.9651	1.0000		
5	0.6708	0.5775	0.4732	0.5347	1.0000	
6	0.6455	0.5277	0.3119	0.3566	0.9194	1.0000

NB DE VARIABLES DE LIAISON 6
ORDRE DE CES 6 VARIABLES

	1	2	3	4	5	6						
			(a)		(b)		(c)	(d)				
	1		0.452362		0.927938		0.861069	0.861069				
	2		0.427713		0.877376		0.769789	0.769789				
	3		0.388335		0.796599		0.634570	0.634570				
	4		0.413717		0.848665		0.720232	0.720232				
	5		0.401028		0.822636		0.676731	0.676731	(e)		(f)	(g)
λ_1 =	4.20790863&	0		1	4.20790863&	0			0.701318	1	0.701318	0.248823
	6		0.360057		0.738592		0.545518	0.545518				
	1		0.012937		0.013711		0.000188	0.861257				
	2		-0.065518		-0.069440		0.004822	0.774611				
	3		-0.488864		-0.516133		0.268462	0.903032				
	4		-0.443127		-0.469658		0.220578	0.940811				
	5		0.446494		0.473226		0.223943	0.900674				
λ_2 =	1.12332813&	0		2	5.33123676&	0			0.888539	2	0.187221	0.562506
	6		0.600695		0.636659		0.405335	0.950853				
	1		-0.472758		-0.351797		0.123761	0.985018				
	2		-0.621056		-0.462152		0.213584	0.988195				
	3		0.383205		0.285157		0.081315	0.984347				
	4		0.283320		0.210830		0.044449	0.985260				
	5		0.349793		0.260295		0.067753	0.968427				
λ_3 =	5.53741824&	-1		1	5.38697864&	0			0.980830	3	0.092290	0.782417
	6		0.203268		0.151259		0.022879	0.973732				

2	-0.237657	-0.052871	0.003953	0.992147					
3	0.325057	0.085992	0.007395	0.991741					
4	-0.187854	-0.049696	0.002470	0.987729					
5	-0.646182	-0.170943	0.029222	0.997649					
6	0.575342	0.152203	0.023166	0.996898					
λ_4 = 6.998325248 -2		2	5.954961898	0	0.992494	4	0,011664	0.834336	
1	-0.317427	-0.049820	0.002482	0.991279					
2	0.167082	0.026223	0.000688	0.992835					
3	-0.491753	-0.077180	0.005957	0.997698					
4	0.647555	0.101633	0.010329	0.998059					
5	-0.312863	-0.049103	0.002411	1.000060					
6	0.335101	0.052594	0.002766	0.999664					
λ_5 = 2.463291478 -2		8	5.979594808	0	0.996599	5	0.004105	0.856897	
1	-0.645708	-0.093385	0.008721	1.000000					
2	0.585368	0.084658	0.007167	1.000002					
3	0.340432	0.049235	0.002424	1.000122					
4	-0.312414	-0.045183	0.002041	1.000100					
5	-0.044957	-0.006502	0.000042	1.000102					
6	0.157758	0.022816	0.000521	1.000185					
λ_6 = 2.091608458 -2		1	6.000510898	0	1.000085	6	0.003486	0.876624	

Rappelons que la composante principale d'ordre l :

$$Z_l = \sum_{j=1}^{p} a_{lj} \frac{X_j - M_j}{\sigma_l}$$

colonne (a) : cosinus directeurs a_{lj} (j=1 à p)

colonne (b) : coefficients de corrélation ρ_{lj}

colonne (c) : ρ^2_{lj}

colonne (d) : coefficient de corrélation multiple entre la variable j et les K premières C.P.,

$$R^2_{X_j/Z_1, Z_2, \ldots, Z_K} = \sum_{l=1}^{K} \rho^2_{lj}$$

λ_l : valeur propre associée à Z_l, variance de cette C.P.

élément (e) : % de variance expliquée par les K premières composantes,

$$\sum_{l=1}^{K} \frac{\lambda_l}{p}$$

élément (f) : % de variance expliquée par la C.P. Z_l soit $\frac{\lambda_l}{p}$

élément (g) : entropie, $\sum_{l=1}^{K} \left(- \frac{\lambda_l}{p} \log_e \frac{\lambda_l}{p} \right)$

6 23 composantes principales

Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
-0.356	-0.328	3.176	-1.064	0.137	0.389
-1.232	0.056	1.732	0.023	-0.917	-1.070
0.027	-0.269	-1.155	-0.748	-2.011	0.755
1.123	-0.568	-0.012	1.018	-0.951	0.657
-0.600	0.365	0.220	0.488	-1.077	0.429
-0.185	2.696	0.072	1.592	-0.791	-0.398
0.657	-1.000	0.505	-1.343	-0.450	-1.371
-2.473	0.648	0.109	-0.291	-0.002	0.600
2.108	-0.383	1.018	0.771	0.847	0.421
-0.044	0.225	-1.595	-1.200	-0.974	-1.825
1.153	1.147	-0.058	-1.004	-0.210	-0.230
0.200	0.546	-0.855	1.524	1.367	-0.564
-0.474	0.396	-0.807	-0.265	2.027	-0.732
-1.019	0.072	-0.733	0.096	0.235	2.079
1.254	2.285	0.415	-0.909	0.558	0.421
-0.104	-1.035	0.215	1.990	-0.013	0.659
0.958	-0.504	-0.005	0.752	-0.174	-0.783
-0.332	-1.086	-0.913	-0.167	-0.355	-0.555
-0.065	-1.207	-0.373	-0.297	1.344	0.548
0.482	-0.134	-0.515	-1.487	1.225	0.560
-0.667	-0.937	0.206	1.109	0.334	-1.978
0.716	-0.747	-0.479	-0.118	-1.206	1.411
-1.126	-0.238	-0.171	-0.450	1.057	0.574

$$Y_j = \left\{ \sum_{i=1}^{6} a_{ij} \frac{X_i - M_i}{S_i} \right\} \frac{1}{\sqrt{\lambda_j}}$$

$(j = 1, 2, \ldots, 6)$

STATIONS PLUVIOMETRIQUES "LOIRE - CEVENNES"

DOUX - EYRIEUX

I Lamastre
2 St Agrève
3 Le Cheylard
4 St Pierreville
5 Vernoux

LOIRE SUPERIEURE

6 Ste Eulalie
7 Issanlas
8 La Palisse
9 Fay - sur - Lignon
IO Lac d'Issarlès

LOIRE MOYENNE

II Le Monastier
I2 Sanssac l'Eglise
I3 Fix St Geneys
I4 Le Puy Chadrac
I5 Chomelix
I6 Retournaguet
I7 Pont de Lignon
I8 Moulas
I9 Tarentaise
20 Mazet St Voy
2I Tence
22 Versilhac

ARDECHE - CHASSEZAC

23 Mayres
24 Montpezat
25 Antraigues
26 Loubaresse
27 Valgorge
28 Villefort
29 Vals - - les Bains
30 Aubenas
3I Joyeuse

ALLIER SUPERIEUR

32 St Etienne de Lugdarès
33 Langogne
34 St Sauveur de Ginestoux
35 Gandrieu
36 Monistrol d'Allier
37 Saugues
38 Faulhac en Margeride
39 Les Uffernets

CEZE

40 Malons
4I Genolhac
42 Collet de Dèze
43 St Maurice de Ventalon
44 St Etienne - vallée française
45 St André de Valborgne

HERAULT - ORB

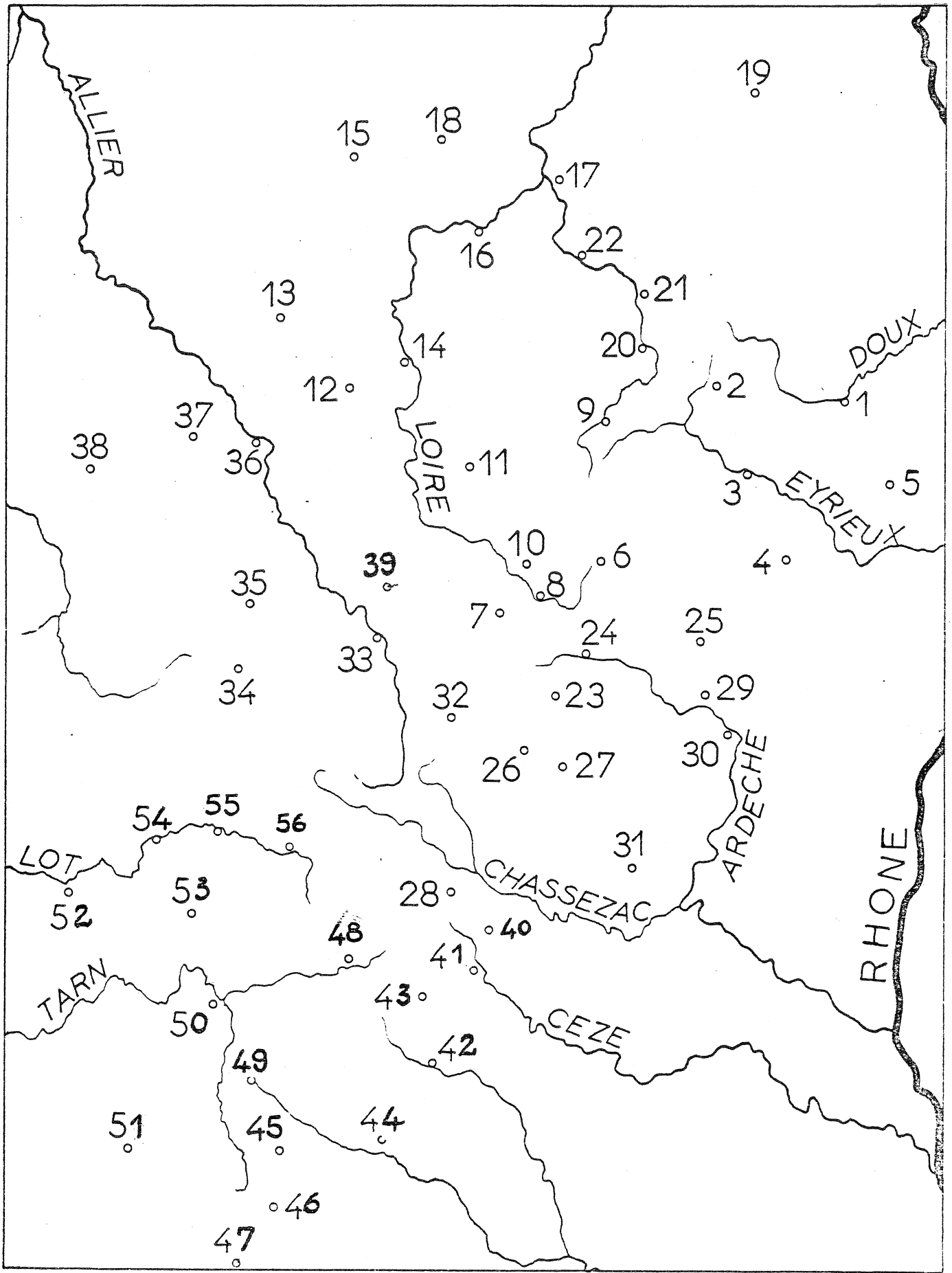
46 Mont Aigoual
47 Puechagut

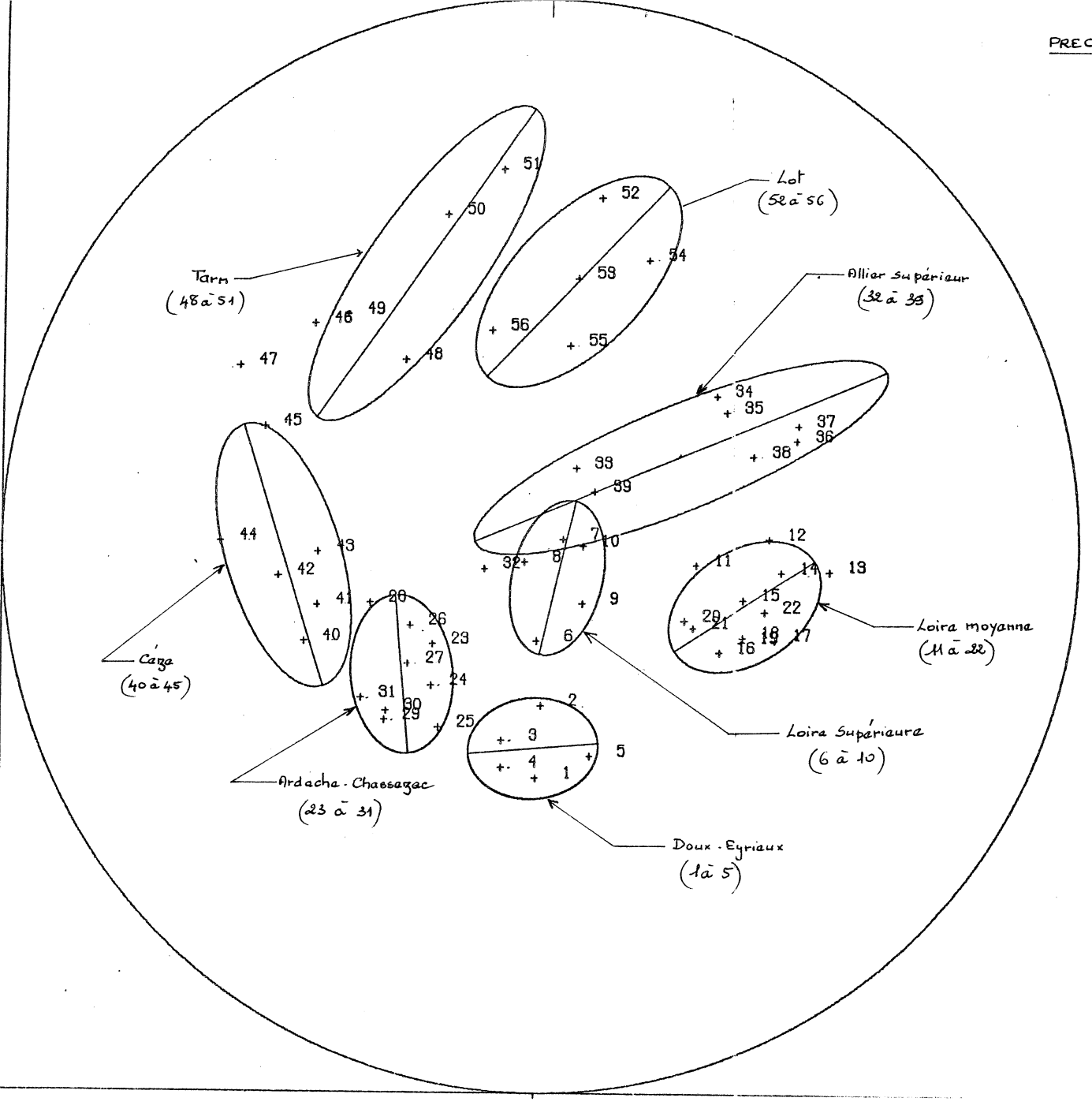
TARN

48 Pont de Montvert
49 Barre de Cévennes
50 Florac
5I Meyrueis

LOT

52 Chanac
53 Montmirat
54 Mende
55 Bagnols les Bains
56 Le Bleymard


ALLIER
LOIRE
DOUX
EYRIEUX
ARDECHE
CHASSEZAC
RHONE
CEZE
TARN
LOT
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56



PRECIPITATIONS OCTOBRE - NOVEMBRE

22 années (1951 à 1972)

56 stations

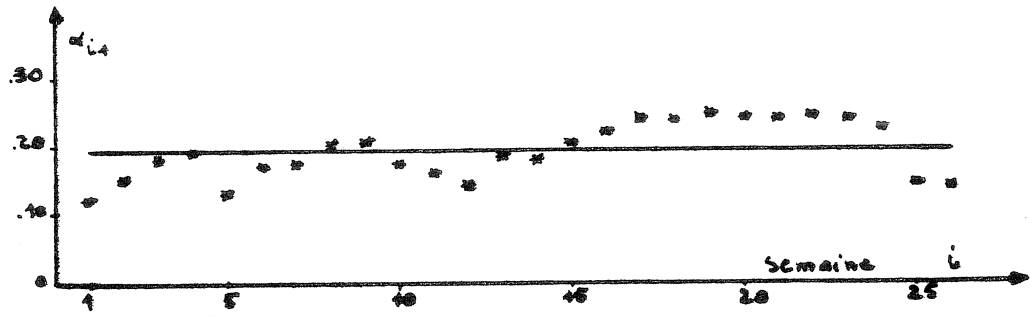
Doux - Eyriaux	5	stations
Loira Supériaura	5	"
Loira Moyanna	12	"
Ardache - Chassezac	9	"
Allier supériaur	8	"
Caze	6	"
Hérault - Orb	2	"
Tarn	4	"
Lot	5	"

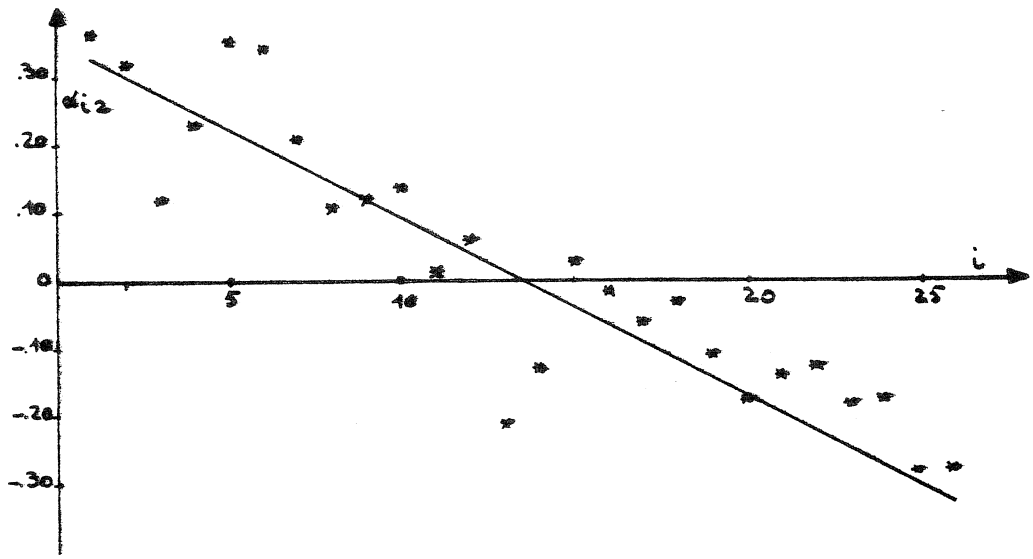
VI 40

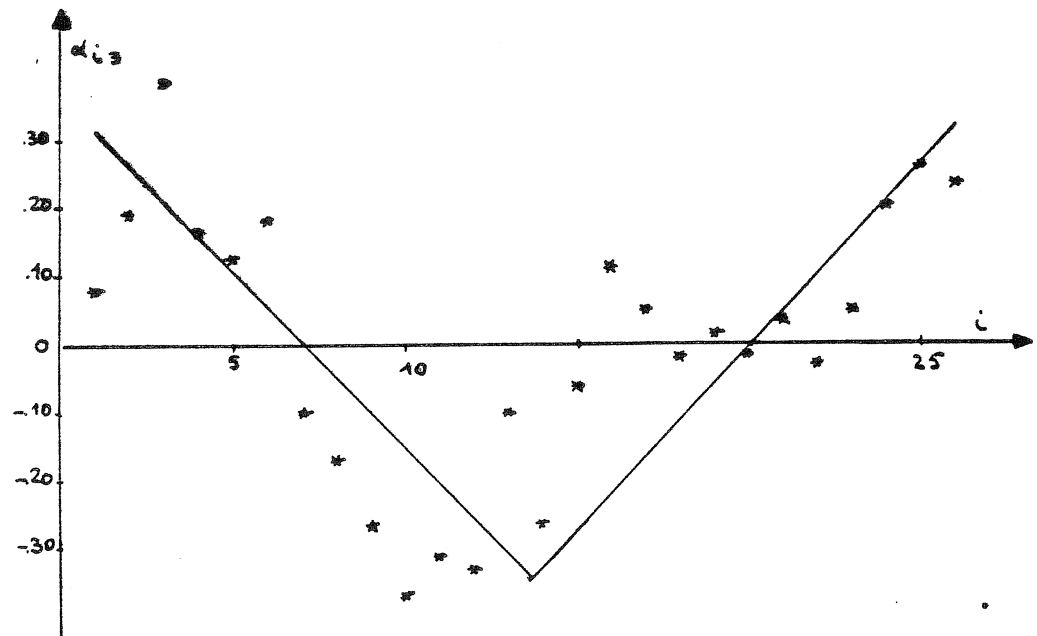
2

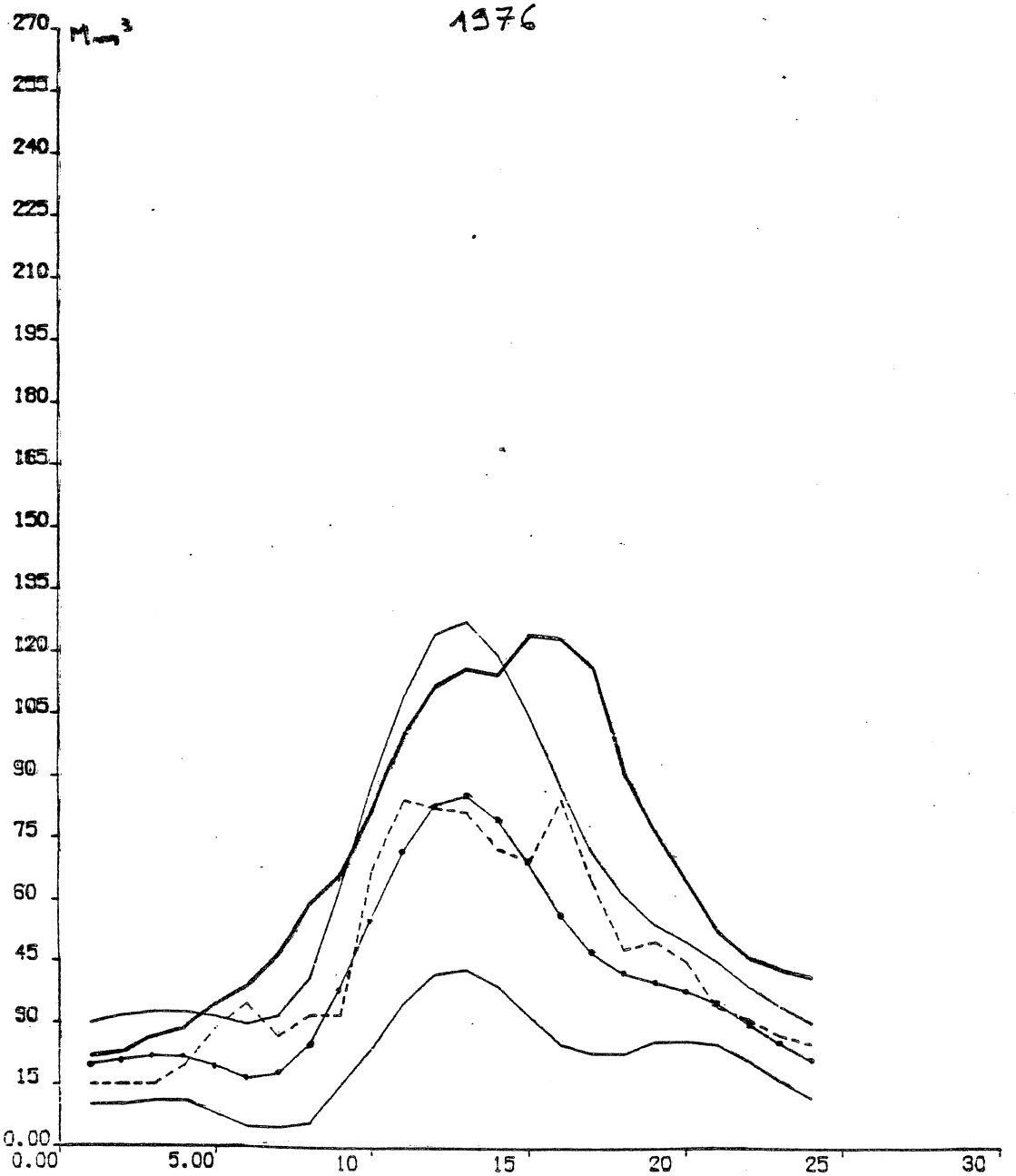
Exemple d'application correspondant aux calculs du § III analyse en composantes principales des apports hebdomadaires des 26 semaines d'été (1er mars-31 août) de la DURANCE à SERRE-PONCON.

- * cosinus directeurs empiriques obtenus avec le programme MUVAR
- — relations linéaires permettant d'obtenir les cosinus directeurs rectifiés.



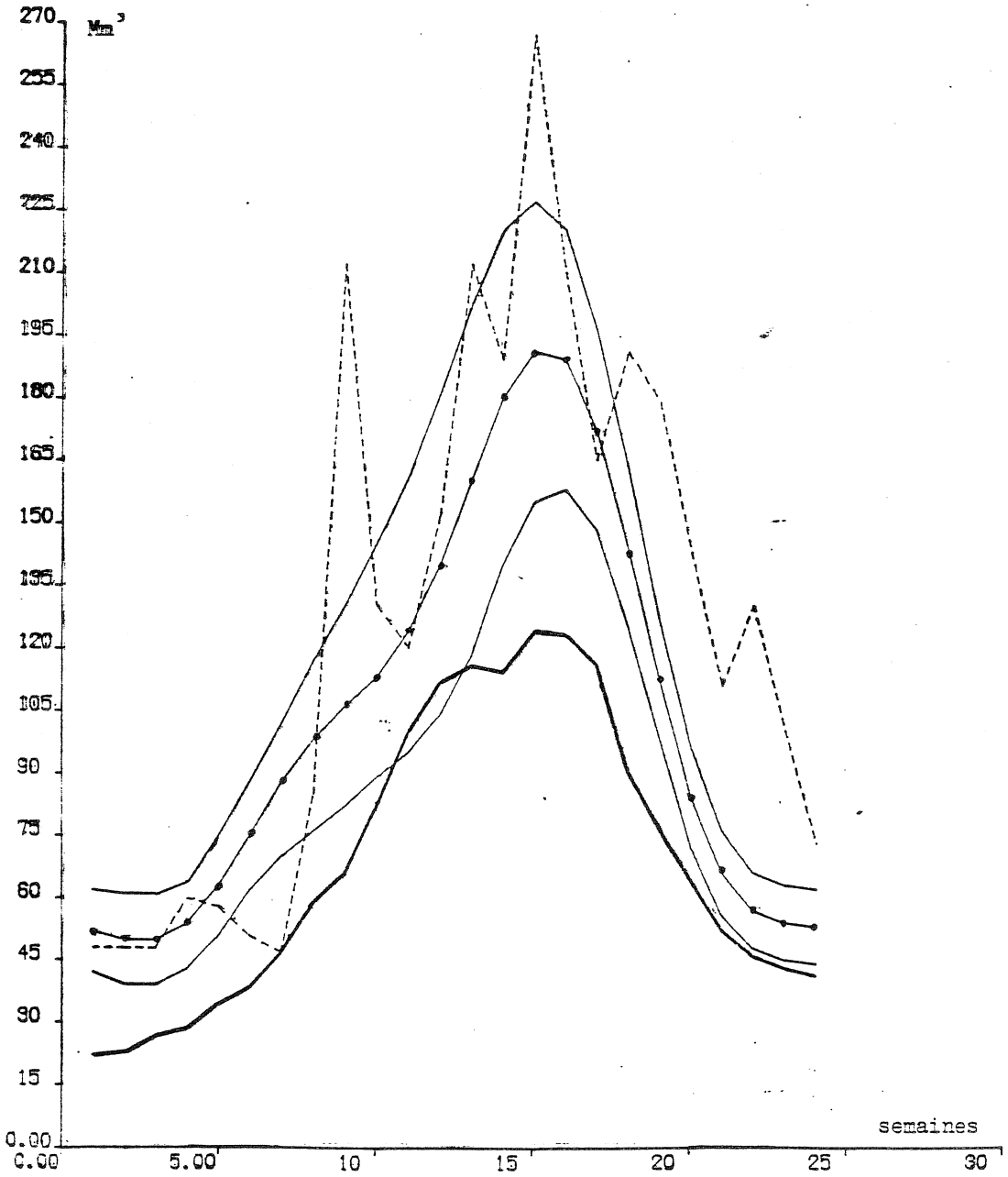





1976
Mm³
270
255
240
225
210
195
180
165
150
135
120
105
90
75
60
45
30
15
0.00
0.00
5.00
10
15
20
25
30

DURANCE à SERRE-PONCON

1977



VII - FONCTIONS DE TRANSFERT LINEAIRES

7.1 - Fonction de transfert pluie-débit sur des bassins versants de l'ordre de 1 000 km² -

Cette méthode a été mise au point pour calculer en temps réel des prévisions de débits de crue d'après les pluies mesurées, pour des bassins versants de l'ordre du millier de kilomètres carrés ($100 \leqslant BV \leqslant 3000$ km²).

Contrairement au processus classique qui consiste d'abord à réduire la pluie brute en pluie efficace puis ensuite à identifier la réponse impulsionnelle (cf DE MARSILY 1971, SINGH 1977, et al ...), démarche qui fait implicitement dépendre le second résultat du premier, nous commençons par déterminer la fonction de transfert linéaire pluie efficace-débit ainsi que la pluie efficace, puis ensuite, on cherche à définir la fonction d'abattement à appliquer à la pluie brute pour obtenir cette pluie efficace : Les deux opérations étant alors indépendantes.

Partant d'une relation linéaire entre variation de débit et pluies brutes concomitante et antérieures, nous proposons une méthode itérative de régression multiple permettant de calculer la pluie efficace (P_1, P_2, ..., P_m) ainsi que la fonction de transfert "moyenne" discrétisée (A_1, A_2, ..., A_K) donnant l'hydrogramme de ruissellement (D_1, D_2, ..., D_n) :

$$D_j = \sum_{i=1}^{K} A_i \, P_{j-i+1} \quad ; \quad (1 \leqslant j \leqslant n)$$

Les calculs sont effectués sur les différences premières, pour l'intervalle de temps unitaire choisi, de :

- chaque hydrogramme observé, $\Delta Q_j = Q_j - Q_{j-1} = q_j$, on court-circuite ainsi la délicate opération de séparation des hydrogrammes (de base, hypodermique ...), la variation unitaire de ces derniers est en effet négligeable devant ΔQ_j, sauf pour des crues complexes (en séquence) ;
- la fonction de transfert, soit $\Delta A_i = A_i - A_{i-1} = a_i$ (avec $a_1 = A_1$), on notera DPFT la différence première de la fonction de transfert.

L'avantage de cette procédure de calcul en différences premières est double :

- diminution importante de la corrélation entre débits successifs, ainsi qu'entre "variables" de la DPFT inversée (déconvolution), on assure ainsi une meilleure stabilité des coefficients de régression partielle, ce qui évite d'imposer des contraintes à ces coefficients, donc d'alourdir la procédure de calcul ;
- permet de s'affranchir d'éventuelles dérives de la courbe de tarage.

Toutefois l'inconvénient de la méthode est sa grande sensibilité au bruit, erreurs de mesures sur les débits ou précipitations, par exemple.

7.1.1. - Calcul de la fonction de transfert pluie efficace-débit et de la pluie efficace -

Cette note décrit plus en détail le processus de calcul de la fonction de transfert (discrétisée) pluie efficace - débit de ruissellement de surface d'après les mesures de pluies brutes et de débits.

On dispose de N épisodes pluie-crue ($15 \leqslant N \leqslant 60$) en limitant les cas de crues complexes (crues consécutives rapprochées). On choisira une durée constante pour l'ensemble des hydrogrammes, soit n+1 débits depuis l'origine de chaque épisode pluvieux (non comme contrainte, mais facilité de calcul). Les épisodes de précipitation peuvent avoir une longueur variable, mais une durée minimale commune.

Notations -

- Variation de débit entre les instants j-1 et j (pas de temps unitaire $0 \leqslant j \leqslant n$) pour la $l^{ième}$ crue ($1 \leqslant l \leqslant N$) :

$$Q_{j,l} - Q_{j-1,l} = \Delta Q_{j,l} = q_{j,l} \qquad (m^3/s)$$

- Précipitation observée (brute), cumulée pendant l'intervalle [j-1, j] pour la $l^{ième}$ crue (moyenne arithmétique simple ou pondérée de plusieurs stations :

$$R_{j,l} \qquad (mm)$$

- Précipitation efficace, ou nette, pour le ruissellement de surface pendant l'intervalle [j-1, j]:

$$P_{l,j}$$

- Coefficients de la Fonction de Transfert (FT) appliqués aux précipitations efficaces pour calculer le débit de ruissellement direct à l'instant j

$$A_1, A_2, \ldots, A_i, \ldots A_m \text{ avec}$$
$$Q_{j,l} = \sum_{i=1}^{m} A_i \; P_{j-i+1,l}$$

- Coefficients de la Différence Première de la Fonction de Transfert appliqués aux précipitations efficaces pour calculer la variation de débit entre j-1 et j pour la $l^{ième}$ crue :

$$a_1, a_2, \ldots a_m$$

avec $\begin{cases} A_i - A_{i-1} = a_i \\ A_1 = a_1 \end{cases}$

et $$q_{j,l} = \sum_{i=1}^{m} a_i \; P_{j-i+1,l}$$

DEFINITION DE LA D P F T

Si l'on note n_i la longueur de chaque hydrogramme (avec le pas de temps unitaire choisi), m_i la longueur de chaque épisode de pluie efficace et r la longueur de la fonction de transfert ($n_i = m_i + r - 1$), l'écriture la plus générale exprimant la relation linéaire pluie efficace-débit est la suivante :

$$\begin{bmatrix} q_{1,1} \\ q_{2,1} \\ \vdots \\ q_{n_1,1} \end{bmatrix} = \begin{bmatrix} P_{1,1} & 0 & 0 & 0 \cdots & 0 \\ P_{2,1} & P_{1,1} & 0 & 0 & \\ P_{3,1} & P_{2,1} & P_{1,1} & 0 & \\ \cdots & \cdots & \cdots & 0 & \\ & & & & 0 \\ 0 & \cdots & \cdots & 0 & P_{m_1,1} \end{bmatrix} \begin{bmatrix} a_{1,1} \\ a_{2,1} \\ \vdots \\ a_{r_1,1} \end{bmatrix}$$

$$\begin{bmatrix} q_{1,2} \\ \vdots \\ q_{n_2,2} \end{bmatrix} = \begin{bmatrix} P_{2,1} & 0 & & \\ & & & \\ 0 & \cdots & \cdots & P_{m_2,1} \end{bmatrix} \begin{bmatrix} a_{1,2} \\ \\ a_{r_2,2} \end{bmatrix}$$

$$\begin{bmatrix} q_{1,N} \\ \vdots \\ q_{n_N,N} \end{bmatrix} \quad \begin{bmatrix} P_{1,N} & 0 & 0 & 0 \\ P_{2,N} & P_{1,N} & 0 & \\ & & & \\ 0 & \cdots & \cdots & P_{m_N,N} \end{bmatrix} \begin{bmatrix} a_{1N} \\ \\ a_{r_n N} \end{bmatrix}$$

Avec les mesures disponibles, on fait l'hypothèse d'une fonction de transfert moyenne (a_1, a_2, ... a_r).

Pour la commodité des calculs, on prend des épisodes pluvieux de longueur constante m, donc des hydrogrammes de longueur constante n; ce qui conduit à tronquer certains épisodes de précipitations ou avoir des 0 pour les dernières pluies efficaces d'autres épisodes. On peut alors écrire la relation linéaire pluie efficace-débit sous les deux formes équivalentes.

.../...

Soit :

(I)

$$\begin{bmatrix} q_{1,1} \\ q_{2,1} \\ \vdots \\ q_{n,1} \\ \\ q_{1,2} \\ \vdots \\ q_{n,2} \\ \vdots \\ q_{1,N} \\ \vdots \\ q_{n,N} \end{bmatrix} = \begin{bmatrix} P_{1,1} & 0 & 0 & \cdots & 0 \\ P_{2,1} & P_{1,1} & 0 & \cdots & 0 \\ P_{3,1} & P_{2,1} & P_{1,1} & & \\ \cdots & & & & \\ 0 & \cdots & & & P_{m,1} \\ & & & & \\ P_{1,2} & 0 & 0 & \cdots & 0 \\ P_{2,2} & P_{1,2} & 0 & & \\ \cdots & & & & \\ 0 & \cdots & & & P_{m,2} \\ \cdots & & & & \\ P_{1,N} & 0 & 0 & \cdots & 0 \\ P_{2,N} & P_{1,N} & 0 & \cdots & \\ & & & & \\ 0 & \cdots & & & P_{m,N} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_r \end{bmatrix}$$

Soit le produit matriciel $[q]_{nN,1} = [P]_{nN,r} \ [a]_{r,1}$

ou encore N produits matriciels :

(II)

$$\begin{bmatrix} q_{1,1} \\ q_{2,1} \\ \vdots \\ q_{n,1} \end{bmatrix} = \begin{bmatrix} a_1 & 0 & 0 & \cdots & 0 \\ a_2 & a_1 & 0 & \cdots & 0 \\ & & & & \\ a_m & \cdots & & & a_1 \\ a_{m+1} & \cdots & & & a_2 \\ 0 & \cdots & & & a_r \end{bmatrix} \begin{bmatrix} P_{1,1} \\ P_{2,1} \\ \vdots \\ P_{m,1} \end{bmatrix}$$

Soit le produit matriciel $[q_1]_{n,1} = [A]_{n,m} * [P_1]_{m,1}$

$$\begin{bmatrix} q_{1,2} \\ q_{2,2} \\ \vdots \\ q_{n,2} \end{bmatrix} = \begin{bmatrix} a_1 & 0 & 0 & \cdots & 0 \\ a_2 & a_1 & 0 & \cdots & 0 \\ \cdots & & & & \\ a_m & \cdots & & & a_1 \\ \cdots & & & & \\ 0 & \cdots & & & a_r \end{bmatrix} \begin{bmatrix} P_{1,2} \\ P_{2,2} \\ \vdots \\ P_{m,2} \end{bmatrix}$$

$$[q_2]_{n,1} = [A]_{n,m} * [P_2]$$

$$\begin{bmatrix} q_{1,N} \\ q_{2,N} \\ \vdots \\ q_{n,N} \end{bmatrix} = \begin{bmatrix} a_1 & 0 & 0 & \cdots & 0 \\ a_2 & a_1 & 0 & \cdots & 0 \\ \cdots & & & & \\ a_m & \cdots & & & a_1 \\ \cdots & & & & \\ 0 & \cdots & & & a_r \end{bmatrix} \begin{bmatrix} P_{1,N} \\ P_{2,N} \\ \vdots \\ P_{m,N} \end{bmatrix}$$

$$[q_N]_{n,1} = [A]_{n,m} * [P_N]_{m,1}$$

On a prolongé la fonction de transfert de r à n (car $n > r$) puisque en fait, a_n tend asymptotiquement vers 0.

En utilisant alternativement les formulations (I) et (II), et en initialisant le calcul avec les pluies brutes (observées), on va calculer une estimation de la DPFT [a] ainsi que des pluies efficaces [P] selon le procédé décrit dans la suite.

On calcule une première estimation des coefficients (a_i^*) de la DPFT tronquée à K, lorsque la différence seconde s'annule ($1 \leqslant i \leqslant K < r$), par la méthode des moindres carrés appliquée au système de relation linéaire suivant :

$$\begin{bmatrix} q_{1,1} \\ q_{2,1} \\ \vdots \\ q_{n,1} \\ q_{1,2} \\ \vdots \\ q_{n,2} \\ \vdots \\ q_{1,N} \\ \vdots \\ q_{n,N} \end{bmatrix} = \begin{bmatrix} R_{1,1} & 0 & 0 & \cdots & 0 \\ R_{2,1} & R_{1,1} & 0 & & 0 \\ \cdots & & & & \\ R_{n,1} & \cdots & & & R_{n-K+1,1} \\ R_{1,2} & 0 & 0 & & \\ R_{2,1} & R_{2,2} & 0 & & \\ \cdots & & & & \\ R_{n,2} & \cdots & & & R_{n-K+1,2} \\ \cdots & & & & \\ R_{1,N} & 0 & \cdots & & 0 \\ R_{2,N} & R_{1,N} & & & 0 \\ \cdots & & & & \\ R_{n,N} & \cdots & & & R_{n-K+1,N} \end{bmatrix} \begin{bmatrix} a^*_1 \\ a^*_2 \\ \vdots \\ a^*_K \end{bmatrix} + \begin{bmatrix} \varepsilon^*_{1,1} \\ \varepsilon^*_{2,1} \\ \\ \varepsilon^*_{n,1} \\ \varepsilon^*_{1,2} \\ \\ \varepsilon^*_{n,2} \\ \\ \varepsilon^*_{1,N} \\ \\ \varepsilon^*_{n,N} \end{bmatrix}$$

soit en écriture matricielle :

$$[q]_{nN,1} = [R]_{nN,K} * [a^*]_{K,1} + [\varepsilon^*]_{nN,1}$$

ε étant l'écart résiduel entre la variation observée de débit et la variation calculée par une combinaison linéaire des précipitations concomitantes et antérieures; on calcule [a] par la méthode des moindres carrés, en résolvant $\frac{\delta}{\delta [a]} [\varepsilon^{*\prime} \varepsilon^*] = 0$.

Ayant obtenu cette estimation des K premièrs coefficients a^*_i, on extrapole cette suite pour $i > K$, en effectuant un lissage exponentiel sur les coefficients A^*_i de la FT pour obtenir $A_i = A_K \exp [-d (K-i)]$ pour $i > K$. On calcule alors, crue par crue, les corrections [e] à appliquer aux précipitations brutes [R] pour obtenir une première estimation des pluies efficaces correspondantes $[u_1] = [R_1] + [e_1^*]$ avec $u \geqslant 0$ $(1 \leqslant l \leqslant N)$.

On utilise la formulation (II) , par exemple pour la première crue :

$$\begin{bmatrix} q_{1,1} \\ q_{2,1} \\ \vdots \\ q_{n,1} \end{bmatrix} = \begin{bmatrix} a^*_1 & 0 & 0 & \cdots & 0 \\ a^*_2 & a^*_1 & 0 & & 0 \\ \cdots & & & & \\ a^*_n & & & & a^*_{n-m+1} \end{bmatrix} \begin{bmatrix} R_{1,1} + e^*_{1,1} \\ R_{2,1} + e^*_{2,1} \\ \cdots \\ R^*_{m,1} + e^*_{m,1} \end{bmatrix} + \begin{bmatrix} \theta^*_{1,1} \\ \theta^*_{2,1} \\ \vdots \\ \theta^*_{n,1} \end{bmatrix}$$

$$\text{Soit } [q_1] = [A^*]_{n,m} * [R_1 + e^*_1]_{m,1} + [\theta^*_1]_{n,1}$$

où $[\theta]$ est un bruit aléatoire.

$$\text{Soit encore } [\varepsilon^*_1]_{n,1} = [A^*]_{n,m} * [e^*_1]_{m,1} + [\theta^*_1]_{n,1}$$

ou

$$\begin{bmatrix} \varepsilon^*_{1,1} \\ \varepsilon^*_{2,1} \\ \vdots \\ \varepsilon^*_{n,1} \end{bmatrix} = \begin{bmatrix} a^*_1 & 0 & 0 & \cdots & 0 \\ a^*_2 & a^*_1 & 0 & & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a^*_n & & & & a^*_{n-m+1} \end{bmatrix} \begin{bmatrix} e^*_{1,1} \\ \vdots \\ e^*_{m,1} \end{bmatrix} + \begin{bmatrix} \theta^*_{1,1} \\ \theta^*_{2,1} \\ \vdots \\ \theta^*_{n,1} \end{bmatrix}$$

On obtient les valeurs des coefficients $[e^*_1]$ pour $1 \leqslant l \leqslant m$ et ainsi une première estimation des pluies efficaces $[u^*_1]$. Ces valeurs vont servir à calculer une nouvelle estimation de la DPFT, par l'intermédiaire de la formulation (I) soit :

$$[q]_{nN,1} = [U^*]_{nN,K} * [a^{**}]_{K,1} + [\varepsilon^{**}]_{nN,1}$$

puis l'on calcule une nouvelle série de corrections $[e^{**}_1]$ pour chaque crue et en utilisant la formulation (II), ce qui permet d'obtenir une nouvelle estimation des pluies efficaces :

$$[u^{**}_1] = [u^*_1] + [e^{**}_1] \quad \text{avec tout élément } u^{**} \geqslant 0.$$

COMMENTAIRES -

1°/- Trois itérations suffisent en moyenne à obtenir un coefficient de corrélation, entre variations de débits observées et variations calculées par combinaison linéaire des pluies efficaces estimées, compris entre 0.90 et 0.98.

Si ce n'est pas le cas, par exemple après 3 itérations ce coefficient de corrélation multiple ne dépasse pas 0.80, cela provient généralement d'un ou deux épisodes de crue dus à une configuration spatiale particulière des averses, dont le calcul par régression est inconsistant (voir analyse des résidus ε^{**} correspondants), on supprime ces crues de l'échantillon et l'on reprend les calculs au début.

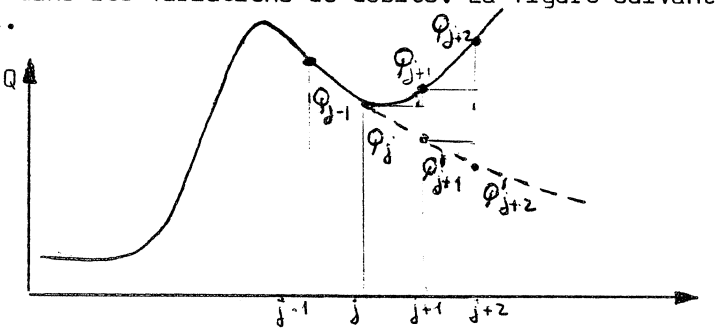
2°/- Il est préférable de calculer la pluie efficace P, qui servira à caler la relation non linéaire P = f(R), en déconvoluant directement les hydrogrammes de crue, après avoir obtenu une estimation stable des coefficients [a] de la DPFT; pour cela on utilisera la formulation (II) :

$$[q_1]_{n,1} = [A]_{n,m} * [P_1]_{m,1} + [\theta_1]_{n,1}$$

on calculera les coefficients P_1 de cette régression multilinéaire, par la méthode des moindres carrés appliquée à chaque crue séparément.

.../...

3°/- Eviter de prendre des séquences de crues consécutives trop rapprochées car on introduit un biais dans les variations de débits. La figure suivante illustre cet inconvénient.



$$\begin{cases} \Delta Q_{i+1} = (Q_{j+1} - Q_j) = \text{variation apparente de débit} \\ \Delta Q'_{i+1} = (Q_{j+1} - Q'_{j+1}) = \text{variation réelle de débit} \end{cases}$$

$$\begin{cases} \Delta Q_{j+2} = (Q_{j+2} - Q_{j+1}) = \text{variation apparente de débit} \\ \Delta Q'_{j+2} = (Q_{j+2} - Q_{j+1}) + (Q'_{j+1} - Q'_{j+2}) = \text{variation réelle} \end{cases}$$

4°/- Enfin, la DPFT ou la FT que l'on calcule ainsi n'est qu'une forme moyenne de réponse du bassin versant aux précipitations, sauf rares cas : elle résulte généralement de la composition de deux, trois ou plusieurs fonctions de transfert propres aux sous-bassins versants homogènes qui composent la mosaïque du bassin versant principal.

Si l'échantillon des averses utilisées pour le calage de la DPFT comprend uniquement des épisodes pluvieux répartis régulièrement sur tout le bassin versant, on aura une bonne définition des pluies efficaces et cela facilitera l'ajustement de la fonction d'abattement $P = \frac{R^2}{R+b}$ (l'homogénéité spatiale des pluies sera également une garantie de bonnes prévisions de débit en exploitation opérationnelle).

Par contre si l'échantillon comprend des épisodes partiellement localisés à un sous-bassin, on risque de converger très lentement dans le calcul de la DPFT. De plus, pour obtenir les pluies efficaces on déconvolue chaque hydrogramme avec une fonction de transfert moyenne, alors qu'il aurait été nécessaire de le convoluer avec la fonction de transfert du sous-bassin concerné : on obtient ainsi des pluies efficaces factices, qui rendront difficile tout calage objectif de la fonction d'abattement.

Or, compte tenu de la modestie des échantillons pluie-crue pour la plupart des bassins, on ne peut faire une sélection basée sur l'homogénéité des pluies spatiales, par conséquent tous les échantillons pluie-brute-pluie efficace déconvoluée sont affectés d'un bruit non négligeable, qui rend illusoire la recherche d'une méthode numérique sophistiquée pour caler les paramètres de la fonction d'abattement de l'ensemble du bassin. De plus on peut imaginer que chaque sous-bassin possède non seulement sa FT propre, mais également une relation $P = \frac{R^2}{R+b}$ spécifique.

Ces remarques montrent également qu'il est illusoire de rechercher des pondérations sophistiqués pour calculer la "vraie" lame d'eau reçue par le bassin total.

7.1.2 - Relation entre pluie brute et pluie efficace -

Dans ce paragraphe, nous présentons un tableau récapitulatif et condensé des essais effectués pour établir une relation non linéaire simple entre la précipitation efficace ou nette calculée par la méthode précédente et la précipitation brute.

Bien que l'imagination des hydrologues soit fertile dans ce domaine, nous nous sommes limités à deux relations simples :

$$P_j = R_j - b\left(1 - \exp\left(-\frac{R_j}{b}\right)\right) \qquad ③$$

$$P_j = R_j - b\,R_j\,(R_j + b)^{-1} \qquad ④$$

(en notant R_j la pluie observée et P_j la pluie efficace pendant l'intervalle unitaire j-1, j).

Ainsi, dans les fortes valeurs, la pluie efficace tend à égaler la pluie brute moins la rétention du bassin versant "b".

Ce terme b, témoin de la rétention, n'est d'ailleurs pas une constante, c'est le produit d'un paramètre saisonnier (fonction de la date dans l'année) et d'une fonction inverse des pluies et (ou) débits antérieurs :

$$b = c\left(H\,(Q,R)\right)^{-1} \qquad ⑤$$

Pour caractériser l'évolution de l'état de saturation des couches supérieures du sol du bassin versant pendant la crue, par une approche empirique, nous avons essayé diverses définitions de la fonction $H(Q,R)$:

- indice des débits antérieurs :

$$H(Q) = \left(IQA_{j-1}\right)^{\beta} = \left(\lambda\, Q_{j-1} + 1-\lambda)\; IQA_{j-2}\right)^{\beta} \qquad \text{avec } 0 \leqslant \lambda \leqslant 1 \text{ et } .8 \leqslant \beta \leqslant 1$$

- indice des précipitations antérieures

$$H(R) = IRA_j = \theta R_j + (1-\theta)\; IRA_{j-1} \qquad \text{avec } .05 \leqslant \theta \leqslant .50$$

- indices combinés des précipitations et débits antérieurs

$$H(Q,R) = (IQA_{j-1})^{\beta} * (IRA_j)$$

$$H(Q,R) = Q_0^{P} * (Q_{j-1})^{1-P} * IRA_j \qquad \text{avec } 0 \leqslant p \leqslant 1$$

Pour chaque essai, on a calculé et graphiqué la corrélation entre la pluie efficace déduite des débits, et la pluie efficace reconstituée d'après les relations ③ et ④ .

Ce travail a été effectué pour plusieurs bassins versants de géomorphologie variée et d'alimentation très différente (pluies d'origine océanique, pluies d'origine méditerranéenne).

En définitive, nous avons retenu le modèle suivant :

$$P_j = R_j^2 \, (R_j + b)^{-1}$$

$$b = c \left(H \, (Q,R) \right)^{-1}$$

$$H \, (Q,R) = \begin{cases} IQA_{j-1} & \text{pour le Massif Central, Forez, Morvan} \\ Q_0^P \, (Q_{j-1})^{1-P} \, IRA_j & \text{pour l'Ardèche, les Cévennes, les Alpes du Sud} \end{cases}$$

Pour le calage du paramètre saisonnier c, on a d'abord pris une valeur constante pour l'année, puis, en établissant la distribution statistique des écarts entre variation de débit observée et variation calculée, mois par mois, on a ajusté par moindres carrés la fonction :

$$c = b_0 + b_1 \cos \frac{2\pi t}{365} + b_2 \cos \frac{4\pi t}{365} + d_1 \sin \frac{2\pi t}{365} + d_2 \sin \frac{4\pi t}{365}$$

avec le jour calendaire $1 \leqslant t \leqslant 365$

CALCUL DU DEBIT Q_j A L'INSTANT j

On peut l'obtenir de deux façons :

- soit en calculant la variation du débit en 4 heures entre j-1 et j, avec les coefficients a_i de la DPFT que l'on applique aux pluies efficaces antérieures, à laquelle on rajoute le débit à l'instant Q_{j-1} :

$$Q_j = Q_{j-1} + \sum_{i=1}^{n} a_i \, P_{j-i+1} \quad ,$$

- soit par calcul direct du débit en appliquant la FT aux pluies efficaces, pendant toute la crue, en partant d'un état initial de débit, avec possibilité de ne pas se recaler sur les débits réels en cours de crue :

$$Q_j = IQA_{j-1} + \sum_{i=1}^{n} A_i \, P_{j-i+1} \quad ,$$

l'index des débits antérieurs traduit la saturation progressive des couches profondes.

EXEMPLES D'APPLICATION

- Le BUECH aux CHAMBONS (BV 723 km²) (figure 1)

Cet affluent, en rive droite de la DURANCE, a un bassin versant situé entre 700 et 2700 m d'altitude; les précipitations importantes qu'il reçoit sont essentiellement d'origine méditerranéenne. La moyenne arithmétique de trois stations (Serre - Lus-la-Croix-Haute - St-Etienne-en-Dévoluy) représente la précipitation reçue par le bassin. L'intervalle de temps unitaire est 2 h pour les pluies et débits; une trentaine d'épisodes précipitation - crue ont été dépouillés pour le calage et 6 ont servi à tester le modèle.

Trois itérations ont été nécessaires pour établir une DPFT stable (Tableau I, figure 2).

La relation entre pluie brute et pluie efficace se calcule d'après

$$P_j = R_j^2 \, (R_j + b)^{-1}$$

avec :
$$b = \begin{cases} c(Q_0 \, IRA_j)^{-1} & \text{de mai à novembre} \\ c \, (Q_0 \, Q_{j-1})^{-1/2} \, (IRA_j)^{-1} & \text{en hiver} \end{cases}$$

$$c = 86 \, (9 - 3.8 \cos \frac{2\pi t}{365} - \cos \frac{4\pi t}{365} - 5.1 \sin \frac{2\pi t}{365} - \sin \frac{4\pi t}{365}) \text{ (Figure 8)}$$

La fonction de transfert entre pluie efficace en millimètres et débits en m^3/s est définie dans le Tableau II. On a représenté graphiquement un exemple d'application de cette procédure pour la crue-test du 8 décembre 1977, la plus importante observée depuis 20 ans (figures 3 et 4) ainsi que la crue du 12 octobre 1976 (fig. 5 & 6).

On note que la dispersion des écarts variation observée - variation ajustée (figure 7) est moindre en cumulant 2 intervalles de temps unitaires il y a effet de lissage.

REFERENCES

De MARSILY G. (1971 - La relation pluie-débit sur le bassin versant expérimental de l'Hallue. Rapport LHM/R/71/15, Laboratoire d'Hydrologie Mathématique - E.N.S. des Mines de Paris.

DISKIN M.H. and BONEH A. (1974). - The Kernel function of linear nonstationary Surface runoff systems, Water Resour. Res., Vol. 10, n° 4, pp. 753-761

NATALE L. and TODINI E. (1976). - A stable estimator for linear models part 1 et 2, Water Resour. Res., Vol. 12, n° 4, pp. 667-676.

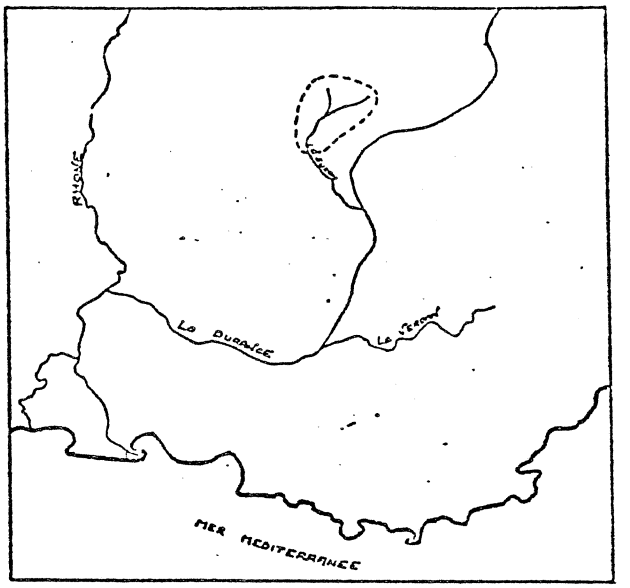
NEUMAN S.P. and DE MARSILY G. (1976). - Identification of linear systems response by parametric programming, Water Resour. Res., Vol. 12, n° 2, pp. 253-262.

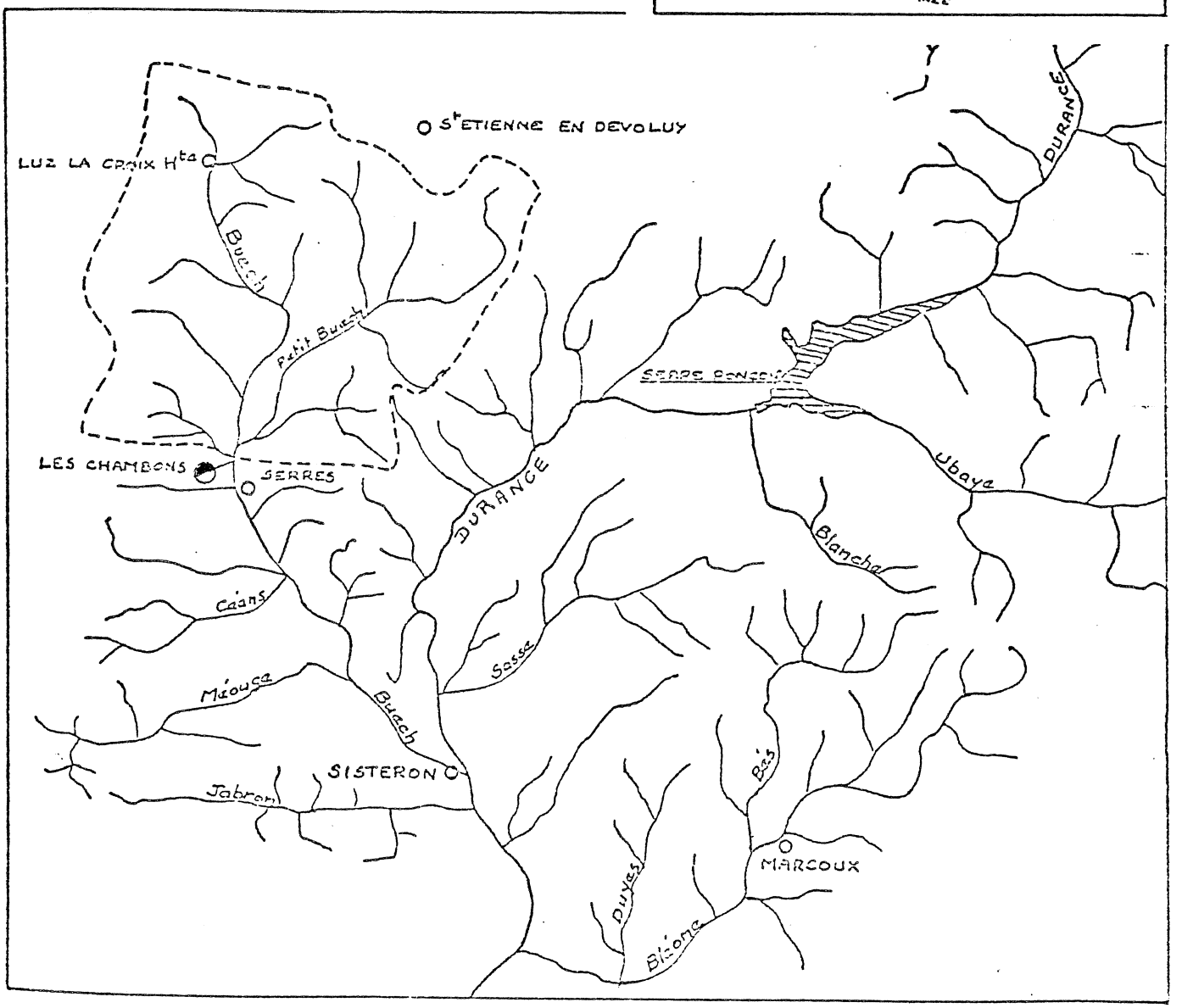
NEWTON D.W. and VINYARD J.W (1967). - Computer determined unit hydrograph from floods, Journal of Hydraul. Div. Amer. Soc. Civil Eng., Vol. 93 (HY5), pp. 219-235.

SINGH V.P. (1977). - Studies on rainfall-runoff modeling, WRRI, Report n° 091, New Mexico Water Resources Research Institute, New Mexico, USA.

WILSON C.B., VALDES J.B. and RODRIGUEZ-ITURBE I. (1979). - On the influence of the Spatial Distribution of Rainfall on Storm runoff, Water Resour. Res.,

BASSIN VERSANT DU BUECH





REMULCA 1 (AOUT 1971) 06/01/78 10/25/12

14 VARIABLES
513 OBSERVATIONS

LE HUECH AUX CHAMBONS

	1	2	3	4	5	6	7	8
M	2.50	2.62	2.75	2.78	2.79	2.79	2.78	2.78
S	5.20	5.24	5.29	5.28	5.28	5.28	5.29	5.29
I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
H	52.20	52.20	52.20	52.20	52.20	52.20	52.20	52.20

	9	10	11	12	13	14
M	2.78	2.74	2.70	1.85	1.61	0.246
S	5.30	5.31	5.32	15.55	15.17	3.799
I	0.00	0.00	0.00	-39.00	-33.72	-23.532
H	52.20	52.20	52.20	121.00	122.65	19.154

R

	1	2	3	4	5	6	7	8
1	1.0000							
2	0.4656	1.0000						
3	0.1887	0.4483	1.0000					
4	0.1266	0.1751	0.4331	1.0000				
5	0.1389	0.1138	0.1629	0.4308	1.0000			
6	0.0183	0.1283	0.1037	0.1599	0.4267	1.0000		
7	-0.0517	0.0073	0.1172	0.1031	0.1581	0.4263	1.0000	
8	-0.0588	-0.0632	-0.0039	0.1155	0.1020	0.1584	0.4278	1.0000
8								
9	-0.1132	-0.0776	-0.0738	-0.0053	0.1161	0.1031	0.1611	0.4287
9	1.0000							
10	-0.1351	-0.1218	-0.0790	-0.0726	-0.0025	0.1196	0.1088	0.1646
10	0.4312	1.0000						
11	-0.1342	-0.1405	-0.1285	-0.0778	-0.0711	0.0009	0.1244	0.1120
11	0.1480	0.4334	1.0000					
12	-0.2766	-0.2645	-0.2844	-0.2960	-0.2626	-0.1622	-0.1602	-0.0180
12	0.3495	0.2715	0.4944	1.0000				
13	-0.2674	-0.3011	-0.3032	-0.3111	-0.2732	-0.1750	-0.1689	-0.0196
13	0.3565	0.2945	0.5078	0.9697	1.0000			
14	-0.0647	0.1194	0.0466	0.0303	0.0159	0.0348	0.0188	0.0045
14	0.0075	-0.0042	-0.0033	0.2216	-0.0234	1.0000		

NOMBRE DE VARIABLES UTILISEES : 12
ORDRE DE CES VARIABLES
12 1 2 3 4 5 6 7 8 9 10 11

	RR	R	T	RR		
12						
1	-0.3546	-0.3136	-8.5	-0.10	P_{j-10}	P_{j-9}
2	-0.0637	-0.0572	-1.4	-0.02	P_{j-8}	P_{j-7}
3	-0.2919	-0.2671	-6.8	-0.09	P_{j-6}	P_{j-5}
4	-0.2612	-0.2347	-6.1	-0.08	P_{j-4}	P_{j-3}
5	-0.3499	-0.3257	-8.4	-0.11	P_{j-2}	P_{j-1}
6	-0.3499	-0.3251	-8.4	-0.11	P_j	
7	-0.4650	-0.4560	-11.8	-0.16		
8	-0.2874	-0.2592	-6.7	-0.09		
9	0.1494	0.1308	3.4	0.04		
10	0.9402	2.3862	61.8	0.81		
11	0.4042	0.3528	9.0	0.12		
	A =	0.1208				

$q_j = Q_j - Q_{j-1}$

précipitation entre j-1 et j

R2 = 0.9410 R = 0.9701 SL = 3.7756 R2 BRUT = 0.9423 F = 747.7498

NOMBRE DE VARIABLES UTILISEES : 0

TABLEAU I

COEFFICIENTS DE REGRESSION MULTIPLE

$q_j = Q_j - Q_{j-1}$	1ère itération	2ème itération	3ème itération
P_{j-10}	- .52	- .38	- .31
P_{j-9}	- .23	- .08	- .06
P_{j-8}	- .16	- .24	- .27
P_{j-7}	- .31	- .26	- .23
P_{j-6}	- .32	- .32	- .33
P_{j-5}	.01	- .16	- .33
P_{j-4}	- .22	- .45	- .46
P_{j-3}	.08	- .14	- .26
P_{j-2}	.60	.21	.13
P_{j-1}	2.26	2.31	2.39
P_j	.21	.30	.35
Coefficient de corrélation multiple	.678	.938	.970

TABLEAU II

LE BUECH aux CHAMBONS

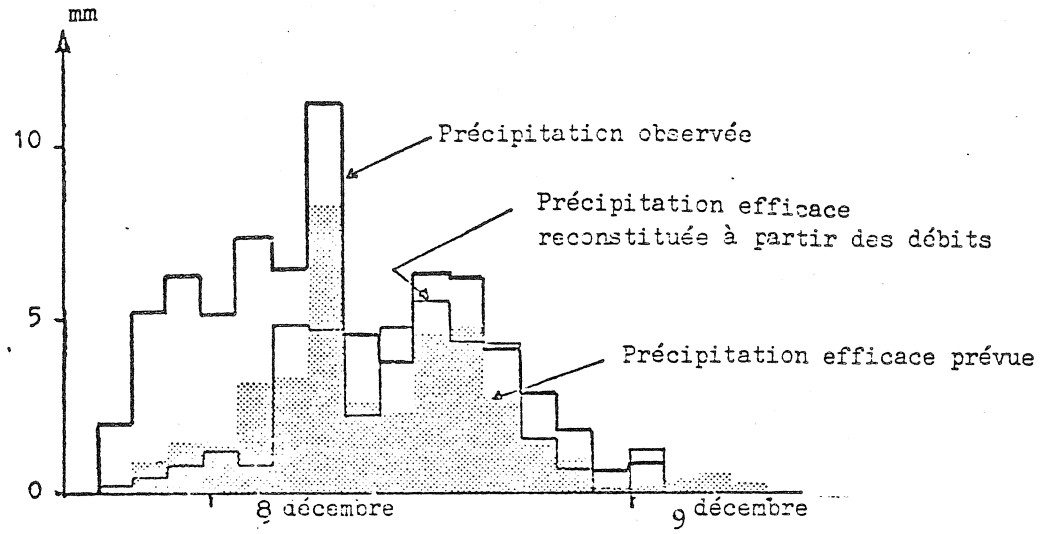
a_i = coefficient de la DPFT

A_i = coefficient de la FT

i	a_i	A_i	i	a_i	A_i
1	1.6	1.6	11	-.7	3.5
2	11.0	12.6	12	-.5	3.0
3	.6	13.2	13	-.5	2.5
4	-1.2	12.0	14	-.4	2.1
5	-1.9	10.1	15	-.3	1.8
6	-1.6	8.5	16	-.3	1.5
7	-1.4	7.1	17	-.3	1.2
8	-1.1	6.0	18	-.2	1.0
9	-1.0	5.0	19	-.2	.8

LE BUECH AUX CHAMBONS

Crue-test du 8 décembre 1977



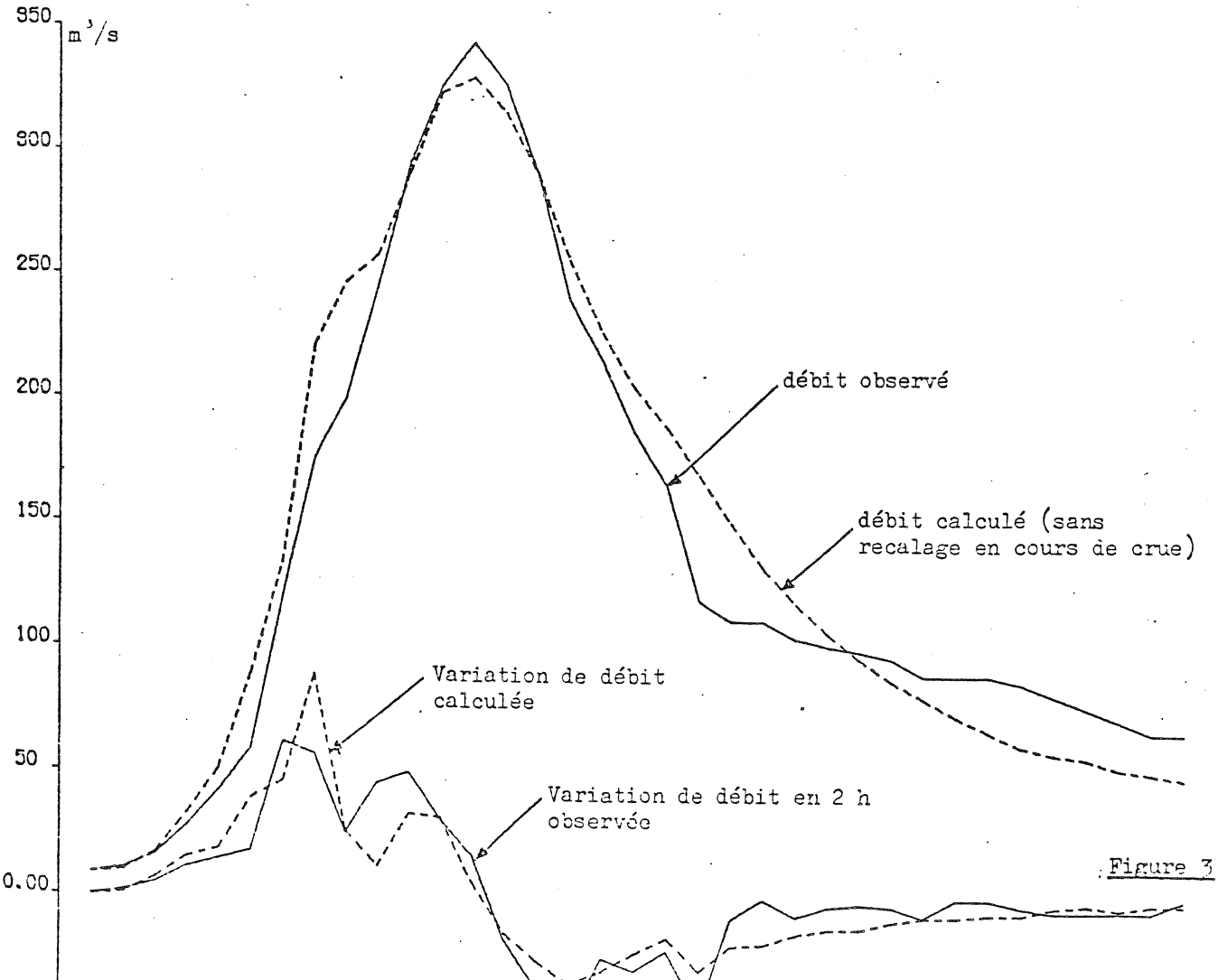


Figure 3

EPISODE PLUIE-CRUE (BUECH aux CHAMBONS)
du 8 décembre 1977

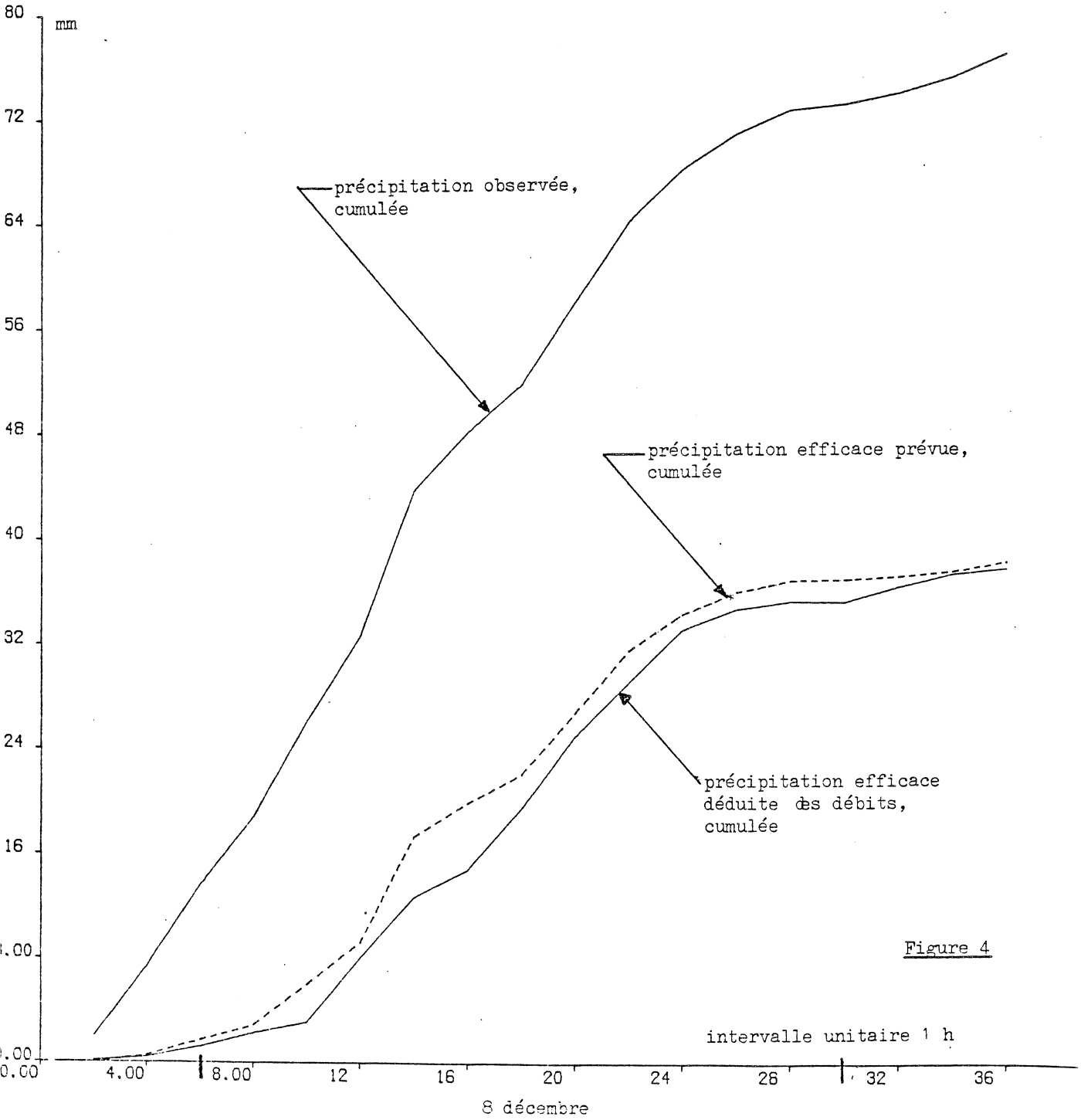


Figure 4

LE BUECH AUX CHAMBONS

crue du 12 octobre 1976

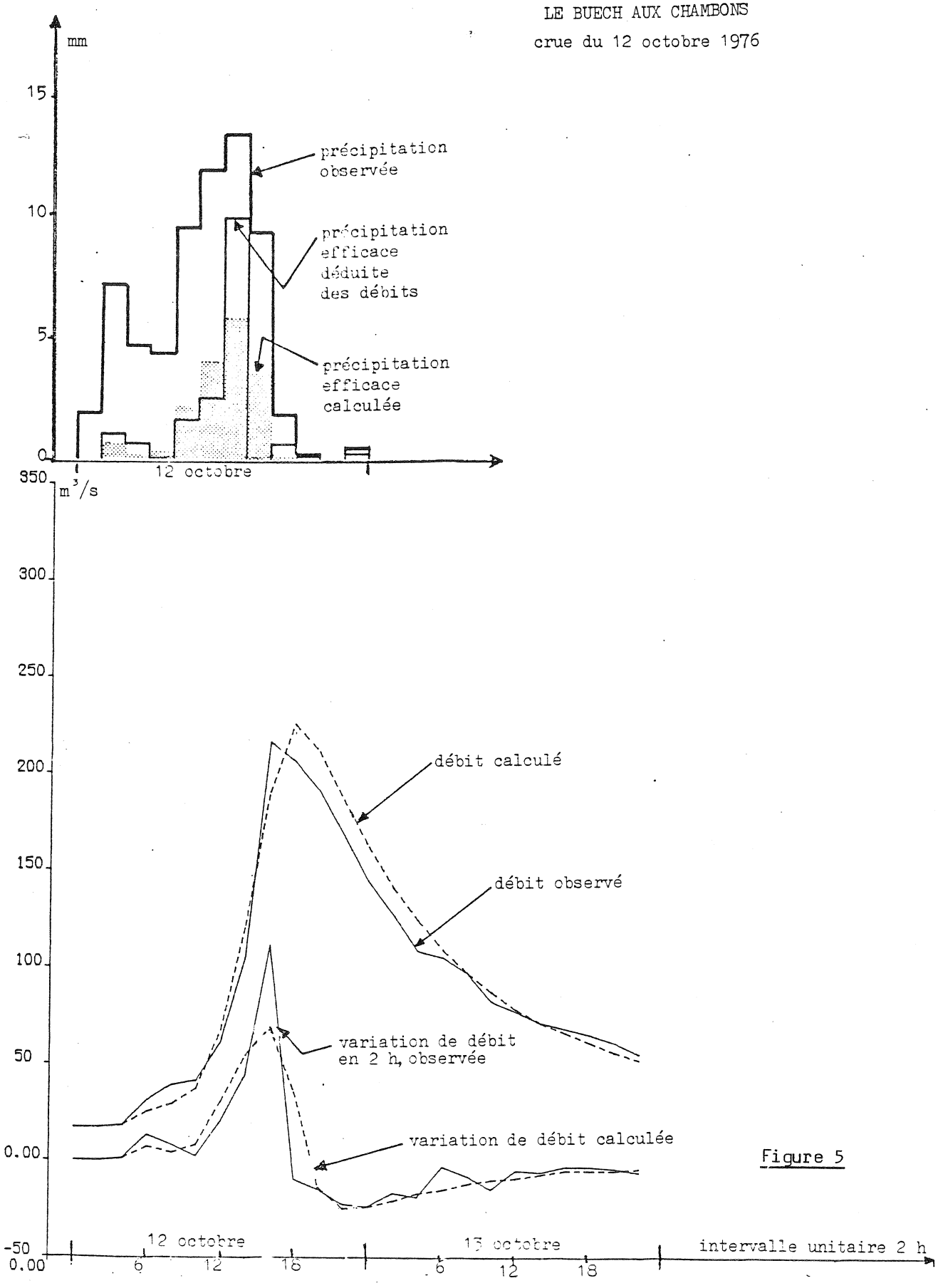


Figure 5

EPISODE PLUIE-CRUE (BUECH aux CHAMBONS)

du 12 octobre 1976

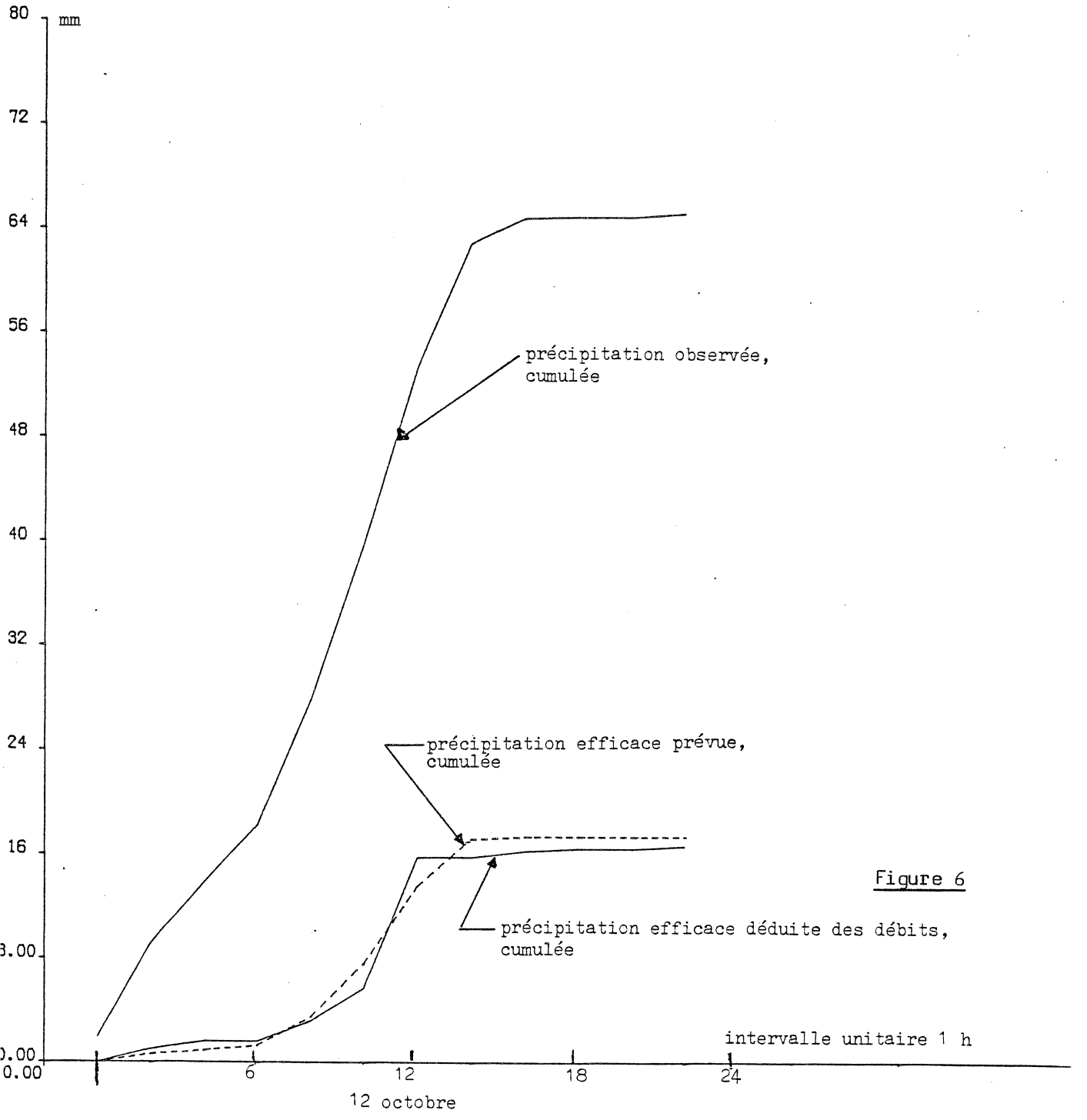
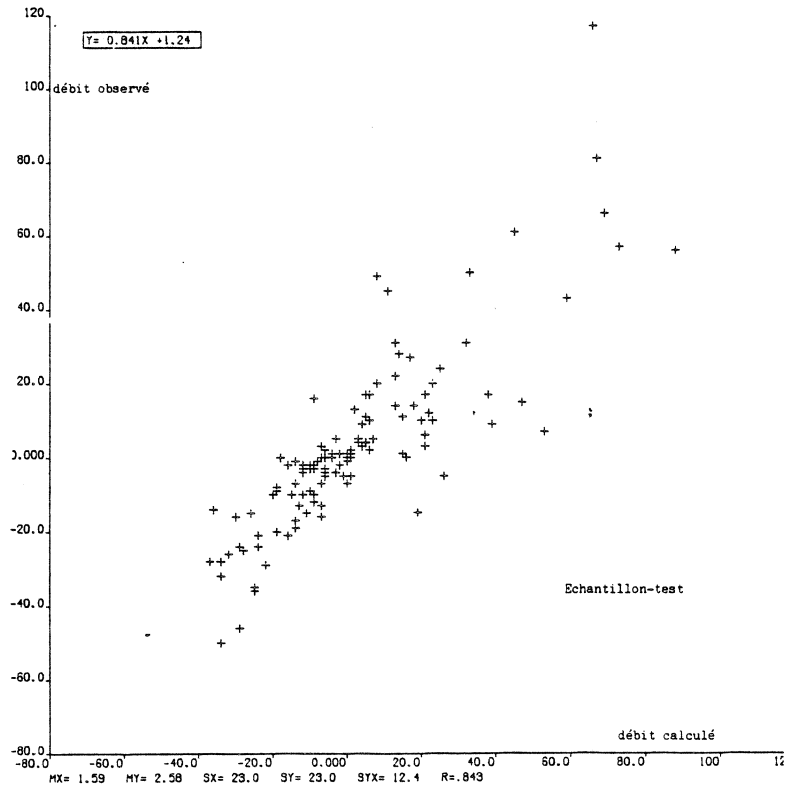


Figure 6



igure 7 – Le Buëch aux Chambons. Crues des 4/5/77, 11/2/79, 10/1)70, 8/12/77. Variations de débits en 2 heures.

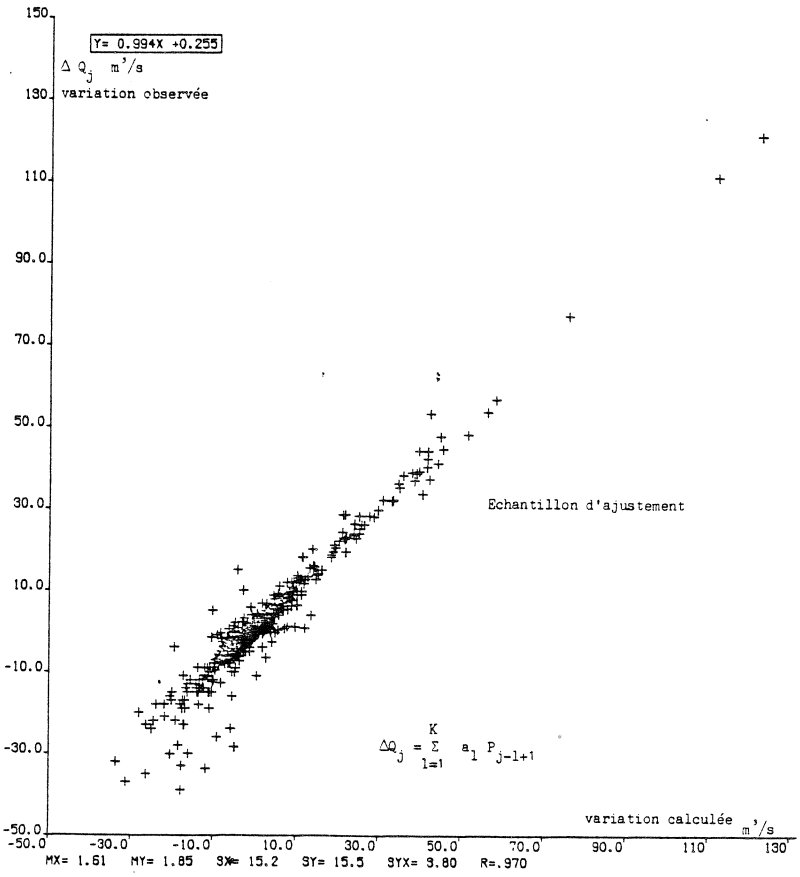


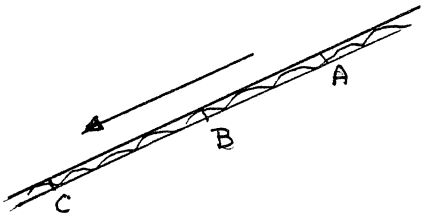
Figure 8 – Le Buëch aux Chambons. Calcul des variations de débit en 2 heures d'après les pluies nettes.

7.2 - Fonction de transfert débit-débit ou propagation de crues

(Dans les applications suivantes, les opérateurs linéaires utilisés sont des approximations suffisantes).

On considère le cas le plus simple où l'on dispose de 3 stations de mesure des débits sur une même rivière (A-B-C, d'amont en aval) avec enregistrement continu des niveaux (limnigrammes, cassettes magnétiques) et l'on suppose que les apports intermédiaires entre A et B, B et C, sont peu importants ou négligeables. On peut établir la fonction de transfert linéaire des débits entre A et B, B et C, A et C par un calcul de régression multiple entre la variation de débit pendant l'unité de temps choisie (1, 2, 3, 4, 5, 6 heures) à la station aval et les variations de débit concomitantes et décalées dans le temps à la station amont, d'après quelques mois d'enregistrement :

$$\Delta Q_h^{AV} = \sum_{i=1}^{k} a_i \, \Delta Q_{h-i+1}^{AM} + \varepsilon_h^{AV}$$



avec :
- $\Delta Q_h^{AV} = Q_h^{AV} - Q_{h-1}^{AV}$
- $\Delta Q_h^{AM} = Q_h^{AM} - Q_{h-1}^{AM}$
- ε_h = écart entre la variation de débit observée et la variation calculée d'après la relation multilinéaire.

Rappelons que les coefficients a_i se calculent par la méthode des moindres carrés, en minimisant $\sum_{h=1}^{n} \varepsilon_h^2$, n étant le nombre total d'observations.

Le test de Student appliqué aux coefficients a_i ou le test de Fisher appliqué aux coefficients de corrélation partielle entre Q_h^{AV} et Q_{h-i+1}^{AM}, permet d'éliminer les variations dont l'influence n'est pas significativement différente de zéro ; on s'appuiera aussi sur la continuité des coefficients.

La présentation des calculs sous forme matricielle s'effectue ainsi :

- on notera $\Delta Q_h^A = q_h^A$, $\Delta Q_h^B = q_h^B$, $\Delta Q_h^C = q_h^C$

$$q_h^B = \sum_{i=1}^{k} b_i \; q_{h-i+1}^A + \varepsilon_h^B$$

$$\begin{bmatrix} q_K^B \\ q_{K+1}^B \\ \vdots \\ q_h^B \\ \vdots \end{bmatrix} = \begin{bmatrix} q_K^A & q_{K-1}^A & \cdots & q_1^A \\ q_{K+1}^A & q_K^A & \cdots & q_2^A \\ \vdots & \vdots & \vdots & \vdots \\ q_h^A & q_{h-1}^A & \cdots & q_{h-K+1}^A \\ \vdots & & & \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_K \end{bmatrix} \qquad (1)$$

Soit :

$$\left[q^B\right] = \left[q^A\right]\left[b\right]$$

de même on obtiendrait :

$$q_h^C = \sum_{i=1}^{l} c_i \, q_{h-i+1}^B + \varepsilon_h^C$$

d'après :

$$\begin{bmatrix} q_1^C \\ q_{l+1}^C \\ \\ q_h^C \\ \vdots \\ q_n^C \end{bmatrix} = \left[\begin{array}{cccc} q_1^B & q_{l-1}^B & \cdots & q_1^B \\ q_{l+1}^B & q_l^B & \cdots & q_2^B \\ \vdots & \vdots & & \vdots \\ q_h^B & q_{h-1}^B & \cdots & q_{h-l+1}^B \\ \vdots & \vdots & & \vdots \\ q_n^B & q_{n-1}^B & \cdots & q_{n-l+1}^B \end{array}\right] \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_l \end{bmatrix} \qquad (2)$$

$$\left[q^C\right] = \left[q^B\right]\left[c\right]$$

on obtiendrait de même $q_h^C = \sum_{i=1}^{m} a_i \, q_{h-i+1}^A + \varepsilon_h^C$

$$\begin{bmatrix} q_m^C \\ \vdots \\ q_h^C \\ \vdots \\ q_n^C \end{bmatrix} = \left[\begin{array}{cccc} q_m^A & q_{m-1}^A & \cdots & q_1^A \\ \cdots & \cdots & \cdots & \cdots \\ q_h^A & q_{h-1}^A & \cdots & q_{h-m+1}^A \\ \cdots & \cdots & \cdots & \cdots \\ q_n^A & \cdots & \cdots & q_{n-m+1}^A \end{array}\right] \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} \qquad (3)$$

Soit $\left[q^C\right] = \left[q^A\right]\left[a\right]$

On peut calculer les coefficients a_i à l'aide des coefficients b_i et c_i ; pour cela on prendra l'égalité des durées des FT, K=l=m, certains coefficients pouvant être nuls.

La relation (1) peut s'écrire sous une autre forme :

./.

$$\begin{bmatrix} q_1^B \\ q_2^B \\ \vdots \\ q_K^B \\ \vdots \\ q_n^B \end{bmatrix} = \begin{bmatrix} b_1 & 0 & \cdots & 0 \\ b_2 & b_1 & 0 & \cdots & 0 \\ \hline b_K & b_{K-1} & \cdots & b_1 & 0 & \cdots & 0 \\ 0 & b_K & & & b_1 & 0 & \cdots & 0 \\ \hline 0 & 0 & \cdots & b_K & \cdots & b_1 \end{bmatrix} \begin{bmatrix} q_1^A \\ q_2^A \\ \vdots \\ q_K^A \\ \vdots \\ q_n^A \end{bmatrix} \qquad (4)$$

Soit :

$$[q^B] = [B]\,[q^A]$$

De même on peut exprimer la relation (2) par :

$$\begin{bmatrix} q_1^C \\ q_2^C \\ \vdots \\ q_K^C \\ \vdots \\ q_n^C \end{bmatrix} = \begin{bmatrix} c_1 & 0 & 0 & \cdots & 0 \\ c_2 & c_1 & 0 & \cdots & 0 \\ \hline c_K & c_{K-1} & \cdots & c_1 & 0 & \cdots & 0 \\ 0 & c_K & \cdots & & c_1 & 0 & \cdots & 0 \\ 0 & \cdots & c_K & \cdots & c_1 \end{bmatrix} \begin{bmatrix} q_1^B \\ q_2^B \\ \vdots \\ q_K^B \\ \vdots \\ q_n^B \end{bmatrix} \qquad (5)$$

soit : $[q^C] = [C]\,[q^B]$

On peut de même représenter la relation (3) par :

$$\begin{bmatrix} q_1^C \\ q_2^C \\ q_3^C \\ \vdots \\ q_K^C \\ \vdots \\ q_n^C \end{bmatrix} = \begin{bmatrix} a_1 & 0 & 0 & 0 & \cdots & 0 \\ a_2 & a_1 & 0 & 0 & \cdots & 0 \\ a_3 & a_2 & a_1 & 0 & \cdots & 0 \\ \hline a_K & a_{K-1} & \cdots & a_1 & 0 & \cdots & 0 \\ 0 & a_k & \cdots & & a_1 & 0 & \cdots & 0 \\ \hline 0 & \cdots & 0 & a_K & \cdots & a_1 \end{bmatrix} \begin{bmatrix} q_1^A \\ q_2^A \\ \vdots \\ q_K^A \\ \vdots \\ q_n^A \end{bmatrix} \qquad (6)$$

Soit : $[q^C] = [A]\,[q^A]$

./.

On voit d'après (4) et (5) que

$$[q^C] = [C][B][q^A]$$

Soit en comparant avec (6) :

$$[A] = [C][B]$$

Soit :

$$[A] = \begin{bmatrix} b_1c_1 & 0 & 0 & 0 \\ b_1c_2 + b_2c_1 & b_1c_1 & 0 & 0 \\ b_1c_3 + b_2c_2 + b_3c_1 & b_1c_2 + b_2c_1 & b_1c_1 & 0 \\ \text{-----} & & & \\ b_1c_K + b_2c_{K-1} + \cdots + b_K c_1 & \cdots & & \cdots \end{bmatrix} \quad (7)$$

d'où :

$$\begin{cases} a_1 = b_1 c_1 \\ a_2 = b_1c_2 + b_2c_1 \\ a_3 = b_1c_3 + b_2c_2 + b_3c_1 \\ \text{-----------} \\ a_K = b_1c_K + b_2c_K + b_2c_{K-1} \text{----} + b_Kc_1 \end{cases}$$

<u>Remarque 1</u> : Tous les calculs précédents sont effectués dans l'hypothèse où les coefficients des fonctions de transfert sont indépendants de l'importance du débit. Ce qui semble le cas pour les crues non exceptionnelles. Dans certains cas, il pourra cependant être utile de calculer les fonctions de transfert par classes de débit pour s'assurer soit de l'invariance de la fonction de transfert, soit de sa déformation en fonction de l'importance des débits

a_i = constante ou $a_i = f(Q)$

<u>Remarque 2</u> : On peut envisager de calculer, d'après la relation (7), une fonction de transfert sans passer par les relations (1), (2), (3).

En effet, supposons que A et C soient à égale distance de B et les tronçons de rivière homogènes :

- si l'on connait les b_i, sachant que $b_i = c_i$, on peut calculer les a_i

$a_1 = b_1^2$, $a_2 = 2\, b_1b_2$

- si l'on ne connait que les a_i on peut calculer $b_i (b_i = c_i)$ d'après :

$$a_1 = b_1^2 \longrightarrow b_1 = \sqrt{a_1}$$

$$a_2 = 2\, b_1b_2 \longrightarrow b_2 = \frac{a_2}{2\sqrt{a_1}}$$

$$a_3 = 2\, b_1b_3 + b_2^2 \longrightarrow b_3 = \left(a_3 - \frac{a_2^2}{4a_1} \right) \frac{1}{2\sqrt{a_1}}$$

<u>Remarque 3</u> : On peut évidemment généraliser les calculs précédents, lorsqu'il y a plusieurs affluents amonts :

$$\Delta Q_h^{AV} = \Sigma a_{i1}\, \Delta Q_{h-i+1}^{AM1} + \Sigma a_{i2}\, \Delta Q_{h-i+1}^{AM2} + \ldots + \mathcal{E}_h^{AV}$$

TABLEAU I

Fonctions de transfert pour calculer
le débit horaire à LR 49 (Meylan)

	RP	B	T	BR		
30					q_h	L R 49
10	0.1862	0.0472	11.1	0.08	q_{h-5}	(
11	0.6569	0.2251	50.8	0.40	q_{h-4}	)
12	0.8089	0.3555	80.2	0.64	q_{h-3}	(LE CHEYLAS
13	0.5086	0.1472	34.4	0.26	q_{h-2}	)
	A =	-0.0116				

R2 = 0.8211 R = 0.9062 SL = 5.3761 R2 BRUT = 0.8213 F =

Exemple : $q_h = .147\ q_{h-2} + .355\ q_{h-3} + .225\ q_{h-4} + .047\ q_{h-5}$

Fonctions de transfert pour calculer le débit horaire

à GRENOBLE (Marius Gontard)

	RP	B	T	BR		
45					q_h	GRENOBLE
27	0.2749	0.1681	16.7	0.17	q_{h-3}	(
28	0.6749	0.6685	53.3	0.69	q_{h-2}	) L R 49
29	0.2733	0.1670	16.6	0.17	q_{h-1}	(
	A =	-0.0250				

R2 = 0.8134 R = 0.9019 SL = 5.3557 R2 BRUT = 0.8136 F =

	RP	B	T	BR		
45					q_h	GRENOBLE
7	0.0655	0.0210	3.3	0.04	q_{h-8}	)
8	0.1938	0.0651	11.5	0.12	q_{h-7}	(
9	0.5594	0.2258	39.3	0.42	q_{h-6}	) LE CHEYLAS
10	0.6619	0.2958	51.5	0.54	q_{h-5}	)
11	0.4197	0.1522	27.0	0.28	q_{h-4}	(
12	0.1281	0.0413	7.5	0.08	q_{h-3}	)
	A =	-0.0403				

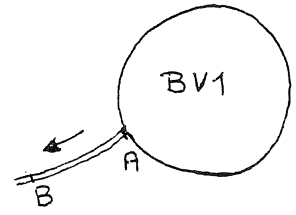
R2 = 0.6965 R = 0.8346 SL = 6.8304 R2 BRUT = 0.6971 F =

7.3 - Composition de la relation pluie efficace-débit et de la relation de propagation -

En fait, la propagation fait partie intégrante de la relation pluie nette-débit, quel que soit le type de bassin, mais on ne sait généralement pas l'identifier.

Cependant nous nous proposons d'illustrer sur deux exemples fictifs la conjugaison d'une transformation pluie nette-débit avec le transfert débit-débit.

Soit Q^A les débits engendrés par les pluies efficaces P à l'éxutoire A d'un petit bassin versant (BV1) circulaire et à forte pente et Q^B les débits résultants de la propagation entre A et B.



En variations unitaires :

$q_j = Q_j - Q_{j-1}$, on aura les relations suivantes

$$q^A_j = \sum_{i=1}^{m} a_i \, P_{j-i+1}$$

$$q^B_j = \sum_{i=1}^{K} b_i \, q^A_{j-i+1}$$

Soit : .700, .187, .067, .027, .012, .006 les coefficients de la fonction de transfert pluie-débit et .1, .7, .2 les coefficients de la fonction de transfert débit-débit.

On peut donc écrire, sous forme matricielle, cet exemple :

$$\begin{vmatrix} q^A_1 \\ q^A_2 \\ q^A_3 \\ \vdots \\ q^A_{11} \end{vmatrix} = \begin{vmatrix} .700 & 0 & 0 & 0 & 0 \\ -.513 & .700 & 0 & & \vdots \\ -.120 & -.513 & .700 & & \vdots \\ -.040 & -.120 & -.513 & & \vdots \\ -.015 & -.040 & -.120 & & \vdots \\ -.006 & -.015 & -.040 & & \vdots \\ 0 & -.006 & -.015 & & \vdots \\ 0 & 0 & -.006 & & \vdots \\ 0 & 0 & 0 & & \vdots \\ \vdots & \vdots & \vdots & & \end{vmatrix} \begin{vmatrix} P_1 \\ P_2 \\ P_3 \\ \vdots \\ P_9 \end{vmatrix} \quad (1)$$

soit : $[q^B] = [B][P]$

./.

et

$$\begin{vmatrix} q_1^B \\ q_2^B \\ q_3^B \\ \vdots \\ q_{11}^B \end{vmatrix} = \begin{vmatrix} .1 & 0 & 0 & 0 & \cdots & 0 \\ .7 & .1 & 0 & 0 & \cdots & \vdots \\ .2 & .7 & .1 & 0 & \cdots & \vdots \\ 0 & .2 & .7 & .1 & \cdots & \vdots \\ 0 & 0 & .2 & .7 & & \vdots \\ 0 & 0 & 0 & .2 & & \vdots \\ 0 & 0 & 0 & 0 & & \\ 0 & 0 & 0 & 0 & & \\ \vdots & \vdots & \vdots & \vdots & & \\ \vdots & \vdots & \vdots & \vdots & & \end{vmatrix} \begin{vmatrix} q_1^A \\ q_2^A \\ q_3^A \\ \vdots \\ q_{11}^A \end{vmatrix} \quad (2)$$

soit : $\left[q^B\right] = \left[B\right]\left[q^A\right]$

En effectuant le produit des deux matrices de coefficients, on obtient :

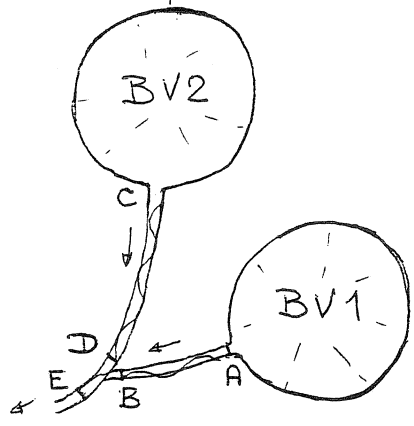
$\left[q^B\right] = \left[B\right]\left[A\right]\left[P\right]$ soit

$$\begin{vmatrix} q_1^B \\ q_2^B \\ q_3^B \\ \vdots \\ q_{11}^B \end{vmatrix} = \begin{vmatrix} .0700. & 0 & 0 & 0 & \cdots & 0 \\ .4387 & .0700 & 0 & & & \vdots \\ -.2311 & .4397 & .0700 & & & \vdots \\ -.1906 & -.2311 & .4397 & & & \\ -.0535 & -.1906 & -.2311 & & & \\ -.0191 & -.0535 & -.1906 & & & \\ -.0072 & -.191 & -.0595 & & & \\ -.0040 & -.0072 & -.0191 & & & \\ 0 & -.0040 & -.0072 & & & \\ 0 & 0 & -.0040 & & & \\ \vdots & \vdots & \vdots & & & \\ \vdots & \vdots & \vdots & & & \end{vmatrix} \begin{vmatrix} P_1 \\ P_2 \\ P_3 \\ \vdots \\ P_9 \end{vmatrix} \quad (3)$$

Soit également $q_j^B = \sum_{i=1}^{9} c_i \, P_{j-i+1}$

De même considérons un second bassin versant (BV2) contigu au précédent et ayant la même fonction de transfert pluie efficace-débit que

$$q_j^C = \sum_{i=1}^{m} a_i \, P_{j-i+1}$$



mais dont la fonction de transfert du débit entre C et D est différente

$$q_j^D = \sum_{i=1}^{m} d_i \, q^C_{j-i+1}$$

Considérons 3 cas de figure pour les coefficients d_i :

	d_1,	d_2,	d_3,	d_4,	d_5,	d_6,	d_7,	d_8
*	.15,	.40,	.25	.15	.05	0	0	0
**	0 ,	.15,	.40,	.25,	.15,	.05,	0,	0
***	0 ,	0,	.15,	.40,	.25,	.15,	.05,	0

En effectuant les mêmes calculs que précédemment (1), (2), (3) pour chacune de ces hypothèses, on obtient les D P F T qui permettent de calculer le débit au point D en fonction des pluies efficaces sur le BV2, et par conséquent les coefficients des F.T suivantes :

(4) * .1050, .3081, .2599, .1826, .0924, .0319, .0137, .008

(5) ** 0, .1050, .3081, .2599, .1826, .0924, .0319, .0137, .008

(6) *** 0, 0, .1050, .3081, .2599, .1826, .0924, .0319, .0137, .008

Si l'on compose chacune de ces fonctions de transfert avec celle obtenue pour le BV1 au point B, on obtient la fonction de transfert pluie net-débit au point E confluent des deux rivières BV1 + BV2 :

(7) * .1750, .8168, .5375, .2696, .1259, .0463, .0209, .0110, .006

(8) ** .0700, .6137, .5857, .3469, .2161, .1068, .0391, .0167, .008, .004

(9) *** .0700, .5087, .2776, .1920, .3416, .2743, .1898, .0954, .0319, .0137, .008, .0

A l'aide de cet exemple nous allons mettre en évidence les erreurs que l'on commet sur le calcul de la plus efficace, lorsqu'on déconvolue un hydrogramme avec une fonction de transfert qui n'est pas correcte (avec même rendement de pluie brute).

On considère une pluie efficace de 10 mm pendant un intervalle de temps unitaire correspondant à une pluie brute de 50 mm et les répartitions spatiales suivantes

./.

— elle n'affecte que le bassin BV1 et génère en E la crue de ruissellement

.70, 5.087, 2.776, .870, .335, .144, .072, .030,

si l'on déconvolue avec la D P F T correspondant au cas (7), on obtient les pluies efficaces :

$$\hat{P}_1 = 6.7,\ \hat{P}_2 = 0,\ \hat{P}_3 = 0,\ \hat{P}_4 = .1,\ \hat{P}_5 = .3$$

si l'on déconvolue avec la D P F T correspondant au cas (8) on obtient les pluies efficaces :

$$\hat{P}_1 = 8.7,\ \hat{P}_2 = 0,\ \hat{P}_3 = .2,\ \hat{P}_4 = 0,\ \hat{P}_5 = .6$$

si l'on déconvolue avec la D P F T correspondant au cas (9) on obtient les pluies efficaces :

$$\hat{P}_1 = 10.2,\ \hat{P}_2 = 0,\ \hat{P}_3 = 0,\ \hat{P}_4 = 0,\ \hat{P}_5 = 1,5$$

le cadrage de la pluie efficace est correct mais l'intensité peut être sous-estimée.

—elle n'affecte que le bassin BV2 et génère en E la crue de ruissellement (

0, 1.05, 3.081, 2.599, 1.826, .924, .319, .137, .08, .06

si l'on déconvolue avec la D P F T correspondant au cas (8) on obtient les pluies efficaces :

$$\hat{P}_1 = 0,\ \hat{P}_2 = 3.2,\ \hat{P}_3 = 2.3,\ \hat{P}_4 = 2.3\ ,\ \hat{P}_5 = 1.6$$

—elle n'affecte que le bassin BV2 et génère en E la crue de ruissellement(6

0, 0, 1.05, 3.081, 2.599, 1.826, .924, .319, .137, .08, .06

si l'on déconvolue avec la D P F T correspondant à (9) on obtient les pluies efficaces

$$\hat{P}_1 = 0,\ \hat{P}_2 = 0,\ \hat{P}_3 = 3.9,\ \hat{P}_4 = 4.1,\ \hat{P}_5 = 2.1$$

On remarque que pour les deux derniers cas de figures, si la somme des pluies efficaces est à peu près égale à 10 mm, les répartitions sont très différentes de l'épisode réel :

$$P_1 = 10,\ P_2 = 0,\ P_3 = 0,\ P_4 = 0,\ P_5 = 0$$

Ces exemples mettent en évidence la difficulté que l'on rencontre pour caler la relation pluie brute pluie efficace, puisque généralement on identifie une fonction de transfert pluie-efficace-débit moyenne pour un bassin versant sans pouvoir désagréger en sous bassins ayant chacun leur propre fonction de transfert

FONCTIONS DE TRANSFERT PLUIE EFFICACE-DEBIT