PRESENTATION DU LOGICIEL DIXLOI

T. LEBEL

1. INTRODUCTION

DIXLOI est un logiciel écrit en FORTRAN 77, permettant d'ajuster une sélection de distributions statistiques continues sur des échantillons de valeurs observées. La version présentée ici est celle de Mai 1987 qui est une refonte complète d'un programme écrit précédemment par M. BRUNET-MORET (ORSTOM).

Le logiciel a été développé dans la perspective d'assurer un véritable dialogue avec l'utilisateur. Le choix d'un langage de programmation standard a permis d'assurer sa portabilité sur plusieurs types d'ordinateurs munis d'un compilateur FORTRAN 77, en particulier les micro-ordinateurs compatibles IBM équipés d'un coprocesseur mathématique. Les modules d'ajustement d'une part et de représentation graphique d'autre part ont été séparés de telle manière que le programme puisse s'adapter à d'autres supports graphiques que ceux utilisés initialement (écrans graphiques IBM PC et APOLLO, ou table traçante HEWLETT PACKARD).

Cette note a pour objectif de décrire les fonctionalités actuelles et futures du logiciel. Pour une présentation plus détaillée des procédures interactives et des flux de données, le lecteur est invité à se reporter à la notice utilisateur disponible auprès du laboratoire d'hydrologie de l'ORSTOM.

2. FONCTIONS REALISEES

2.1 GENERALITES

On étudie la distribution d'une variable aléatoire (V.A) Z continue, dont on a observé N réalisations supposées indépendantes. Conformément à la démarche classique, on procède en deux étapes :

1) Caractérisation de la distribution expérimentale ; c'est à dire calcul de la moyenne, de la variance, du mode, de la médiane et des coefficients de dissymétrie (Fisher) et d'aplatissement. Les valeurs sont classées et on leur attribue une fréquence expérimentale selon la formule :

$$F_{obs} = (i\text{-}0,5)/N$$

Cette expression est un cas particulier d'une formulation plus générale : $(i+a)/(N+b)$. De nombreuses publications existent à ce sujet (eg. Adamovski, 1981, Cunnane, 1978) préconisant d'adopter différentes valeurs pour le couple (a,b) selon la distribution théorique dont on suppose que sont extraites les données, mais en règle générale le binome (-0,5 ; 0) constitue une approximation qui s'adapte bien aux distributions traitées ici. Il est à noter que le calcul de la fréquence empirique n'a de toute manière aucune influence sur le calcul des paramètres des distributions théoriques, lorsqu'il est effectué par la méthode des moments ou par celle du maximum de vraisemblance.

2) Ajustement d'une distribution théorique dont les paramètres $(\theta_j ; j = 1,m)$ sont calculés de manière à satisfaire un critère donné. Le critère peut être de nature statistique ou empirique. Les deux critères statistiques les plus couramment employés sont les suivants (voir e.g Mood *et al.* 1974) :

   a) Les deux (resp. trois) premiers moments de la distribution théorique doivent être égaux aux deux (resp. trois) premiers moments de la distribution expérimentale. C'est la méthode des moments, applicable aux lois à deux (resp. trois) paramètres.

   b) On maximise la "vraisemblance" de l'échantillon, sous l'hypothèse qu'il est tiré de la distribution théorique que l'on cherche à ajuster. Cette vraisemblance est le produit des probabilités des valeurs observées :

$$L\ (\theta_1,\theta_2 \ldots \theta_m) = \prod_{i=1}^{N} P_Z(z_i\ ;\ \theta_1, \theta_2 \ldots \theta_m)$$

où P_z est la distribution théorique que l'on cherche à ajuster, et (z_i ; i = 1, m) sont les valeurs observées.

Les m paramètres sont inconnus. On leur attribue les valeurs qui maximisent la vraisemblance que l'échantillon traité soit effectivement issu de la distribution théorique P_z. Ces paramètres sont en conséquence les estimateurs du maximum de vraisemblance. La procédure d'estimation consiste à annuler les dérivés partielles de $L\ (\theta_1, \theta_2 \ldots \theta_m)$ par rapport à chacun des paramètres :

$$\frac{\delta L(\theta_1, \theta_2 \ldots \theta_m)}{\delta \theta_j} = 0 \qquad j = 1, m$$

ou

$$\frac{\delta Ln\left[L\left(\theta_1, \theta_2 \ldots \theta_m\right)\right]}{\delta \theta_j} = 0 \qquad j = 1, m$$

puisque L étant le produit de m termes P_z (z_i ; θ_1, θ_2 ...θ_m), la transformation Log, permet de transformer ce produit en une somme de n termes.

On obtient ainsi m équations à m paramètres. Dans le cas fréquent où il est impossible d'expliciter analytiquement chaque paramètre, la résolution du système est numérique.

2.2 ESTIMATION DES PARAMETRES

Lorsque l'échantillon est réellement issu de la population mère ajustée, la méthode du maximum de vraisemblance est celle qui fournit les estimateurs efficaces (c'est à dire de variance minimum) des paramètres à estimer. C'est donc cette méthode d'ajustement qui a été retenue ici. Une exception a été faite concernant la loi de Gumbel, pour laquelle on peut opter pour un ajustement par la méthode du maximum de vraisemblance, ou par la méthode des moments, ou éventuellement par les deux méthodes. Cette particularité tient au fait que dans ce cas la méthode des moments fournit des estimateurs plus robustes des paramètres de la loi, lorsque la population mère n'est pas exactement distribuée selon la loi de Gumbel (Lebel, 1983).

2.3 LES DISTRIBUTIONS THEORIQUES

Même si on se limite aux distributions continues, il existe encore un choix très vaste qu'on ne peut explorer entièrement lorsqu'il s'agit de sélectionner la distribution la mieux adaptée à

l'échantillon étudié. Bien qu'il n'existe a priori aucune autre contrainte sur la V.A que l'hypothèse de continuité, DIXLOI a été développé par des chercheurs travaillant sur des phénomènes naturels tels que la pluie ou les débits de rivière par exemple, c'est à dire auxquels on ne peut pas fixer de borne supérieure. Le nombre de distributions disponibles a donc été limité à 10 :

1.GAUSS (normale)
2.GUMBEL (valeurs extrêmes de type I)
3.GALTON (log normale)
4.PEARSON III (gamma incomplète)
5.PEARSON V (gamma incomplète en 1/X)
6.GOODRICH (exponentielle généralisée en X)
7.FRECHET (valeurs extrêmes de type II)
8.LOG.GAMMA (de 1e espèce)
9.Loi des FUITES
10.Loi de POLYA.

Mise à part la loi normale qui sert d'étalon, toutes ces distributions sont à dissymétrie positive (ou droite), ce qui est en accord avec le comportement de beaucoup de processus géophysiques qui prennent leurs valeurs sur l'intervalle $[0,\infty[$. Le fait que quelques unes de ces lois ne soient pas bornées inférieurement n'est généralement pas gênant, car la probabilité de valeurs négatives est très faible.

Bien souvent ces lois ont des comportements voisins dans la partie centrale de la distribution observée, mais elles diffèrent dans leur comportement asymptotique lorsque Z tend vers l'infini. Pour mieux analyser le comportement d'une loi donnée pour les fortes valeurs de Z on peut procéder à la transformation :

$$U(P_z) = -\text{Log}(-\text{Log}\, P_z)$$

La fonction $U(P_z)$ croît de façon monotone de $-\infty$ à $+\infty$ et présente l'avantage d'une forte dilatation de l'échelle dans la partie supérieure de la distribution. Il devient alors immediat de comparer les formes des lois en extrapolation, comme cela est fait sur la figure 1 ci-dessous.On peut distinguer quatre grandes classes de comportement :

1. $Z \simeq U^{1/n}$ $(n>1)$
   * Loi Normale
   * GOODRICH $(\beta<1)$

2. $Z \simeq U$ (Décroissance asymptotiquement exponentielle)

* GUMBEL
* PEARSON III
* GOODRICH ($\beta = 1$)
* Loi des FUITES

3. $Z \simeq U^n$ $(n>1)$

* GOODRICH ($\beta > 1$)

4. $Z \simeq \exp(U^n)$ $(n>0)$ (Lois de type log.)

* GALTON
* PEARSON V
* FRECHET
* LOG GAMMA

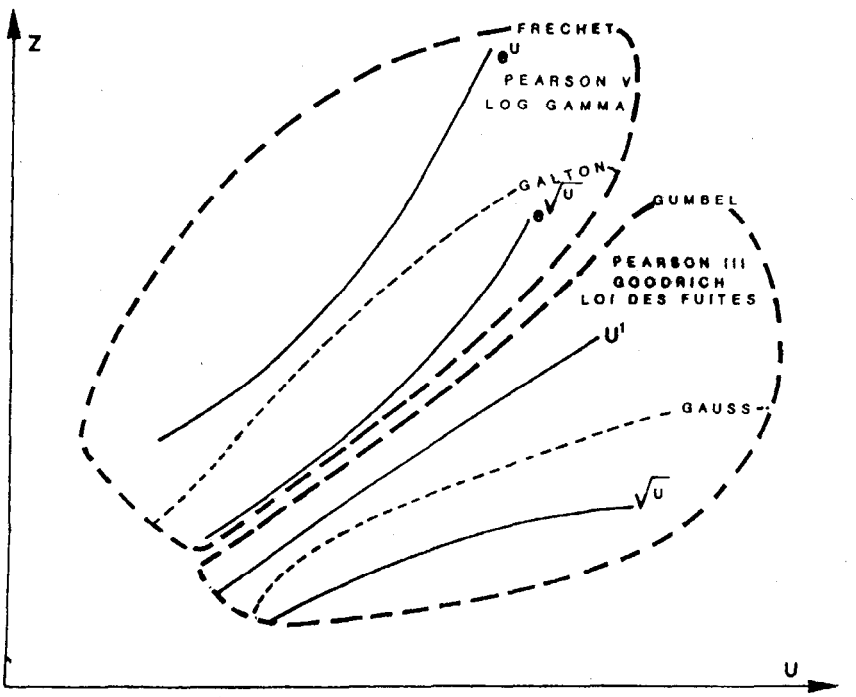


Figure 1 : Comportement asymptotique ($z \to \infty$) des lois utilisées.

DENSITES DE PROBABILITE ($z' = z - z_0$)	PARAMETRES			MOMENTS				REMARQUES
	Position	échelle	Forme	Moyenne (μ)	Variance (σ^2)	C.V	γ Fisher	
Gauss $]-\infty, +\infty[$ $f(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}$	μ	$\sigma(>0)$	/	μ	σ^2	σ/μ	0	Loi de référence (N) Dissymétrie nulle
Gumbel $]-\infty, +\infty[$ $f(z) = \frac{1}{\alpha} \exp - \left\{ -\frac{z-z_0}{\alpha} + \exp(-\frac{z-z_0}{\alpha}) \right\}$	z_0 (mode)	$\alpha(>0)$	/	$z_0 + .577\alpha$	$\Pi^2\alpha^2/6$ ou $\alpha = .78\sigma$		1,298	$F(z) = \exp\{-\exp-[(z-z_0)/\alpha]\}$ Dissymétrie positive et constante
Galton $]z_0, +\infty[$ $f(z) = \frac{1}{\beta s\sqrt{2\pi}} \frac{1}{z} e^{-\frac{(\log z'/s)^2}{2\beta^2}}$	z_0 (borne inf)	s (>0)	β (>0)	$z_0 + s\exp\left(\frac{\beta^2}{2}\right)$	$\mu^2(e^{\beta^2}-1)$	$(e^{\beta^2}-1)^{\frac{1}{2}}$	$e^{3\beta^2}+3e^{2\beta^2}$	log s = moyenne de log z β : e.type de log z borne inf. z_0 fixée
Pearson III $]z, \infty[$ $f(z) = \frac{z'^{\beta-1} e^{-\alpha z'}}{\alpha^\beta \Gamma^*(\beta)}$	z_0	$\alpha(\neq 0)$	$\beta(>0)$	$\frac{\beta}{\alpha}$ $(+ z_0)$	$\beta\alpha^2$	$1\sqrt{\beta}$ $(z_0 = 0)$	$2/\sqrt{\beta}$	Σ de β lois expo. de param. α β grand : $\rightarrow$ loi normale On note : $F(z) = FG(z/\alpha, \beta)$
Pearson V $f(z) = \frac{z'^{-1-\beta} \alpha^\beta e^{-\alpha/z'}}{\Gamma(\beta)}$	z_0	$\alpha(>0)$	$\beta(>0)$	$\frac{\alpha}{\beta-1}$ $(+ z_0)$	$\frac{\alpha^2}{(\beta-1)^2(2\beta-1)}$	$\frac{1}{2\beta-1}$ $(z_0 = 0)$		$F(z) = 1 - FG(\alpha/z' ; \beta)$ Variance faible
Goodrich Frechet $f(u) = \frac{1}{\lvert\alpha\beta\rvert} u^{1/\beta-1} \exp(-u^{1/\beta})$ $]z_0, \infty[$ $(u = \frac{z-z_0}{s})$	z_0 z_0	$s(>0)$ $s(>0)$	$\beta(>0)$ $\beta(<0)$	moment non centré d'ordre i : $m_i = \Gamma(i\beta + 1)$ $\gamma < 0$ pour $0 < \beta < 0,28$ $\gamma > 0$ ailleurs				$\beta \rightarrow 0$: loi de Gumbel $\beta \sim 0,28$: symétrique $\beta = 1$: Gumbel en extrap. Fisher Tippet II ou log u suit Gumbel
Log Gamma $f(u) = \frac{u^{\beta-1} e^{-u/\alpha}}{\alpha^\beta \Gamma(\beta)} \frac{1}{z} \left\{ u = \frac{\log(z/\alpha)}{\beta^*} \right\}$	α	α^*	β β^*	$\alpha\left\{1 + 1/(1-\beta^*)^\beta\right\}$	$\alpha^2 SM(5)^+$	.	.	log z/α suit Pearson III
Fuites $]0, \infty[$ $f(u) = ke^{-k} e^{-u} \frac{I_1^{**}(2\sqrt{ku})}{\sqrt{ku}}$ avec $u = \frac{z}{\alpha}$	/	$\alpha(>0)$	k	$k\alpha$	$2k\alpha^2$	$1/\sqrt{2k}$		Proche de Pearson III avec dissymétrie plus faible
Polya $P_K = P_{K-1} \frac{m + (k-1)d}{k(1+d)}$	μ	α	/	μ	$\mu(\alpha+1)$	$\left(\frac{\alpha+1}{\mu}\right)^{\frac{1}{2}}$	$\frac{(1+3\alpha+2\alpha^2)\mu^2}{\mu(1+\alpha)^3}$	Description des états secs et pluvieux (Persistance)

* Γ: fonction GAMMA **: fonction de BESSEL d'ordre 1. + $SM(5) = 1/(1-2\beta)\beta \cdot 1/(1-\beta)^2\beta$

La figure 1 montre que la plus grande prudence s'impose lorsque l'on cherche à ajuster une distribution qui sera utilisée en extrapolation. Les lois de type log notamment peuvent conduire à de valeurs exagérément fortes pour les fréquences rares.

Remarque : compte tenu que la loi de Gumbel peut-être ajustée par 2 méthodes différentes, 11 ajustements sont a priori possibles. La sélection des ajustements à calculer effectivement se fait dans le fichier de départ (voir section 3) mais peut être modifiée en cours d'exécution dans la version interactive du logiciel.

2.4 TESTS D'AJUSTEMENT

La méthode du maximum de vraisemblance permet d'obtenir pour chaque loi un jeu de paramètres unique. Il reste alors à choisir parmi ces lois, quelle est celle qui parait la mieux adaptée à l'échantillon traité. Ce choix peut reposer sur un test statistique ou sur une évaluation empirique des qualités respectives des ajustements lorsque l'on superpose la loi ajustée au nuage des points observés. Le logiciel DIXLOI offre à l'utilisateur différents critères de sélection, grâce d'une part au calcul de la valeur de deux tests statistiques et d'autre part à la visualisation réalisée dans le module graphique.

a) Le test du CHI2

C'est le test le plus couramment utilisé pour juger de l'adéquation d'un ajustement. La puissance de ce test est assez faible (voir Haan , 1982 p 178 pour ce qui concerne les applications hydrologiques), lorsque l'on traite des échantillons très dissymétriques, car il accorde un poids prépondérant aux valeurs pour lesquelles la densité de probabilité est la plus forte.

Le CHI2 se calcule de la manière suivante :

*l'intervalle de variation de la V.A est divisé en K classes équiprobables (au sens de la loi théorique ajustée).

*le nombre observé de valeurs par classe est alors : $N_{tk} = N/K$

*à l'aide de l'ajustement testé, on recalcule les fréquences théoriques de chaque valeur observée.

*On range ces fréquences dans chacune des k classes, ce qui permet de calculer un effectif N_k qui doit être le plus proche possible de l'effectif théorique N_{tk} pour chacune des classes

* on obtient la valeur du CHI2 à l'aide de la formule ci dessous :

$$\chi^2 = \sum_{k=1}^{K} \frac{(N_k - N_{tk})^2}{N_{tk}}$$

b) Le test de BRUNET-MORET (1978)

Ce test a été imaginé pour remédier à la mauvaise adéquation du CHI2 lorsque l'on traite des échantillons géophysiques. On calcule la distance entre d'une part la ligne brisée reliant les points observés, et d'autre part la fonction ajustée. Afin d'assurer un plus grand poids aux écarts mesurés sur les fréquences rares on adopte une distance Normale centrée réduite pour l'axe Y des fréquences :

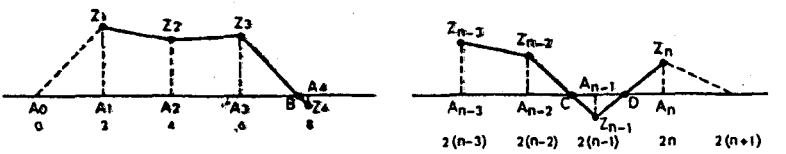
$$y = P^{-1}_{Gauss}(z)$$

$$avec \qquad P_{Gauss}(z) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt,$$

où μ et σ sont la moyenne et l'écart-type de la population considérée.

Sur l'axe des X, on porte simplement les rangs d'observations i ($i = 1,n$), et non les valeurs de la variable. Il est alors simple de montrer que la surface des triangles (figure 2) est la somme des écarts absolus ($y_{i+1} - y_i$, ($i = 1, n-1$)).

BRUNET-MORET (1978) a en outre étudié par simulation la distribution de ce test. Pour la loi normale, il semble que cette distribution, dont on ne peut trouver l'expression analytique exacte, puisse être approchée par une loi Gamma incomplète. Cette approximation a été étendue aux autres lois, ce qui n'est a priori pas légitime. Aussi, bien que la fréquence du dépassement du test soit calculée et stockée par le programme, cette valeur n'est pas prise en compte comme critère de sélection du meilleur ajustement.



La figure ci-dessus explicite les surfaces élémentaires prises en compte et élevées au carré :

pour le premier point :	A_0	Z_1	A_1
deuxième	A_1	Z_2	A_2
troisième	A_2	Z_3	B
$(n-2)^{ième}$	A_{n-3}	Z_{n-2}	C
$(n-1)^{ième}$	C	Z_{n-1}	D
$n^{ième}$	D	Z_n	A_{n+1}

On voit que lorsque la ligne brisée joignant les Z_i coupe l'axe des abscisses, les surfaces élémentaires prises en compte sont bien diminuées par rapport à celles prises en compte lorsque la ligne brisée ne coupe pas l'axe des abscisses.

Figure 2 : Calcul du test de Brunet-Moret (d'après Brunet-Moret, 1978)

c) Remarques comparatives

Malgré les critiques dont il fait fréquemment l'objet, le test du CHI2 a le mérite de fournir des valeurs graduées en probabilité (tables du CHI2, c'est-à-dire qui sont comparables entre elles quel que soit l'échantillon d'origine et les distributions qui lui sont ajustées. Son inconvénient majeur est que, comme la presque totalité des tests statistiques sa puissance (probabilité de rejeter une hypothèse fausse) est faible.

Le test de BRUNET-MORET, en théorie mieux adapté aux échantillons à forte dissymétrie positive doit être utilisé avec prudence :

- les résultats ne sont pas comparables sur deux échantillons de taille différente.
- sa graduation en probabilité ne peut en toute rigueur être appliquée qu'à l'ajustement de la loi Normale.

d) Visualisation graphique

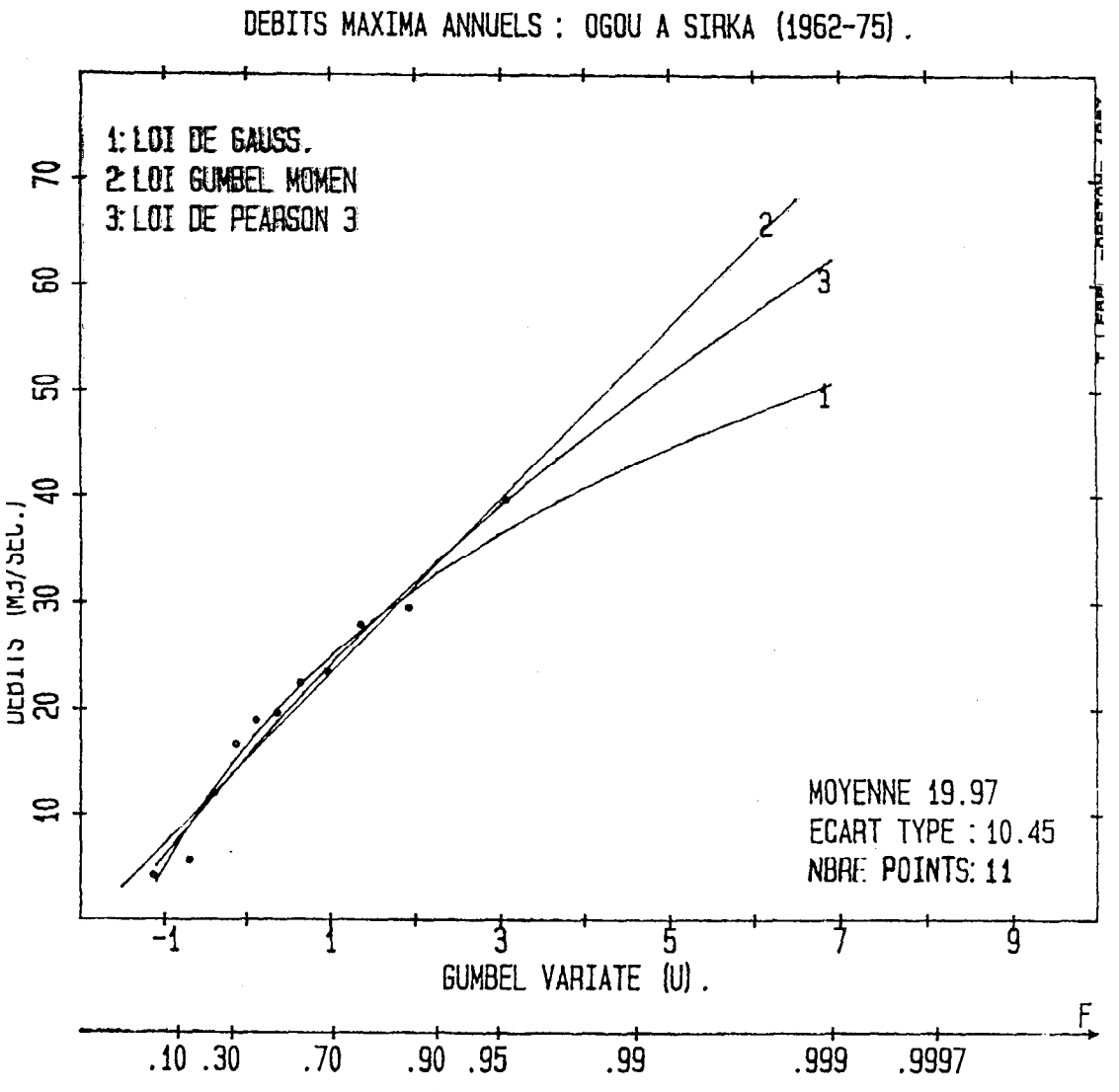
Compte tenu des insuffisances des tests statistiques, toute décision concernant la sélection d'un ajustement doit se fonder sur un examen conjoint du tracé de la loi ajustée et de la distribution expérimentale.

Le test du CHI2 n'est pas calculé lorque $N < 25$. Celui de Brunet-Moret ne l'est que si $N < 200$. Dans l'intervalle (25,200) où les deux tests peuvent être calculés conjointement, on peut s'appuyer sur les résultats comparés des deux tests pour opérer une premiére sélection des lois à tracer. En dehors de cet intervalle, la valeur du test non ajusté est mise à O. par convention.

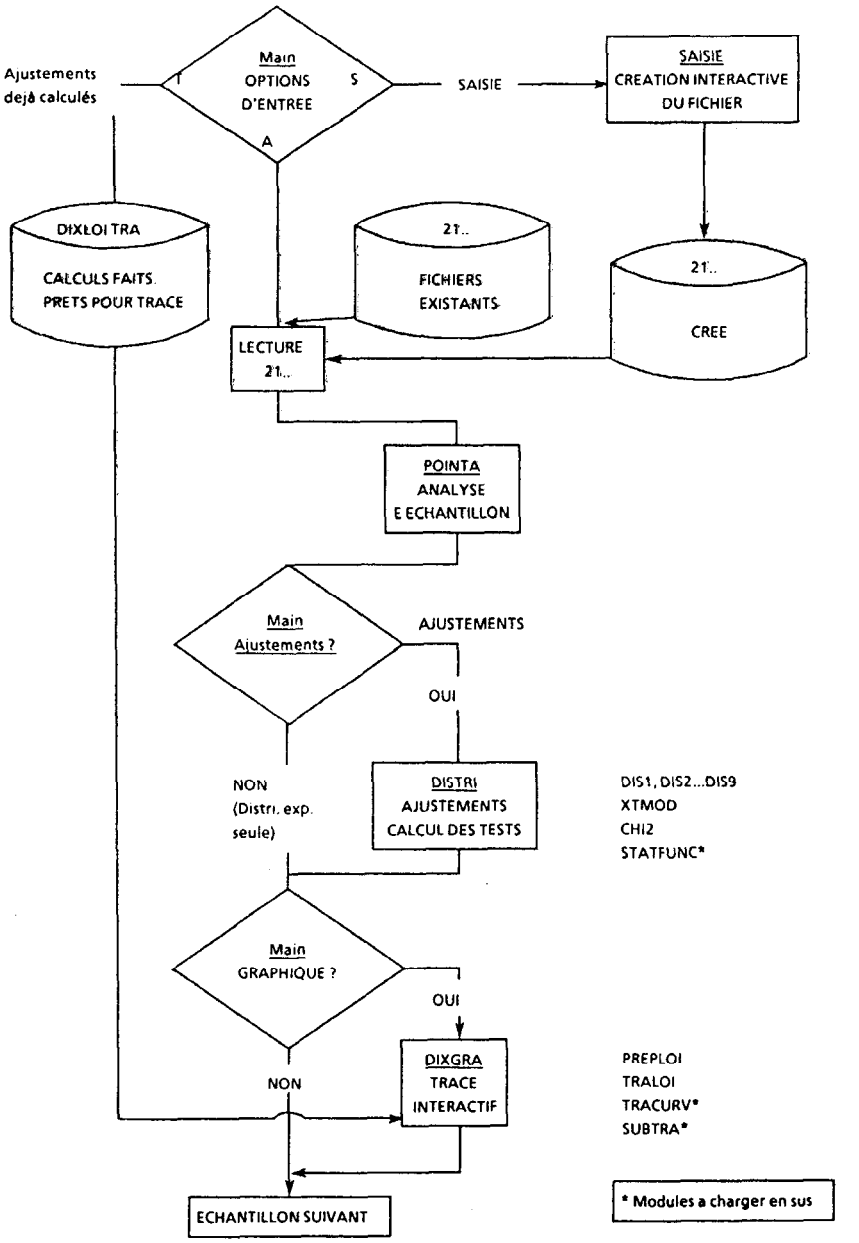
Remarque : Si $N < 10$, l'échantillon n'est pas traité.

La forte incertitude dont sont entachés les tests d'ajustement rend pratiquement obligatoire un examen visuel de la qualité des ajustements. L'oeil est dans bien des cas le meilleur juge, car il permet d'accorder l'importance voulue à la gamme de fréquences pour laquelle on désire obtenir le meilleur modèle. Le module graphique de DIXLOI permet de superposer jusqu'à trois distributions théoriques différentes sur la distribution empirique. Le tracé peut être effectué en coordonnées arithmétiques naturelles (Z, fréquences), ou en coordonnées de GAUSS (Y_{gauss}, Z) ou de GUMBEL (U_{gumbel}, Z).

c) Table traçante Hewlett Packard.



3. STRUCTURE GENERALE DU PROGRAMME



4. EVOLUTIONS PREVUES

4.1. COMPLEMENTS STATISTIQUES

Ces compléments sont de deux ordres :

- Calcul de statistiques,
- présentation des résultats.

Dans la première catégorie figurent :

- le calcul des intervalles de confiance,
- la possibilité de travailler sur des variables transformées (1/X ; Log X;X^n).

Dans la deuxième catégorie :

- le calcul des périodes de retour en années lorsque cela a un sens,
- calcul de la probabilité théorique affectée aux valeurs X observées par les lois ajustées,
- le tracé des histogrammes,
- les possibilité de choisir son découpage en classes pour le test du Chi 2.

4.2. AMELIORATION DU DIALOGUE AVEC L'UTILISATEUR

Cette partie du travail demandé concerne essentiellement la version micro-ordinateur de DIXLOI qui est celle diffusée à l'extérieur du Laboratoire. Les deux premiers objectifs à atteindre sont les suivants :

- mettre au point une saisie des données en utilisant une grille d'écran,
- parvenir à une véritable gestion interactive des sorties graphiques.

BIBLIOGRAPHIE

ADAMOVSKI, K. 1981. Plotting formula for flood frequency, Water Resources Bulletin, vol. 17, n° 2.

BRUNET-MORET, Y. 1978. Recherche d'un test d'ajustement. cah. ORSTOM, sér. Hydrol., vol. XV n° 3, pp 261, 280.

CUNNANE, C. 1978. Unbiased plotting positions. A review. Journal of Hydrology, vol 37 pp 205-222.

HANN, CH. T., 1982. Statistical methods in hydrology. The Iowa state University Press. Ames (Iowa).

LEBEL, T. 1983. Le problème des pluies extrèmes. Méthodes d'estimation et régionalisation. Séminaire INPG "crues et Précipitations Intenses". Grenoble

MOOD, A.M., F.A. GRAYBILL and D.C. BOES 1974. Introduction to the theories of statistics. Third Edition, Mc Graw-Hill, New York.

ANNEXE : Liste des sous-programmes

NOM	FONCTION	MODULE SOURCE *
SAISIE	Saisie en ligne du fichier entrée	MAIN
POINTA	calcul des statistiques expérimentales	POINTA
DISTRI	calcul des ajustements	DISTRI
XTMOD	calcul du test Brunet-Moret	"
CHI2	calcul du test du CHI2	"
DIS1	ajustement loi de Gauss	"
DIS2	ajustement loi de Gumbel (MV)	"
DIS3	ajustement loi de Gumbel (moments)	"
DIS4	ajustement loi de Galton	"
DIS5	ajustement loi de Pearson III (K = 5)	"
	ou de Pearson V (K = 6)	
DIS6	ajustement loi de Goodrich (K = 7)	"
	ou de Frechet (K = 8)	"
DIS7	ajustement loi log Gamma	"
DIS8	ajustement loi des fuites	"
DIS9	ajustement loi de Polya	"
DIXGRA	module graphique	MAIN
SELOUT	selection du terminal graphique	MAIN
HARCO	test de la table	MAIN
PREPGAU	transformation en coordonnées de Gauss	MAIN
PREPGUM	transformation en coordonnées de Gumbel	MAIN
TRALOIS	tracé des lois statistiques	MAIN
TRACURV	tracé d'une courbe	TRACURV
ERROUT	sortie des codes erreurs	ERROUT