DEA DE BIOSTATISTIQUE

Université de MONTPELLIER II
Ecole Nationale Supérieure Agronomique de MONTPELLIER

DIFFERENTES APPROCHES DES PROBLEMES DE SEUIL EN PLUVIOMETRIE

GOUDET Christophe
Juin 1992

ORSTOM - MONTPELLIER

Responsables de stage: A. BERLINET
H. LUBES

Remerciements:

A l'issue de ce stage, je tiens à remercier toutes les personnes qui m'ont aidé dans mes recherches. Ainsi je remercie vivement M. BERLINET et Melle LUBES pour la qualité de leur encadrement. Les discussions avec M. ELGUERO et Mme HOLMES m'ont permis d'éclaircir de nombreux problèmes, ce dont je leur suis particulièrement reconnaissant.

Pour tous les problèmes informatiques, la patience et la disponibilité de Catherine, Martine et Thierry m'ont été d'une grande utilité, je les en remercie.

Table des matières

Introduction

L'estimation des risques est un problème auquel sont confrontés de nombreux hydrologues. La connaissance des fréquences d'apparition d'évènements rares permet de dimensionner de nombreux ouvrages hydrauliques. On cherche ici une méthodologie permettant de déterminer la période de retour décennale des pluies journalières, c'est-à-dire la hauteur d'eau journalière qui tombe en moyenne une fois tous les dix ans. Dans le cadre de cette étude, on dispose de deux fichiers renfermant les relevés pluviométriques du Sénégal et d'Abidjan, pendant la saison humide, pour des durées respectives de 69 et 12 ans.

Les contraintes liées aux mesures modifient le problème: des phénomènes d'évaporation peuvent avoir lieu au sein du pluviomètre, ainsi des jours seront comptabilisés sans pluie alors qu'une certaine quantité d'eau était tombée, le seuil de sensibilité de l'appareil est de 0.1 mm, les précipitations inférieures à ce seuil ne sont pas mesurées.

On considère que ces phénomènes ne concernent que les faibles valeurs. Cette réflexion a conduit certains auteurs à distinguer les valeurs entachées de fortes incertitudes relatives de celles considérées comme "bonnes". Après avoir décrit sommairement les principes de la méthode proposée par BRUNET-MORET, qui repose sur la considération de ces contraintes, on développera les fondements d'autres méthodes permettant une approche différente du problème.

Chapitre 1

Présentation de la méthode développée par BRUNET-MORET

1.1 Présentation

Sur l'échantillon des valeurs non nulles, on procède à l'ajustement d'une loi de probabilité tronquée (sur les séries étudiées, une loi Log-Normale) à un seuil So en-deçà duquel les données sont considérées comme "douteuses". Les paramètres de cette cette loi sont déterminés par la méthode du maximum de vraisemblance. Cette opération est faite pour diverses valeurs de So, So variant de 0.09 à 14.09mm. Le seuil So retenu est celui qui fournit le meilleur ajustement graphique des données sur la loi Log-Normale, dont les paramètres ont été préalablement calculés. Les fréquences calculées sont corrigées par une estimation de la probabilité de valeurs nulles en fonction des paramètres de l'ajustement.

1.2 Discussion

Les questions posées par cette méthode sont multiples: on constate en effet, qu'il y a non stabilité des paramètres σ et s de la loi Log-Normale lorsque So varie.

Dans ces conditions, le critère de sélection de So est-il pertinent? Ce seuil est introduit à partir de considérations physiques réelles. Mais peut-on considérer que l'ajustement de données sur une loi reflète la réalité physique du seuil So. A savoir, ce seuil ainsi retenu, distingue-t-il vraiment les données "douteuses", de celle considérées comme bonnes, ou devient-il seulement un paramètre d'ajustement?

Le choix de la loi Log-Normale comme loi d'ajustement, n'est-il pas fait seulement pour simplifier les calculs?

Comment s'explique la variabilité des résultats obtenus dans la détermination de la hauteur d'eau décennale [LUBES]. Cette dernière sera notée par la suite H_{dec}:

$\hat{H_{dec}}$=168 mm sans troncature

H_{dec}=110 pour So=14.09 mm

$\hat{H_{dec}}$=125 pour So=4.09 mm (So optimal)

Est-ce parce que l'ajustement est réalisé sur les données d'Abidjan (soit 12 ans de relevés) ou est-ce dû à la méthode employée?

Quel est l'importance des valeurs faibles vis-à-vis de l'objectif recherché? Notons que cette méthode ne fournit pas d'intervalle de confiance.

On va, par des approches différentes du problème, essayer de répondre à ces questions.

Chapitre 2

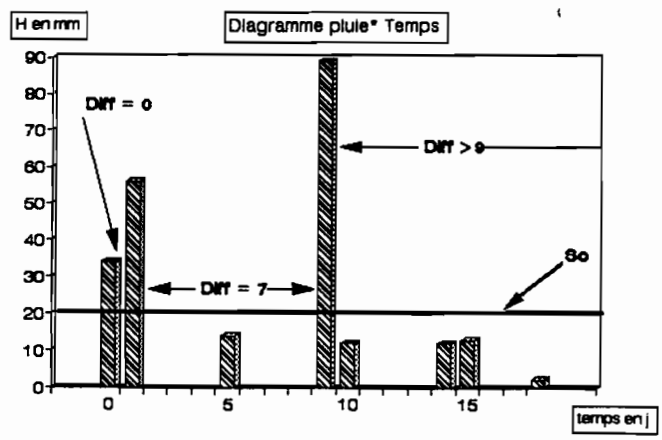
Evaluation par bootstrap

2.1 Résultats obtenus sur l'échantillon d'origine

Méthode utilisée: Régression linéaire

Pour l'échantillon du Sénégal (soixante neuf ans de données), dont les observations sont classées et numérotées par ordre d'apparition chronologique (soit deux variables: la hauteur d'eau et la date), on trace le graphique hauteur d'eau*date. Dans une seconde étape, on sélectionne les données qui dépassent un certain seuil So. Ensuite, on mesure l'écart (variable diff), en nombre de jours, qui sépare deux pluies supérieures au seuil So (cf diagramme).

La moyenne des écarts est ensuite calculée. L'idée initiale de cette méthode est qu'en faisant varier le seuil So, on doit parvenir à déterminer le seuil auquel est associé un écart moyen de dix ans, c'est-à-dire la période de retour recherchée. La connaissance des relevés pluviométriques sur une grande période permettrait de définir précisement,par cette méthode, la période recherchée. Mais la connaissance de 69 années de précipitations n'est pas suffisante pour déterminer avec précision la hauteur d'eau décennale, d'ou l'idée d'une étude de l'existence d'une relation entre le seuil fixé et la moyenne des écarts observés (variable Mean). On trouve la relation suivante:



$$ln(Mean) = a * (H) + b + \varepsilon$$

Ou H=hauteur d'eau journalière et Mean la moyenne des écarts associée à H.
$\hat{a} = 0.047 \quad \hat{\sigma} = 0.005$
$\hat{b} = 1.66 \quad \hat{\sigma} = 0.03$
R^2=0.988
N=140 observations

De l'équation précédente, on déduit la hauteur d'eau décennale, on fixe mean=1530 (une année de mesure comporte 153 valeurs):

$$\hat{H}_{dec} = \frac{ln(mean) - \hat{b}}{\hat{a}}$$

Soit $\hat{H}_{dec}$=120.63 mm de pluie

Cette valeur de 120.63 mm est voisine de 128 mm (cette valeur de 128 mm ayant été calculée à partir de l'échantillon du Sénégal , pour un seuil de troncature de 14 mm) trouvée par la méthode de Brunet-Moret.

Cependant, l'obtention d'une telle valeur n'a d'intérêt que si on peut lui associer une incertitude. Les variances de $\hat{H}_{dec}$,R^2, $\hat{a}$, $\hat{b}$ vont être déterminées par bootstrap.

2.2 Calcul de l'incertitude de H_{dec} par la technique du bootstrap

2.2.1 Description de la méthode [EFRON]:

On tire au hasard et avec remise dans l'échantillon de départ des valeurs de hauteur d'eau. A chaque tirage, on affecte un numéro d'ordre. Cette opération est réalisée 10557 fois de manière à recréer un échantillon de taille identique à l'original. L'indépendance des hauteurs d'eau a été vérifiée à l'aide du test de Wald et Wolfowitz.

Note:

Il y a indépendance des hauteurs d'eau, mais l'indépendance jours sans pluie et jours pluvieux n'est pas vérifiée. La loi du nombre d'épisodes pluvieux consécutifs suit une loi de Pareto en $\frac{1}{x^2}$ [JOHNSON and KOTZ], et non une loi géométrique, comme cela aurait été le cas si les observations avaient été indépendantes.

2.2.2 Résultats

Les mêmes opérations de calcul que celles décrites précédemment sont faites 250 fois. Les valeurs de $\hat{a}$, $\hat{b}$, R^2 et $\hat{H}_{dec}$ sont stockées. On touve:

$\overline{H}_{dec} = 117.25\ mm \quad \hat{\sigma}_{H_{dec}} = 7.419\ mm$
$\overline{a} = 0.0494 \quad \hat{\sigma}_a = 0.004$
$\overline{b} = 1.563 \quad \hat{\sigma}_b = 0.108$
$\overline{R^2} = 0.973 \quad \hat{\sigma}_{R^2} = 0.031$
B=250 rééchantillonnages

2.3 Incertitude de $\overline{R^2}$

La connaissance de $\overline{R^2}$ permet de valider à postériori le modèle retenu. L'écart-type associé à R^2 peut être estimé par:

$$\hat{\sigma}_{R^2} = \left(\frac{\sum_{b=1}^{B}(R_b^2 - \overline{R^2})^2}{B-1}\right)^{\frac{1}{2}}$$

Soit $\hat{\sigma}_{R^2}$=0.031.

En pratique, on préfère connaître l'intervalle , compris entre 0 et 1, contenant 95% des observations. L'histogramme de répartition des R^2 ne suit pas une loi Normale (Cf annexe 1). La répartion de fréquences est dissymétrique car R^2 est borné par 1. Cette particularité ne permet pas de déterminer avec précision les percentiles.

Une transformation des données peut-être utilisée pour déterminer l'intervalle de confiance. Celle-ci est:

$$\hat{\Phi} = tan(R^2) = 0.5 * ln(\frac{1+R^2}{1-R^2})$$

Les données transformées sont classées par ordre croissant et numérotées. Par ce calcul on augmente la dispersion des points de l'échantillon. Les percentiles peuvent alors être determinés de manière précise. Sur l'échantillon des données transformées on retire les 2.5% premières valeurs et les 2.5% dernières. Les valeurs $\hat{\Phi}$ des première et dernière observations déterminent l'intervalle de confiance, on note x et y ces valeurs de $\hat{\Phi}$.

On en déduit que: $R^2 \in [tanh(x), tanh(y)]$

Soit $R^2 \in [0.8812, 0.996]$

Le choix du modèle de régression utilisé précédemment n'est pas remis en cause par cet intervalle de confiance.

2.4 Influence du seuil

L'incertitude pesant sur les petites valeurs a-t-elle une influence sur la détermination de H_{dec}? Dans la méthode proposée par Brunet-Moret, le choix d'un seuil entre les valeurs douteuses et celles considérées comme bonnes faisait varier notablement H_{dec}.

La technique du bootstrap est employée pour déterminer si le seuil a une influence notable. Pour chaque rééchantillonnage, le calcul des paramètres de la droite de régression est réalisé pour deux seuils, on distingue deux échantillons:

-Tous les couples (H Mean) sont conservés, soit So=0 mm.

-Seul les couples (H Mean) dont Mean>20mm sont utilisés dans le calcul de la droite de régression, soit So=20 mm.

Les résultats sont les suivants:

B=250 rééchantillonnages

So= 0 mm

$\overline{R^2} = 0.9743$ $\hat{\sigma}_{R^2}$=0.0302

$\overline{H}_{dec} = 11.43\ mm$ $\hat{\sigma}_{H_{dec}}$=7.36 mm

So=20 mm

$\overline{R^2} = 0.954$ $\hat{\sigma}_{R^2}$=0.061

$\overline{H}_{dec} = 120.96\ mm$ $\hat{\sigma}_{H_{dec}}$=10.19 mm

Le seuil fait varier légèrement la valeur de H_{dec}, mais cette dernière reste dans l'intervalle de confiance.

2.5 Importance de la taille de l'échantillon

Le programme de bootstrap a été testé sur les données d'Abidjan (douze ans de valeurs). Les résultats obtenus sont les suivants:

$\overline{H}_{dec} = 121.35mm$ σ=26.20
$\overline{a} = 0.0462$ σ=0.0085
$\overline{b} = 1.406$ σ=0.148
$\overline{R^2} = 0.887$ σ=0.06
N=200 rééchantillonnages

Les résultats montrent que la dispersion augmente, mais l'estimation de H_{dec} est relativement proche de la valeur précédemment calculée. L'importance de l'écart-type s'explique par la difficulté qu'il y a de chercher une période de retour décennale avec seulement douze ans de données.

2.6 Critique de la méthode utilisée:

Cette méthode a été construite à partir d'une approche empirique. La moyenne des écarts est calculée pour chaque seuil de pluviométrie, les écarts ainsi mesurés ne sont pas indépendants des précédents. Les variables sont toutes corrélées les unes aux autres. Donc il existe des limitations à de la droite de régression. De plus, la nature du modèle de régression n'est pas forcément ubiquiste.

Une autre méthode consisterait à s'affranchir du modèle de régression. Il faudrait reprendre l'idée initiale, à savoir la détermination de l'écart moyen dont la valeur correspond à la période de retour recherchée, le tout sur des échantillons créés par bootstrap.

L'échantillon sur lequel les études ont été faites est d'une taille peu commune. Bien souvent les séries pluviométriques sont plus courtes et l'application d'une telle méthode sur des échantillons de taille plus restreinte pourrait s'avèrer décevant.

D'un point de vue pratique, cette méthode est facile à mettre en oeuvre. Aucune hypothèse sur la nature de la loi de distribution des hauteurs d'eau n'a été faite; toutefois les temps de calcul sont importants, 36 heures sont nécessaires pour effectuer 250 rééchantillonnages et les calculs des droites de régression sur une station de travail SUN.

Chapitre 3

Méthode du renouvellement

3.1 Présentation

Cette méthode a déja fait l'objet d'étude en hydrologie, notamment dans l'estimation des probabilités de crues, des amplitudes de marées ...

On combine deux évènements:

-nombre d'épisodes pluvieux dépassant un certain seuil chaque année, soit n_1,n_2,..,n_{N_A} valeurs.

-pour chaque pluie, on note la hauteur d'eau mesurée, soit $H_1, H_2, .., H_{N_P}$ soit N_P valeurs.

Avec N_A le nombre d'années et N_P le nombre de pluies.

On calcule F(H) pour qu'une pluie, H* la plus forte de l'année (ou de la période de référence choisie), ne dépasse pas la valeur H.

Cela revient à chercher:

$$P[H^* < H] = \sum_{k=0}^{+\infty} P[\exists\, k \ pluies \ > So \ et \ toutes < H]$$

On note:

$$P(k) = Prob[\exists k \ pluies > So \ au \ cours \ de \ l'année]$$

et

$$G(H) = Prob[H^* < H \ / \ H^* > So]$$

On en déduit:

$$F(H) = \sum_{k=0}^{+\infty} P(k).G(H)^k$$

Dans la littérature [J. Miquel], G(H) est une loi de type exponentielle, alors que P(k) est souvent modélisée par une loi de Poisson, voire une loi Binomiale négative. Ces deux hypothèses s'interprètent facilement si on considère que les épisodes pluvieux étudiés, sont rares tant dans leurs amplitudes que dans leurs fréquences.

3.2 Choix et calcul des paramètres de P(k) et G(H)

On suppose que les lois P(k) et G(H) suivent respectivement des lois de Poisson et de Weibull.

Soit:

$$P(k) = \frac{e^{-\mu}\mu^k}{k!} \quad et \quad G(H) = 1 - e^{-\rho(H-So)^p}$$

La détermination des paramètres s'effectue en maximisant la vraisemblance.

3.2.1 Calcul de la vraisemblance

On suppose que les évènements sont indépendants.
La vraisemblance s'écrit:

$$V = Prob(avoir\ n_1\ pluies\ la\ 1^{ère}\ année)*$$

$$\vdots$$

$$Prob(avoir\ n_{N_A}\ pluies\ la\ N_A^{ème}\ année)*$$

$$Prob(la\ 1^{ère}\ pluie = H_1) = \frac{\partial G(H_1)}{\partial H}*$$

$$\vdots$$

$$Prob(la\ N_p^{ème}\ pluie = H_{N_p}) = \frac{\partial G(H_{N_p})}{\partial H}$$

Avec:

$$\frac{\partial G}{\partial Q} = \rho e^{-\rho(H-So)^p}$$

$$V = \lambda e^{\mu N_A}\, \mu^{N_p}\, p^{N_p}\, \rho^{N_p} [\prod_{i=1}^{N_p}(H_i - So)]^{p-1}\, e^{-\rho\sum_{i=1}^{N_p}(H_i-So)^p}$$

$(\hat{\rho}, \hat{\mu}, \hat{p})$ s'obtiennent en résolvant:

$$\frac{\partial V}{\partial \mu} = 0 \quad \frac{\partial V}{\partial \rho} = 0 \quad \frac{\partial V}{\partial p} = 0$$

On obtient le système suivant:

$$\begin{cases} \frac{N_p}{\mu} - N_A = 0 \\ \frac{N_p}{\rho} - \sum_{i=1}^{N_p}(H_i - So)^p = 0 \\ \frac{N_p}{p} + \sum_{i=1}^{N_p} ln(H - So) - \rho\sum_{i=1}^{N_p}(H_i - So)^p ln(H - So) = 0 \end{cases}$$

3.2.2 Choix du seuil So

On a supposé que P(k) suit une loi de Poisson. Mais les hypothèses faites sur la distribution des évènements n'ont pas été vérifiées. Pour divers seuils So (So variant de 30 mm à 100 mm), on trace l'histogramme du nombre d'années ayant k pluies dépassant ce seuil (k varie de 0 à l'infini). Pour chaque seuil, l'hypothèse de distribution Poissonnière est vérifié. Le critère d'ajustement retenue est le test du χ^2. Le seuil So choisi est celui qui minimise le χ^2, vérifiant par la même occasion l'hypothèse de distribution Poissonnière. Dans cette méthode le choix de So n'est pas fait pour distinguer les valeurs "bonnes" des douteuses, mais plutôt pour constituer un échantillon d'évènements rares homogène.

Sous ces contraintes, on obtient So=51mm.

La résolution des équations donne:

$$\hat{\rho} = 0.0384$$

$$\hat{p} = 1.043$$

$$\hat{\mu} = 2.59$$

3.2.3 Vérification des hypothèses

Loi de Poisson

Le seuil So a été choisi en fonction du χ^2, celui-ci est de 10.59, le nombre de ddl est de 6, et la limite pour refuser l'ajustement au risque de 5% est de 12.59.

Loi de Weibull

L'ajustement de la fonction de répartion empirique G(H) par une loi de Weibull dont les paramètres ont été calculés précédemment est concluant. Le χ^2 est de 56 avec 48 ddl, soit une limite de refus de l'ajustement au risque de 5% de 65.34.

La méthode du maximum de vraisemblance nous donne une valeur de p=1.043. Cette valeur est proche de 1, mais l'ajustement par une loi exponentielle (loi de Weibull avec p=1) n'a pas été retenu selon le critère du χ^2, le modèle ne peut donc pas être simplifié.

Remarque: Plusieurs seuils ont été essayés, il s'est avéré que l'ajustement par une loi de Poisson n'est pas toujours concluant selon le critère du χ^2. La modélisation de P(k) par une loi Binomiale négative n'a pas donné de meilleur résultats.

3.3 Calcul de la hauteur d'eau décennale

On a:

$$F(H) = \sum_{k=0}^{+\infty} P(k).G(H)^k$$

Si H est considéré comme un évènement rare (on peut considérer que H_{dec} est un évènement rare), on a $G(H_{def}) \simeq 1$, et:

$$F(H) \simeq \sum_{k=0}^{+\infty} P(k)(1 - k(1 - G(H))$$

$$F(H) \simeq 1 - \hat{\mu}(1 - G(H))$$

$$\hat{H_{dec}} = So + (\frac{ln\hat{\mu}T}{\hat{\rho}})^{\frac{1}{\hat{p}}} \quad T \text{ période de retour}$$

Soit $\hat{H}_{dec}$=121.57mm

3.4 Estimation de l'incertitude

3.4.1 Calcul des variances de $\hat{\mu}$, $\hat{\rho}$, $\hat{p}$

Soit W=ln(V). Ecrivons le développement limité de W à l'ordre 2 et autour de $\hat{\mu}, \hat{\rho}, \hat{p}$

$$W = W(\hat{\mu}\hat{\rho}, \hat{p}) + \frac{\partial W}{\partial \mu}(\hat{\mu},\ \hat{\rho},\ \hat{p})(\mu - \hat{\mu}) + \ ie \text{ pour } \rho \text{ et } p$$

$$+\frac{\partial^2 W}{\partial \mu^2}(\hat{\mu},\ \hat{\rho},\ \hat{p})(\mu - \hat{\mu})^2 + \ ie \text{ pour } \rho \text{ et } p$$

$$+\frac{\partial^2 W}{\partial \mu \partial \rho}(\hat{\mu},\ \hat{\rho},\ \hat{p})(\mu - \hat{\mu})(\rho - \hat{\rho}); ie \text{ pour } \rho \text{ et } p$$

$$+...$$

$$Et \quad \frac{\partial V}{\partial \mu}(\hat{\mu},\ \hat{\rho},\ \hat{p}) = 0 \implies \frac{\partial W}{\partial \mu}(\hat{\mu}, \hat{\rho}, \hat{p}) = 0 \quad ie \text{ pour } \rho \text{ et } p$$

$$D'où \quad V = V(\hat{\mu}, \hat{\rho}, \hat{p}).exp(\frac{\partial^2 W}{\partial \mu^2}(\hat{\mu}, \hat{\rho}, \hat{p})(\mu - \hat{\mu})^2 + ...)$$

Les variables aléatoires $\hat{\mu}$, $\hat{\rho}$, $\hat{p}$ suivent une loi Normale dont les variances et covariances sont issues de la matrice des derivées secondes de V [NORMAN JOHNSON]. On obtient une loi Normale à trois dimensions.

Le résultat approché est [J. MIQUEL]:

$$Var\ \rho = \frac{A_{22}.A_{33} - A_{32}^2}{Det} \qquad Var\ \mu = \frac{A_{11}.A_{33} - A_{31}^2}{Det}$$

$$Var\ p = \frac{A_{11}.A_{22} - A_{12}^2}{Det} \qquad Cov\ \rho\mu = -\frac{A_{12}.A_{33} - A_{32}.A_{13}}{Det}$$

$$Cov\ \rho\ p = -\frac{A_{12}.A_{32} - A_{22}.A_{13}}{Det} \qquad Cov\ \mu\ p = -\frac{A_{11}.A_{32} - A_{12}.A_{13}}{Det}$$

Det : déterminant de la matrice des $A_{i,j}$

Avec:

$$A_{11} = \frac{N_p}{\hat{\rho}^2} \qquad A_{22} = \frac{N_p}{\hat{\mu}^2} \qquad A_{12} = 0$$

$$A_{33} = \frac{N_p}{\hat{p}^2} + \sum_{i=1}^{N_p} (H_i - So)^{\hat{p}} ln^2(H_i - So) \qquad A_{32} = 0$$

$$A_{13} = \sum_{i=1}^{N_p} (H_i - So)^{\hat{p}} ln(H_i - So)$$

L'application numérique donne:
$Var\ \rho = 9.50 * 10^{-6}$ $Var\ \mu = 0.0432$ $Var\ p = 1.85 * 10^{-5}$
$Cov\ \rho\ \mu = -0.04323$ $Cov\ \rho\ p = -2.52 * 10^{-6}$ $Cov\ \mu\ p = 0$

3.4.2 Variance de $\hat{H}_{dec}$

Le calcul de la variance de $\hat{H_{dec}}$ s'effectue en réalisant le développement limité de H_{dec} au voisinage de $\hat{H}_{dec}$.

$$H_{dec} = So + (\frac{ln\hat{\mu}T}{\hat{\rho}})^{\frac{1}{\hat{p}}}$$

$$Posons \quad \hat{x_T} = H_{dec} - So = (\frac{ln\hat{\mu}T}{\hat{\rho}})^{\frac{1}{\hat{p}}}$$

$$H_{dec} - \hat{H}_{dec} = -\frac{\hat{x_T}}{\rho\ p} d\rho + \frac{\hat{x_T}^{1-p}}{\rho\mu p} d\mu - \frac{\hat{x_T} ln(\hat{x_T})}{p} dp + ...$$

On élève au carré, et on prend l'espérance :

$$var\hat{H}_{dec} \simeq \frac{(\hat{H}_{dec} - So)^2}{\hat{\rho}^2\hat{p}^2} Var\ \rho \ + \frac{(\hat{H}_{dec} - So)^{2(1-\hat{p})}}{\hat{\rho}^2 p^2 \hat{\mu}^2} Var\ \mu$$

$$+ \frac{(\hat{H}_{dec} - So)^2 ln^2(\hat{H}_{dec} - So)}{\hat{p}^2} Var\ p \ - 2\frac{(\hat{H}_{dec} - So)^{2(1-\hat{p})}}{\hat{\rho}^2\hat{p}^2\hat{\mu}} Cov\ \rho\mu$$

$$+2\frac{(\hat{H_{dec}} - So)^2 ln(\hat{H_{dec}} - So)}{\hat{\rho}\hat{p}^2} Cov\ \rho\ p \ - 2\frac{(\hat{H}_{dec} - So)^2 ln(\hat{H}_{dec} - So)}{\hat{\rho}\hat{\mu}\hat{p}^2} Cov\ \mu p$$

Les calculs faits, on trouve:
$Var(\ \hat{H}_{dec}) = 45.69$ *soit* $\sigma = 6.76mm$

3.5 Vérification des hypothèses

Répartion des pluies uniformément au cours du temps: cf annexe 2
Le test du χ^2 est concluant.
Qualité de l'ajustement: cf diagramme
-Ajustement de P(k) par une loi de Poisson: cf annexe 2
-Ajustement de G(k) par une loi de Weibull: cf annexe 3

3.6 Discussion

On constate que les valeurs de $\hat{H}_{dec}$ et $Var\ \hat{H}_{dec}$ sont proches de celles trouvées par la méthode du bootstrap. Cette méthode nécessite peu de calculs, mais elle introduit des hypothèses fortes sur la nature et le comportement du phénomène étudié. Elle peut-être utilisée pour calculer des périodes supérieures à dix ans. Ainsi en hydrologie, ces méthodes permettent de déterminer l'amplitude des crues millénaires.

Chapitre 4

Estimation non paramétrique de H_{dec}

4.1 Propriétés de l'estimateur à noyau

On cherche à construire un estimateur à noyau de la densité. Tout d'abord on choisit comme estimateur, celui de PARZEN-ROSENBLATT:

$$f_n(x) = \frac{1}{n\,\delta_n} \sum_{i=1}^{n} K\left[\frac{X_i - x}{\delta_n}\right]$$

Où n est le nombre d'individus de l'échantillon et δ_n la fenêtre ou paramètre de lissage, X_i une variable aléatoire et K le noyau.

Le choix d'un estimateur à noyau se justifie si on considère ces différentes propriétés:

outre sa simplicité de programmation, il s'avère que f_n converge vers f au sens L_1 pour toute densité dès que $\frac{1}{h}$ et nh tendent vers l'infini [BERLINET]. On a l'inégalité suivante:

$$P(\int_I |f_n - f| > \varepsilon) \leq exp(-\frac{1}{3} n\varepsilon^2)$$

Pour tout $n \geq n_0$, où n_0 dépend de δ_n, f, K et ε

L'erreur qu'on ne peut éviter vaut:

$$inf\ E \int_I |f_n - f| \geq \frac{1}{\sqrt{528n}}$$

Cette erreur ne peut être minorée, quelque soit le choix de f, δ_n et K.

4.1.1 Choix du noyau:

Les estimations réalisées avec les noyaux habituels (Unité, Epanechnikov, Normal ...) ont des efficacités similaires [DEHEUVELS]. Pour des commodités de programmation ainsi que certaines propriétés le noyau normal est utilisé. Il est de la forme suivante:

$$K(y) = \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}}$$

4.1.2 Choix de la fenêtre:

Si le choix du noyau ne revêt pas une très grande importance, le choix de δ_n détermine fortement la qualité de l'estimation.
La recherche de la fenêtre optimale à donné lieu a deux approches:
-La fenêtre est déterminée à partir de résultats théoriques.
-Le choix de δ_n est fait par itérations successives.

Première approche [DEHEUVELS]:

On considère le comportement asymptotiques des estimations qui définissent le critère du *M.I.S.E* (minimun integrated squared error). On cherche δ_n qui minimise le critère du *M.I.S.E*:

$$On\ minimise\ la\ quantité \quad M^2 = \int_{-\infty}^{+\infty} E((f_n(x) - f(x))^2)dx$$

$$M^2 = B_1^2 + B_2^2 \quad ou \quad B_1^2 = \int_{-\infty}^{+\infty} (E(f_n(x)) - f(x))^2 dx \quad B_2^2 = \int_{-\infty}^{+\infty} V(f_n(x))dx$$

Le développement de Taylor à l'ordre 1 de B_1^2 et B_2^2 (DEHEUVELS), s'écrit:

$$B_1^2 \simeq \frac{1}{4}\delta_n^4 \left[[y^2K]\right]^2 \left[\int_I (f'(x))^2 dx\right] \qquad B_2^2 \simeq \frac{1}{n\delta_n}\left[[K^2]\right]$$

$$Avec \quad \left[[K^2]\right] = \int_I K^2 \quad et \quad \left[[y^2K]\right]^2 = (\int_I t^2 K)^2$$

Note: Ces approximations sont possibles car K est un noyau normal.

On a une équation de la forme $M^2 = A\delta_n^4 + \frac{B}{n\delta_n}$ à minimiser. Le minimum est obtenu en dérivant l'équation précédente et en annulant sa dérivée.
On trouve $\delta_n^5 = \frac{B}{4An}$, d'où la valeur de δ_n:

$$\delta_n = \left[\frac{\left[[K^2]\right]}{\left[[y^2K]\right]^2 \int_I (f''(x))^2 dx}\right]^{\frac{1}{5}} n^{-\frac{1}{5}}$$

Dans cette formule, sont connus $\left[[K^2]\right]$, $\left[[y^2K]\right]^2$, et $n^{-\frac{1}{5}}$.
Seul le terme $\int_I (f''(x))^2 dx$ reste à déterminer.

Choix de f:

Il est déterminé à la suite des résultats obtenus en estimation paramétrique. Dans la méthode proposée par Brunet-Moret, l'ajustement est fait par une loi Log-Normale. Or, il s'avère que $\int_I (f''(x))^2 dx$ ne converge pas en zéro.
Une deuxième étape consiste à travailler sur des données transformées; si X suit une loi Log-Normale, alors $U = \delta + \gamma lnX$ suit une loi Normale. L'intégrale recherchée peut alors être calculée et une valeur approchée de:

$$\delta_n = \left[\frac{\left[[K^2]\right]}{\left[[y^2K]\right]^2 \int_I (f''(x))^2 dx}\right]^{\frac{1}{5}} n^{-\frac{1}{5}}$$

peut être calculée. Si K est un noyau Normal et f suit une loi Normale alors [BOSQ]:

$$\delta_n = \frac{S_n}{(\frac{3}{8}\sqrt{\pi})} \quad ou \ S_n^2 \ est \ la \ variance \ empirique$$

Toutefois, le choix d'une telle transformation s'avère décevant, la loi Log-Normale adoptée est trop dissymétrique, son mode est proche du seuil de sensibilité du pluviomètre. En conséquence la transformation adoptée ne restitue qu'une partie de la loi Normale.

En fait, le choix de la loi Log-Normale répondait à deux conditions:

-L'ajustement à l'échantillon tronqué était correct.

-Cette loi est nécessaire pour la réalisation des calculs.

Toutefois, le comportement de cette loi pour les faibles valeurs de précipitation ne reflète pas la réalité du phénomène. C'est pourquoi, elle n'est utilisée qu'avec les échantillons tronqués.

L'ajustement peut-être fait par une loi Gamma, de fonction densité:

$$f(x) = \frac{(x-\gamma)^{\alpha-1}}{\beta^\alpha \Gamma(\alpha)} exp(-\frac{(x-\gamma)}{\beta}) \qquad \alpha, \ \beta, \quad \gamma > 0$$

Détermination des coefficients

Les paramètres α, β, γ sont déterminés par les moments, on a les relations suivantes:

$$\hat{\alpha} = 4\frac{m_2^3}{m_3^2} \quad \hat{\beta} = \frac{m_3}{2m_2} \quad \hat{\gamma} = \overline{X} - 2\frac{m_2^3}{m_3}$$

Avec:

$$\overline{X} = \sum_{i=1}^{n} X_i \qquad m_2 = \frac{1}{n}\sum i = 1n(X_i - \overline{X})^2 \quad m_3 = \frac{1}{n}\sum i = 1n(X_i - \overline{X})^3$$

Les calculs faits, on obtient:

$$\hat{\alpha} = 0.189 \qquad \hat{\beta} = 28.321 \qquad \hat{\gamma} = 0.6742$$

D'où $\int_I (f''(x))^2 dx = 0.44, \quad soit \quad \delta_n = 1517$

Autre approche

La détermination de δ_n se fait par itérations successives. On retient le δ_n qui minimise $\int_I |f_n - f| dx$. Les avantages d'une telle méthode sont multiples:

-Il n'y a pas d'hypothèse sur la nature de la fonction densité.

-L'estimation se fait directement sur l'échantillon et non sur une loi de densité. Cet aspect est intéressant si la fonction f est difficilement quantifiable.

Par ailleurs, cette méthode comporte quelques inconvenients:

-Les temps de calcul sont importants (12 heures sur une station SUN).

-Le pas de l'integrale n'est pas aussi précis qu'il serait souhaitable.

4.1.3 Qualité de l'ajustement.

Première méthode

L'intégrale $\int_I |f_n - f| dx$ est calculée. La fonction f utilisée est la loi Gamma définie précédemment. L'intégration numérique est faite par la méthode des rectangles.
Note: Cette méthode d'intégration donne de meilleurs résultats que la méthode de Simpson en raison du pas d'intégration et de la quantité intégrée. En effet, sur une partie du graphe, $|f_n - f|$ est petit et le pas d'intégration pas suffisamment fin pour obtenir de bons résultats avec cette méthode.
La distance L_1 a été calculée pour plusieurs seuils So, So variant de 0 à 90 mm. Les valeurs trouvées ont été pondérées par la valeur de $\int_{So}^{\infty} f_n dx$. On trouve ainsi un seuil So optimal pour l'ajustement de f par f_n

Seconde méthode

Pour δ_n, δ_n variant de 0.01 à 5 par pas de 0.01, la distance L_1 est calculée. La fenêtre optimale est obtenue pour L_1 minimun. L'ajustement est fait sur les données. On trouve δ_n=0.47 et $\int_I |f_n - f| dx$=0.03633.

4.1.4 Résultats:

Première méthode

Si on considère l'échantillon dans son entier, l'ajustement est médiocre, car la loi Gamma s'ajuste mal sur les données pour les faibles précipitations. Lorsque l'échantillon est tronqué $\int_I |f_n - f| dx$ diminue jusqu'à un minimum de 0.076, atteint pour un seuil So de 15 mm.

Seconde méthode

On trouve δ_n=0.47 et $\int_I |f_n - f| dx$=0.03633. Ce résultat est obtenu à partir de l'échantillon du Sénégal tronqué en 0.
Les différences de résultats obtenu, s'expliquent en partie par l'emploi d'une loi Gamma. La modélisation de f par cette loi n'est pas parfaite.

4.2 Calcul de la période de retour décennale

4.2.1 Théorème de Scheffé:

$$sup_{B \in \mathcal{B}} |\mu_n(B) - \mu(B)| = sup| \int_B f_n - \int_B f| = \frac{1}{2} \int |f_n - f|$$

Dès que la distance L_1 tend vers zéro, μ_n la probabilité associée à f_n, tend vers μ. Pour B appartenant à l'ensemble des boréliens, une mesure empirique basée sur l'estimateur f_n de la densité est:

$$\mu_n(B) = \int_B f_n$$

4.2.2 Calcul de $\hat{H}_{dec}$

On cherche B tel que $\mu_n(B) = \frac{1}{1530}$, qui peut s'écrire:

$$\int_{\hat{H}_{dec}}^{+\infty} = \frac{1}{1530}$$

Résultats:

-Première méthode:
$\hat{H}_{dec}$=128 mm
-Deuxième méthode:
$\hat{H}_{dec}$=131 mm
Les résultats obtenus sont du même ordre de grandeur que ceux trouvés prédemment.

4.2.3 Discussion

L'estimateur à noyau ne donne pas d'écart-type mais une convergence vers la vraie valeur. Cette notion de convergence est diifficilement utilisable en pratique où les intervalles de confiance sont très largement utilisés.

Cette méthode fait la synthèse des avantages et défauts des précédentes:

-Les temps de calcul ne sont pas très importants (12 heures pour déterminer la fenêtre optimale par itérations successives).

-Les hypothèses faites sur la nature du phénomène ne sont pas contraignantes.

Conclusion

Les méthodes proposées procurent des résultats voisins, d'où la difficulté de choisir celle qui donne le meilleur résultat.

L'étude qui a été faite donne peu de renseignements quand à la robustesse de ces méthodes. Les critères de sélection peuvent être la facilité de mise en oeuvre ou alors l'absence d'hypothèses fortes sur la nature des fonctions de répartition. Ces méthodes offrent diverses configurations de travail, à l'utilisateur de déterminer celle qui lui convient.

Bibliographie

[1] A. BERLINET et L. DEVROYE Estimation d'une densité: Un point sur la méthode du noyau. *Statistique et Analyse des Données*, vol. 14 n°1, pp 1-32, 1989.

[2] J. MIQUEL et J. BERNIER Sûreté des centrales et état de la mer *Bulletin de la direction des études et recherches série à nucléaire, hydraulique, thermique*, n°2, pp73-78, 1981.

[3] D. BOSQ et J.P. LECOUTRE *Théorie de l'estimation fonctionnelle*, Economica, Paris, 1987.

[4] V. TE CHOW et *al* Applied hydrology, 1988.

[5] B. EFRON and G. GONG A leisurely look at the bootstrap, the jacknife and cross-validation. *The American Statistician*, vol. 37 n°1, pp 36-48, 1983.

[6] P. DEHEUVELS Estimation non paramétrique de la densité par histogrammes généralisés. *Revue de Statistique Appliquée*, vol. 25 pp 5-42, 1977.

[7] P. DIACONIS B. EFRON Méthodes de calculs statistiques intensifs sur ordinateurs. *La Recherche*

[8] JOHNSON and KOTZ Pareto distribution *Continuous Univariate Distribution-1*, pp 233-249, 1972.

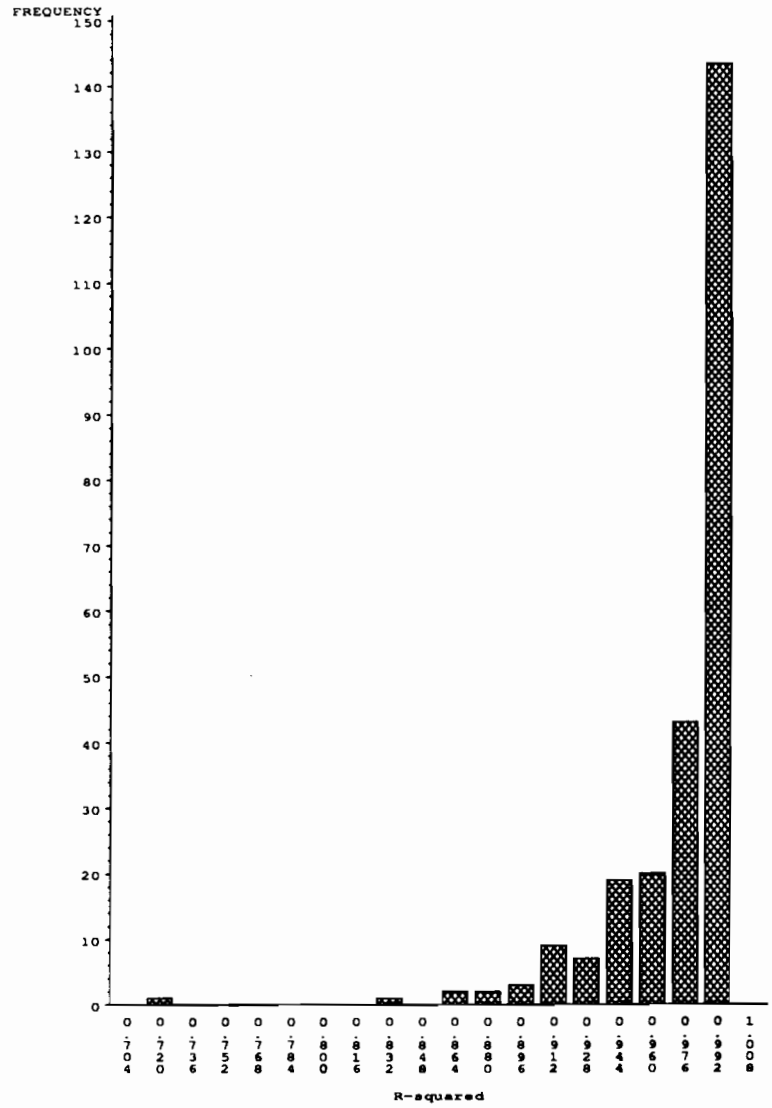
[9] H. LUBES Fitting of statistical distributions on hydrological data samples with zero values. Statistics in public ressources and utilities and in care of the environement. 7-10 April 1992 Lisbon.

[10] J. MIQUEL CRUE: Un modèle d'estimation des probabilités de crues. *La houille blanche*, n°2, 1983.

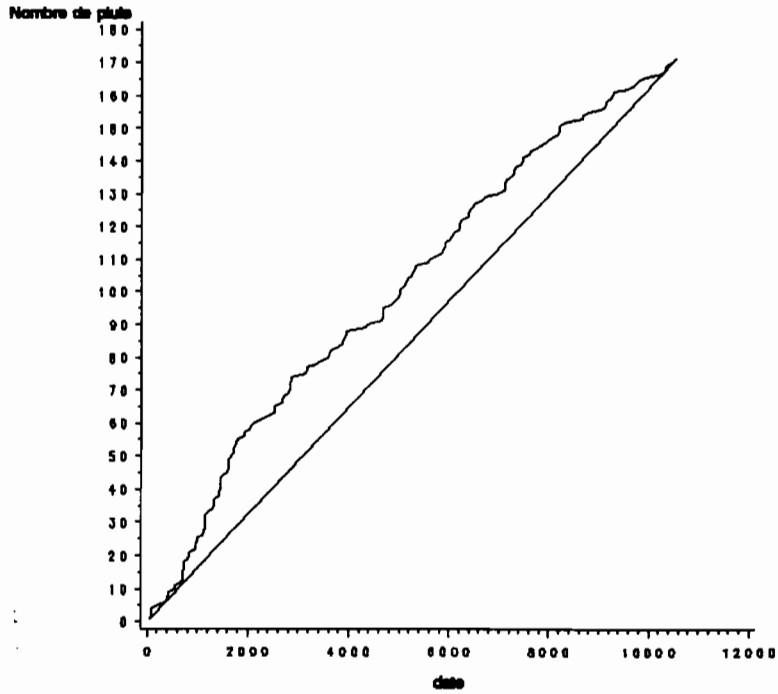
[11] J. MIQUEL *Guide pratique d'estimation des probabilités de crues.* Editions Eyrolles, 1983.

ANNEXES

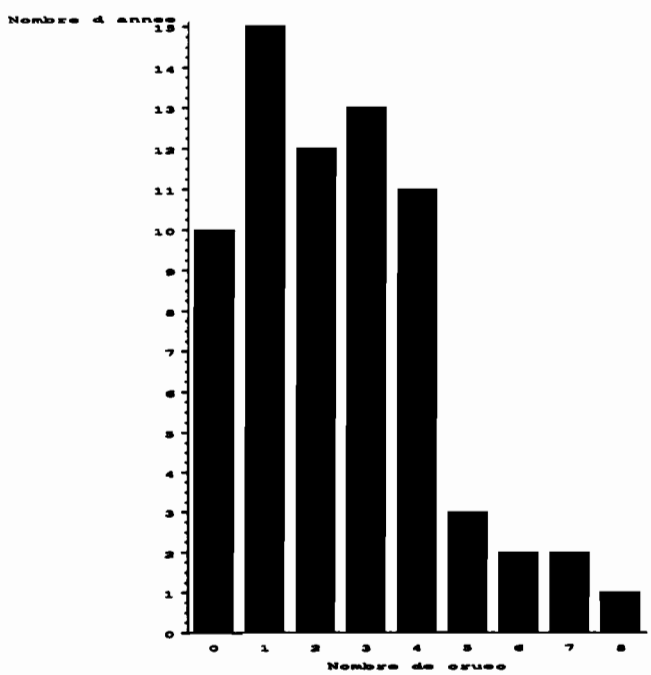
Repartition des frequences



Test de repartition uniforme des dates

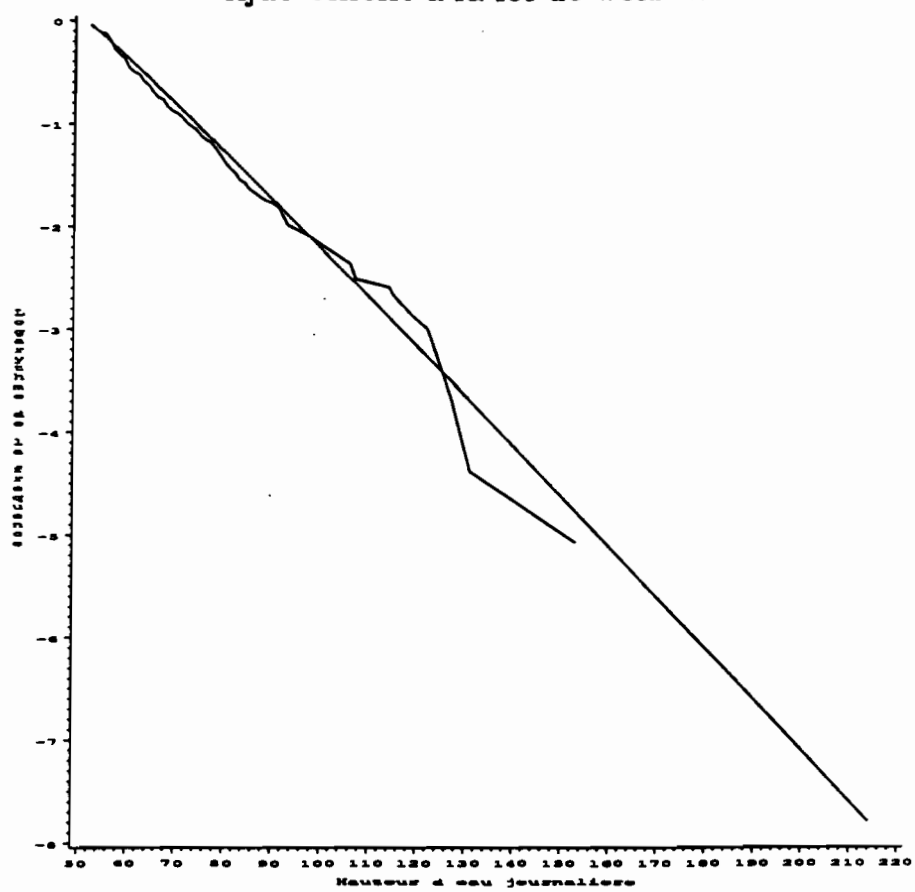


Repartition de Poisson



seuil = 81 mm

Ajustement a la loi de Weibull



Modelisation non parametrique