Un problème parmi d'autres dans l'analyse des distributions des variables hydrologiques: Les horsains (outliers).

J.M. Masson[a]

Introduction

L'étude des risques associés à un événement hydrologique (crue, sécheresse) passe toujours par l'analyse d'observations faites dans le passé. Le projeteur en bureau d'étude, qui utilise des formules empiriques régionales ou des abaques, n'échappe pas à la régle: Pour établir les formules, il a fallu procéder à l'analyse d'observations faites dans le passé. De ce point de vue, les méthodes des hydrologues se rapprochent de celles des historiens et il arrive à ces derniers de compléter les estimations faites par les hydrologues (DESBORDES et *al.*1989).
L'analyse statistique des observations du passé a pour objectif d'obtenir des informations sur la population d'où elles sont tirées. Ces informations sont ensuite utilisées pour énoncer des probabilités concernant l'avenir. Cette démarche, *qui suppose la stabilité de la population,* s'effectue en trois étapes :

- Sélection d'observations en rapport avec le phénomène étudié. L'échantillon des observations ainsi constitué doit avoir certaines qualités pour qu'on puisse en extraire des informations concernant la population.
- Ajustement d'une loi de probabilité théorique à la distribution de fréquence de cet échantillon. Il existe un éventail très large de lois théoriques de probabilité et de méthodes d'ajustement.
- Utilisation des résultats de l'étape précédente pour énoncer des probabilités concernant l'avenir, en y associant, si possible, un intervalle de confiance.

Comme aucune théorie n'est encore capable de déterminer la loi de probabilité suivie par la plupart des variables hydrologiques et que les tests d'adéquation des ajustements donnent seulement des indications et non des certitudes, *les lois de probabilité ne sont rien d'autre que des modèles à qui on demande d'être à la fois descriptifs et prédictifs.* On peut constater très rapidement que ces objectifs sont contradictoires.
Il est relativement facile de bien ajuster le modèle à l'échantillon des observations: Il suffit de multiplier le nombre de paramètres du modèle. Cependant, on sait que

[a]Laboratoire d'hydrologie et modélisation U.S.T.L. 34095 Montpellier cedex 05

le seul hasard peut produire des échantillons très particuliers où certaines valeurs fortes ou faibles sont sous-représentées ou sur-représentées, et ceci d'autant plus facilement que l'échantillon a un faible effectif. Une fonction de répartition munie de nombreux paramètres va parfaitement décrire les sinuosités de la distribution, mais risque de conduire à des extrapolations éloignées de la réalité, à cause justement des courbures qui permettent à la fonction de répartition de passer par les valeurs extrêmes observées.
Un bon modèle prédictif est un modèle robuste, peu sensible aux fluctuations d'échantillonnage. Il donne des résultats voisins avec des échantillons qui, bien que provenant de la même population, présentent des particularités différentes. Les lois de probabilité comportant peu de paramètres sont les plus robustes.

1 Que faire en présence de horsains (outliers) dans l'échantillon des observations ?

Prenons un cas précis:- la hauteur maximale de pluie en 6 heures observée chaque année à NIMES-COURBESSAC. Alors que de 1946 à 1987, les 42 valeurs observées sont comprises entre 22 et 132 millimètres, il apparaît, en 1988, la valeur 228 millimètres. Si on ajuste aux observations une des lois de probabilité utilisées pour ce genre de variables, on s'aperçoit (figure 1) que la loi de JENKINSON (G.E.V.distribution) convient mieux que la loi de GUMBEL (E.V.1 distribution). La valeur observée en 1988, qui est 1.7 fois plus élevée que la plus forte des valeurs observées en 42 ans, est à coup sûr due à un événement exceptionnel. Celui qui s'est produit à NIMES le 3 octobre 1988 fut catastrophique.
Dans la population des hauteurs maximales en 6 heures observées chaque année, la valeur 228 millimètres constitue ce que les statisticiens de langue anglaise appellent un "outlier" et que je traduirai par un mot français utilisé surtout en Normandie pour désigner celui qui n'est pas né au pays: un horsain (ALEXANDRE 1988). Cette traduction me semble justifiée par le fait qu'on cherche à identifier une population parent (Houghton 1978).
Cependant, avant de regarder particulièrement ces événements exceptionnels, il n'est peut-être pas inutile de rappeler quelques propriétés statistiques de la plus forte des 43 valeurs d'un échantillon aléatoire tiré d'une loi de probabilité couramment utilisée.

2 Comportement statistique de la plus forte valeur d'un échantillon aléatoire de 43 individus.

Pour une loi normale, les tables de PEARSON et HARTLEY (1969) donnent, en valeur centrée réduite, l'espérance mathématique de cette plus forte valeur:

$$E[u_{43}] = 2.19,$$

ce qui correspond à une probabilité de non dépassement de:

$$F(u_{43}) = 0.9837$$

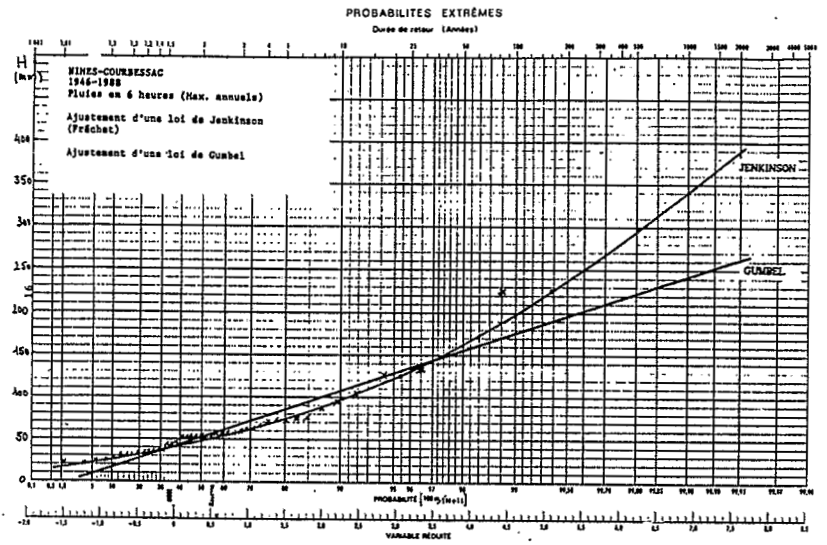

Figure 1:

En hydrologie on s'intéresse souvent aux événements (hauteur de pluie en 6 heures par exemple) tels que $X > x$. Au lieu de parler de probabilité *annuelle* de dépassement $\Pr(X > x)$, on parle d'une période de retour T en années de l'événement.

$$T = 1/\Pr(X > x)$$

T est en moyenne la durée d'observation qui voit la réalisation d'un événement tel que $X > x$, sans que cette notion implique une quelconque périodicité.
L'espérance mathématique de la plus forte valeur d'un échantillon d'effectif 43, tiré d'une loi normale, a donc une période de retour de:

$$T(u_{43}) = 1/(1 - 0.9837) \approx 70 \text{ ans}$$

Si on a affaire à une loi de GUMBEL, la fonction de répartition de la plus forte valeur d'un échantillon de 43 individus s'écrit, en variable réduite:

$$F(y_{43}) = F(y)^{43} = e^{-e^{-(y - \ln(43))}}$$

Après avoir dérivé cette expression, on peut calculer l'espérance mathématique de y_{43}. On obtient:

$$E(y_{43}) = 0.5772 + \ln(43) = 4.3384$$

Cette valeur de la variable réduite de GUMBEL correspond à une probabilité de non dépassement de 0.987 et à une période de retour au dépassement de 77 ans.

Dans le cas de la loi de GUMBEL, on constate que la plus forte valeur d'un échantillon suit aussi une loi de GUMBEL, c'est à dire que 43% des échantillons auront leur plus forte valeur supérieure à son espérance mathématique !
Pour estimer la fréquence expérimentale au non dépassement des différentes valeurs de l'échantillon, on les classe par valeurs croissantes:

$$x_1 \leq x_2 \leq \ldots \leq x_r \leq \ldots \leq x_n$$

Les expressions recommandées pour calculer la fréquence de non dépassement de la valeur x_r de rang r, prennent généralement en compte son espérance mathématique, laquelle dépend du type de loi. C'est la raison pour laquelle il ne faut pas utiliser l'expression dite de WEIBULL:

$$F(x_r) = r/(n+1),$$

qui correspond à une loi uniforme, très éloignée des distributions habituellement rencontrées en hydrologie. Une bonne expression (BRUNET-MORET 1973) est celle dite de HAZEN:

$$F(x_{43}) = (r - 0.5)/n,$$

qui est mieux adaptée aux lois rencontrées en hydrologie. Avec cette expression, la plus forte valeur d'un échantillon d'effectif 43 a une probabilité de non dépassement de:

$$F(x_r) = 42.5/43 = 0.98837,$$

soit une période de retour au dépassement de 86 ans.

De ce paragraphe nous pouvons retenir deux choses:

- Il convient de s'inscrire en faux contre une tradition qui veut que la plus forte valeur d'un échantillon n'ait pas une période de retour supérieure à la durée des observations ;

- La plus forte valeur d'un échantillon a une très grande variabilité. On montre, par un calcul très simple, qu'avec 43 années d'observation, on a une probabilité de: $1 - 0.99^{43} = 0.35$ d'avoir au moins un événement centennal (un ou plus). Si on pouvait obtenir de nombreux échantillons indépendants d'effectif 43, un sur trois possèderait au moins un événement centennal. D'autre part, il existe au moins une bonne dizaine d'expressions pour calculer les fréquences expérimentales de non dépassement. Si leurs résultats sont très voisins en ce qui concerne les valeurs centrales, il n'en est pas de même pour les valeurs extrêmes; pour le constater il suffit de comparer les résultats des expressions de WEIBULL et HAZEN. Comme la qualité d'un ajustement se juge par comparaison des probabilités théoriques avec les fréquences expérimentales, il ne paraît donc pas judicieux de donner un poids important aux valeurs extrêmes comme le propose BRUNET-MORET (1978). De ce point de vue, le test χ^2 d'ajustement, qui ne prend pas en compte les valeurs des variables, mais leur nombre dans des classes de valeurs, est plus satisfaisant, malgré sa faible puissance.

3 Traitement d'un échantillon comportant des horsains.

Dans ce paragraphe nous examinerons quelques attitudes possibles face à la présence de horsains. Il s'agit tout d'abord de détecter leur présence. BOBEE et ASHKAR (1991) proposent un test qui permet cette détection, à condition que la distribution soit proche d'une loi Log-normale. Ce test, dit de GRUBBS et BECK, calcule des limites inférieures et supérieures au-delà desquelles on a affaire à des horsains. Ces limites sont:

$$X_s = e^{\bar{X}+K_n\sigma}$$

$$X_i = e^{\bar{X}-K_n\sigma}$$

où $\bar{X}$ et σ sont la moyenne et l'écart type des variables après transformation en logarithmes Népériens. K_n est lu dans des tables ou calculé par une approximation polynomiale en fonction de n, l'effectif de l'échantillon. On trouve $K_{43} = 2.71$. En appliquant ce test aux pluies maximales annuelles de 6 heures à NIMES, qui par ailleurs suivent bien une loi Log-normale (ajustement accepté par différents tests au seuil 90%) , dont les paramètres sont:

$$\bar{X} = 3.9542 \text{ et } \sigma = 0.5$$

on trouve X_s=202.2 millimètres. D'après ce test, la hauteur de pluie en 6 heures observée à NIMES le 3 octobre 1988 est bien un horsain !
ROSSI, FIORENTINO et VERSACE (1984) détectent les horsains à partir des valeurs prises sur l'échantillon par l'estimateur du coefficient de dissymétrie g_1 ou par la valeur de y_n, variable réduite de GUMBEL correspondant à la plus forte valeur de l'échantillon, les paramètres de la loi étant calculés par la méthode du maximum de vraisemblance. Ils montrent que les valeurs de g_1, calculées sur 39 séries de débits maximaux instantanés annuels de rivières italiennes, sont souvent bien supérieures à celles qu'on pourrait attendre d'un tirage aléatoire dans une loi de GUMBEL. Par simulation ils ont déterminé des limites d'acceptation sur g_1 et y_n. Ils en déduisent si la ou les plus fortes valeurs d'un échantillon sont ou non des horsains.
Cependant la partie la plus intéressante du travail de ces auteurs concerne le traitement appliqué en présence de horsains.
On peut représenter la succession des hauteurs de pluie maximales en 6 heures Z des épisodes pluvieux (Z dépassant un seuil H_s judicieusement choisi), par un processus de Poisson.

- Si λ est le paramètre de ce processus (nombre moyen annuel de dépassements du seuil),
- Si $F_z(z)$ est la fonction de répartition des hauteurs maximales en 6 heures des épisodes pluvieux, hauteurs supérieures à H_s, on montre que la fonction de répartition du maximum annuel en 6 heures X a pour expression:

$$F_x(x) = e^{-\lambda(1-F_z(x))}$$

- Si la loi de Z est exponentielle:

$$F_z(z) = 1 - e^{-z/\theta}$$

alors la loi du maximum annuel en 6 heures X est une loi de GUMBEL :

$$F_x(x) = e^{-\lambda e^{-x/\theta}} = e^{-e^{-(x - x_0)/\theta}} \quad \text{avec : } x_0 = \theta . \ln(\lambda)$$

Le processus est décrit comme la loi de GUMBEL par 2 paramètres λ et θ, avec : θ = hauteur moyenne des valeurs de Z dépassant le seuil H_s.
L'originalité de la méthode proposée par ROSSI, FIORENTINO et VERSACE (1984) est de combiner deux processus Poissoniens:

- Un processus de base concernant les événements ordinaires, de paramètres λ_1 et θ_1
- Un processus spécial pour les horsains, de paramètres λ_2 et θ_2

La loi combinée des deux processus est une loi des valeurs extrêmes à deux composantes (Two Component Extreme Value = T.C.E.V. distribution), qui s'écrit:

$$F_x(x) = e^{-\lambda_1 e^{-x/\theta_1} - \lambda_2 e^{-x/\theta_2}}$$

Sur une série de maxima annuels, les paramètres de cette loi peuvent s'estimer par la méthode du maximum de vraisemblance. Appliquée à la série de NIMES, la méthode donne aux paramètres les valeurs présentées au tableau 1. L'ajustement

Tableau 1:

	Evénements ordinaires	Horsains
Nombre moyen annuel	12.8	0.39
Hauteur moyenne (mm)	15.9	58.7

obtenu est présenté sur la figure 2 La probabilité de non dépassement associée à la valeur 228 millimètres est, pour cette loi, de 0.992, soit une période de retour de 125 ans. La loi de JENKINSON que nous avons utilisée (fig. 1), nous donnait une probabilité de non dépassement de 0.99356, soit une période de retour de 155 ans. Les résultats donnés par les deux lois sont du même ordre de grandeur.
La loi à deux composantes a un intérêt supplémentaire: Elle se prête particulièrement bien à la régionalisation. Les paramètres λ_1 et θ_1 sont des paramètres locaux ajustés sur chaque station après élimination des horsains. Ces paramètres sont utilisés pour transformer toutes les variables locales en variables adimensionnelles et le mélange des variables adimensionnelles de toutes les stations permet le calcul des valeurs régionales de λ_2 et θ_2 La simulation a montré que ce modèle reproduisait correctement la réalité des crues italiennes.
En ce qui concerne les horsains pluvieux, leur caractère régional ne fait aucun doute. Ce sont des pluies extrêmement intenses qui sont dues à des situations météorologiques particulières qui se produisent généralement en automne.
Dans ces situations, dites "Cévenoles", une "goutte froide" au nord bloque l'arrivée d'air chaud et humide en provenance de la Méditerrannée. Sur la ligne de front se forment des cumulonimbus très actifs où l'air chaud et humide s'élève avec

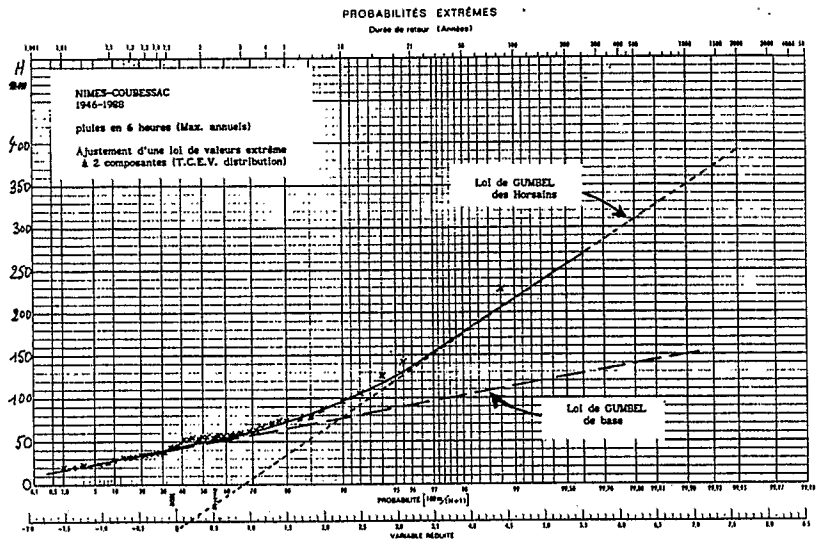

Figure 2:

des vitesses considérables (10 à 20 m/s). Habituellement ces orages intenses touchent les Cévennes, mais quelquefois ils se déplacent plus au nord (crue de la Haute Loire en septembre 1980), ou au contraire restent sur la plaine Languedocienne sans qu'on sache bien pourquoi (P.A. ROCHE 1989). Ce sont des phénomènes susceptibles de toucher une région étendue sur la bordure septentrionale de la mer Méditerrannée.

4 Informations régionales et historiques se complètent.

Compte tenu de l'aspect régional associé à l'occurrence des horsains pluvieux, nous avons essayé de compenser la faible ancienneté des observations de pluie à NIMES en regardant ce qui s'était passé autour de NIMES. On peut estimer à 13 les horsains pluvieux qui se sont produits dans la plaine Languedocienne, entre Rhône et Aude, depuis la dernière guerre. Leur période de retour régionale est donc comprise entre 3 et 4 ans. A partir de ces données, un schéma rudimentaire nous a permis d'estimer une période de retour pour l'événement de NIMES :

- La plaine Languedocienne est un rectangle de 200 X 40 kilomètres.

- Les orages horsains ont une forme circulaire et touchent entre 300 et 500 kilomètres carrés; leur rayon R a une distribution uniforme dans l'intervalle :

$$\sqrt{(300/\pi)} \leq R \leq \sqrt{(500/\pi)}$$

- Les coordonnées du centre de l'orage (x,y) sont distribuées de manière uniforme dans le rectangle:

$$0 \leq x \leq 200Km$$

$$0 \leq y \leq 40Km$$

La simulation de 1000 orages horsains permet d'estimer la probabilité qu'un bassin versant de 49 Km^2 (NIMES et ses cadereaux) soit touché à 80% par un événement. Deux simulations ont donné respectivement 31 et 25 atteintes, soit une période de retour comprise dans l'intervalle:

$$(1000 * 3)/31 = 97 \text{ ans} < T < (1000 * 4)/25 = 160 \text{ ans}$$

Des recherches historiques concernant les écrits relatant des pluies importantes ayant provoqué des inondations dans la ville de NIMES, ont permis de remonter jusqu'en 1334 (PEY 1988). Depuis le 15° siècle, on recense 5 catastrophes majeures (dont celle du 3 octobre 1988) et 5 évènements secondaires qui auraient pu occasionner des dégâts importants dans le NIMES d'aujourd'hui compte tenu du développement récent de son urbanisation.
La période de retour des événements majeurs est donc de l'ordre de 120 ans, mais elle n'est plus que de 60 ans si on prend en compte aussi les événements secondaires.

5 Chaos ou Déterminisme?

Dans un article récent, LADOY, LOVEJOY et SCHERTZER (1991), présentent *les cascades multifractales*, outil mathématique permettant l'analyse des systèmes chaotiques, comme un moyen puissant d'étude de la variabilité des données climatologiques. Cependant leur étude des fréquences de dépassement des différences des hauteurs des pluies journalières successives entre 1949 et 1988 à NIMES reste du domaine de la statistique classique. Portée sur un graphique log-log, la distribution des fréquences de dépassement montre un comportement asymptotique hyperbolique d'exposant 0.3 environ. De là à en déduire que le plus fort événement observé a une période de retour de l'ordre de 100 ans, il y a un pas que nous n'avons pas su franchir. Cependant, comme le disent ces auteurs, on peut penser que ce sont des méthodes non classiques (fractales ou autres), qui permettront de mieux estimer les probabilités d'occurrence des horsains. La statistique classique trouve en effet assez vite ses limites. On connait bien le cas de la sécheresse au Sahel. On pourrait voir se produire un phénomène analogue sur le Sud-Ouest de la France, qui risque de vivre cette année sa troisième année consécutive de sécheresse. La sécheresse atmosphérique de 1989 fut décennale à TOULOUSE-BLAGNAC (LAMBERT *et al.* 1990); comme, en 1990, elle fut au moins aussi sévère, si celle qui se profile pour 1991 s'avérait également décennale, la probabilité de cette succession serait de $(0.1)^3 = 0.001$, assez difficile à admettre dans un contexte Gaussien. N'y aurait-il pas du déterminisme dans l'entêtement des anticyclones à vouloir se décaler de leur position habituelle pendant plusieurs années consécutives ?

Références bibliographiques

ALEXANDRE (B.), 1988.- Le Horsain. Vivre et survivre en Pays de Caux.- Terre Humaine, 554p.

BOBEE (B.), ASHKAR (F.), 1991.- The Gamma family and derived distribution applied in hydrology.- Water Resources Publications, 203p.

BRUNET-MORET (Y.), 1973.- Statistique de rang.- Cahier ORSTOM, série hydrologie, vol. 10, N° 2, pp 133-151.

BRUNET-MORET (Y.), 1978.- Recherche d'un test d'ajustement. - Cahier ORSTOM, série hydrologie, vol. 12, N° 3, pp 261-280.

DESBORDES (M.), DUREPAIRE (P.), GILLY (J.C.), MASSON (J.M.), MAURIN (Y.), 1989.- 3octobre 1988, inondations sur Nîmes et sa région. Manifestations, causes et conséquences.- Lacour Editeur, Nîmes, 93p.

HOUGHTON (J.C.), 1978.- Birth of a parent: The Wakeby distribution for modeling flood flows.- Water Resources Research, vol 14, N° 6, pp 1105-1109.

LADOY (P.), LOVEJOY (S.), SCHERTZER (D.), 1991.- Extreme variability of climatological data: Scaling and intermittency.- Non linear variability in Geophysics, Kluwer Academic Publishers, pp 241-250.

LAMBERT (R.), LAMI (J.M.), SENGES (F.), 1990.- La sécheresse de 1989 dans le bassin de la Garonne à l'amont de Mas d'Agenais.- Ministère de l'environnement, Délégation de bassin Adour-Garonne et Université de Toulouse-Mirail, Institut de Géographie, 75 p. et annexes 79 p.

PEARSON (E.S.), HARTLEY (H.O.), 1969.- Biometrika tables for statisticians. Cambridge University Press, 3° édition, 270 p.

PEY (J.), 1988.- Nîmes et ses cadereaux, principales dates des crues et inondations 1334-1988.- Musée Archéologique de Nîmes, 28 p.

ROCHE (P.A.),1989.- Les inondations: L'exemple de Nîmes.- Supplément de La Recherche, N° 212, pp 17-21.

ROSSI (F.), FIORENTINO (M.), VERSACE (P.), 1984.- Two component extreme value distribution for flood frequency analysis.- Water Resources Research, vol. 20, N° 7, pp 847-856.