GUIDE POUR LES MÉTHODES DU SYSTÈME D'AIDE À LA DÉCISION (SAD) DE HYFRAN-PLUS

El Adlouni, Salaheddine, Bernard Bobée et Ouejdene Samoud

bernard.bobee@ete.inrs.ca; el_adlouni@yahoo.com; ouejdene.samoud.1@ulaval.ca

1. Introduction au SAD

HYFRAN-PLUS permet d'ajuster un nombre important de distributions statistiques à une série de données qui vérifient les hypothèses d'indépendance, d'homogénéité et de stationnarité (cf. Aide de HYFRAN-PLUS). Un **S**ystème d'**A**ide à la **D**écision (SAD) a été développé pour permettre de choisir la classe de distributions la plus adéquate pour estimer le quantile Q_T de période de retour T élevée tel que $\Pr[Q \geq Q_T] = \frac{1}{T}$. En effet, une classification des lois par rapport à la queue droite de la distribution, permet de distinguer trois principales catégories dans lesquelles on peut classer les dix distributions les plus utilisées en hydrologie pour représenter les débits maximums annuels :

- la classe C (distribution à variations régulières) : Fréchet (EV2), Halphen B Inverse (HIB), Log-Pearson type 3 (LP3), Gamma Inverse (GI).
- la classe D (distributions sub-exponentielles) : Halphen type A (HA), Halphen type B (HB), Gumbel (EV1), Pearson type 3 (P3), Gamma (G).
- la classe E (loi exponentielle).

La queue droite d'une distribution de la classe C est plus lourde que celle d'une loi de la classe D qui est elle-même plus lourde que celle d'une loi de la classe E (Figure 1). On peut en déduire une relation équivalente pour les quantiles estimés à partir de ces lois. En effet, pour un échantillon donné, les quantiles de période de retour T estimés à partir de trois lois des classes C, D et E, respectivement, données par $Q_T(C)$, $Q_T(D)$ et $Q_T(E)$, vérifient la relation théorique suivante : $Q_T(E) < Q_T(D) < Q_T(C)$.

La loi Log-normale (LN) n'appartient strictement à aucune des classes C et D. Elle a un comportement asymptotique qui se situe à la frontière des classes C et D. En effet, la queue droite de la loi LN est plus légère (respectivement, plus lourde) que celle d'une loi de la classe C (respectivement de la classe D). En

effet les quantiles Q_T estimés à partir des distributions appartenant aux classes C, D et la loi LN, vérifient la relation d'ordre suivante : $Q_T(D) < Q_T(LN) < Q_T(C)$ (Figure 1).

Il est donc, important de déterminer si la loi la plus adéquate pour représenter un échantillon est la loi Log-normale ou appartient à la classe C ou D. La version 2.1 du SAD permet de tester initialement l'hypothèse de Log-normalité de la distribution des observations.

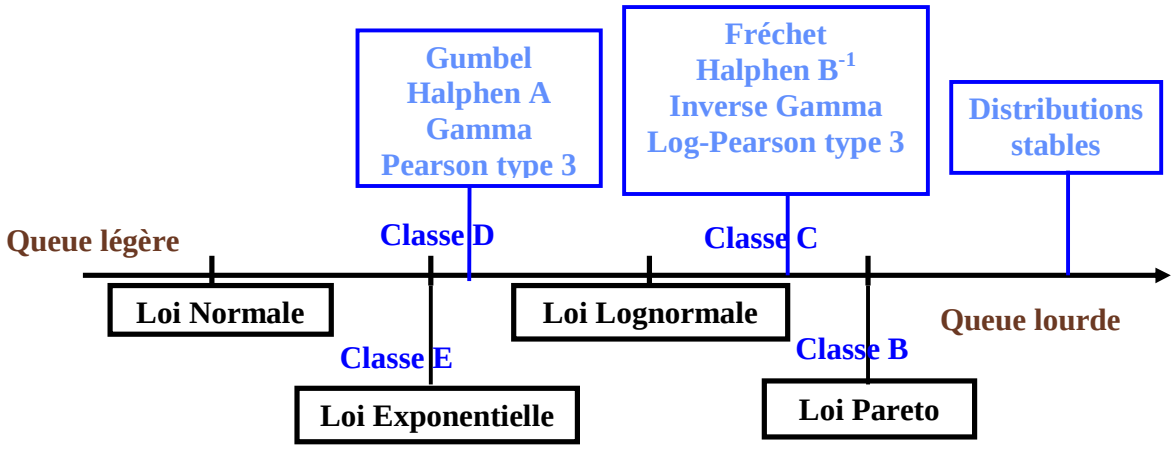

Figure 1 : Distributions ordonnées par rapport à leurs queues droites (El Adlouni et al., 2008)

Les méthodes développées dans le SAD permettent d'identifier la classe la plus adéquate pour l'ajustement d'un échantillon donné. Ces méthodes sont (Figure 2):

- le test de Jarque-Bera : considéré pour tester la Log-normalité avec une sélection a priori basée sur le diagramme (Cv,Cs) (cf. Section 2);
- le graphique Log-Log : utilisé pour discriminer d'une part la classe C et d'autre part les classes D et E (cf. Section 3),
- la fonction moyenne des excès (FME) : utilisée pour discriminer les classes D et E (cf. Section 4).
- deux statistiques : le rapport de Hill (cf. Section 5) et la statistique de Jackson (cf. Section 6) qui peuvent être utilisées pour effectuer une analyse confirmatoire des conclusions suggérées à partir des deux précédentes méthodes (graphique Log-Log et FME).

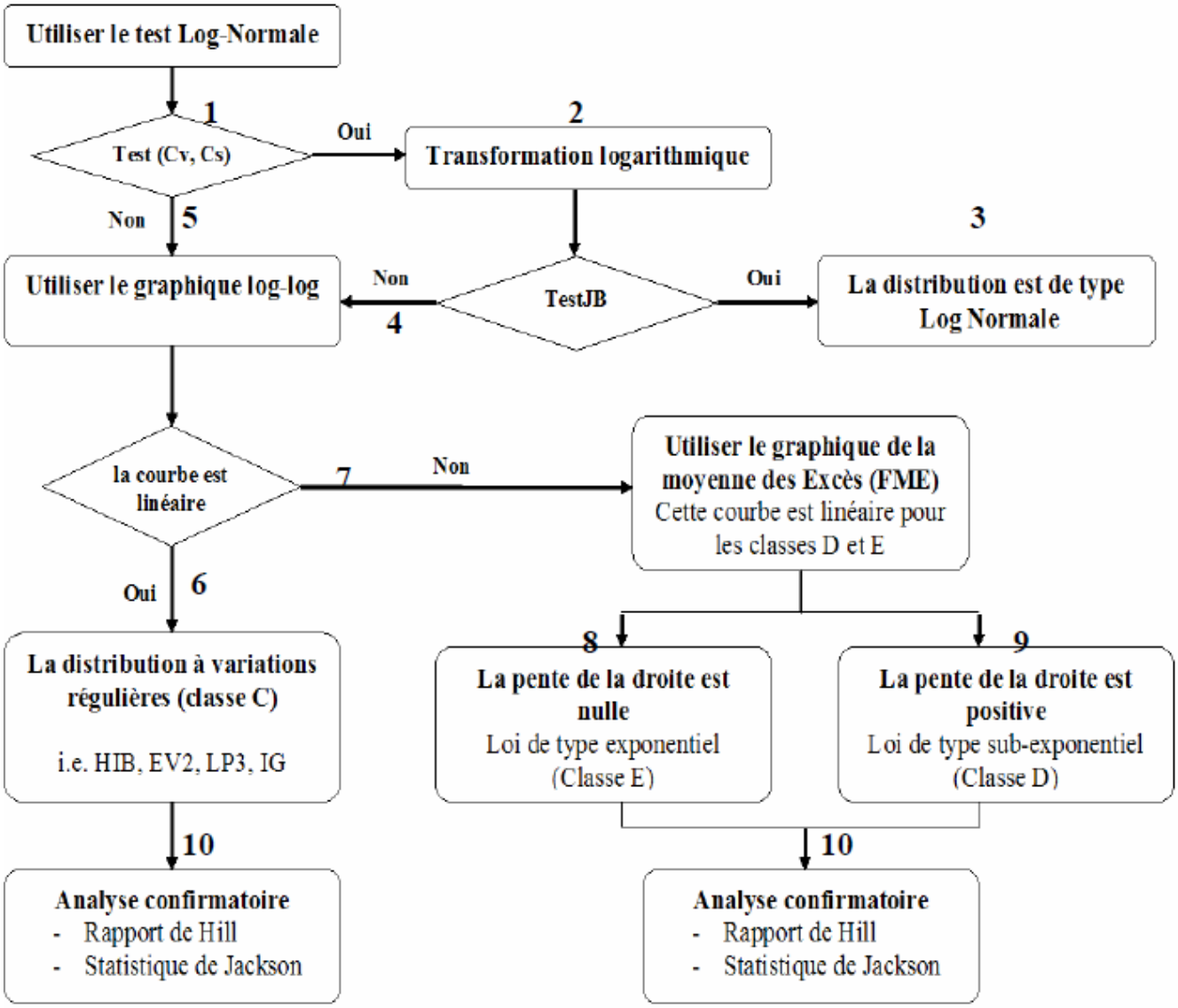

Figure 2 : Diagramme des critères de choix entre les classes C, D et E

Plus de détails théoriques concernant cette classification et les critères de choix entre les différentes classes sont disponibles dans El Adlouni, Bobée et Ouarda (2008) et Martel, El Adlouni et Bobée (2011). Cet article est disponible comme fichier attaché lors de l'installation de HYFRAN-PLUS.

2. Test de Log-normalité

2.1. Diagramme (Cv,Cs)

Pour tester la Log-normalité on considère des tests de normalité appliqués sur la série des données initiales transformées par la fonction logarithme (Yi=Log(Xi), où Xi est la variable initiale). De récents travaux ont été effectués par Martel, El Adlouni et Bobée (2011), pour définir une procédure basée sur un ou plusieurs tests permettant de discriminer la loi Log-normale des lois des deux autres classes, C et D. Cinq tests (test d'Anderson-Darling (AD), Shapiro-Wilk (SW), Lilliefors (Lf), Jarque-Bera (JB) et Filliben (FB)) ont été comparés par simulation pour examiner leurs puissances. Notons que les cinq tests permettent de tester la normalité et ont été utilisés sur les séries transformées par la fonction logarithme, pour tester la Log-normalité. Martel, El Adlouni et Bobée (2011) [Article disponible comme fichier attaché lors de l'installation de HYFRAN-PLUS] ont montré que lorsque la série étudiée a certaines caractéristiques (représentées par les coefficients de variation Cv et d'asymétrie Cs), le test de Jarque-Bera (JB) a une puissance satisfaisante pour tester la Log-normalité ou l'appartenance à une des classes C ou D. En effet, si les coefficients d'asymétrie et de variation de la série étudiée sont représentés par un point de la zone correspondant à la loi de Halphen type Inverse B (HIB) (Figure 3), la puissance du test JB est très satisfaisante pour tester l'hypothèse de Log-normalité et nous en recommandons l'utilisation.

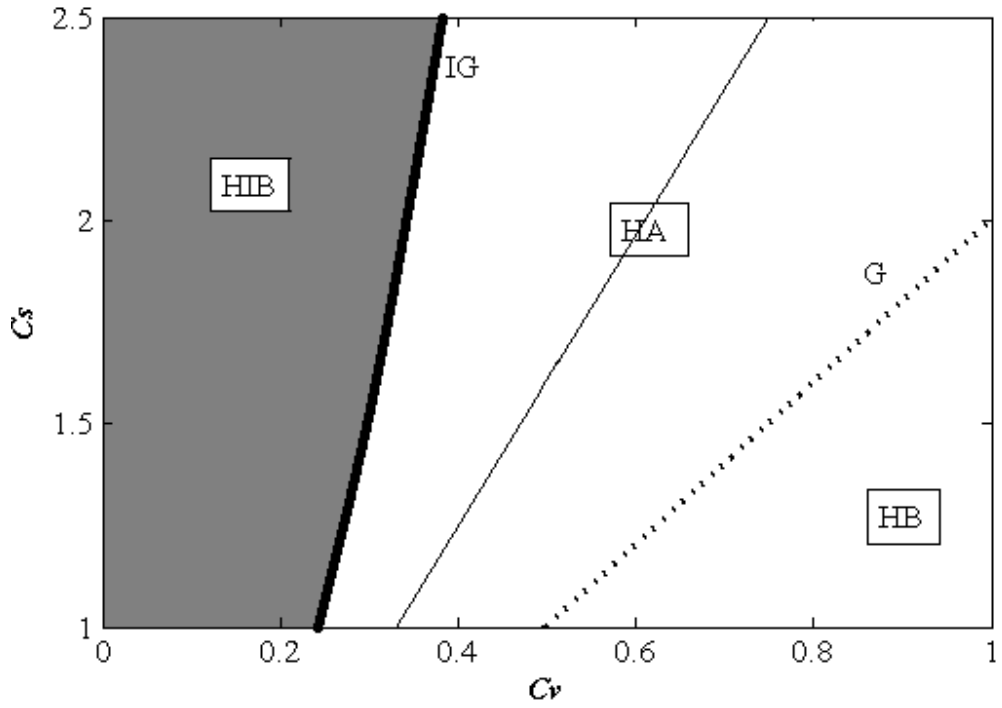

Figure 3 : La zone HIB dans le diagramme (Cv,Cs)

On propose donc comme première étape dans le SAD l'utilisation du test JB basé sur le diagramme *(Cv,Cs)*. Les parties de cette étape sont résumées par l'algorithme suivant :

1. Calcul des Cv et Cs de l'échantillon [Étape 1];
2. Si l'échantillon appartient à la zone Halphen type Inverse B (HIB) [Étape 2] (Figure 3) : Appliquer le test JB sur l'échantillon transformé (la série des logarithmes) pour vérifier si l'ajustement par une loi LN est acceptable ou non.
3. Si l'hypothèse de Log-normalité est acceptée au niveau de signification de 5%(1), on recommande la distribution Log-normale pour représenter les données [Étape 3].
4. Si l'hypothèse de Log-normalité n'est pas acceptable, est acceptée au niveau de signification de 5%, on passe à l'étape suivante (Diagramme Log-Log, Section 3) [Étape 4].
5. Si l'échantillon n'appartient pas à la zone HIB : on ne considère pas la loi LN pour l'ajustement et on passe à l'étape suivante (Diagramme Log-Log, Section 3) [Étape 5].

2.2. Test Jarque-Bera [JB; Jarque et Bera, 1980]

Le test JB (Jarque and Bera, 1980) utilise des fonctions des 3e et 4e moments de l'échantillon, soient les coefficients d'asymétrie (Cs) et d'aplatissement (Ck). Dans le cas d'une loi normale N(0,1), ces coefficients sont respectivement égaux à 0 et 3. Le test JB combine ces deux coefficients en une statistique unique :

$$\mathrm{JB} = N\left\{\frac{1}{6}(Cs)^2 + \frac{1}{24}(Ck-3)^2\right\}$$

où N est la taille de l'échantillon, avec la condition $N > 7$. La statistique JB suit alors une distribution Chi-deux, χ_2^2, avec $\nu = 2$ degrés de liberté. On compare alors la valeur JB calculée à partir de

(1) Le niveau utilisé par défaut dans le SAD est basé sur les travaux de Martel, El Adlouni et Bobée (2011).

l'échantillon avec la valeur critique au niveau de signification de 5% ; $\chi^2_{2,0.95}=5.99$. La règle de décision est la suivante :

- Si $JB<\chi^2_{2,0.95}$ on accepte H0 (la série transformée est normale) et on accepte donc l'hypothèse de Log-normalité [Étape 3];
- Si $JB>\chi^2_{2,0.95}$ on rejette H0 (la série transformée n'est pas normale) et on rejette donc l'hypothèse de Log-normalité [Étape 4].

Il est à noter qu'il s'agit en théorie d'un test asymptotique, donc généralement peu efficace pour les petits échantillons ($N<30$). Nous avons décidé, dans le cadre du SAD, de l'utiliser pour les échantillons tels que les coefficients Cv, Cs appartiennent à la région HIB, pour laquelle on a montré que la puissance du test est acceptable même pour des échantillons de petite taille Martel, El Adlouni et Bobée (2011).

3. Diagramme Log-Log

Pour les distributions de type sub-exponentiel (Classe D) de moyenne θ, la fonction de survie (probabilité au dépassement) $\bar{F}(u)=P(X>u)$ est donnée par : $\bar{F}(u)=P(X>u)=e^{-u/\theta}$, et pour une distribution à variations régulières (Classe C, asymptotiquement de type puissance) on a : $\bar{F}(u)=P(X>u)\approx C\int_u^{\infty}\frac{1}{x^{\alpha}}dx=C\left[\frac{x^{-\alpha+1}}{1-\alpha}\right]_u^{\infty}=C_1u^{-\alpha+1}$ (pour $\alpha>1$, qui est équivalente à la condition d'existence de la moyenne).

En considérant $\log\left(P\left(X>u\right)\right)$, on obtient respectivement pour les deux types de distributions $-\frac{u}{\theta}$ et $\log(C_1)-(\alpha-1)\log(u)$. Ainsi, en portant sur un graphique les valeurs de $\log\left(P\left(X>u\right)\right)$ en fonction de $\log u$, on devrait obtenir une courbe linéaire pour une distribution de la classe C, et concave pour une distribution qui n'est pas à variations régulières (Figure 4). C'est-à-dire une distribution de type sub-exponentiel (Classe D) ou exponentiel (Classe E).

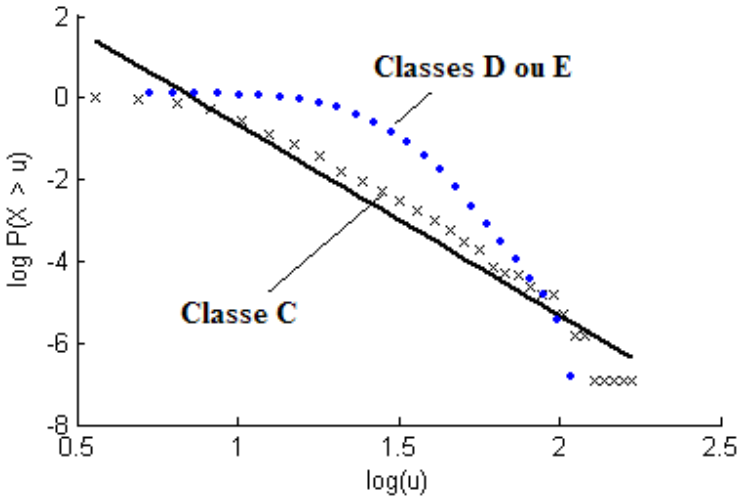

Figure 4 : Illustration du graphique Log-Log pour la caractérisation des lois de la classe C.

Ce diagramme est donc, linéaire [Étape 6] pour les échantillons distribués selon une loi de la classe C (distribution à variations régulières), i.e. Fréchet (EV2), Halphen type B Inverse (HIB), Log-Pearson type 3 (LP3), Gamma Inverse (GI)). Lorsque le diagramme n'est pas linéaire [Étape 7] on suggère l'emploi de la méthode FME (Fonction moyenne des excès) pour discriminer les classes D et E (Figure 2).

Pour vérifier l'hypothèse de linéarité dans le diagramme log-log, on calcule le coefficient de corrélation associé à la courbe représentée dans ce diagramme. Des études de simulation nous ont permis d'obtenir pour des niveaux de signification 5% et 1%, les valeurs critiques (rc(5%) et rc(1%)) pour tester l'hypothèse H0 : L'ÉCHANTILLON EST ISSU D'UNE LOI DE LA CLASSE C (i.e. LA COURBE EST LINÉAIRE). Ces valeurs critiques sont calculées, par simulation, en fonction de la taille N de l'échantillon ($30 \leq N \leq 200$). ***Notons que les décisions affichées par le SAD sont basées par défaut sur le niveau de signification 5%.***

Ainsi, si le coefficient de corrélation observé (ro) est supérieur à la valeur critique (rc) au niveau de signification 5%, alors le coefficient de corrélation n'est pas significativement différent de 1 au niveau de signification 5% et l'hypothèse H0 de linéarité est acceptée à ce niveau (Figure 5); le choix le plus adéquat correspond alors à une loi de la classe C des distributions à variations régulières : Halphen type B Inverse (HIB), Fréchet (EV2), Log-Pearson type 3 (LP3), Gamma Inverse (GI).

Lorsqu'il y a rejet de l'hypothèse H0, au niveau de signification 5%, on suggère l'emploi d'un graphique basé sur la fonction moyenne des excès (méthode FME). ***Cependant, les valeurs critiques au niveau de***

significations 1% sont données pour avoir plus de flexibilité et pour permettre à l'utilisateur de prendre éventuellement une autre décision que celle suggérée sur la base du niveau de signification 5%.

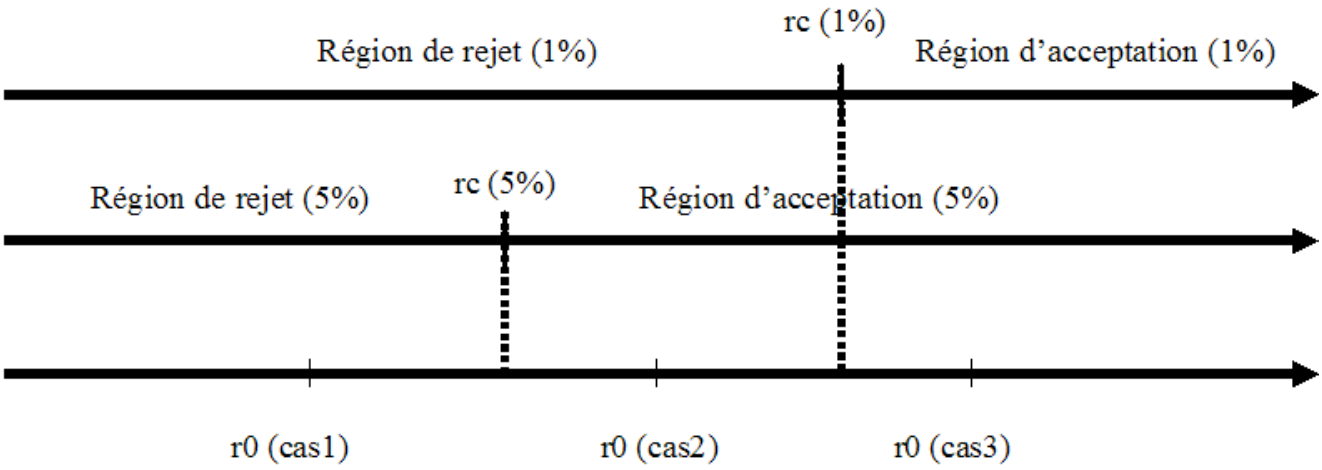

Figure 5 : Illustration de la décision d'un test unilatéral de l'hypothèse H0.

Notons que dans ce cas l'hypothèse nulle est H0 : L'ÉCHANTILLON EST ISSU D'UNE LOI DE LA CLASSE C.

La règle de décision est tirée à partir des situations suivantes :

- (cas 1) Si r0 < rc(5%) On rejette H0 (la série n'est pas à variations régulières).
- (cas 2) Si rc(5%) < r0 < rc(1%) On rejette H0 à 5%, mais on l'accepte à 1%.
- (cas 1) Si r0 > rc(1%) On rejette H0 pour les deux niveaux de signification.

4. Diagramme de la Fonction Moyenne des Excès (FME)

Cette méthode est basée sur la moyenne des excès $e(u) = E[X - u \mid X > u]$ pour un seuil u donné. Cette fonction est constante pour les distributions de type exponentiel (classe E). Pour un échantillon $x_{[1]}, \ldots, x_{[k]}, \ldots, x_{[N]}$ ordonné dans un ordre décroissant, et où $x_{[1]} > x_{[2]} \ldots > x_{[k]}$ sont les dépassements au seuil u, une estimation de $e(u)$ est donnée par $\hat{e}(u) = \frac{1}{k}\sum_{i=1}^{k}\left(x_{[i]} - u\right)$.

L'utilisation de la méthode FME (fonction moyenne des excès) permet, donc, de discriminer entre la classe D (des distributions sub-exponentielles) et la classe E (loi Exponentielle). En effet, si en plus du

fait que la courbe de la fonction FME est linéaire pour les distributions des classes D et E (au niveau des observations les plus élevées), lorsque la pente :

- est nulle, la loi la plus adéquate appartient à la classe E (loi Exponentielle, Figure 6-a) [Étape 8].
- est strictement positive, la distribution la plus adéquate appartient à la classe D des distributions sub-exponentielles : Halphen type A (HA), Gumbel (EV1), Halphen type B (HB), Pearson type 3 (PIII), Gamma (G) (Figure 5-b) [Étape 9].

Notons que dans le SAD, cette méthode doit logiquement être utilisée après la méthode log-log (Figure 2). En effet, si l'hypothèse H0 de la méthode log-log est rejetée (c'est-à-dire si la distribution n'est pas à variations régulières) la méthode FME permet de tester si la distribution est sub-exponentielle ou exponentielle.

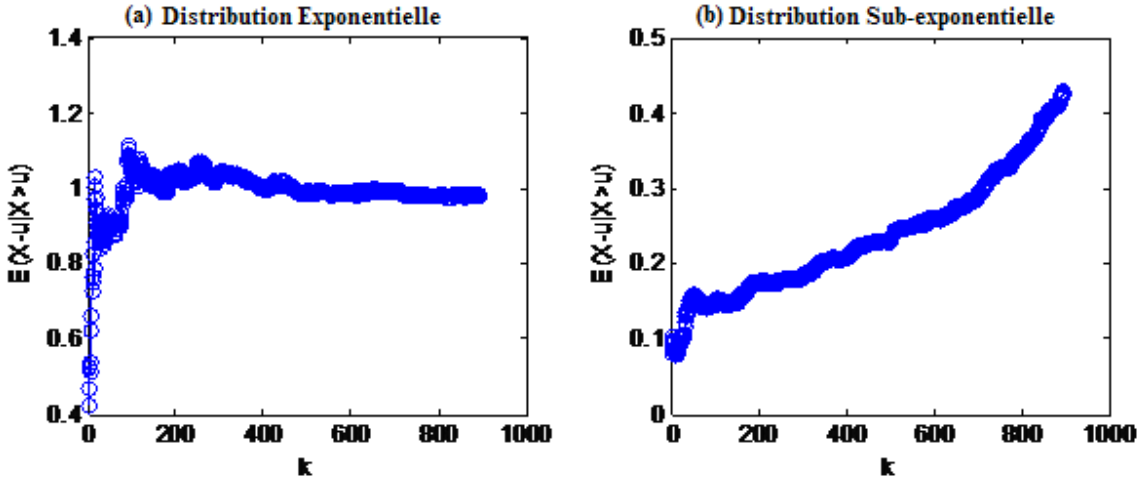

Figure 6 : Fonction moyenne des excès pour une loi exponentielle une loi de la classe Sub-exponentielle

Pour savoir si la loi exponentielle est la plus adéquate, classe E, on calcule la pente de la courbe FME associée aux valeurs de l'échantillon qui dépassent la médiane (50% des valeurs les plus élevées de l'échantillon).

Des études de simulation nous ont permis d'obtenir pour des niveaux de signification 5% et 1%, des valeurs critiques de la pente pour tester l'hypothèse (test unilatéral) H0 : LA PENTE DE LA FME EST NULLE contre l'hypothèse alternative, H1 : LA PENTE DE LA FME EST STRICTEMENT POSITIVE. Ces valeurs critiques sont calculées en fonction de la taille N de l'échantillon ($30 \leq N \leq 200$). ***Notons que les décisions affichées par le SAD sont basées par défaut sur le niveau de signification 5%.***

Ainsi,

- Si l'hypothèse H0 est acceptée au niveau de signification de 5%, on suggère la distribution exponentielle pour représenter les données [Étape 8].
- Si l'hypothèse H0 est rejetée, au niveau de signification 5%, on suggère une distribution de la classe D (HA, EV1, HB, PIII, G) [Étape 9].

Les régions d'acceptation ou de rejet de cette hypothèse nulle pour les deux niveaux de signification (5% et 1%), peuvent être illustrées d'une manière similaire à celle de la méthode Log-log (Figure 5). ***Cependant, les valeurs critiques au niveau de signification 1% sont données pour avoir plus de flexibilité et pour permettre à l'utilisateur de prendre éventuellement une autre décision que celle suggérée sur la base du niveau de signification 5%.***

4. Rapport de Hill [Hill, 1975]

Soit le rapport de Hill :

$$a_N(x_k) = \frac{\sum_{i=1}^{N} I(X_i > x_k)}{\sum_{i=1}^{N} \log(X_i / x_k) * I(X_i > x_k)}$$

où $I(X > x) = \begin{cases} 1 & \text{si } X > x \\ 0 & \text{sinon} \end{cases}$.

$X_1, \ldots, X_N$ sont les valeurs de la variable X et x_k est la k$^{\text{ième}}$ plus grande valeur de X. Sur le graphique de $a_N(x_k)$ en fonction de x_k, on cherche une région stable pour déterminer un estimateur de l'indice des extrêmes (mesure pour la queue de la distribution). Cette statistique est utilisée en pratique dans le SAD pour confirmer le choix d'une distribution appartenant soit à la classe C soit aux classes D et E.

- Si la courbe **converge vers une valeur constante différente de zéro** (Figure 7-a), la distribution étudiée appartient à la classe C (distribution à variations régulières). On suggère alors les lois de la classe C : Fréchet (EV2), Halphen type B Inverse (HIB), Log-Pearson type 3 (LP3), Gamma Inverse (GI).

- Si la courbe **<u>décroit vers zéro</u>** (Figure 7-b), la distribution appartient aux classes : sub-exponentielle (classe D : Halphen type A, Gamma, Pearson type III, Halphen type B, Gumbel et exponentielle (classe E : loi Exponentielle).

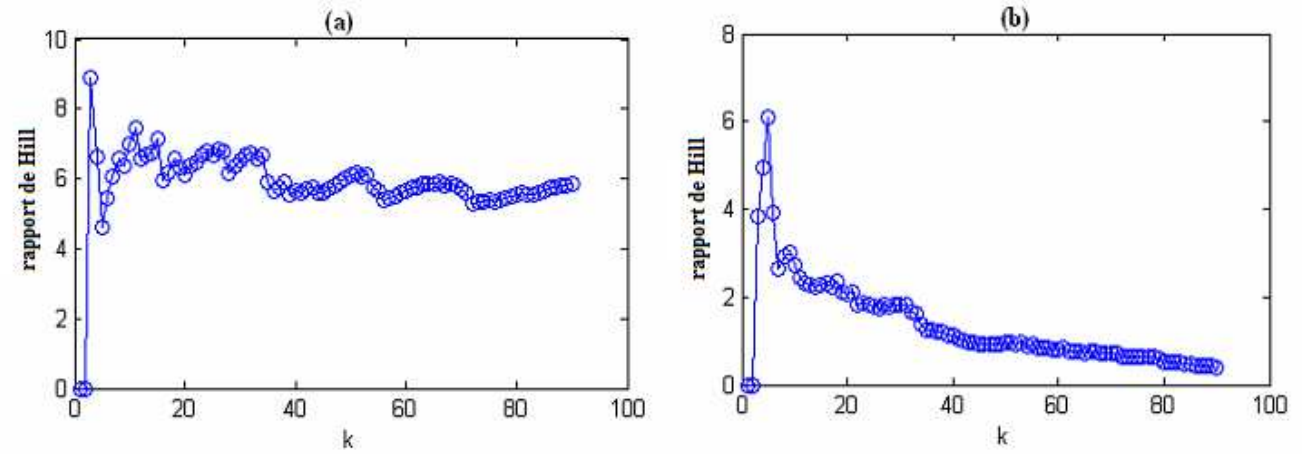

Figure 7 : Rapport de Hill pour (a) une loi à variations régulières (classe C) et (b) de type sub-exponentiel (classe D).

Rappelons que (cf. section 4) pour discriminer entre les classes D et E on peut considérer la méthode FME (Fonction Moyenne des Excès). Le rapport de Hill est utilisé dans un but de confirmation de la conclusion obtenue par la FME.

6. Statistique de Jackson [Jackson, 1967]

Cette statistique est utilisée en pratique dans le SAD pour confirmer le choix d'une distribution appartenant soit à la classe C soit aux classes D et E; Beirlant et al. (2006) ont présenté cette procédure pour caractériser les distributions qui ont strictement un comportement de Pareto (type puissance). Rappelons que les distributions de la classe C (distribution à variations régulières) ont un comportement asymptotique de type puissance, et on peut ainsi utiliser la statistique de Jackson pour examiner l'appartenance de la distribution des observations à classe C. Pour plus de détails sur cette approche voir El Adlouni, Bobée et Ouarda (2008).

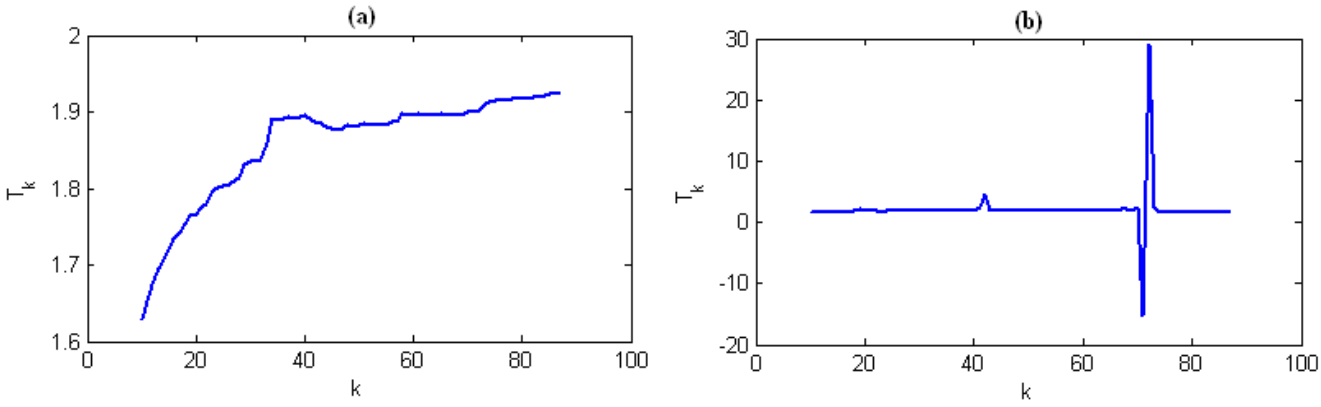

Figure 8 : Statistique de Jackson pour (a) une loi à variations régulières (classe C) et (b) de type sub-exponentiel (classe D).

Ainsi [Étape 10]:

- Si la courbe converge clairement et régulièrement vers 2 (Figure 8-a), la distribution étudiée appartient à la classe C (distribution à variations régulières). On suggère alors les lois de la classe C : Fréchet (EV2), Halphen type B Inverse (HIB), Log-Pearson type 3 (LP3), Gamma Inverse (GI);

- Si la courbe présente des irrégularités et ne converge pas vers 2 (Figure 8-b), la distribution appartient à la classe sub-exponentielle (classe D : Halphen type A, Gamma, Pearson type 3, Halphen type B, Gumbel), ou exponentielle (classe E : loi Exponentielle, Figure 8-b).

Rappelons que (cf. section 4) pour discriminer entre les classes D et E on peut considérer la méthode FME (Fonction Moyenne des Excès).

Références :

Beirlant, J., de Wet, T., Goegebeur, Y., (2006). A goodness-of-fit statistic for Pareto-type behaviour. Journal of Computational and Applied Mathematics, 186, 99-116.

El Adlouni, S., Bobée, B. et Ouarda, T. B.M.J (2008). On the tails of extreme event distributions in Hydrology. Accepté au Journal of Hydrology.

Hill, B.M. (1975). A Simple General Approach to Inference about the Tail of a Distribution. The Annals of Statistics 3 (5), 1163-1174.

Jackson, O.A.Y., (1967). An analysis of departures from the exponential distribution. Journal of the Royal Statistical Society B, 29, 540-549.

Jarque, C.M. et A.K. Bera (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Economics Letters 6 (3): 255–259.

Martel, B., S. El Adlouni et B. Bobée (2011). Comparison of the power of Log-normality tests with different right tail alternative distributions. Soumis au JHE (ASCE).