

GUIDE FOR THE USE OF THE DECISION SUPPORT SYSTEM (DSS)*

*Note: In French SAD (Système d'Aide à la Décision)

1. Introduction to the DSS

Eighteen statistical distributions are available in HYFRAN-PLUS software to fit data sets that are independent, homogenous and stationary. A Decision Support System (DSS) is developed to support selection of the most appropriate class of distributions, with respect to extreme values. Distributions that are usually used in flood frequency analysis can be grouped in three main classes:

- Class C (regularly varying distributions): Fréchet (EV2), Halphen IB (HIB), Log-Pearson (LP3), Inverse Gamma (IG).
- Class D (sub-exponential distributions): Halphen type A (HA), Halphen type B (B), Gumbel (EV1), Pearson type 3 (P3), Gamma (G).
- Class E (Exponential distribution).

Figure 1 presents exponential (E), sub-exponential (D) and regularly varying (C) distributions. Distributions are ordered from light tailed (from the left) to heavy tailed (to the right). The limiting cases (bottom squares) represented by distributions in the limits of classes. The tail of the class C distributions is heavier than that of the class D distributions, which is heavier than that of the class E. Thus, estimated quantiles can be ordered equivalently. Indeed, for a given sample, the T-event corresponds to the quantile of the probability of non-exceedance $p = 1 - 1/T$ estimated by distributions of the classes C, D and E, are QT (C), QT (D) and QT (E) respectively, which verify the following relation: $QT (E) < QT (D) < QT (C)$.

~ In Business Since 1971 ~



Water Resources Publications, LLC.

P.O. Box 630026 • Highlands Ranch, CO 80163-0026, USA
E-mail: info@wrpllc.com • <http://www.wrpllc.com>

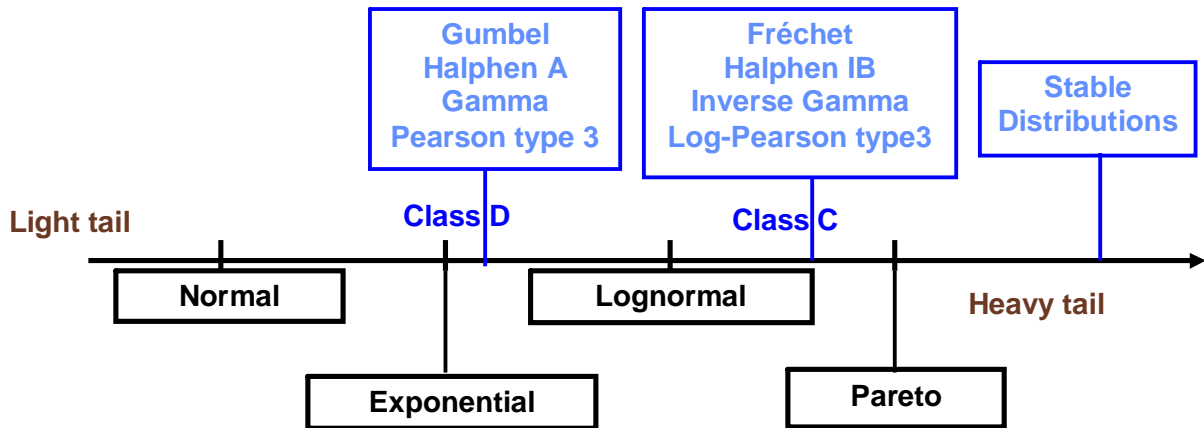


Figure 1: distributions ordered with respect to their right tails (El Adlouni et al., 2008).

The methods developed in the DSS allow the identification of the most adequate class of distribution to fit a given sample, especially for extremes. These methods are (cf. Diagram):

- The Log-Log plot : used to discriminate between on the one hand the class C and on the other hand the classes E and D;
- The mean excess function (MEF) to discriminate between the classes D and E; and
- Two statistics: Hill's ratio and modified Jackson statistic, for confirmatory analysis of the conclusions suggested by the previous two methods.

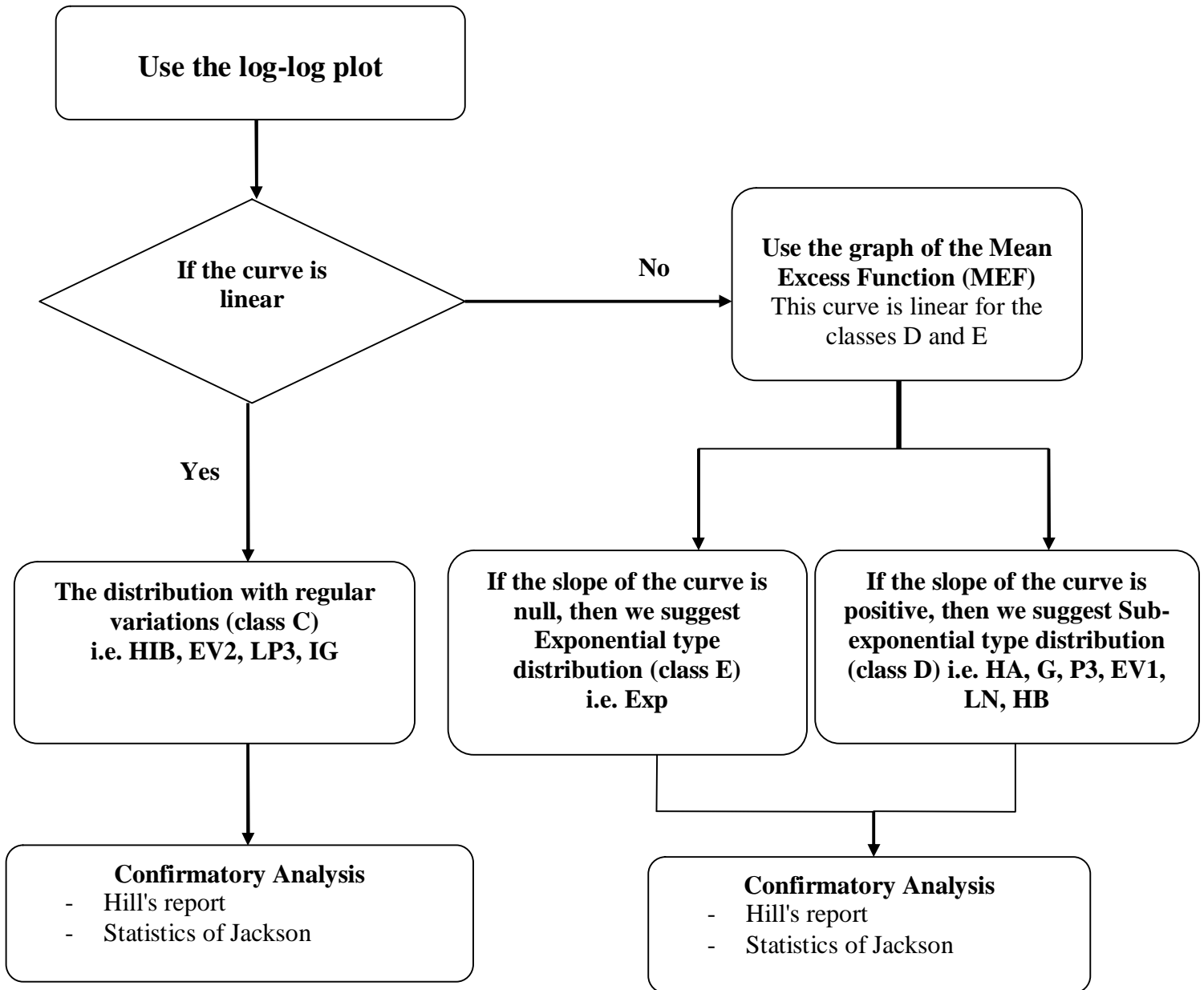


Figure 2: Diagram for class selection used in the DSS

More theoretical details of this classification and the criteria are available in El Adlouni et al. (2008). This article is available as attachment in the HYFRAN-PLUS setup.

2. Log-Log plot

The log-log plot is based on the fact that the survival function $\bar{F}(u) = P(X > u)$, is given by $\bar{F}(u) = P(X > u) = e^{-u/\theta}$ for exponential tail with mean θ , and for regularly varying distribution with tail index α , \bar{F} is equivalent to (for large quantile) :

$$\bar{F}(u) = P(X > u) \approx C \int_u^\infty \frac{1}{x^\alpha} dx = C \left[\frac{x^{-\alpha+1}}{1-\alpha} \right]_u^\infty = C_1 u^{-\alpha+1} \text{ (with } \alpha > 1, \text{ which is equivalent to finite mean).}$$

Therefore, taking the logarithm we have regularly varying distributions $\log[P(X > u)] \approx \log C_\alpha - (\alpha - 1)\log(u)$. This suggests that, for the log-log plot, the tail probability is represented by a straight line for power-law (or regularly varying distributions, class C) but not for the other sub-exponential or exponential distributions (class D or E).

As illustrated in figure 3, the curve represented in the Log-Log plot corresponds to a straight line for the distributions of the class C i.e. Fréchet (EV2), Halphen type IB (HIB), Log-Pearson type 3 (LP3) and Inverse Gamma (IG), but not for sub-exponential or exponential type tails (class D or E). When the diagram is not linear we suggest the use of the Mean Excess Function (MEF) to discriminate between the classes D and E.

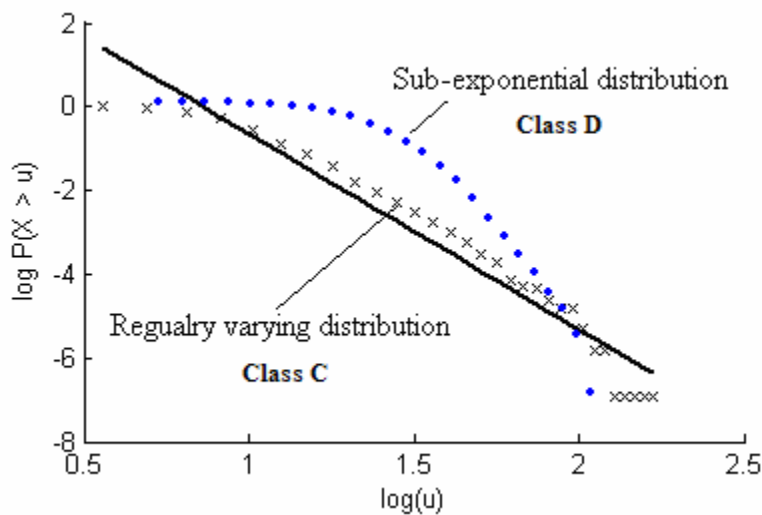


Figure 3: Illustration of the Log-Log plot to characterize the regularly varying distributions

To check the linearity of the curve in the log-log diagram, a test on the associated correlation coefficient is considered. Simulation studies allow the determination of critical values corresponding to significance levels of 5 % and 1 %, to test the HYPOTHESIS H0: THE DATA FOLLOW A DISTRIBUTION OF

THE CLASS C (i.e. THE CURVE IS LINEAR). These critical values are calculated according to the size N of the sample ($30 \leq N \leq 200$). *Note that the decisions given by the DSS are based, by default, on the significance level 5 %.*

If the hypothesis H_0 is rejected, at the significance level 5 %, we suggest the use of the mean excess function plot (MEF). *However the critical values at the significance level 1 % are given for more flexibility and to allow the user to make another decision than that based on the significance level 5 %.*

Indeed, if the observed correlation coefficient (r_0) is greater than critical value (r_c) at the significance level 5 %, then we conclude that it is not significantly different from 1 at the significance level 5 % and the hypothesis H_0 of linearity is accepted at this level (Figure 4). In this case, the most adequate choice corresponds to the class C of regularly varying distributions (power-law type): Halphen type IB (HIB), Fréchet (EV2), Log-Pearson type 3 (LP3), Inverse Gamma (IG).

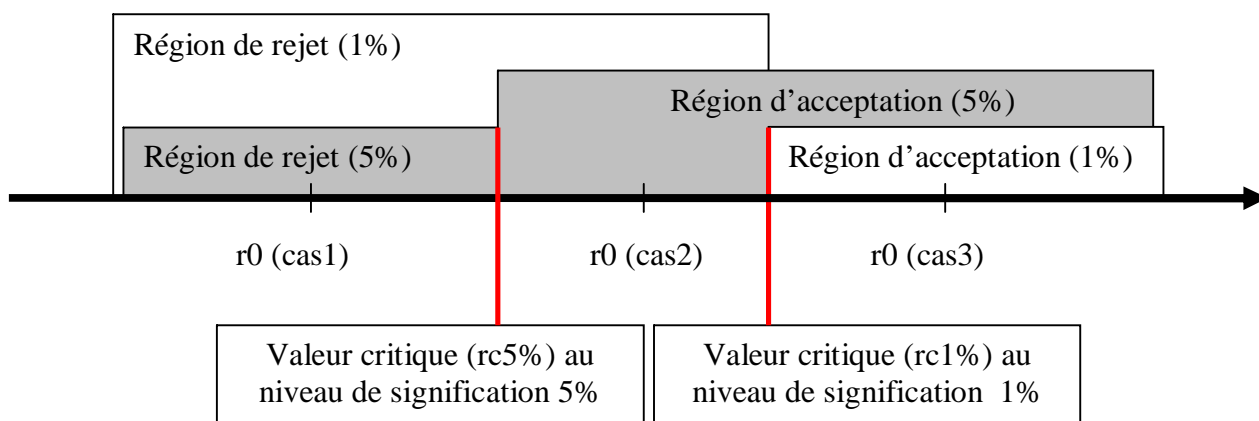


Figure 4 : Illustration de la décision d'un test unilatéral de l'hypothèse H_0 .

Figure 4 shows, in general, the decision rule for an unilateral test related to two significance levels 1% and 5%. The critical values corresponding to each significance level are, respectively, $r_{c1\%}$ and $r_{c5\%}$. These two critical values are obtained by Monte Carlo simulations generated from regularly varying distributions. For a given dataset, we calculate the correlation coefficient r_0 . To illustrate the use of this test, three cases are considered such as the correlation coefficients verify: $r_0(\text{cas1}) < r_{c5\%} < r_0(\text{cas2}) < r_{c1\%} < r_0(\text{cas3})$. The hypothesis H_0 (case1) is rejected for the significance levels 1% and 5%. Indeed, $r_0(\text{cas1}) < r_{c5\%}$ and $r_0(\text{cas1}) < r_{c1\%}$. In this case the distribution is not regularly varying (the curve is not linear). For case2, the hypothesis H_0 is rejected at the significance level 1%, but it is accepted at the significance level of 5%. Indeed, $r_0(\text{cas2}) > r_{c5\%}$ and $r_0(\text{cas2}) < r_{c1\%}$. For this case, the hypothesis H_0 is

accepted by the SAD and the use of regularly varying distribution is suggested (based on the significance level 5%). However, the critical value at the significance level of 1% is presented to give more flexibility to the user. The case 3, corresponds to the case where r_0 is higher than the two critical values ($r_0(\text{cas3}) > r_{c5\%}$ and $r_0(\text{cas3}) > r_{c1\%}$). In this case, and for the two significance levels, the hypothesis H_0 is accepted and the suggested distribution belong to the class C of regularly varying distributions.

3. The Mean Excess Function Diagram (MEF)

The mean excess function method is based on the function $e(u) = E[X - u | X > u]$. This function is constant for exponential tail distributions ($e(u) = \theta$). However, in the case of regularly varying distribution with tail index α ($\alpha > 2$): $e(u) = \frac{u}{(\alpha - 2)}$. The Mean Excess Function (MEF) allows

discriminating between the class D (sub-exponential distributions) and the class E (Exponential distribution). Indeed, the curve presented in the MEF diagram is linear for high observed values for distributions of both classes D and E. If in addition the slope of this curve is (Figure 5):

- Equal to zero, the most adequate distribution belongs to the class E (Exponential law);
- Strictly positive, the most adequate distribution belongs to the class D of sub-exponential distributions: Halphen type A (HA), Gumbel (EV1), Halphen type B (HB), Pearson type 3 (P3), Gamma (G).

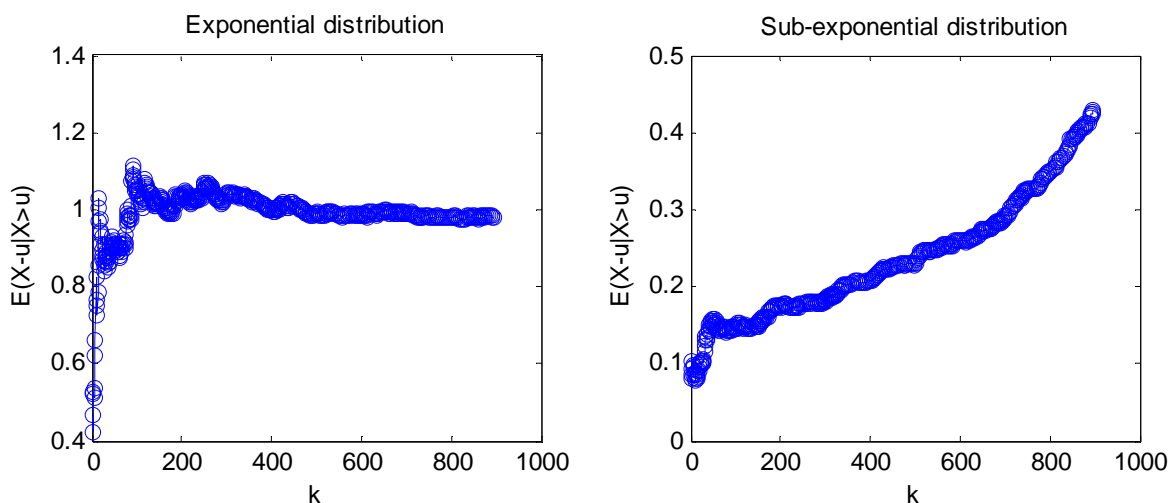


Figure 5: Mean excess function for exponential and sub-exponential distributions.

The use of this diagram in the DSS is based on the slope of the MEF curve for the observations that exceed the median (50 % of the highest observed value of the sample).

Simulation studies allow the determination of critical values corresponding to significance levels of 5 % and 1 %, to test the HYPOTHESIS H0: THE DATA FOLLOW A DISTRIBUTION OF THE CLASS E (i.e. THE SLOPE OF THE MEF IS EQUAL TO ZERO). These critical values are calculated according to the size N of the sample ($30 \leq N \leq 200$). ***Note that the decisions given by the DSS are based, by default, on the significance level 5 %.***

When the hypothesis H0 is accepted we suggest the use of the Exponential distribution (class E). However, when it is rejected at the significance level 5 %, we suggest the use of a distribution of the class D (HA, EV1, HB, P3, G). ***Note that the critical values at the significance level 1 % are given for more flexibility and to allow the user to make possibly another decision than that suggested for the significance level of 5% (Figure 4).***

Remark:

- The Lognormal distribution (LN) doesn't belong to any of these classes. It has an asymptotic behaviour which is in the frontier of the classes C and D. Indeed, the LN tail is lighter (respectively, heavier) than that of a distribution of the class C (respectively, class D). Thus, the quantiles (QT) estimated by a distribution belonging to the classes C, D and the LN, verify the following relation:

$QT(D) < QT(LN) < QT(C)$. Consequently:

- If the parent distribution is regularly varying (class C), and the LN distribution is considered for the fit, thus the estimated quantile, for a fixed return period, will be lower than the real value and there is a risk to underestimate this quantile;
- If the true distribution is sub-exponential (class D), and the LN distribution is considered for the fit, thus the estimated quantile, for a fixed return period, will be higher than the real value and there is a risk to overestimate this quantile.

In the DSS, and to have a safe choice, LN is considered by default as a distribution of the class D. However, the user could make a different decision and associate it to the class C.

4. Hill's ratio plot [for the theoretical details cf. El Adlouni et al. 2008]

The Hill ratio is defined by

$$a_n(x_n) = \frac{\sum_{i=1}^n I(X_i > x_n)}{\sum_{i=1}^n \log(X_i / x_n) I(X_i > x_n)}$$

where $I(X_i > x_n) = \begin{cases} 1 & \text{if } X_i > x_n \\ 0 & \text{if } X_i < x_n \end{cases}$.

This method is based on the fact that a_n is a consistent estimator of α if the tail is regularly varying (Class C) with tail index α (Hill, 1975). In the expression of the Hill ratio, x_n is chosen to be large such that $P(X > x_n) \rightarrow 0$ and $nP(X > x_n) \rightarrow \infty$, and I is the indicator function. The standard Hill estimator, of the tail index, corresponds to the particular case where the observations are ordered $X_{(1)} \leq \dots \leq X_{(n)}$ and $x_n = X_{(k_n+1)}$, where k_n is an integer which tends to infinity as n tends to infinity.

In practice, one plots $a_n(x_n)$ as a function of x_n and looks for some stable region from which $a_n(x_n)$ can be considered as an estimator of α . Figure 4, presents the Hill ratio plot for a sample generated from the regularly varying (a) and Exponential (b) distributions.

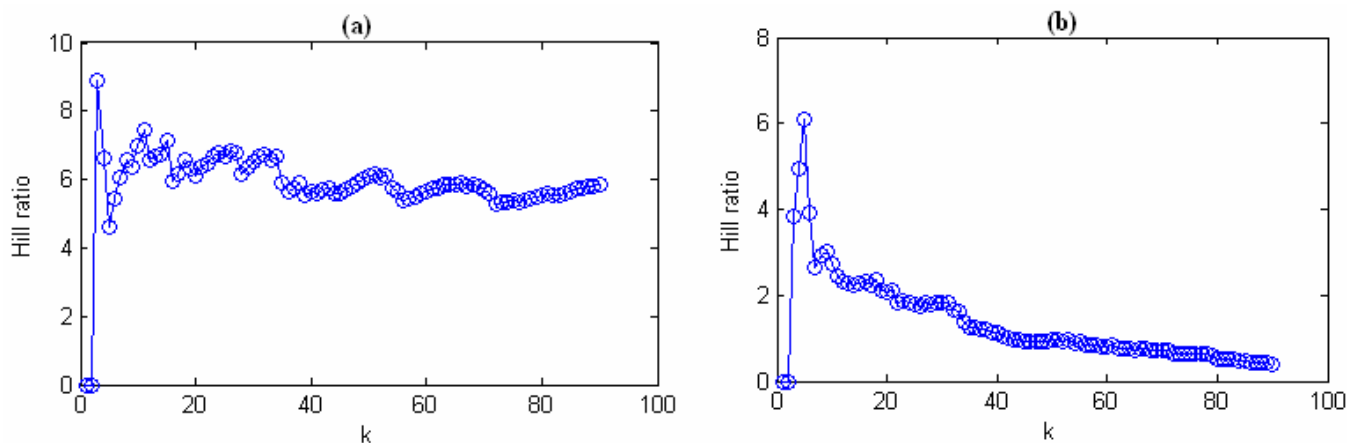


Figure 4: Generalized Hill ratio plot for (a) regularly-varying and (b) sub-exponential distributions.

This statistics is used in the DSS to confirm the suggested choice given by the first two diagrams (the distribution belongs to the class C, D or E).

- If the curve converges to a non-null constant value, the most adequate distribution belongs to the class C (regularly varying distribution). We suggest then the use of a distribution of the class C: Fréchet (EV2), Halphen type B Inverse (HIB), Log-Pearson type 3 (LP3), Inverse Gamma (IG).
- If the curve **decreases to zero**, the distribution belong to the Sub-exponential class (class D: Halphen type A, Gamma, Pearson type 3, Halphen type B, Gumbel); and the Exponential class (class E: Exponential distribution).

Note that (cf. section 3) to discriminate between the classes D and E, we suggest the use of the MEF method.

5. Jackson Statistic [for the theoretical details cf. El Adlouni et al. 2008]

This method is presented by Beirlant et al. (2006) and is based on the Jackson statistic. It allows to test whether the sample is consistent with Pareto type distributions (Class B). Note that the distributions of the class C (regularly varying distribution) have asymptotically the same behaviour as that of the Pareto distribution. Originally the Jackson statistic (Jackson, 1967) was proposed as a goodness-of-fit statistic for testing exponential behaviour, and given the link between the Exponential and the Pareto distribution (if X has a Pareto distribution the logarithmic transformation $Y = \log(X)$ is exponentially distributed) this statistic is used to assess Pareto-type behaviour. The Jackson statistic is further modified by taking into account the second-order tail behaviour of a Pareto-type model. Beirlant et al. (2006) give the limiting distribution of this statistic with corrected bias version for finite size samples. The modified Jackson statistic converges to 2 for regularly varying distribution (Power-law) and has an irregular behaviour for sub-exponential or exponential distributions (Figure 5).

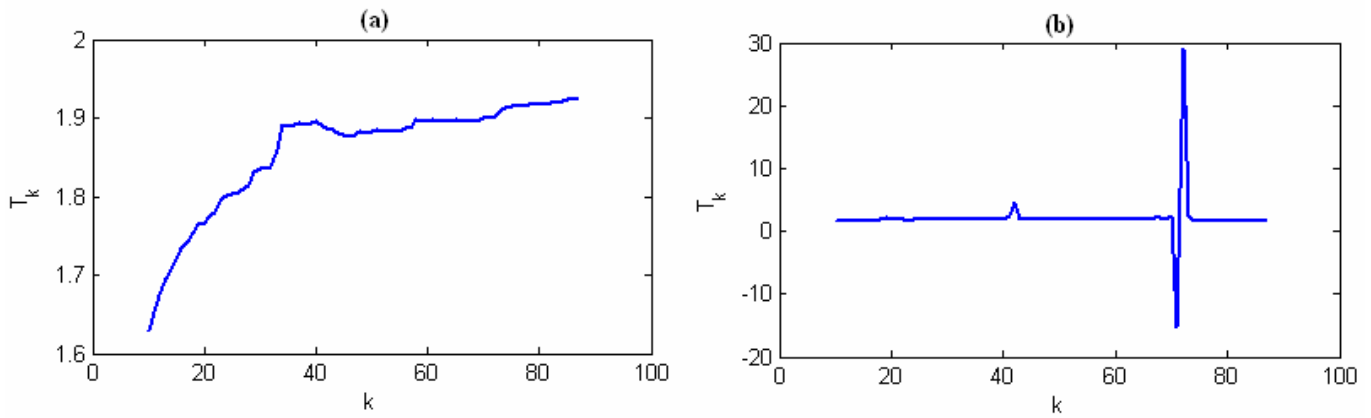


Figure 5: Modified Jackson statistic for (a) regularly varying and (b) sub-exponential distributions.

In the DSS this method is considered as a confirmatory method for suggested decision based on the Log-Log and the MEF. So:

- If the curve converges clearly and regularly to 2, the studied distribution belongs to the class C (regularly varying distribution). We suggest then, the use of: Fréchet (EV2), Halphen type IB (HIB), Log-Pearson type 3 (LP3), Inverse Gamma (IG);
- If the curve presents some irregularities for the distribution tail, than we suggest the sub-exponential class (class D: Halphen type A, Gamma, Pearson type 3, Halphen type B, Gumbel); or exponential (class E: Exponential distribution).

Note that (cf. section 3) to discriminate between the classes D and E, we suggest the use of the MEF method.

Remarque:

Even if the modified Jackson statistic was developed to test Pareto type behaviour, it is used in the DSS to check if the of the studied distribution has similar tail as regularly varying distribution (class C). In deed, distributions of the class C have asymptotically Pareto type tail. In practice, the Generalized Pareto distribution (GPD) is used in the Peaks-over-threshold model (POT). However, the GPD is available in HYFRAN and can be used to fit any data sets that are independent, homogenous and stationary.

Reference:

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