TP SdF N° 26

La Théorie des Valeurs Extrêmes (TVE)

Ce TP porte sur la théorie des valeurs extrêmes qui a pour objet de modéliser une queue de distribution pour laquelle peu d'observations sont disponibles et d'estimer la probabilité d'occurrence d'un évènement rare (une hauteur de crue, une vitesse de vent, une dérive de signal… jamais observées jusqu'alors par exemple).

1 – Théorie des valeurs extrêmes

Présenter la théorie des valeurs extrêmes.

2 – Application à la loi normale

- Simuler 20 échantillons de 100 valeurs suivant une loi normale ($m = 10$, $\sigma = 2$) correspondant par exemple à 20 années d'observation d'une hauteur d'eau.

- Estimer la probabilité d'occurrence d'une valeur supérieure à $m + 3{,}5\,\sigma$ au cours d'un siècle :
    - par la loi théorique (normale),
    - par la loi généralisée des extrêmes (GEV) à partir des valeurs simulées,
    - par la méthode des dépassements (POT) à partir de ces mêmes valeurs.

- Comparer les résultats obtenus.

1 Théorie des valeurs extrêmes

La théorie des valeurs extrêmes propose d'approximer la queue d'une distribution expérimentale par une loi théorique particulière puis de réaliser des estimations à partir de cette dernière.

Deux approches sont envisagées :

- l'analyse des maxima par intervalles de temps fixes (crues maximales décennales par exemple),
- l'analyse des valeurs au-dessus d'un seuil (toutes les crues supérieures à une certaine hauteur par exemple) selon la méthode des dépassements ou POT (Peak Over Threshold)

La loi généralisée des extrêmes ou GEV (Generalized Extreme Value) est utilisée dans le premier cas et la loi généralisée de Pareto ou GPD (Generalized Pareto Distribution) dans la seconde.

L'approche POT est généralement préférée car elle exploite plus d'informations sans avoir la contrainte de devoir relever les valeurs expérimentales à intervalle régulier.

1.1 Loi généralisée des extrêmes (GEV)

Soit X_1, X_2, X_3,…, X_n, un ensemble de valeurs prises par une variable aléatoire de loi inconnue et M_n, la valeur maximale parmi ces n valeurs, le but est de définir la loi que suit Mn.

D'après le théorème de Fischer-Tippett, s'il existe deux réels a ≥ 0 et b tels que $\lim_{n\to\infty} P(\frac{M_n - b}{a} \le x) = G(x)$, alors G est du type de l'une des trois distributions suivantes :

- Gumbel :

$$G = \exp(-\exp(-\frac{x-\mu}{\sigma})) \quad avec -\infty < x < +\infty$$

- Fréchet :

$$G(x) = 0 \text{ si } x \le \mu \qquad G(x) = \exp(-(\frac{x-\mu}{\sigma})^{-\alpha}) \text{ si } x > \mu,\ \alpha > 0 \text{ et } \sigma > 0$$

- Weibull négative :

$$G(x) = \exp(-(-(\frac{x-\mu}{\sigma}))^{\alpha}) \text{ si } x < \mu, \alpha > 0 \text{ et } \sigma > 0 \qquad G(x) = 1 \text{ si } x \ge \mu$$

Von Mises et Jenkinson ont unifié ces 3 lois par la distribution généralisée des valeurs extrêmes (GEV : Generalized Extreme Value) :

$$G(x) = \exp(-(1+\xi(\frac{x-\mu}{\sigma}))^{-\frac{1}{\xi}}) \quad \text{pour } \xi \neq 0 \text{ et } 1+\xi(\frac{x-\mu}{\sigma}) > 0$$

$$G(x) = \exp(-\exp(-\frac{x-\mu}{\sigma})) \quad \text{pour } \xi = 0$$

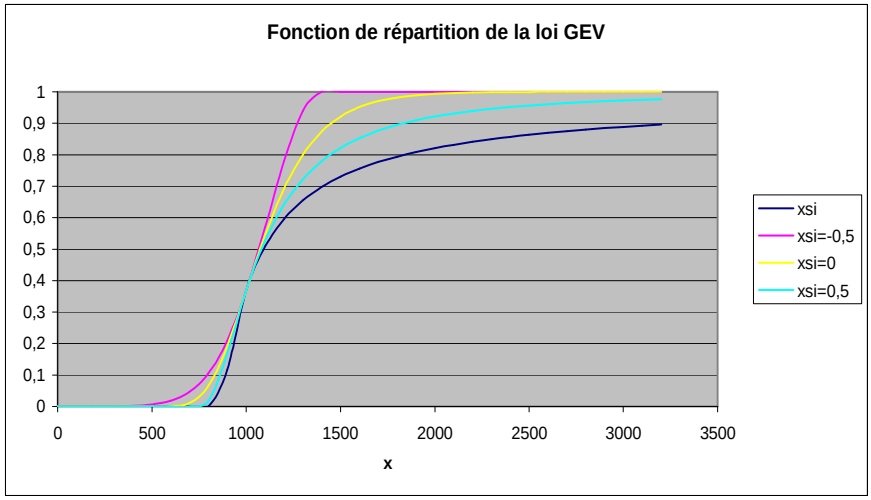
Avec μ le paramètre de localisation, σ le paramètre de dispersion et ξ le paramètre de forme.

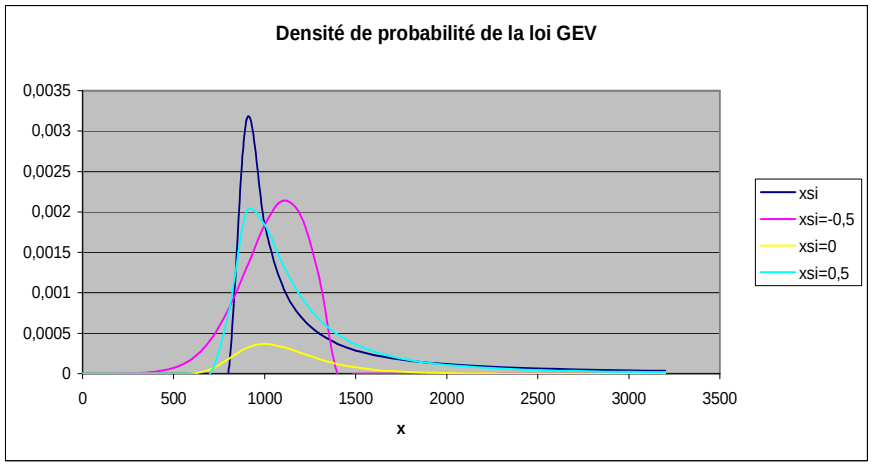
- loi de Gumbel si ξ = 0
- loi de Fréchet si ξ > 0 G(x) = 0 pour x ≤ μ - σ/ξ
- loi de Weibull négative si ξ < 0 G(x) = 1 pour x ≥ μ - σ/ξ

Plus la valeur de ξ est élevée plus la queue de la distribution est épaisse comme le montrent les graphiques suivants.

LOI GEV

mu :	1000
sigma :	200
xsi :	1,2





Les fichiers Excel sont disponibles en cliquant sur les icônes :

Feuille de calcul Microsoft Excel

La loi GEV a pour densité de probabilité :

$$g(x) = \frac{1}{\sigma} \times e^{-(1+\xi(\frac{x-\mu}{\sigma}))^{-\frac{1}{\xi}}} \times (1+\xi(\frac{x-\mu}{\sigma}))^{-1-\frac{1}{\xi}} \quad pour\ \xi \neq 0$$

$$g(x) = \frac{1}{\sigma} \exp(-(\frac{x-\mu}{\sigma})) \times \exp(-\exp(-(\frac{x-\mu}{\sigma}))) \quad pour\ \xi = 0$$

La méthode du maximum de vraisemblance permet d'ajuster la loi GEV en maximisant le produit des densités de probabilité obtenues pour les valeurs expérimentales (ou la somme de leur logarithme).

La valeur d'un quantile peut être calculée par la formule suivante obtenue par inversion de la fonction de répartition G :

$$q_{1-p} = \mu - \frac{\sigma}{\xi}(1-(-\ln(1-p))^{-\xi}) \quad pour\ \xi \neq 0 \quad et \quad q_{1-p} = \mu - \sigma \ln(-\ln(1-p)) \quad pour\ \xi = 0$$

1.2 Méthode des dépassements (POT)

Soit X une variable aléatoire de fonction de répartition F et u une valeur de seuil, la variable aléatoire Y = X – u pour X > u suit la fonction de répartition conditionnelle :

$$G(y) = G(x-u) = \frac{F(x)-F(u)}{1-F(u)} \text{ avec } x > u$$

Pour une grande valeur de seuil u, G suit une loi généralisée de Pareto (GPD : Generalized Pareto Distribution) de la forme :

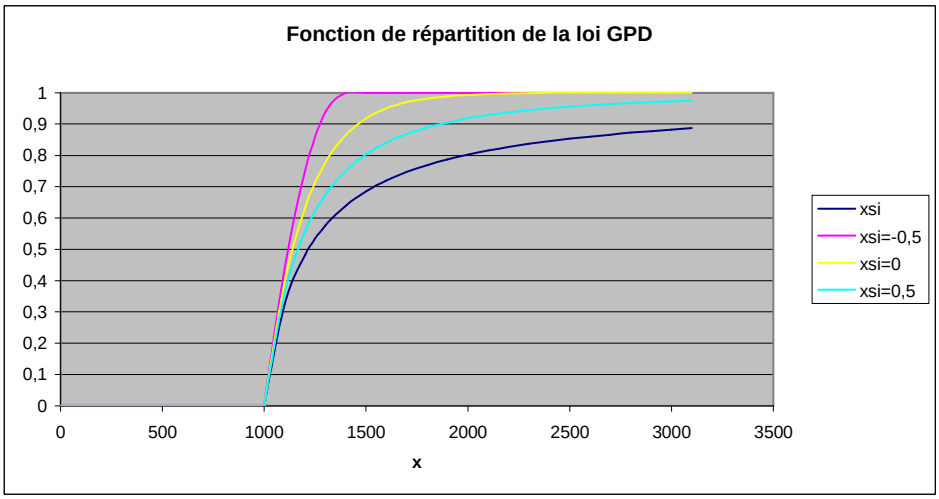
$$G(y) = 1 - \left[1 + \xi\left(\frac{y}{\sigma}\right)\right]^{-\frac{1}{\xi}} \quad \text{pour } \xi \neq 0 \text{ et } 0 \leq y \leq -\sigma/\xi \ (si\ \xi < 0)$$

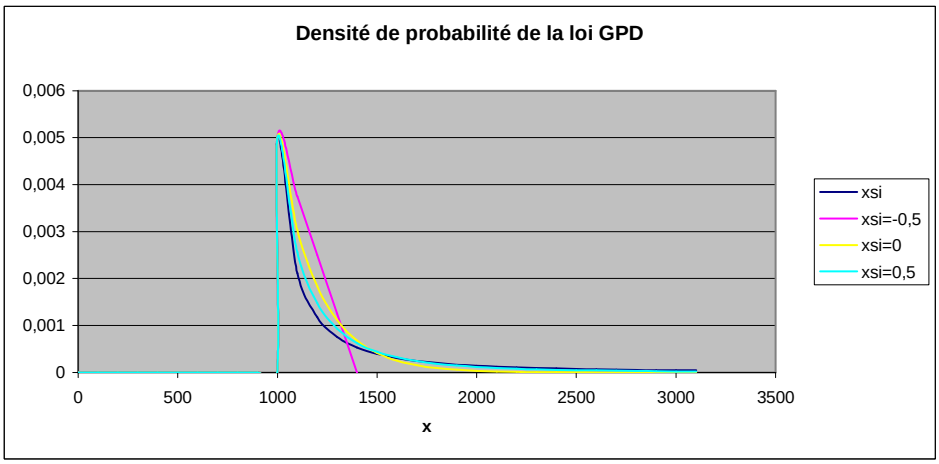
$$G(y) = 1 - \exp\left(-\frac{y}{\sigma}\right) \quad \text{pour } \xi = 0$$

La queue de la distribution est d'autant plus épaisse que la valeur de ξ est élevée :

LOI GPD

u :	1000
sigma :	200
xsi :	1,2





Feuille de calcul Microsoft Excel

La loi GPD a pour densité de probabilité :

$$f(y) = \frac{1}{\sigma}\left(1 + \xi\left(\frac{y}{\sigma}\right)\right)^{-1-\frac{1}{\xi}} \ pour\ \xi \neq 0 \qquad f(y) = \frac{1}{\sigma} \times \exp\left(-\frac{y}{\sigma}\right) \ \text{pour } \xi = 0$$

La méthode du maximum de vraisemblance permet d'ajuster la loi GPD. Une difficulté réside cependant dans le choix de la valeur du seuil u. En effet, un biais va apparaître si le seuil est trop petit et on perdra de l'information si le seuil est trop grand. Aussi, est-il préconisé de choisir comme seuil la valeur pour laquelle la fonction moyenne des excès (FME) devient approximativement linéaire :

$$FME = \frac{1}{n_u} \times \sum_{i=1}^{n_u} (x_i - u)$$ avec u le seuil et nu le nombre de dépassements de u

La valeur d'un quantile peut être calculée par la formule suivante obtenue par inversion de la fonction de répartition G :

$$q_{1-p} = u + \frac{\sigma}{\xi}((np/k)^{-\xi} - 1) \ pour \ \xi \neq 0 \quad et \quad q_{1-p} = u - \sigma \ln(-np/k) \ pour \ \xi = 0$$

avec k le nombre de dépassements parmi n le nombre de valeurs.

De par la définition même de G(x-u), il est également possible de calculer F(x) à partir de F(u) calculé par ailleurs :

$$F(x) = F(u) + G(x-u)[1-F(u)]$$

2 Application à la loi normale (m = 10, σ = 2)

2.1 Estimation par la loi théorique

La probabilité qu'une valeur de la loi normale ne dépasse pas m + 3,5 σ s'obtient par la fonction de répartition directement fournie sous Excel :

= LOI.NORMALE(m+3,25*σ;m; σ;VRAI) soit 0,999767327

A raison de 100 valeurs par an, la probabilité de dépassement au cours d'un siècle est :

$1-(0{,}999767327)^{10000} = 0{,}902412363$

2.2 Estimation par la loi généralisée des extrêmes (GEV)

On simule 20 échantillons de 100 valeurs suivant la loi normale en appliquant la fonction inverse de la fonction de répartition à une valeur tirée aléatoirement entre 0 et 1, soit sous Excel :

= LOI.NORMALE.INVERSE(ALEA();m; σ)

Puis on ajuste la loi GEV par la méthode du maximum de vraisemblance à partir des 20 valeurs maximales de chaque échantillon comme dans l'exemple ci-dessous.

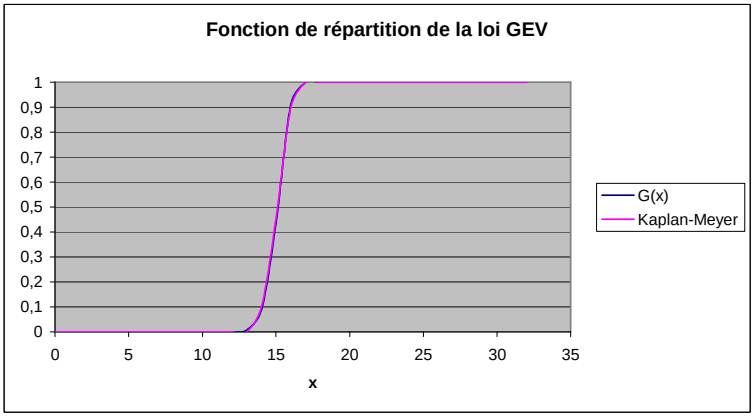
Ajustement de la loi GEV

mu :	14,8654509
sigma :	0,78901228
xsi :	-0,46558745

LN Vraisemblance : -21,6427956

x (moyenne + 3,5 sigma) :	17
G(x) :	1
Probabilité de dépassement au cours d'un siècle :	0
Valeur théorique :	0,90241236

Année	Echantillon	g(x)	Ln (g(x))
1	16,3227467	0,13082293	-2,03391052
2	16,0674791	0,28606559	-1,25153415
3	15,9526568	0,34956825	-1,05105647
4	15,8320351	0,40833931	-0,89565681
5	15,6846081	0,46618634	-0,76316985
6	15,5755092	0,49766777	-0,69782256
7	15,4727321	0,51793441	-0,65790667
8	15,4327107	0,52332907	-0,64754482
9	15,2901834	0,53136403	-0,63230793
10	15,1361207	0,52157079	-0,65091027
11	15,0069431	0,50040926	-0,692329
12	14,8510633	0,4622513	-0,7716466
13	14,7439019	0,43001586	-0,84393319
14	14,6572754	0,40148293	-0,91259025
15	14,6180554	0,3880493	-0,94662289
16	14,5661167	0,3699143	-0,99448393
17	14,4608135	0,33248551	-1,101159
18	14,1505973	0,22568627	-1,48860945
19	13,8797645	0,14751887	-1,91379921
20	13,4752981	0,06748823	-2,695802



Feuille de calcul Microsoft Excel

Dans cet exemple, la probabilité de dépassement de m + 3,5 σ au cours d'un siècle est estimée négligeable, contrairement à la valeur théorique.

2.3 Estimation par la méthode des dépassements (POT)

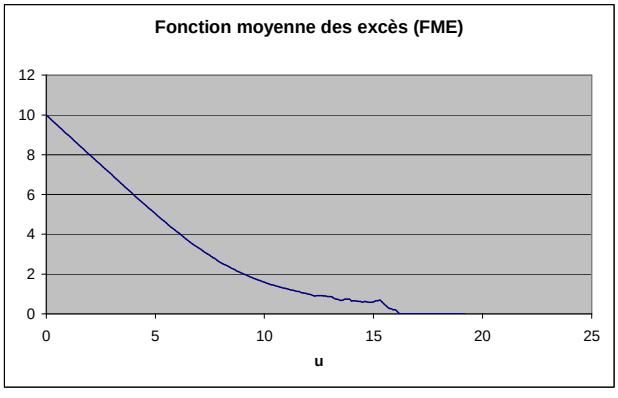
A partir des 2000 valeurs simulées précédemment suivant la loi normale, on utilise la fonction moyenne des excès (FME) pour déterminer le seuil u. La FME peut être calculée sous Excel au moyen de fonctions conditionnelles :

$$FME(u) = \frac{1}{n_u} \times \sum_{i=1}^{n_u} (x_i - u)$$ = SOMME.SI(plage;">u")/NB.SI(plage;">u")-u

Fonction moyenne des excès (FME)

xi	u	FME
9,79979342	0	9,98701554
6,87979476	0,1	9,88701554
6,36588817	0,2	9,78701554
8,30773791	0,3	9,68701554
10,9764312	0,4	9,58701554
10,3410919	0,5	9,48701554
11,3551417	0,6	9,38701554
10,0046144	0,7	9,28701554
10,0448169	0,8	9,18701554
10,8974207	0,9	9,08701554
8,51601661	1	8,98701554
10,9914281	1,1	8,88701554
9,41545958	1,2	8,78701554
8,85319708	1,3	8,68701554
9,54305618	1,4	8,58701554
9,21109969	1,5	8,48701554
8,41926053	1,6	8,38701554
10,4590793	1,7	8,28701554
9,22717533	1,8	8,18701554
8,97387973	1,9	8,08701554
5,7937138	2	7,98701554
7,46416584	2,1	7,88701554
10,3454512	2,2	7,78701554
6,50925442	2,3	7,68701554

$$FME = \frac{1}{n_u} \times \sum_{i=1}^{n_u} (x_i - u)$$



Feuille de calcul Microsoft Excel

La valeur 12 (m + 1 σ) apparaît convenir comme valeur de seuil.

La loi GPD, ci-dessous, a été ajustée par la méthode du maximum de vraisemblance à partir de 303 valeurs dépassant le seuil, parmi les 2000 valeurs simulées.

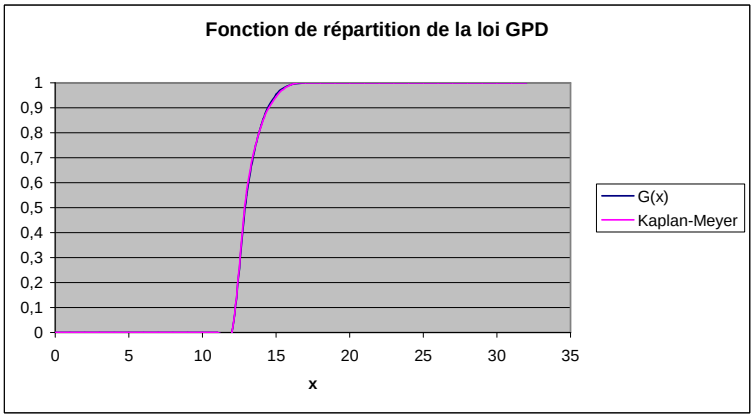
Ajustement de la loi GPD

u :	12
sigma :	1,37168724
xsi :	-0,23466419

LN Vraisemblance : -327,657337

n :	3,5	
x (moyenne + n sigma) :	17	
G(x) :	0,99973617	par siècle
Probabilité de dépassement :	4,18579E-05	0,34202489
Valeur théorique :	0,000232673	0,90241236

Xi	g(x)	Ln (g(x))
16,3227467	-4,70341741	-4,70341741
16,0674791	-4,19790631	-4,19790631
15,9526568	-3,99379014	-3,99379014
15,8320351	-3,79229139	-3,79229139
15,7912745	-3,72692104	-3,72692104
15,6846081	-3,56180729	-3,56180729
15,5755092	-3,4011544	-3,4011544
15,5662766	-3,38791534	-3,38791534
15,4727321	-3,25672349	-3,25672349
15,4327107	-3,20216805	-3,20216805
15,4096661	-3,17116381	-3,17116381
15,2901834	-3,01497357	-3,01497357
15,22897	-2,93776124	-2,93776124
15,1464097	-2,83643614	-2,83643614
15,1361207	-2,82402639	-2,82402639
15,0293848	-2,69800137	-2,69800137
15,0069431	-2,67211252	-2,67211252
14,8589698	-2,50636899	-2,50636899
14,8510633	-2,49774561	-2,49774561
14,7896378	-2,43151676	-2,43151676

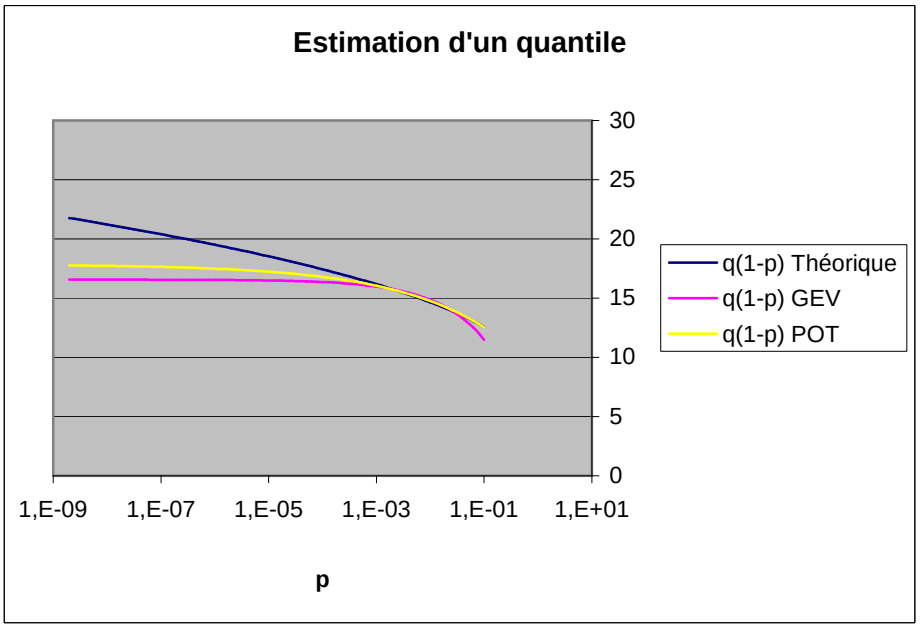


Feuille de calcul Microsoft Excel

Dans cet exemple, l'estimation de la probabilité de dépassement apparaît sensiblement inférieure à la valeur théorique.

2.4 Comparaison des résultats

Le graphique suivant donne les quantiles correspondant à une observation estimés par les différentes méthodes à partir d'un échantillon.



Feuille de calcul Microsoft Excel

Les estimations par la méthode GEV sont très sensibles au jeu de données simulées (ξ positive ou négative).

Conclusions :

- La Théorie des Valeurs Extrêmes permet d'effectuer des interpolations dans la queue des distributions expérimentales mais les extrapolations apparaissent discutables. Aussi doit-on se méfier des estimations effectuées pour des valeurs non encore observées, notamment quand elles concernent des problématiques relatives à la sécurité des personnes et des biens, même si des intervalles de confiance peuvent être calculés par inversion de la matrice de Fisher[1].

- L'ajustement par la méthode du maximum de vraisemblance de la loi généralisée des extrêmes (GEV) et de la loi généralisée de Pareto (GPD) peut s'effectuer de manière très efficace et précise au moyen d'un outil d'optimisation globale pouvant s'affranchir des divers optima locaux (GENCAB dans ces exemples). Diverses méthodes approchées d'ajustement sont proposées dans la littérature scientifique.

Bibliographie :

Balkema A., de Haan L. (1974) - Residual life time at great age - The Annals for Probability, vol. 2, n°5, pp. 792-804

Fisher R.A., Tippett L.H.C. (1928) - Limiting forms of the frequency distribution of the largest or smallest member of a sample - Proc. Cambridge Philos. Soc., 24, pp.180-190

Fréchet M. (1927) - Sur la loi de probabilité de l'écart maximum - Ann. Soc. Math. Polon., vol. 6, pp. 93-116

Gnedenko B.V. (1943) - Sur la distribution limite du terme maximum d'une série aléatoire - Ann. Math., 44, pp. 423-453

Gumbel E.J. (1955) - Statistical theory of extremes values and some practical applications - Journal of the Royal Statistical Society, Serie A, vol. 119, n.1, p. 106

Jenkinson A.F (1955) - The frequency distribution of the annual maximum (or minimum) of meteorological elements - Quart. J. R. Met. Soc. 81, pp.158-171

Mises, R., von (1954). La distribution de la plus grande de n valeurs. Selected papers, Vol. II, p. 271-294. Providence, R. I.: Amer. Math. Soc.

Pickands J. (1975) - Statistical inference using extreme order statistics - Ann. Statist.3, 119-131.

[1] Devrait faire l'objet d'un prochain TP.