Encadrement de l'erreur asymptotique d'estimation des quantiles extrêmes

Clément ALBERT [1], Anne DUTFOY [2] & Stéphane GIRARD [1]

[1] *Equipe MISTIS, Inria Grenoble Rhône-Alpes & Laboratoire Jean Kuntzmann, 655, avenue de l'Europe, Montbonnot, 38334 Saint-Ismier cedex, France.*
[2] *EDF R&D MRI, 1, avenue du Général de Gaulle - 92141 Clamart*
clement.albert@inria.fr, anne.dutfoy@edf.fr, stephane.girard@inria.fr

Résumé. La maîtrise des risques environnementaux est un sujet de préoccupation majeur au sein d'EDF. Dans ce cadre, une méthodologie d'analyse des valeurs extrêmes consiste, à partir d'une loi de valeurs extrêmes ajustée sur des données, à déterminer les quantiles extrêmes de période de retour centenale, millénale voire décamillénale. Ces quantiles extrêmes dépendent du modèle de valeurs extrêmes utilisé ainsi que du nombre de données disponibles. Dans cette communication, nous exposons les expressions des erreurs asymptotiques relatives des quantiles extrêmes issus d'une loi des valeurs extrêmes à trois paramètres. Des éléments de réponses sont ensuite donnés quant à l'encadrement de ces erreurs dans le cas où le paramètre de forme est proche de zéro.

Mots-clés. Théorie des valeurs extrêmes, quantiles extrêmes, normalité asymptotique, risques environnementaux.

Abstract. Risk management is a major concern at EDF. In this context, an analysis methodology of extreme values consists in estimating extreme quantiles - one hundred years return period or more - from an extreme-value distribution adjusted on data. These extreme quantiles depend on the extreme-value model used and on the number of available data. In this communication, we present the expressions of extreme quantile asymptotic errors considering a three parameter extreme-value distribution. Some bounds are provided on these errors when the shape parameter is close to zero.

Keywords. Extreme-value theory, extreme quantiles, asymptotic normality, environmental risks.

1 Introduction

La maîtrise des risques est un sujet de préoccupation majeur au sein d'EDF, et ce aussi bien en hydrologie/météorologie que dans le domaine du nucléaire. Dans ce cadre, la R&D d'EDF utilise la théorie des valeurs extrêmes pour effectuer de nombreuses études statistiques d'évènements extrêmes à partir de relevés de variables météorologiques (température, débit, vitesse de vent, ...). Ces études servent à dimensionner les ouvrages

EDF aux agressions météorologiques, de type inondation, tempête ou encore sécheresse. Elles consistent, à partir d'une loi de valeurs extrêmes ajustée sur des données, à déterminer les quantiles extrêmes de période de retour centenale, millénale voire décamillénale. Ces quantiles extrêmes dépendent du modèle de valeurs extrêmes utilisé ainsi que du nombre de données disponibles pour estimer les paramètres dudit modèle (Renard et al (2013)). Il est donc important de pouvoir quantifier ces sensibilités afin de rendre plus robustes les prises de décision. L'objectif de notre travail est d'étudier les limites d'extrapolation des lois de valeurs extrêmes en proposant des résultats sur la robustesse de ces extrapolations. Ce travail consiste également à comprendre pourquoi, dans la littérature (cf MeteoFrance (2014)), on trouve des références qui préconisent de ne pas extrapoler plus de quatre fois la durée d'observation des données.

Les travaux préliminaires que nous présentons se placent dans le cadre où les données sont issues d'une loi généralisée des valeurs extrêmes (GEV) à trois paramètres (μ, σ, ξ), respectivement de position, d'échelle et de forme.

Dans un premier temps, nous exposons les expressions des erreurs asymptotiques relatives des quantiles extrêmes issus d'une loi GEV, et ce lorsque le paramètre de forme ξ est supposé égal à zéro (cas d'une loi de Gumbel) puis lorsque le paramètre de forme ξ est proche de 0. Nous en profitons pour faire une comparaison de ces différentes erreurs relatives.

Dans un second temps, nous encadrons l'erreur asymptotique relative associée à une loi de Gumbel. Pour un jeu de données de taille n, cet encadrement va nous permettre, moyennant une erreur seuil donnée ϵ_0, de pouvoir déterminer quel est l'ordre du quantile jusqu'où il est possible d'extrapoler garantissant une erreur inférieure à ϵ_0.

2 Comparaison des approches GEV et Gumbel

Soit X une variable aléatoire de loi GEV, de fonction de répartition définie sur l'ensemble $\left\{x : 1 + \xi\frac{x-\mu}{\sigma} > 0\right\}$ par :

$$\begin{aligned} G_{\mu,\sigma,\xi}(x) &= \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \quad si\ \xi \neq 0 \\ G_{\mu,\sigma}(x) &= \exp\left\{-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right\} \qquad si\ \xi = 0. \end{aligned}$$

Le quantile x_q d'ordre $1-q$ est défini par $G_{\mu,\sigma,\xi}(x_q) = 1-q$ (resp $G_{\mu,\sigma}(x_q) = 1-q$). Il est donné par :

$$x_q = \mu - \frac{\sigma}{\xi}[1 - y_q^{-\xi}] \quad si\ \xi \neq 0 \tag{1}$$

$$x_q = \mu - \sigma \log y_q \qquad si\ \xi = 0 \tag{2}$$

avec $y_q = -\log(1-q)$. Par la suite, on supposera que q est proche de 0 et on dira que les quantiles associés sont des quantiles extrêmes.

Sous l'hypothèse que $X_1, ..., X_n$ sont des variables iid de loi GEV et que $1+\xi\left(\frac{x_i-\mu}{\sigma}\right) > 0$ pour $i = 1, ..., n$, la log-vraisemblance en les paramètres (μ, σ, ξ) s'écrit :

$$\ell(\mu,\sigma,\xi) = -n\log(\sigma) - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{n}\log\left(1+\xi\frac{x_i-\mu}{\sigma}\right) - \sum_{i=1}^{n}\left(1+\xi\frac{x_i-\mu}{\sigma}\right)^{-\frac{1}{\xi}} \quad (3)$$

De même, si $X_1, ..., X_n$ sont des variables iid de loi de Gumbel, on suppose $\xi = 0$ et la log-vraisemblance en les paramètres (μ, σ) s'écrit :

$$\ell(\mu,\sigma) = -n\log(\sigma) - \sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right) - \sum_{i=1}^{n}\exp\left(-\frac{x_i-\mu}{\sigma}\right). \quad (4)$$

Les estimateurs du maximum de vraisemblance $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ sont obtenus en maximisant les équations (3) ou (4) en les paramètres (μ, σ, ξ). Puis, en substituant dans les équations (1) et (2), on obtient l'estimateur par maximum de vraisemblance du quantile d'ordre $1-q$: $\hat{x}_q$. Sous la condition $\xi > -0.5$ (Smith (1985)), la loi asymptotique de $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ est une loi normale multivariée de moyenne (μ, σ, ξ) et de matrice de covariance l'inverse de la matrice d'information de Fisher.

Des intervalles de confiance sur le quantile extrême x_q avec probabilité $1-\alpha$ se déduisent à l'aide d'une delta-méthode (Coles (2001)) :

$$x_q \in \left[\hat{x}_q - u_\alpha\frac{\sigma_{\hat{x}_q}}{\sqrt{n}}; \hat{x}_q + u_\alpha\frac{\sigma_{\hat{x}_q}}{\sqrt{n}}\right],$$

où u_α est le quantile d'ordre $1-\frac{\alpha}{2}$ d'une loi normale centrée réduite, $\sigma_{\hat{x}_q} = \sqrt{Var(\hat{x}_q)}$, avec :

$$Var(\hat{x}_q) = \hat{\nabla}^t_{x_q}\hat{\Sigma}\hat{\nabla}_{x_q},$$

Σ est la matrice de covariance asymptotique calculée pour $n = 1$ de $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ et

$$\begin{aligned}
\nabla^t_{x_q} &= \left[\frac{\partial x_q}{\partial\mu}, \frac{\partial x_q}{\partial\sigma}, \frac{\partial x_q}{\partial\xi}\right] \\
&= \left[1, -\frac{1}{\xi}\left[1-y_q^{-\xi}\right], \frac{\sigma}{\xi^2}\left[1-y_q^{-\xi}\right] - \frac{\sigma}{\xi}y_q^{-\xi}\log(y_q)\right] \quad si\ \xi \neq 0 \\
&= [1, -\log y_q] \quad si\ \xi = 0.
\end{aligned}$$

L'erreur asymptotique relative est alors définie par :

$$\epsilon := \frac{u_\alpha\sigma_{\hat{x}_q}}{\hat{x}_q\sqrt{n}}.$$

Notre premier résultat établit les expressions analytiques de ces erreurs dans les cas GEV et Gumbel.

Lemme 1 *L'expression de l'erreur relative asymptotique associée à une modélisation de type Gumbel est donnée par :*

$$\begin{aligned}\epsilon_{Gum}\left(q,n,\frac{\hat{\mu}}{\hat{\sigma}}\right) &= \frac{u_\alpha\sqrt{P_2(\log y_q)}}{\sqrt{n}\left(\frac{\hat{\mu}}{\hat{\sigma}}-\log y_q\right)},\\ P_2(t) &:= \frac{1}{\pi^2}\left[\pi^2+6\left(1-\gamma-t\right)^2\right].\end{aligned}$$

Celle associée à une modélisation de type GEV lorsque $\xi \longrightarrow 0$ est donnée par :

$$\epsilon_{GEV}\left(q,n,\frac{\hat{\mu}}{\hat{\sigma}}\right) \overset{\xi\to 0}{\sim} \frac{u_\alpha\sqrt{P_4(\log y_q)}}{\sqrt{n}\left(\frac{\hat{\mu}}{\hat{\sigma}}-\log y_q\right)},$$

$$\begin{aligned}P_4(t) &:= \frac{3}{2}\left\{t^4\left[60\pi^2\right]\right.\\ &+ 240t^3\left[6\zeta(3)+\pi^2(\gamma-1)\right]\\ &+ 24t^2\left[\pi^4+5\pi^2(3\gamma^2-6\gamma+4)+180\zeta(3)(\gamma-1)\right]\\ &+ 48t\left[\pi^4(\gamma-1)+5\pi^2(\gamma^3-3\gamma^2+4\gamma-2-\zeta(3))+30\zeta(3)(3\gamma^2-6\gamma+4)\right]\\ &+ 9\pi^6+4\pi^4(6\gamma^2-12\gamma+1)+60\pi^2\left(\gamma^4-4\gamma^3+8\gamma^2-4\gamma(\zeta(3)+2)+4(\zeta(3)+1)\right)\\ &+ \left.1440\zeta(3)\left(\gamma^3-3\gamma^2+4\gamma-(\zeta(3)+2)\right)\right\}\\ &/ \left(11\pi^6-2160\zeta(3)^2\right).\end{aligned}$$

avec $\gamma \approx 0.577$ la constante d'Euler et $\zeta(3) \approx 1.202$ la constante d'Apéry.

Ces deux expressions nous permettent de remarquer Figure 1 que l'erreur relative asymptotique associée à la loi de Gumbel croît en $\log^2 y_q$ alors que celle associée à la loi GEV - qui prend en compte la variabilité du paramètre ξ contrairement à la loi de Gumbel - croît en $\log^4 y_q$. Ainsi, dès lors que l'on va considérer des quantiles extrêmes, l'erreur relative asymptotique associée à la loi de Gumbel sera inférieure à celle associée à la loi GEV quand le paramètre de forme ξ est proche de 0.

Par ailleurs, ces deux expressions ne dépendent plus des paramètres de leur loi respective qu'au travers du rapport $\hat{\mu}/\hat{\sigma}$, $P_2(\cdot)$ et $P_4(\cdot)$ étant indépendants de μ, σ et ξ.

3 Contrôle de l'erreur asymptotique

Nous nous intéressons aux erreurs relatives données par le Lemme 1. L'idée va être de donner un encadrement de ces erreurs en q, pour une taille de jeu de données n fixée. On pourra ainsi justifier le fait de ne pas extrapoler plus loin que le quantile d'ordre $1-q$ par

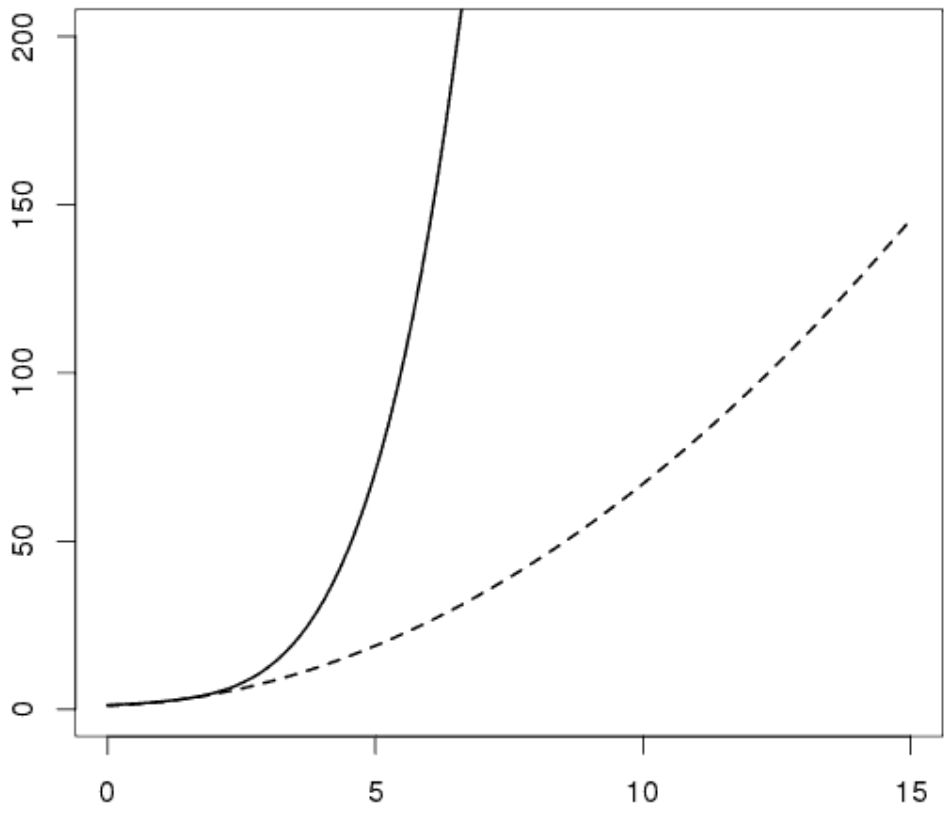

FIGURE 1 – Graphes de $P_4(-t)$ (ligne continue) et de $P_2(-t)$ (tirets) pour $t \in [0, 15]$.

le fait que l'erreur relative asymptotique associée est supérieure à une erreur seuil donnée ϵ_0.

Le but est alors de trouver $q_0\left(n, \frac{\hat{\mu}}{\hat{\sigma}}\right)$ tel que $\epsilon\left(q_0\left(n, \frac{\hat{\mu}}{\hat{\sigma}}\right), n, \frac{\hat{\mu}}{\hat{\sigma}}\right) = \epsilon_0$. Par construction de $q_0\left(n, \frac{\hat{\mu}}{\hat{\sigma}}\right)$, on aura alors que :

$$\forall\ 1-q \leq 1-q_0\left(n, \frac{\hat{\mu}}{\hat{\sigma}}\right), \epsilon\left(q, n, \frac{\hat{\mu}}{\hat{\sigma}}\right) \leq \epsilon_0.$$

Déterminer $q_0(n, \frac{\hat{\mu}}{\hat{\sigma}})$ nécessite alors d'étudier les fonctions :

$$\begin{aligned} t \to \Phi_\beta^{(2)}(t) &:= \frac{u_\alpha \sqrt{P_2(-t)}}{\sqrt{n}\,(\beta + t)}, \\ t \to \Phi_\beta^{(4)}(t) &:= \frac{u_\alpha \sqrt{P_4(-t)}}{\sqrt{n}\,(\beta + t)}, \end{aligned}$$

sur $\mathbb{R}^+$, avec $\beta > 0$.

4 Perspectives

Après avoir illustré les encadrements obtenus sur des simulations, on s'appliquera à chercher des encadrements similaires pour une modélisation de type GEV lorsque le paramètre de forme ξ est quelconque. Ces méthodes pourront ensuite être appliquées à des cas réels de mesures de variables environnementales dont EDF possède les séries chronologiques : chroniques de débit, de température, de vitesse instantanée de vent.

Références

[1] B.Renard, K.Kochanek, M.Lang, F.Garavaglia, E.Paquet, L.Neppel, K.Najib, J.Carreau, P.Arnaud, Y.Aubert, F.Borchi, J.M.Soubeyroux, S.Jourdain, J.M.Veysseire, E.Sauquet, T.Cipriani et A.Auffray. Data-based comparison of frequency analysis methods : A general framework. *Water Resources Research*, **49(2)**, 825-843, 2013.

[2] S.Coles. *An introduction to statistical modeling of extreme values*. Springer, 2001.

[3] MétéoFrance, Direction de la Climatologie. Durées de retour de précipitations extrêmes. `http://pluiesextremes.meteo.fr/media/doc/Cartes_reseau/Fiche_methode_durees_retour.pdf`, 2014.