Control of the estimation error on extreme quantiles.

Clément ALBERT

Supervisors : S. Girard (INRIA, LJK), A. Dutfoy (EDF)

1^{st} year-PhD
May 2016

Outline

Outline

Risks Management : Which methodology ?

Roselend dam

- Risk management is a major concern at EDF
- Use of the extreme value theory (EVT) to perform many statistical studies of extreme events from weather variables statements
- These studies are used to size the EDF works to weather attacks

Nuclear power plant of Nogent

- These studies mainly consist in identifying extreme quantile of 100-year return period or more (Renard et al 2013)
- These extrapolations depend on the extreme-value model used and on the number of available data

Goal : Provide tools to assess the veracity of extrapolations using extreme value theory.

Outline

Study of the maximum : The GEV distribution

Theorem (Extremal Types Theorem)

Let $X_1, X_2, ..., X_n$ be a sample of iid random variables and $M_n = max(X_1, X_2, ..., X_n)$. Suppose there exist sequences $\{a_n > 0\}$ and $\{b_n\}$ such as :

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) \to G(x) \text{ when } n \to \infty,$$

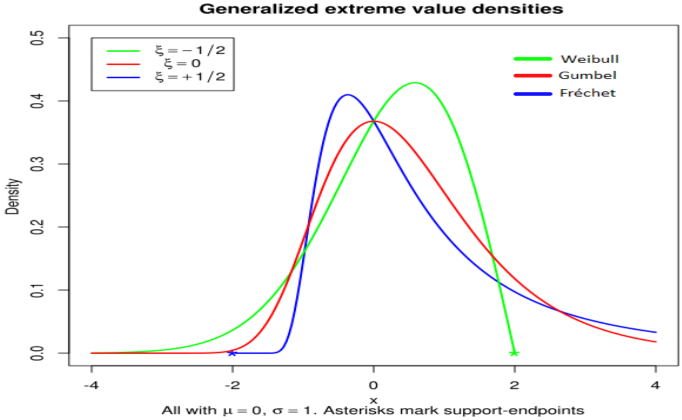
with G a non-degenerated distribution. Then,

$$G(x) = \exp\left\{-\left[1 + \xi\frac{x-\mu}{\sigma}\right]_+^{-\frac{1}{\xi}}\right\},$$

with $\mu \in \mathbb{R}$ (position), $\sigma > 0$ (scale) and $\xi \in \mathbb{R}$ (shape). G is called the generalized extreme value distribution (GEV).

Study of the maximum : The GEV distribution

$G(x)$ can be classified into three types, depending on the value of the extreme value index ξ : Weibull ($\xi < 0$), Fréchet ($\xi > 0$) and Gumbel ($\xi = 0$).



In the following, we distinguish between $\xi \neq 0$ and $\xi = 0$.

Expressions of G

Let X be a random variable distributed according to a GEV distribution. Provided that $\left\{x : 1+\xi\frac{x-\mu}{\sigma} > 0\right\}$, the distribution function of X is :

$$G_{\mu,\sigma,\xi}(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \quad \textit{if } \xi \neq 0$$

$$G_{\mu,\sigma}(x) = \exp\left\{-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right\} \quad \textit{if } \xi = 0.$$

The $1-q$ extreme quantile of the GEV distribution, x_q, is then given by :

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

$$x_q = \mu - \sigma \log y_q \quad \textit{if } \xi = 0$$

with $y_q = -\log(1-q)$ and q small ($q \in [0, 1-e^{-1}]$).

Maximum likelihood estimators

Let $x_1, ..., x_n$ be an iid sample from a GEV distribution. The maximum likelihood (ML) estimators of μ, σ and ξ are obtained by maximizing the log-likelihood ℓ given (respectively for $\xi \neq 0$ and $\xi = 0$) by :

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

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

The ML estimators are :

- not explicit;
- asymptotically Gaussian under the condition $\xi > -0.5$ (Smith 1985).

Confidence interval for x_q

Using the Delta method and Slutsky lemma lead to (Coles 2001) :

$$\frac{1}{\sigma_{\hat{x}_q}}\sqrt{n}(\hat{x}_q - x_q) \xrightarrow{d} N(0,1)$$

which permit to deduce an $1-\alpha$ confidence interval for x_q :

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

where $\hat{x_q}$ is obtained by plug-in of the ML estimators, u_α is the $1-\alpha$ quantile of the standard normal distribution and $\sigma^2_{\hat{x}_q} = \hat{\nabla}^t_{x_q}\hat{\Sigma}\hat{\nabla}_{x_q}$, with Σ the asymptotic covariance matrix of $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ and

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

Outline

Estimation error

We measure the variability of x_q due to the estimation of μ and σ thanks to the quantity $\epsilon\left(q, n, \hat{\mu}, \hat{\sigma}, \hat{\xi}\right)$ such that :

$$\mathbb{P}\left(\left|\frac{x_q}{\hat{x}_q} - 1\right| \leq \epsilon\left(q, n, \hat{\mu}, \hat{\sigma}, \hat{\xi}\right)\right) \rightarrow 1 - \alpha \textit{ when } n \rightarrow +\infty,$$

where the estimation error is given by :

$$\epsilon\left(q, n, \hat{\mu}, \hat{\sigma}, \hat{\xi}\right) = \frac{u_\alpha \sigma_{\hat{x}_q}}{\hat{x}_q \sqrt{n}}$$

In what follows, we give expressions of the estimation error associated with the extreme quantile of the GEV x_q when $\xi = 0$ and ξ is near zero. Then, we give bounds on the estimation error in the Gumbel case.

Estimation error - Gumbel case

The estimation error associated with a Gumbel distribution is given by :

$$\epsilon_{Gum}\left(q, n, \frac{\hat{\mu}}{\hat{\sigma}}\right) = \frac{u_\alpha \sqrt{P_2(\log y_q)}}{\sqrt{n}\left(\frac{\hat{\mu}}{\hat{\sigma}} - \log y_q\right)}$$

$$P_2(t) := \frac{1}{\pi^2}\left[\pi^2 + 6\left(1 - \gamma - t\right)^2\right].$$

with $\gamma \approx 0.577$ the Euler-constant and $y_q = -\log(1-q)$. See also Cunnane (1973).

$\rightsquigarrow$ It depends on the parameters of the distribution only through the ratio $\hat{\mu}/\hat{\sigma}$.

Estimation error - GEV case, $\xi \to 0$

We show that the estimation error associated with a GEV distribution with $\xi \longrightarrow 0$ is given by

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

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

with $\zeta(3) \approx 1.202$ the Apéry-constant.

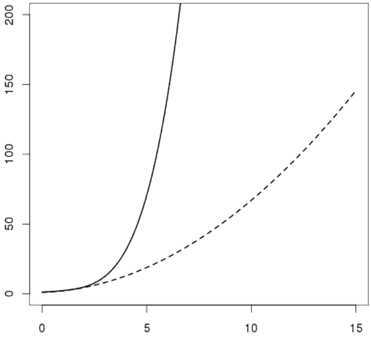
Comparison of the two previous estimation errors

Both expressions :

$$\epsilon_{Gum}\left(q, n, \frac{\hat{\mu}}{\hat{\sigma}}\right) = \frac{u_\alpha \sqrt{P_2(\log y_q)}}{\sqrt{n}\left(\frac{\hat{\mu}}{\hat{\sigma}} - \log y_q\right)}$$

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

- depend on the parameters of their distribution only through the ratio $\hat{\mu}/\hat{\sigma}$.



Graphs of $P_4(-t)$ (solid line) and $P_2(-t)$ (dashed line) for $t \in [0, 15]$.

- have their numerator increasing with respect to $-\log(y_q)$, to the power 2 in the Gumbel case and to the power 4 in the GEV case.

We obtain uniform bounds on the estimation error associated with a Gumbel distribution for all $q \in [0, 1-e^{-1}]$:

Proposition 1

Let $\beta := \hat{\mu}/\hat{\sigma}$ *and* $q \in [0, 1-e^{-1}]$. *Then :*

$$\begin{aligned}
\text{if} \quad 0 < \beta < \beta_1 \quad &, \quad \epsilon_{Gum}(q, n, \beta) \in [\epsilon_2(n), \epsilon_3(\beta, n)]; \\
\text{if} \quad \beta_1 < \beta < \beta_2 \quad &, \quad \epsilon_{Gum}(q, n, \beta) \in [\epsilon_1(\beta, n), \epsilon_3(\beta, n)]; \\
\text{if} \quad \beta_2 < \beta < \beta_3 \quad &, \quad \epsilon_{Gum}(q, n, \beta) \in [\epsilon_1(\beta, n), \epsilon_2(n)]; \\
\text{if} \quad \beta > \beta_3 \quad &, \quad \epsilon_{Gum}(q, n, \beta) \in [\epsilon_3(\beta, n), \epsilon_2(n)],
\end{aligned}$$

... with 3 universal constants :

$$\beta_1 := (1-\gamma) \approx 0.42,$$

$$\beta_2 := \sqrt{\frac{\pi^2 + 6(1-\gamma)^2}{6}} \approx 1.35,$$

$$\beta_3 := \frac{6(1-\gamma)^2 + \pi^2}{6(1-\gamma)} \approx 4.31,$$

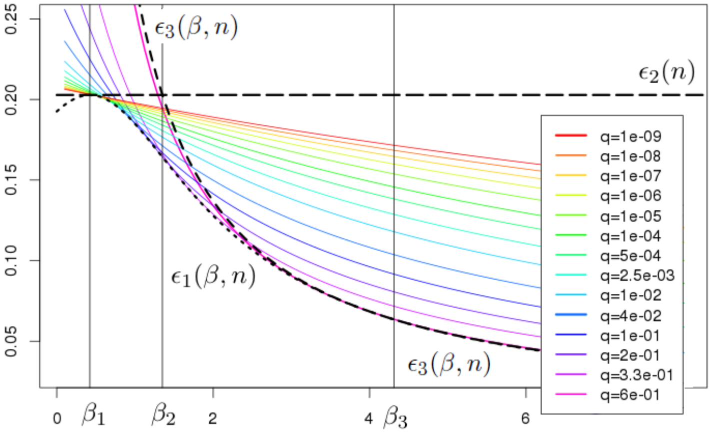
and the following 3 error functions :

$$\epsilon_1(\beta, n) := \sqrt{\frac{6u_\alpha^2}{n\left(\pi^2 + 6(\beta - (1-\gamma))^2\right)}},$$

$$\epsilon_2(n) := \sqrt{\frac{6u_\alpha^2}{n\pi^2}},$$

$$\epsilon_3(\beta, n) := \sqrt{\frac{u_\alpha^2(\pi^2 + 6(1-\gamma)^2)}{n\beta^2\pi^2}}.$$

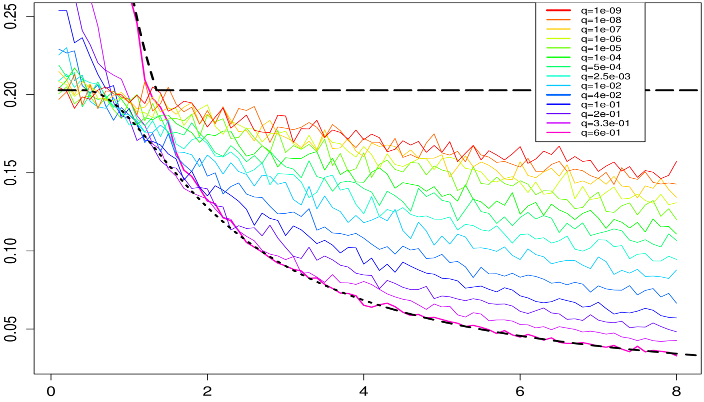
Illustration of $\epsilon_{Gum}(q, n, \beta)$ with $n = 40$ and $\alpha = 0.1$:



Control of the estimation error in the Gumbel case

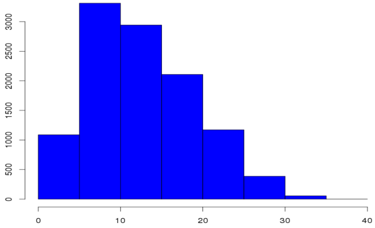
Simulated data

Empirical validation on 1000 replications, with $n = 40$, $\alpha = 0.1$.

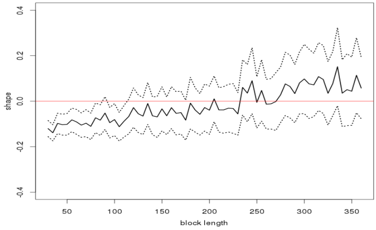


In practice : wind speed measures (Orange)

By choosing annual blocks, we obtain $n = 31$ maxima. By imposing $\xi = 0$, the ML estimators give $\hat{\mu} \approx 32$, $\hat{\sigma} \approx 1.45$ and then $\beta \approx 22$. Thus, Proposition 1 entails that, with probability 90% , regardless of the extrapolation level q, the estimation error is between 1.4% and 23%.



Histogram of data : N=11 077 daily wind speeds (in m/s) from 1981 to 2011.



ML estimations of ξ with block length.

Outline

Work in progress

- Research of bounds for the error in the case of a GEV distribution for any ξ (to begin with ξ near zero)

- Take into account the fact that extremes are not perfectly iid from a Gumbel/GEV distribution in practice (cf Ferreira and de Haan (2015))

  $\rightsquigarrow$ Additional approximation error

Bibliography

- **S.Coles (2001)**, *An introduction to statistical modeling of extreme values*, Springer.
- **A.Ferreira and L.de Haan (2015)**, On the block maxima method in extreme value theory : PWM estimators, *The Annals of Statistics*, volume 43, number 1, 276–298.
- **RL.Smith (1985)**, Maximum likelihood estimation in a class of nonregular cases, *Biometrika*, volume 72, number 1, 67–90.
- **B.Renard et al (2013)**, Data-based comparison of frequency analysis methods : A general framework, *Water Resources*, volume 49, doi:10.1002/wrcr.20087.