Extrapolation limits of extreme-value methods for return-levels estimation

Clément ALBERT

Anne Dutfoy (EDF), Stéphane Girard (INRIA)

3rd year-PhD
April 2018

Outline

Outline

Introduction

For safety authorities, major challenge is to identify return level of 100-year return period or more. However, what we have observed may be less exceptional than the event of interest. This requires to extrapolate outside the sample range.

AUTORITÉ DE SÛRETÉ NUCLÉAIRE

But, up to which order ? How far one should extrapolate ? In this talk, we quantify the extrapolation limits, using mathematical tools issued from the convergence analysis towards extreme value distributions.

Extreme quantile estimation

The study takes place in the context of extreme quantiles estimation.
Suppose $X_1, ..., X_n$ iid from a distribution F.
An extreme quantile of F is a $(1 - p_n)$th quantile $q(p_n)$ defined by

$$\begin{aligned} \overline{F}(q(p_n)) &= p_n, \\ p_n &\ll 1/n, \end{aligned}$$

where $\overline{F}$ is the survival distribution function.

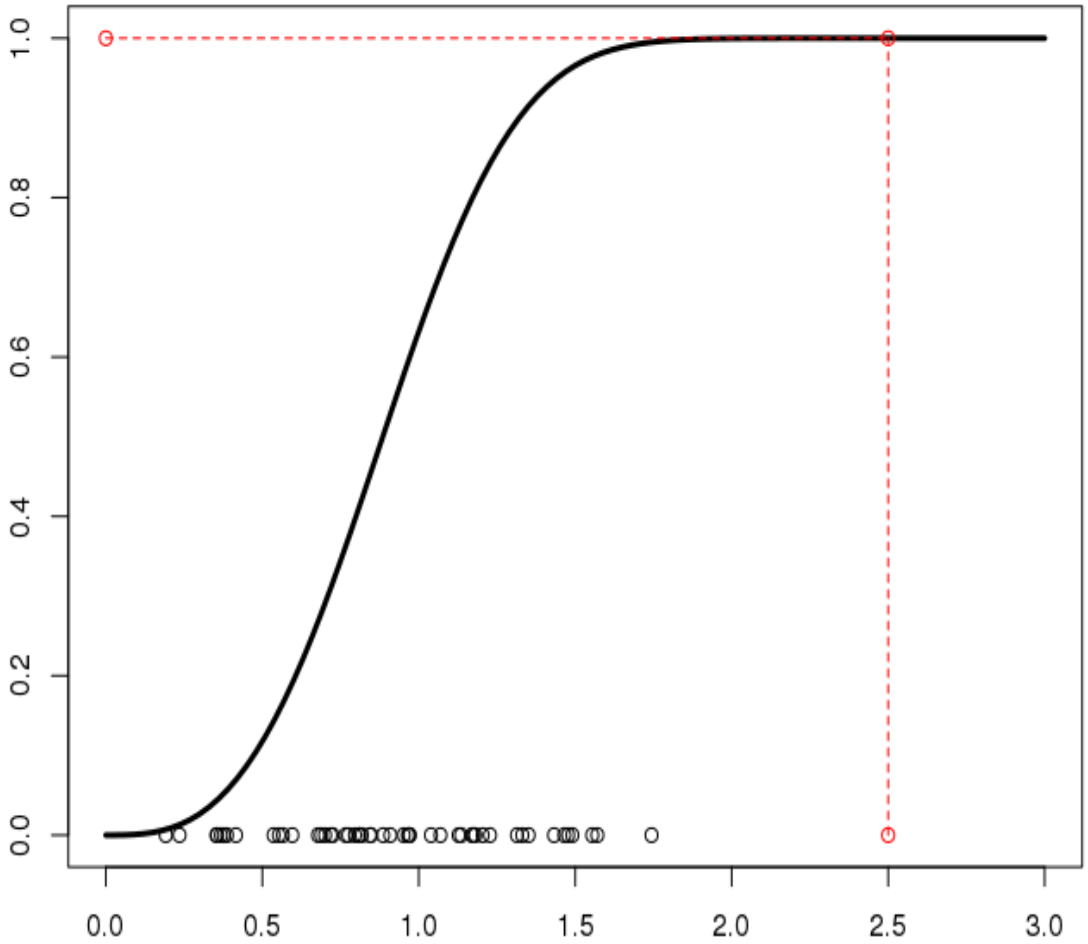


Figure: Extreme quantile estimation

Extreme quantile estimation

The excesses above u_n are defined as $Y_i = X_i - u_n$ for all $X_i > u_n$.

[Pickands, 1975] says that the distribution of excesses $\overline{F}_{u_n}$ can be approximated by a Generalized Pareto Distribution (GPD) :

$$\overline{F}_{u_n}(x) \approx \begin{cases} \left(1 + \frac{\gamma_n x}{\sigma_n}\right)^{-1/\gamma_n} & , \gamma_n \neq 0 \\ \exp\left(-\frac{x}{\sigma_n}\right) & , \gamma_n = 0 \end{cases}$$

where σ_n and γ_n are the scale and shape parameters of the GPD distribution.

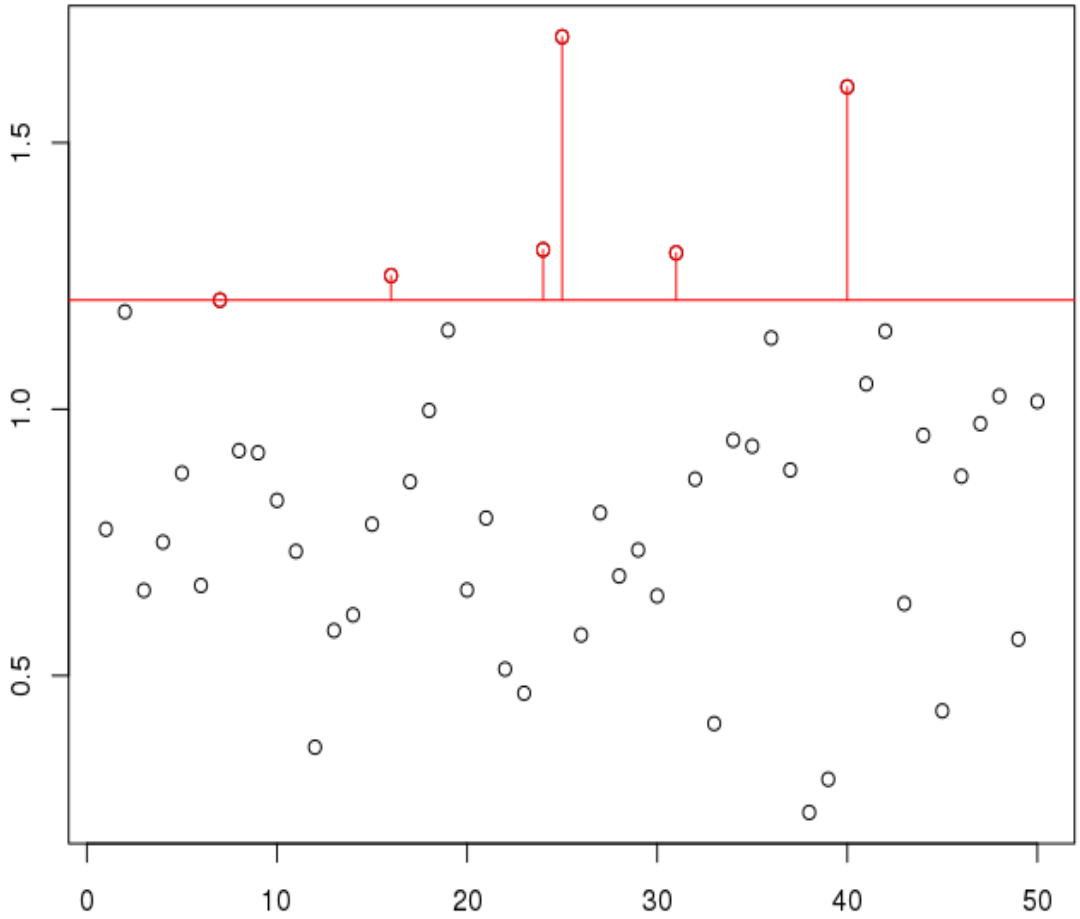


Figure: Definition of excesses

Extreme quantile estimation

Remark

$$\begin{aligned}\overline{F}_{u_n}(x) &= \mathbb{P}(Y \geq x | X \geq u_n), \\ &= \frac{\overline{F}(x+u_n)}{\overline{F}(u_n)}.\end{aligned}$$

so that

$$\overline{F}(x+u_n) = \overline{F}(u_n)\overline{F}_{u_n}(x)$$

Let $v_n = x + u_n$, with u_n a threshold such that $u_n = q(\alpha_n)$:

$$\overline{F}(v_n) \approx \begin{vmatrix} \alpha_n \left(1 + \gamma_n \frac{v_n - u_n}{\sigma_n}\right)^{-1/\gamma_n} \\ \alpha_n \exp\left(-\frac{v_n - u_n}{\sigma_n}\right) \end{vmatrix}$$

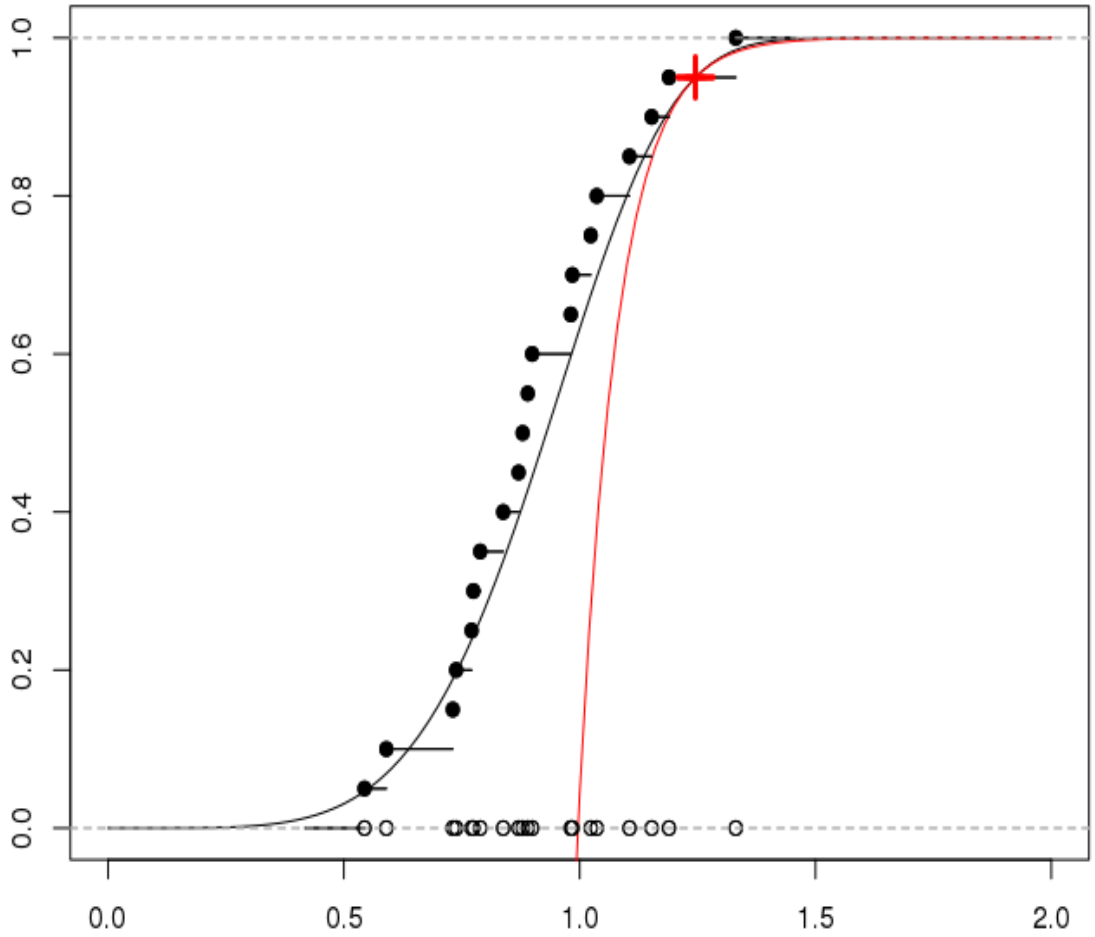


Figure: Tail approximation

Extreme quantile estimation

As a consequence, $q(p_n)$ can be in turn approximated by the deterministic term :

$$q(p_n) \approx \begin{cases} q(\alpha_n) + \dfrac{\sigma_n}{\gamma_n}\left[\left(\dfrac{\alpha_n}{p_n}\right)^{\gamma_n} - 1\right] \\ q(\alpha_n) + \sigma_n \log\left(\dfrac{\alpha_n}{p_n}\right) \end{cases}$$

In the following, we focus on the case $\gamma_n = 0$ ($F \in MDA(Gumbel)$) and we note this second approximation $\tilde{q}_{\mathrm{ET}}(p_n; \alpha_n)$.

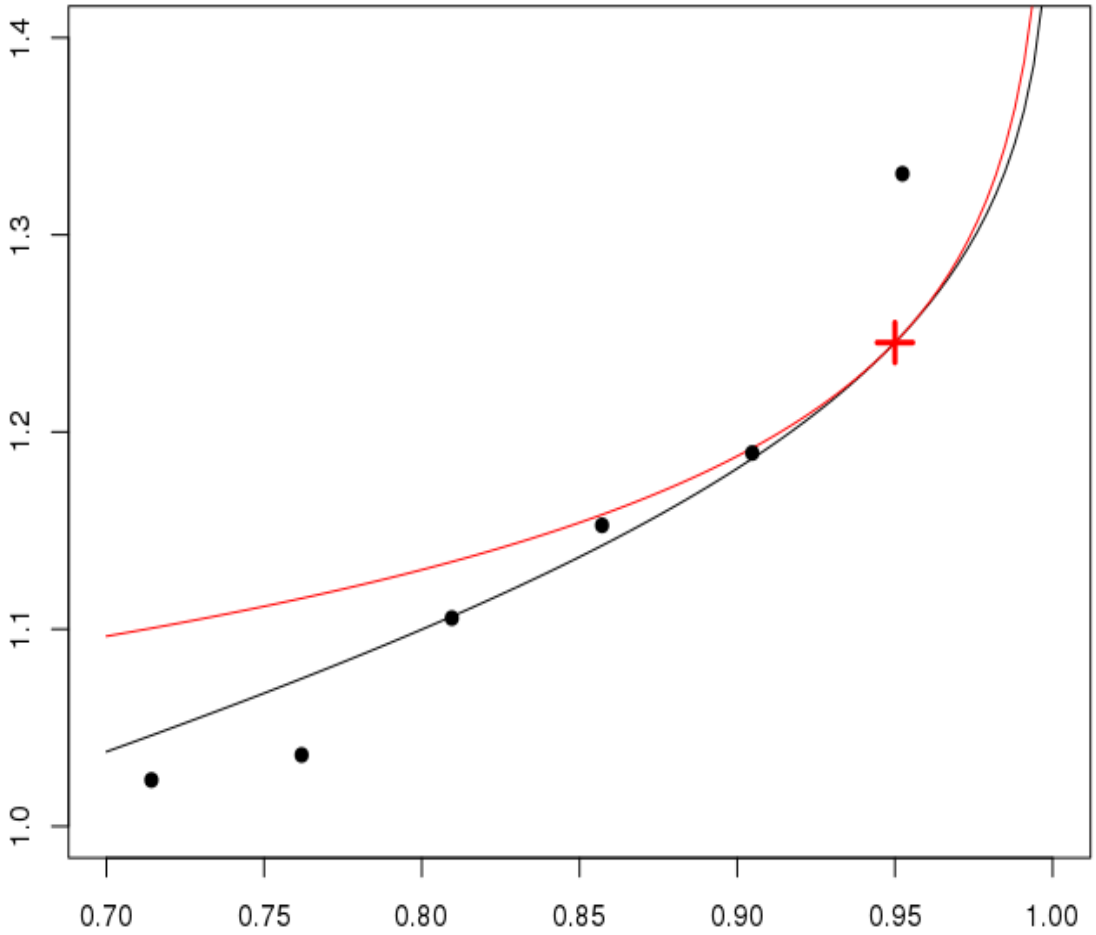


Figure: Quantile approximation

Extrapolation error

Let us define the relative extrapolation error :

$$\varepsilon_{\mathsf{ET}}(p_n;\alpha_n) := (q(p_n) - \tilde{q}_{\mathsf{ET}}(p_n;\alpha_n))/q(p_n).$$

Theorem (Necessary and sufficient conditions for $\varepsilon_{\mathsf{ET}}(p_n;\alpha_n) \to 0$)

As $n \to \infty$, one has, under some technical assumptions :

$$\varepsilon_{ET}(p_n;\alpha_n) \to 0 \quad \Longleftrightarrow \quad \delta_n^2 K_2(y_n) \to 0.$$

1. δ_n is a term which measures the proximity between p_n and α_n. The closer δ_n to zero, the closer p_n and α_n to $1/n$.
2. $K_2(y_n)$ is a term which is specific to the tail of F.

- [1] **Albert, C., Dutfoy, A., & Girard, S. (2018)**, *Asymptotic behavior of the extrapolation error associated with the estimation of extreme quantiles*, submitted, hal-01692544v2.

Example of distributions

Table: K_2, $\lim_{x \to +\infty} K_2(x)$.

Distributions	K_2	$\lim_{x \to +\infty} K_2(x)$
Exponential	0	0
Gamma $(a > 0)$	$\frac{1-a}{x}(1+o(1))$	0
Weibull $(\beta \neq 1)$	$\frac{1-\beta}{\beta^2}$	$\frac{1-\beta}{\beta^2}$
Gaussian	$-\frac{1}{4}+o(1)$	$-\frac{1}{4}$
Lognormal $(\sigma > 0)$	$\frac{\sigma^2}{2}x(1+o(1))$	$+\infty$
LogWeibull $(\beta > 1)$	$\frac{1}{\beta^2}x^{2/\beta}(1+o(1))$	$+\infty$

Hierarchy of distributions

The previous theorem exhibits a hierarchy of distributions :

1. First of all, one has the Exponential and Gamma distributions, for which $\lim_{x \to +\infty} K_2(x) = 0$.
   Extrapolation is not limited from an asymptotic point of view : the relative approximation error tends to zero when n tends to infinity (in the case of the exponential distribution, $K_2 = 0$ and $\varepsilon_{\mathsf{ET}}(p_n; \alpha_n) = 0$).

2. Then the Weibull($\beta \neq 1$), Gaussian distributions, for which $\lim_{x \to +\infty} K_2(x) = c \neq 0$.
   Extrapolation is limited.

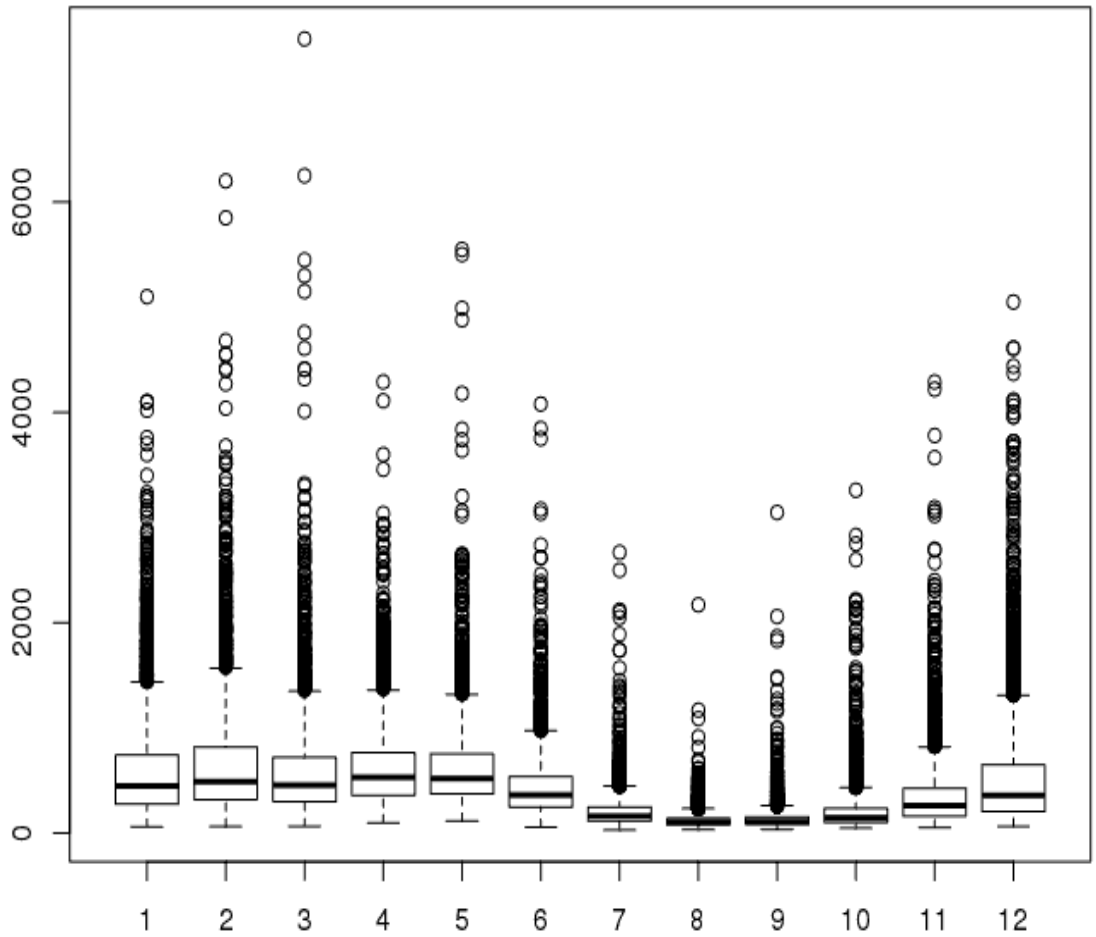
3. Finally, the Log-normal distribution, for which $\lim_{x \to +\infty} K_2(x) = +\infty$. Extrapolation is greatly limited.

Outline

The Data

Figure: Left figure : first rows of the dataset. Right figure : Boxplot representing the twelve months of the year (January to December).

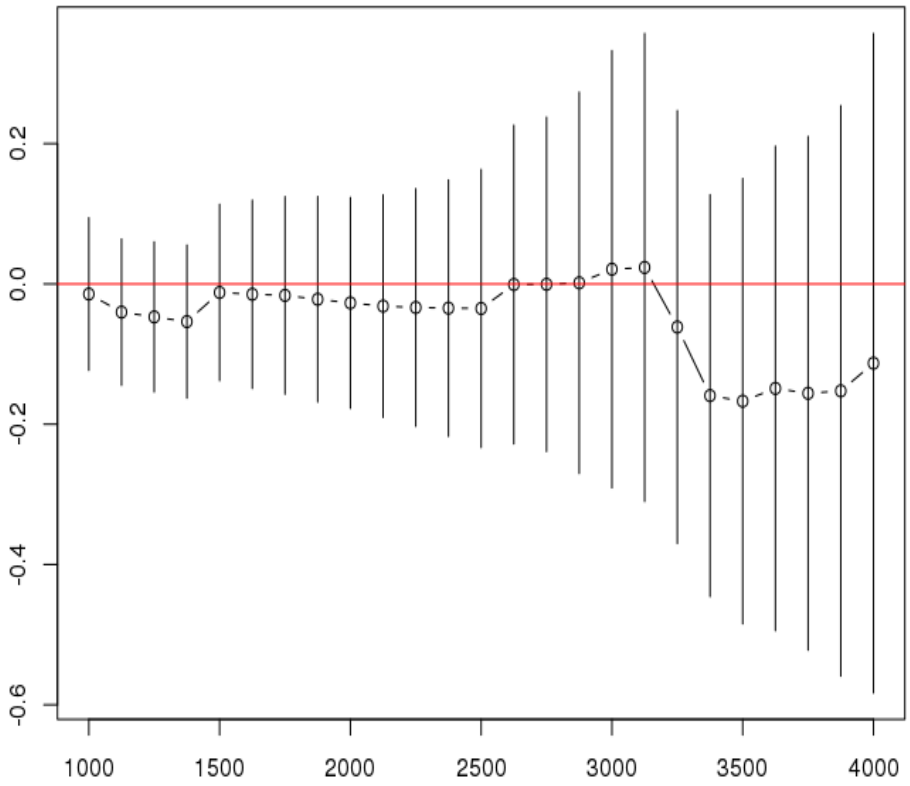
Date	Debit
1915-01-01	540
1915-01-02	865
1915-01-03	1140
1915-01-04	1330
1915-01-05	1750
1915-01-06	2310
1915-01-07	1920
1915-01-08	1470
1915-01-09	1230
1915-01-10	1560
1915-01-11	1830
1915-01-12	2570
1915-01-13	4020
1915-01-14	2700
1915-01-15	2260
1915-01-16	1720

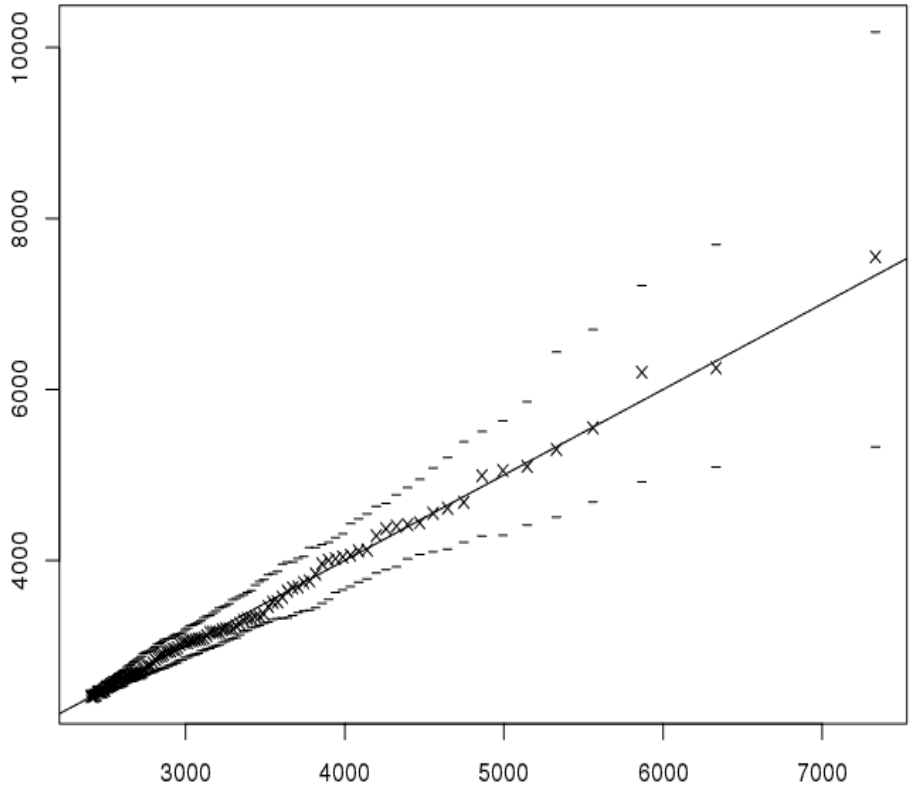


We consider daily river flow measures, in m^3/s of the Rhône from 1915 to 2013. Due to seasonality aspect, only flows from December 1 to May 31 are retained leading to $n = 18043$ measures.

Assumptions verification

Figure: Left figure : extreme value index estimation (maximum likelihood) as a function of the threshold. Right figure : Quantile plot after GPD fitting of the excesses when $u = 2400\text{m}^3/\text{s}$, $\gamma = 0$ and weekly clusters of exceedences are considered.

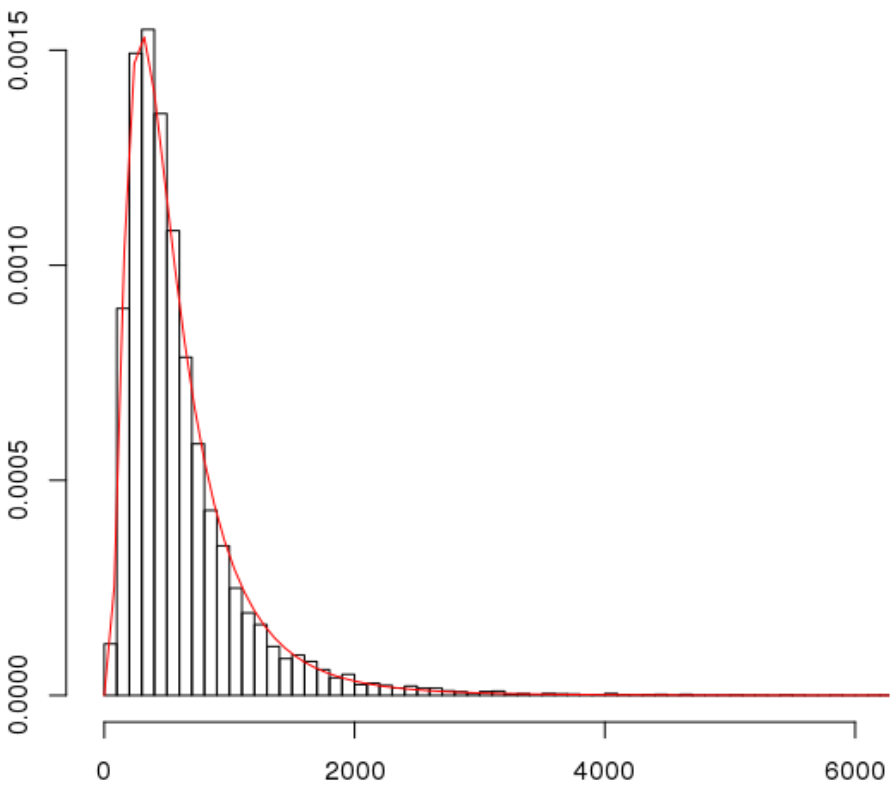


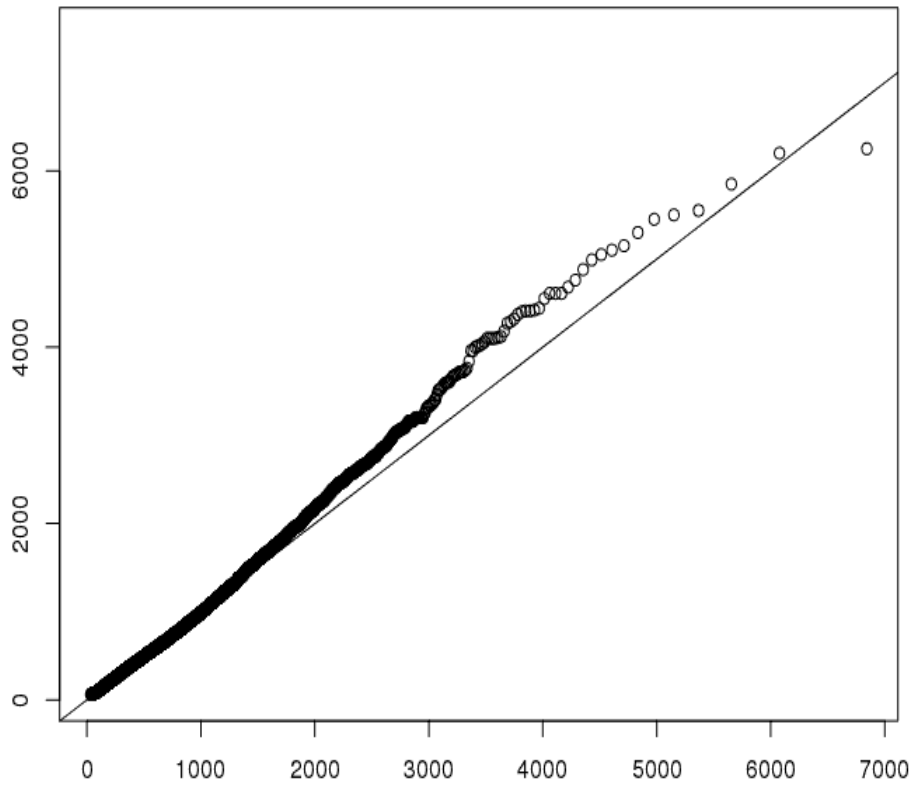


Former studies suggest to take $u = 2400\text{m}^3/\text{s}$. Using this threshold yields : $\hat{\gamma} = -0.0016 \pm 0.0995$ and $\hat{\sigma} = 919 \pm 127$. Supposing that $F \in MDA(Gumbel)$ seems thus reasonable. This is confirmed by both Figures.

Lognormal fitting

Figure: Left figure : histogram of the data and lognormal fitting. Right figure : empirical quantiles against lognormal one

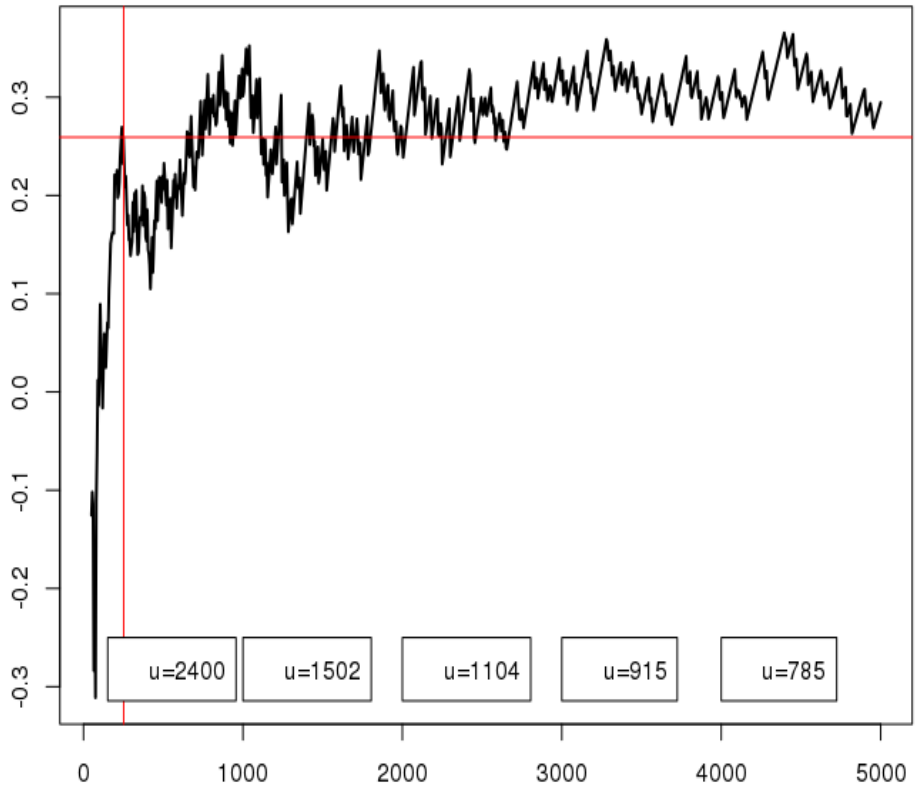


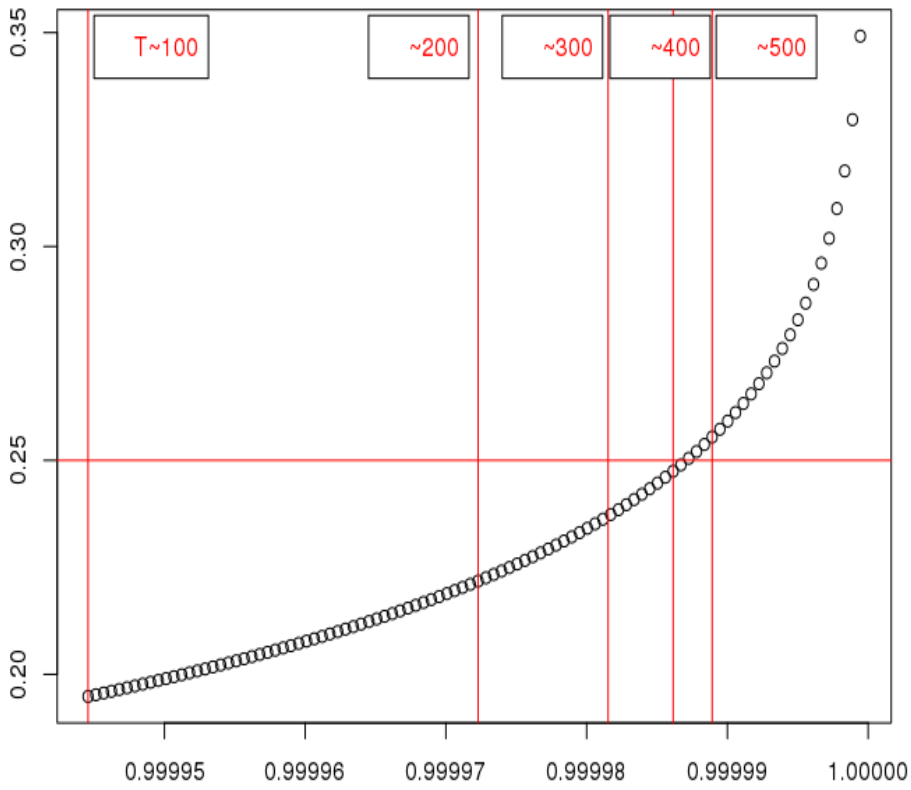


Note that observations seem issued from a lognormal distribution.

Estimation of the extrapolation error

Figure: Left figure : Estimation of $\varepsilon_{\mathsf{ET}}(p_n; \alpha_n)$ as a function of k_n (i.e the threshold), when $T = 400$ years. Right figure : Estimation of $\varepsilon_{\mathsf{ET}}(p_n; \alpha_n)$ as a function of p (i.e T the return period in years), when $u = 2400\mathrm{m}^3/\mathrm{s}$.





We can estimate the relative extrapolation error and indeed, extrapolation is greatly limited : Both figures suggests a 25% error for extrapolation beyond $T = 400$ years.

Conclusions

The approximation of $q(p_n)$ by $\tilde{q}_{\mathsf{ET}}(p_n; \alpha_n)$ has important consequences in the sense that the relative difference between the real quantile and its approximation can grow with n, even if $F \in MDA(Gumbel)$ ($\gamma_n = 0$).

In this talk, we have studied the extrapolation error associated with the previous approximation. We have concluded that the rate of convergence towards zero of the extrapolation error limits the way one extrapolates.

We showed that, the further F from the exponential distribution, the more stringent the extrapolation limits.