

# extremefit : package to estimate the conditional probabilities and extreme quantiles

Kévin Jaunâtre

based on a joint work with  
Gilles Durrieu, Ion Grama, Khoai-Quang Pham and Jean-Marie Tricot

Université de Bretagne Sud

Rencontres R, Toulouse, 2016



- 1 Formulation of the problem
- 2 Estimation of the parameters
- 3 Other extreme value theory packages
- 4 Example of utilisation of extremefit

# Formulation of the problem

Let be  $F_t(x) = P(X \leq x | T = t)$  a family of conditional distributions with  $t \in [0, T_{\max}]$ ,  $T_{\max} > 0$ .

- Assume that  $F_t \in \text{DA}$  of the Fréchet law with parameter  $1/\gamma_t$  :

$$\Phi_{1/\gamma_t}(x) = \exp\left(-x^{-1/\gamma_t}\right), \quad x \geq 0.$$

- We observe independent variables

$$X_{t_1}, \dots, X_{t_n} \text{ where } 0 \leq t_1 < \dots < t_n \leq T$$

and each  $X_{t_i}$  has the distribution function  $F_{t_i}$ .

## The goal :

- For a given large  $x$  to provide a pointwise estimate of the (extreme) tail probability process :  $S_t(x) = 1 - F_t(x)$ ,  $t \in [0, T_{\max}]$ .
- For a given  $p \in (0, 1)$  to provide a pointwise estimate of the (extreme)  $p$ -quantile process  $F_t^{-1}(p)$ ,  $t \in [0, T_{\max}]$ .

# Theoretical background

- The tail of  $F_t$  will be represented by its excess d.f. over the threshold  $\tau > 0$  :

$$F_{t,\tau}(x) = \mathbb{P}(X_t \leq x\tau | X_t > \tau) = 1 - \frac{1 - F_t(x\tau)}{1 - F_t(\tau)}, \quad x \geq 1.$$

- Consider the Pareto d.f. :  $P_\theta(x) = 1 - x^{-1/\theta}$ ,  $x \geq 1$ ,  $\theta > 0$ .

## Theorem (Fisher-Tippet-Gnedenko theorem)

- $F_t \in DA$  of the Fréchet law with parameter  $1/\gamma_t$  **iff** for any  $x \geq 1$ ,

$$F_{t,\tau}(x) \rightarrow P_{\gamma_t}(x) \quad \text{as } \tau \rightarrow \infty.$$

- Equivalent condition :  $1 - F_t(x) = x^{-1/\gamma_t} L_t(x)$ , with  $L_t(x)$  slowly varying perturbation.

F-T-G theorem suggests to approximate  $F_{t,\tau}(x) \approx P_{\gamma_t}(x)$  with estimated  $\gamma_t$  for large  $\tau$ .

# Our approach

Approaching  $F_t$  with a family of semi-parametric models :

$$F_{t,\tau,\theta}(x) = \begin{cases} F_t(x) & x \in [0, \tau], \\ 1 - (1 - F_t(\tau)) (1 - P_\theta(\frac{x}{\tau})) & x > \tau, \end{cases}$$

We fit the tail of  $F_t(x)$  by a Pareto model where :

- $\tau$  is the unknown threshold
- $F_t$  and  $\theta$  are unknown

## Pointwise estimation in $t$

For each time  $t$  we use obs. in the window  $[t - h, t + h]$  of size  $h$  :

- 1  $F_t(\cdot)$  is estimated by the weighted empirical distribution function, for  $x \in [0, \tau]$
- 2  $\theta$  is estimated by maximizing the weighted likelihood.
- 3  $\tau$  is replaced by the adaptive threshold  $\hat{\tau}$  explained later.

# Estimation of $\theta$

Given the time  $t$  and the bandwidth  $h$ , we define the weights :

$$0 \leq W_{t,h}(t) = K\left(\frac{t_i - t}{h}\right) \leq 1,$$

where  $K(\cdot)$  is a kernel function (for ex.  $K(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)$ ).

**The weighted quasi-log-likelihood function is :**

$$\mathcal{L}_{t,h}(\tau, \theta) = \sum_{i=1}^n W_{t,h}(t_i) \log \frac{dF_{t,\tau,\theta}}{dx}(X_{t_i}).$$

Maximum quasi-likelihood estimator with fixed threshold  $\tau$  :

$$\hat{\theta}_{t,h,\tau} = \frac{1}{\hat{n}_{t,h,\tau}} \sum_{i=1}^n W_{t,h}(t_i) 1_{\{X_{t_i} > \tau\}} \log \left( \frac{X_{t_i}}{\tau} \right),$$

here  $\hat{n}_{t,h,\tau} = \sum_{i=1}^n W_{t,h}(t_i) 1_{\{X_{t_i} > \tau\}}$  is the weighted nb. of obs. beyond  $\tau$ .

# Choice of the threshold $\tau$

Consider the case of i.i.d. observations :  $F_{t_i} = F$   
(no choice of the bandwidth  $h$ ).

We propose a procedure to choose the threshold  $\tau = \tau_n$  in two steps :

- **Propagation step :**  
given a sequence of embedded Pareto models, it consists in determining the largest one to fit the data.
- **Selection step :**  
determine the fitted model by penalized maximum likelihood.

# Propagation step

We choose as thresholds the order statistics :  $\tau_1 \equiv X_{(1)} \geq \dots \geq \tau_n \equiv X_{(n)}$ .

- ① Fix  $\tau_{m_0}$  a starting threshold ( $m_0 \geq 3$ ).
- ② For  $m = m_0, \dots, n$ , test sequentially :
  - **null hypothesis**  $\mathcal{H}_{\tau_m} : F_{\tau_m}$  has a **Pareto** distribution  $P_\theta$  against
  - **alternative**  $\tilde{\mathcal{H}}_{\tau_m} : F_{\tau_m}$  has a **Pareto CP** distribution  $P_{\theta_1, \theta_2, \nu}$  with some change-point  $\nu$
- ③ Stop when the null hypothesis is first rejected (say  $\hat{m}$ ).  
**The output** is the **break point**  $\hat{s} = \tau_{\hat{m}} \equiv X_{(\hat{m})}$ .

To test the null hypothesis  $\mathcal{H}_{\tau_m}$  we use the **likelihood ratio test statistic**

(precisely its max of over all change points  $\nu_l = \tau_l / \tau_m$ ) :

$$TS(\tau_m) = \max_{\delta' k \leq l \leq (1-\delta'')m} \{ \max_{\theta_1, \theta_2} \mathcal{L}_{\tau_m}(P_{\theta_1, \theta_2, \nu_l}) - \max_{\theta} \mathcal{L}_{\tau_m}(P_{\theta}) \}.$$

$\mathcal{H}_{\tau_m}$  is rejected if  $TS(\tau_m) > D$ , where  $D$  is a critical value.



# Selection step

Usually one chooses the last accepted parametric model in the propagation step. **This can introduce a huge bias**, since the last accepted model is close to the rejected one.

The idea : since at the break point  $\hat{s}$  the null hypothesis is rejected, penalize for getting close to  $\hat{s}$ .

Introduce a penalty term  $\text{Pen}(\tau) = \mathcal{L}_\tau \left( P_{\hat{\theta}_\tau} \right), \quad \tau \leq s.$

Choose the **adaptive threshold**  $\hat{\tau}$  by maximizing in  $\tau$  the penalized likelihood :

$$\max_{\theta} \mathcal{L}_\tau(P_\theta) - \text{Pen}(\tau) = \mathcal{L}_\tau \left( P_{\hat{\theta}_\tau} \right) - \mathcal{L}_\tau \left( P_{\hat{\theta}_s} \right) \geq 0$$

**The output :**

- ① the adaptive threshold  $\hat{\tau}$
- ② the adaptive estimator  $\hat{\theta}_{\hat{\tau}} = \hat{\theta}_{n, \hat{\tau}}$ .
- ③ (the break point  $\hat{s}$ )

# Choice of the bandwidth $h$

## Cross Validation (globaly for all $t_i$ ).

- 1 For each bandwidth  $h_m = aq^m$ ,  $m = 1, \dots, M_h$  choose the threshold  $\hat{\tau}_m$  by the adaptive procedure presented before.
- 2 We choose an adaptive  $\hat{h}$  among the values  $h_m$  by minimizing the cross validation function

$$CV(h_m, p) = \frac{1}{M_h \text{card}(T_{grid})} \sum_{h_l} \sum_{t_i \in T_{grid}} \left| \log \frac{\hat{q}_p^{(-i)}(t_i, h_m)}{\hat{F}_{t_i, h_l}^{-1}(p)} \right|,$$

- $\hat{q}_p^{(-i)}(\cdot)$  is the adaptive quantile estimator with  $X_{t_i}$  removed.
- $\hat{F}_{t_i, h_l}^{-1}(p)$  is the empirical quantile in the window  $[t - h_l, t + h_l]$
- $p$  should be not very high :  $p = p_{cv} = 0.99$ .

# R packages for extreme value theory

There are several packages on the extreme value theory with different application domain and methods. The following packages are the most used ones and close to our method

- **evd** package : Univariate and bivariate extreme modelling, DA of Gumbel and Weibull computed
- **evir** package : Univariate modelling. DA of Gumbell computed.

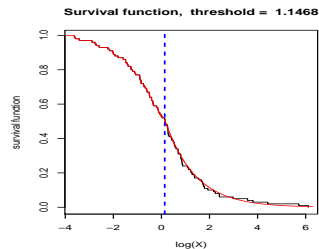
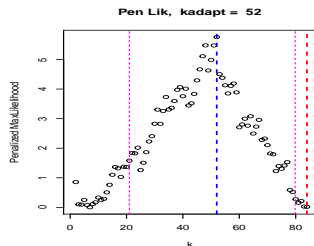
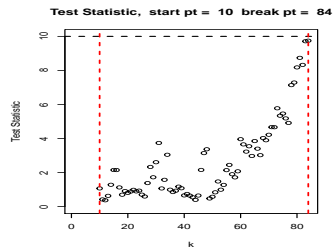
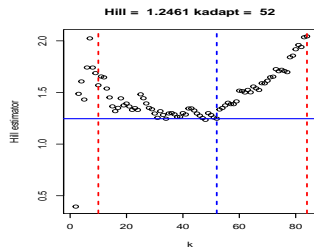
Both of them have functions helping the choice of the threshold.  
The package **evir** allow an easy prediction of the extreme quantiles.

# extremefit package

What new features does the **extremefit** package bring ?

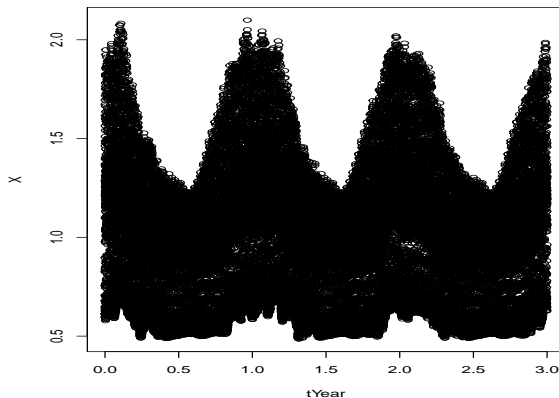
- Adaptive choice of the threshold  $\tau$  in the POT method for computation of the extreme  $p$ -quantile and of the extreme survival probabilities.
- Computation of the extreme  $p$ -quantile and of the extreme survival probabilities varying with the time  $t \in [0, T_{max}]$ .
- Choice of the bandwidth parameter by cross validation method.
- Goodness-of-fit test for testing the adjustment of the tail by a Pareto type d.f.

# Adaptive choice of the threshold $\tau$

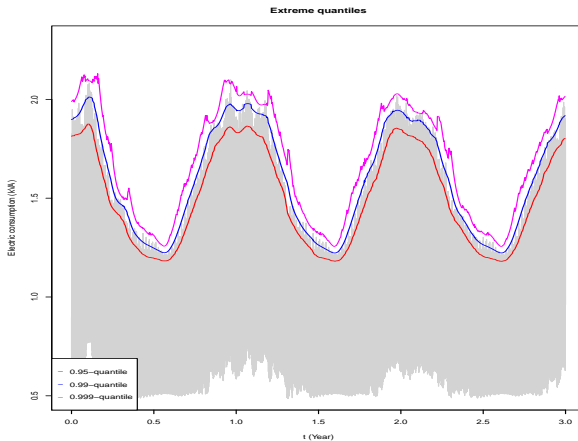


# Example of a data set (ERDF)

The data are the average consumption on a sample of clients with 3 or 6 kVa contract. The data are collected every 30 minutes during 3 years.



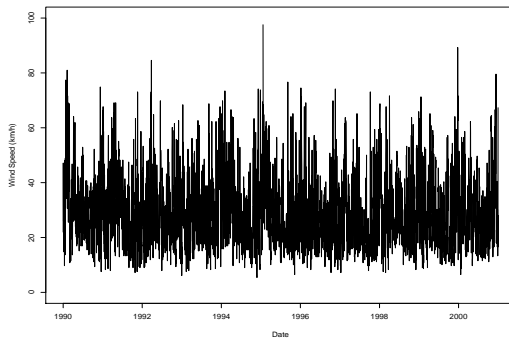
# Computation of quantiles



# An example of a data set (Wind from Brest)

The data are from a study of the wind speed in Brest (France) from 1976 to 2005. The data are collected daily.

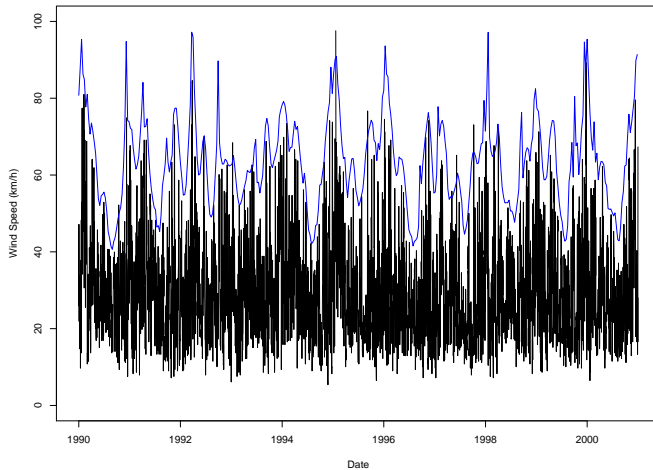
**FIGURE:** Wind speed for the period 1990 to 2001





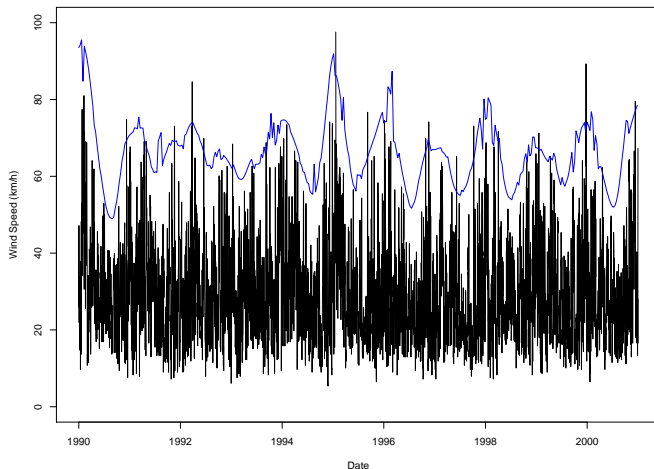
# Computation of the 0.99-quantile

The bandwidth is chosen to be two months.



# Computation of the 0.99-quantile

The bandwidth is chosen by cross-validation (more than 4 months).



Thank you for your attention !