extremefit: An R Package for Extreme Quantiles

Gilles Durrieu, Ion Grama, Kevin Jaunatre, Quang-Khoai Pham, Jean-Marie Tricot

extremefit: An R Package for Extreme Quantiles

Gilles Durrieu
Université de Bretagne Sud

Ion Grama
Université de Bretagne Sud

Kevin Jaunatre
Université de Bretagne Sud

Quang-Khoai Pham
University of Hanoi

Jean-Marie Tricot
Université de Bretagne Sud

Abstract

extremefit is an R package to estimate the extreme quantiles and probabilities of rare events. The idea of our approach is to adjust the tail of the distribution function over a threshold with a Pareto distribution. We propose a pointwise data driven procedure to choose the threshold. To illustrate the method, we use simulated data sets example and three real-world data sets available on the package.

Keywords: Nonparametric estimation, Tail conditional probabilities, Extreme conditional quantile, Adaptive estimation, Application and case study.

1. Introduction

Extreme values study plays an important role in several practical domain of applications, such as insurance, biology and geology. For example, in Buishand, de Haan, and Zhou (2008), the authors study extremes to determine how severe the rainfall periods occur in North Holland. Sharma, Khare, and Chakrabarti (1999) use extreme values procedure to predict violations of air quality standards. Various applications were presented in a lot of areas such as hydrology (Davison and Smith (1990) and Katz, Parlange, and Naveau (2002)), insurance (McNeil (1997) and Rootzén and Tajvidi (1997)) or finance (Danielsson and de Vries (1997), McNeil (1998), Embrechts, Resnick, and Samorodnitsky (1999) and Gencay and Selcuk (2004)). Other applications goes from rainfall data (Gardes and Girard (2010)) to earthquake (Sornette, Knopoff, Kagan, and Vanneste (1996)). The extreme value theory consists using appropriate statistical models to estimate extreme quantiles and probability of rare events.

The idea of the approach implemented in the **extremefit** package is to fit a Pareto distribution to the data over a threshold τ using the Peak-Over-Threshold method. The choice of τ is a challenging problem, a large value can lead to an important variability while a small value may increase the bias. We refer to Hall and Welsh (1985), Drees and Kaufmann (1998), Guillou

and Hall (2001), Huisman, Koedijk, Kool, and Palm (2001), Beirlant, Goegebeur, Teugels, and Segers (2004), Grama and Spokoiny (2008, 2007) and El Methni, Gardes, Girard, and Guillou (2012) where several procedures for choosing the threshold τ have been proposed. Here, we adopt the method from Grama and Spokoiny (2008) and Durrieu, Grama, Pham, and Tricot (2015). The package **extremefit** includes the modeling of time dependent data. The analysis of time series involves a bandwidth parameter h whose data driven choice is non trivial. We refer to Staniswalis (1989) and Loader (1999) for the choice of the bandwidth in a nonparametric regression. For the purposes of extreme value modeling, we use a cross-validation approach from Durrieu *et al.* (2015).

The **extremefit** package for the R system (R Development Core Team (2016)) is based on the methodology described in Durrieu *et al.* (2015). The package performs a nonparametric estimation of extreme quantiles and probabilities of rare events. It proposes a pointwise choice of the threshold τ and, for the time series, a global choice of the bandwidth h and gives graphical representations of the results.

The paper is organized as follows. Section 2 is an overview of several existing R packages dealing with extreme value analysis. In Section 3, we describe the model and the estimation of the parameters, including the threshold τ and the bandwidth h choices. Section 4 contains a simulation study whose aim is to illustrate the performance of our approach. In Section 5, we give several applications on real data sets and we conclude in Section 6.

2. Extreme value packages in R

There exist several R packages dealing with the extreme value analysis. We give a short description of some of them. For a detailed description of these packages, we refer to Gilleland, Ribatet, and Stephenson (2013). A CRAN task view exists in extreme value analysis giving a description of registered packages available in CRAN, see `https://CRAN.R-project.org/view=ExtremeValue`.

Some of the packages have a specific use, such as the package **SpatialExtremes** (Ribatet, Singleton, and team (2011)), which models spatial extremes and provides maximum likelihood estimation, bayesian hierarchical and copula modeling, or the package **fExtremes** (Wuertz *et al.* (2013)) for financial purposes using functions from the packages **evd**, **evir** and others.

The **copula** package (Hofert, Kojadinovic, Maechler, and Yan (2010)) provides tools for exploring and modeling dependent data using copulas. The **evd** package (Stephenson and Ferro (2002)) provides both block maxima and peak-over-threshold computations based on maximum likelihood estimation in the univariate and bivariate cases. The **evdbayes** package (Stephenson and Ribatet (2010)) provides an extension of the **evd** package using bayesian statistical methods for univariate extreme value models. The package **extRemes** (Gilleland (2011)) implements also univariate estimation of block maxima and peak-over-threshold by maximum likelihood estimation allowing non stationarity. The package **evir** (Pfaff, McNeil, and Stephenson (2008)) is based on fitting a generalized Pareto distribution with the Hill estimator over a given threshold. The package **lmom** (Hosking (2009)) is dealing with L-moments to estimate the parameters of extreme value distributions and quantile estimations for reliability or survival analysis. The package **texmex** (Southworth and Heffernan (2010)) provides statistical extreme value modeling of threshold excesses, maxima and multivariate extremes, including maximum likelihood and Bayesian estimation of parameters.

In contrast to previous described packages, the **extremefit** package provides tools for modeling heavy tail distributions without assuming a general parametric structure. The idea is to fit a parametric Pareto model to the tail of the unknown distribution over some threshold. The remaining part of the distribution is estimated nonparametrically and a data driven algorithm for choosing the threshold is proposed in Section 3.2. We also provide a version of this method for analyzing extreme values of a time series based on the nonparametric kernel function estimation approach. A data driven choice of the bandwidth parameter is given in Section 3.3. These estimators are studied in more details in Durrieu *et al.* (2015).

3. Extreme value prediction using a semi-parametric model

3.1. Model and Estimator

We consider $F_t(x) = P(X \leq x|T = t)$ the conditional distribution of a random variable X given a time covariate $T = t$, where $x \in [x_0, \infty)$ and $t \in [0, T_{\max}]$. We observe independent random variables $X_{t_1}, ..., X_{t_n}$ associated to a sequence of times $0 \leq t_1 < ... < t_n \leq T_{\max}$, such that for each t_i, the random variable X_{t_i} has the distribution function F_{t_i}. The purpose of the **extremefit** package is to provide a pointwise estimation of the tail probability $S_t(x) = 1 - F_t(x)$ and the extreme p-quantile $F_t^{-1}(p)$ processes for any $t \in [0, T_{\max}]$, given $x > x_0$ and $p \in (0, 1)$. We assume that F_t, are in the domain of attraction of the Fréchet distribution. The idea is to adjust, for some $\tau \geq x_0$, the excess distribution function

$$F_{t,\tau}(x) = 1 - \frac{1 - F_t(x)}{1 - F_t(\tau)}, \quad x \in [\tau, \infty) \tag{1}$$

by a Pareto distribution:

$$G_{\tau,\theta}(x) = 1 - \left(\frac{x}{\tau}\right)^{-\frac{1}{\theta}}, \quad x \in [\tau, \infty), \tag{2}$$

where $\theta > 0$ and $\tau \geq x_0$ an unknown threshold, depend on t. The justification of this approach is given by the Fisher-Tippett-Gnedenko theorem (Beirlant *et al.* (2004), Theorem 2.1) which states that F_t, is in the domain of attraction of the Fréchet distribution if and only if $1 - F_{t,\tau}(\tau x) \to x^{-1/\theta}$ as $\tau \to \infty$. This consideration is based on the peak-over-threshold (POT) approach (Beirlant *et al.* (2004)). We consider the semi-parametric model defined by:

$$F_{t,\tau,\theta}(x) = \begin{cases} F_t(x) & \text{if } x \in [x_0, \tau], \\ 1 - (1 - F_t(\tau))(1 - G_{\tau,\theta}(x)) & \text{if } x > \tau, \end{cases} \tag{3}$$

where $\tau \geq x_0$ is the threshold parameter. We propose in the sequel to estimate F_t and θ which are unknown in (3).

The estimator of $F_t(x)$ is taken as the weighted empirical distribution given by

$$\widehat{F}_{t,h}(x) = \frac{1}{\sum_{j=1}^{n} W_{t,h}(t_j)} \sum_{i=1}^{n} W_{t,h}(t_i) \mathbb{1}_{\{X_{t_i} \leq x\}}, \tag{4}$$

where, for $i = 1, \ldots, n$, $W_{t,h}(t_i) = K\left(\frac{t_i - t}{h}\right)$ are the weights and $K(.)$ is a kernel function assumed to be continuous, non-negative, symmetric with support on the real line such that $K(x) \leq 1$, and $h > 0$ is a bandwidth.

By maximizing the weighted quasi-log-likelihood function (see Durrieu *et al.* (2015), Staniswalis (1989) and Loader (1999))

$$\mathcal{L}_{t,h}(\tau,\theta) = \sum_{i=1}^{n} W_{t,h}(t_i) \log \frac{dF_{t,\tau,\theta}}{dx}(X_{t_i}) \tag{5}$$

with respect to θ, we obtain the estimator

$$\widehat{\theta}_{t,h,\tau} = \frac{1}{\widehat{n}_{t,h,\tau}} \sum_{i=1}^{n} W_{t,h}(t_i) \mathbb{1}_{\{X_{t_i}>\tau\}} \log\left(\frac{X_{t_i}}{\tau}\right), \tag{6}$$

where $\widehat{n}_{t,h,\tau} = \sum_{i=1}^{n} W_{t,h}(t_i)\mathbb{1}_{\{X_{t_i}>\tau\}}$ is the weighted number of observations over the threshold τ.

Plug-in (4) and (6) in the semi-parametric model (3), we obtain:

$$\widehat{F}_{t,h,\tau}(x) = \begin{cases} \widehat{F}_{t,h}(x) & \text{if } \; x \in [x_0, \tau], \\ 1 - \left(1 - \widehat{F}_{t,h}(\tau)\right)\left(1 - G_{\tau,\widehat{\theta}_{t,h,\tau}}(x)\right) & \text{if } \; x > \tau. \end{cases} \tag{7}$$

For any $p \in (0,1)$, the estimator of the p-quantile of X_t is defined by

$$\widehat{q}_p(t,h) = \begin{cases} \widehat{F}_{t,h}^{-1}(p) & \text{if} \quad p < \widehat{p}_\tau, \\ \tau\left(\frac{1-\widehat{p}_\tau}{1-p}\right)^{\widehat{\theta}_{t,h,\tau}} & \text{otherwise,} \end{cases} \tag{8}$$

where $\widehat{p}_\tau = \widehat{F}_{t,h}(\tau)$.

3.2. Selection of the Threshold

The determination of the threshold τ in model (3) is based on a testing procedure which is a goodness-of-fit test for the parametric-based part of the model. At each step of the procedure, the tail adjustment to a Pareto distribution is tested based on k upper statistics. If it is not rejected, the number k of upper statistics is increased and the tail adjustment is tested again until it is rejected. If the test rejects the parametric tail fit from the very beginning, the Pareto tail adjustment is not significant. On the other hand, if all the tests accept the parametric Pareto fit then the underlying distribution is significantly Pareto. The critical value denoted by D depends on the kernel choice and is determined by Monte-Carlo simulation, using the **CriticalValue** function of the package.

In Table 1, we display the critical values obtained for several kernel functions using **CriticalValue**.

In our package, the Gaussian kernel with standard deviation 1/3 is approximated by the truncated Gaussian kernel $\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right)\mathbb{1}_{|x|\leq 1}$ with $\sigma = 1/3$.

The default values of the parameters in the algorithm yield satisfying estimation results on simulation study without being time-consuming even for large data sets. The choice of these tuning parameters is given in Durrieu *et al.* (2015).

The following commands compute the critical value D for the truncated Gaussian kernel with $\sigma = 1$ (default value) and display the empirical distribution function of the goodness-of-fit test statistic which determines the threshold τ.

Table 1: Critical values associated to kernel functions

Kernel	D	$K(x)$
Biweight	7	$\frac{15}{16}(1-x^2)^2\mathbb{1}_{\lvert x\rvert\leq 1}$
Epanechnikov	6.1	$\frac{3}{4}(1-x^2)\mathbb{1}_{\lvert x\rvert\leq 1}$
Rectangular	10.0	$\mathbb{1}_{\lvert x\rvert\leq 1}$
Triangular	6.9	$(1-\lvert x\rvert)\mathbb{1}_{\lvert x\rvert\leq 1}$
Truncated Gaussian, $\sigma = 1/3$	8.3	$\frac{3}{\sqrt{2\pi}}\exp(-\frac{(3x)^2}{2})\mathbb{1}_{\lvert x\rvert\leq 1}$
Truncated Gaussian, $\sigma = 1$	3.4	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)\mathbb{1}_{\lvert x\rvert\leq 1}$

```
R> library(extremefit)
R> n <- 1000 #Define the sample size
R> NMC <- 500 #Define the number of Monte-Carlo simulated samples
R> CriticalValue(NMC, n, TruncGauss.kernel, prob = 0.99, plot = TRUE)

[1] 3.432665
```

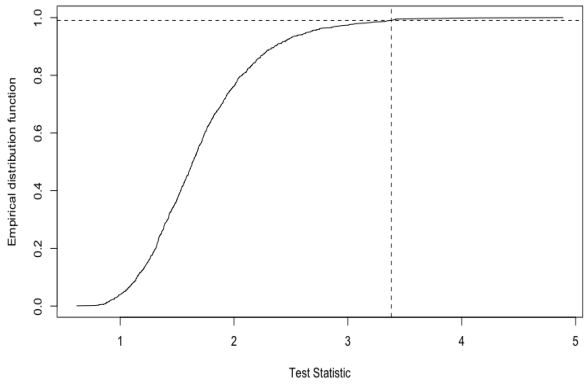

Figure 1: Empirical distribution function of the test statistic for the truncated Gaussian kernel with $N_{MC} = 1000$ Monte-Carlo samples of size $n = 500$. The vertical dashed line represents the critical value ($D = 3.4$) corresponding to the 0.99-empirical quantile of the test statistic.

For a given t, the function **hill.adapt** allows a data-driven choice of the threshold τ and the estimation of θ_t.

3.3. Selection of the bandwidth h

We determine the bandwidth h by cross-validation from a sequence of the form $h_l = aq^l$, $l = 0, \ldots, M_h$ with $q = \exp\left(\frac{\log b - \log a}{M_h}\right)$, where a is the minimum bandwidth of the sequence,

b is the maximum bandwidth of the sequence and M_h is the length of the sequence. The choice is performed globally on the grid $T_{grid} = \{t_1, \ldots, t_K\}$ of points $t_i \in [0, T_{\max}]$, where the number K of the points on the grid is defined by the user. The choice $K = n$ is possible but can be time consuming for large samples. We recommend to use a fraction of n.

We choose h_{cv} by minimizing in h_m, $m = 1, \ldots, M_h$ the cross-validation function

$$CV(h_m, p_{cv}) = \frac{1}{M_h \text{card}(T_{grid})} \sum_{h_l} \sum_{t_i \in T_{grid}} \left| \log \frac{\widehat{q}_{p_{cv}}^{(-i)}(t_i, h_m)}{\widehat{F}_{t_i,h_l}^{-1}(p_{cv})} \right|, \tag{9}$$

where $\widehat{F}_{t_i,h_l}^{-1}(p_{cv})$ is the empirical quantile from the observations in the window $[t_i - h_l, t_i + h_l]$, $\widehat{q}_{p_{cv}}^{(-i)}(t_i, h_m)$ is the quantile estimator inside the window $[t_i - h_m, t_i + h_m]$ defined by (8) with the observation X_{t_i} removed and τ being the adaptive threshold given by the remaining observations inside the window $[t_i - h_m, t_i + h_m]$. The function **bandwidth.CV** selects the bandwidth h by cross-validation.

4. Package presentation on simulation study

The **extremefit** package is written in R, (R Development Core Team (2016)). In this section, we demonstrate the package using its application on two simulated data sets.

The following code displays the computation of the survival probabilities and quantiles using the adaptive choice of the threshold provided by the **hill.adapt** function.

```
R> library(extremefit)
R> set.seed(5)
R> X <- abs(rcauchy(200))
R> n <- 100
R> HA <- hill.adapt(X)
R> predict(HA, newdata = c(3, 5, 7), type = "survival")$p
R> predict(HA, newdata = c(0.9, 0.99, 0.999, 0.9999), type = "quantile")$y
```

A simple use of the method described in Section 3 is given by the following example. With $t_i = i/n$, we consider data $X_{t_1}, \ldots, X_{t_n}$ generated by the Pareto change-point model defined by

$$F_t(x) = \left(1 - x^{-1/2\theta_t}\right) \mathbb{1}_{x \leq \tau} + \left(1 - x^{-1/\theta_t} \tau^{1/2\theta_t}\right) \mathbb{1}_{x > \tau}, \tag{10}$$

where θ_t is a time varying parameter depending on $t \in [0, 1]$ defined by $\theta_t = 0.5 + 0.25 \sin(2\pi t)$ and $\tau = 3$ as described in Durrieu *et al.* (2015). We consider the sample size $n = 50000$. The following commands generate one sample from the model (10).

```
R> library(extremefit)
R> set.seed(5)
R> n <- 50000 ; tau <- 3
R> theta <- function(t){0.5+0.25*sin(2*pi*t)}
R> T <- 1:n/n; Theta <- theta(T); X <- NULL
R> for(i in 1:n){
R>     X[i] <- rparetoCP(1, a0 = 1/(Theta[i]*2), a1 = 1/Theta[i], x1=tau)
R> }
```

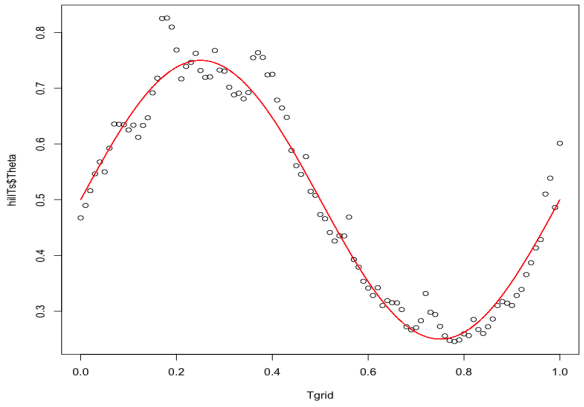

Figure 2: Plot of the θ_t estimate $\widehat{\theta}_{t,h_{cv},\tau}$ (black dots) and the true θ_t (red line) for each $t \in T_{grid}$.

The **extremefit** package provides the estimates of θ_t, $q_p(t,h)$ and $F_t(x)$ for large values of x particularly. We select the bandwidth h_{cv} by minimizing the cross-validation function implemented in **bandwidth.CV**. The weights are computed using the truncated Gaussian kernel ($\sigma = 1$), which is implemented in **TruncGauss.kernel**. This kernel implies $D = 3.4$. To select the bandwidth h_{cv}, we define a grid of possible values of h as indicated in Section 3.3 with $a = 0.005$, $b = 0.05$ and $M_h = 20$. Moreover, we fix the parameter $p_{cv} = 0.99$.

```
R> a <- 0.005 ; b <- 0.05 ; Mh <- 20
R> hl <- bandwidth.grid(a, b, Mh, type = "geometric")
R> Tgrid <- seq(0, 1, 0.02) #define a grid to perform the cross-validation
R> Hcv <- bandwidth.CV(X, T, Tgrid, hl, pcv = 0.99,
+                kernel = TruncGauss.kernel, CritVal = 3.4, plot = FALSE)
R> Hcv$h.cv

[1] 0.02727797
```

For each $t \in T_{grid}$, we determine the data-driven threshold τ and the estimates $\widehat{\theta}_{t,h_{cv},\tau}$ using the function **hill.ts**.

```
R> Tgrid <- seq(0, 1, 0.01)
R> hillTs <- hill.ts(X, T, Tgrid, h = Hcv$h.cv,
+                    kernel = TruncGauss.kernel, CritVal = 3.4)
```

For each $t \in T_{grid}$, we display $\widehat{\theta}_{t,h_{cv},\tau}$ and the true value $\theta_t = 0.5 + 0.25\sin(2\pi t)$ in Figure 2.

```
R> plot(Tgrid, hillTs$Theta)
R> lines(T, Theta, col = "red")
```

The estimates of the quantiles and the survival probabilities are determined using the **predict.hill.ts** function. For instance the estimate of the p-quantile $F_t^{-1}(p)$ of order $p = 0.99$ and $p = 0.999$ are computed with the following R command:

```
R> PredQuant <- predict(hillTs, newdata = c(0.99, 0.999), type = "quantile")
```

Figure 3 displays the true and the estimated quantiles of order $p = 0.99$ and $p = 0.999$ of the Pareto change-point distribution defined by (10). The true quantiles can be accessed by the **qparetoCP** function.

```
R> TrueQuant <- matrix(0, ncol = 2, nrow = n)
R> for(i in 1:n){
+     TrueQuant[i,] <- qparetoCP(c(0.99, 0.999), a0 = 1/(Theta[i]*2),
+                                a1 = 1/Theta[i], x1 = tau)
+   }
R> plot(Tgrid, log(as.numeric(PredQuant$y[1,])), ylim = c(1.5, 6.3),
+            ylab = "estimated log-quantiles")
R> points(Tgrid, log(as.numeric(PredQuant$y[2,])), pch = "+")
R> lines(T, log(TrueQuant[,1]))
R> lines(T, log(TrueQuant[,2]), col = "red")
```

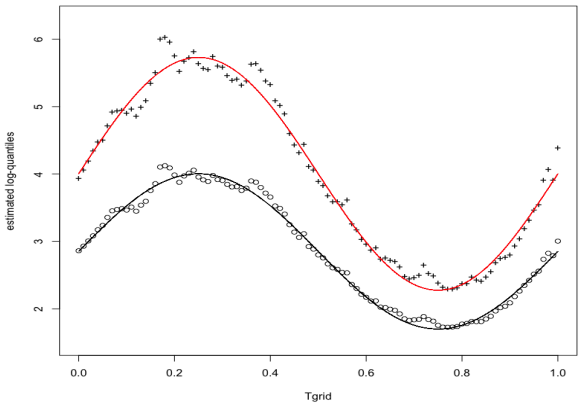

Figure 3: Plot of the true log quantiles $F_t^{-1}(p)$ with $p = 0.99$ (black line) and $p = 0.999$ (red line) and the corresponding estimated quantiles with $p = 0.99$ (black dots) and $p = 0.999$ (black cross) as function of $t \in T_{grid}$.

In the same way, we estimate for each $t \in T_{grid}$ the survival probabilities $S_t(x) = 1 - F_t(x)$. The function **predict.hill.ts** with the option **type = "survival"** computes the estimated survival function for a given x. For each $t \in T_{grid}$, the following commands compute the estimate of $S_t(x)$ for $x = 20$ and $x = 30$.

```
R> PredSurv <- predict(hillTs, newdata = c(20, 30), type="survival")
R> TrueSurv <- matrix(0, ncol = 2, nrow = n)
R> for(i in 1:n){
+      TrueSurv[i,] <- 1 - pparetoCP(c(20, 30), a0 = 1/(Theta[i]*2),
+                                a1 = 1/Theta[i], x1 = tau)
+   }
```

```
R> plot(Tgrid, as.numeric(PredSurv$p[1,]),
+          ylab = "estimated survival probabilities")
R> points(Tgrid, as.numeric(PredSurv$p[2,]), pch = "+")
R> lines(T, TrueSurv[,1])
R> lines(T, TrueSurv[,2], col = "red")
```

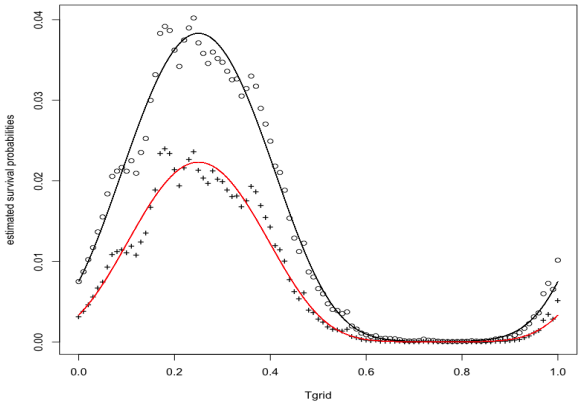

Figure 4: Plot of the true survival probabilities $S_t(x)$ at $x = 20$ (black line) and $x = 30$ (red line) and the corresponding estimated survival probabilities at $x = 20$ (black dots) and $x = 30$ (black cross) as function of $t \in T_{grid}$.

The estimations of the quantile and the survival function displayed in Figures 3 and 4 have been performed for one sample. Now we analyze the performance of the quantile estimator via a Monte-Carlo study. We obtain in Figure 5 satisfying results on a simulation study using $N_{MC} = 1000$ Monte-Carlo replicates. The iteration of the previous R commands on $N_{MC} = 1000$ simulations are not given because of the computation time.

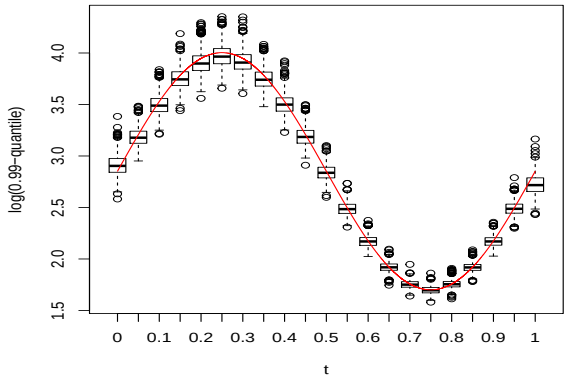

Figure 5: Boxplots of the log $\widehat{q}_{0.99}(t, h_{cv})$ adaptive estimators with bandwidth h_{cv} chosen by cross-validation and $t \in T_{grid}$. The true log 0.99-quantile is plotted as a red line.

5. Real-world data sets

5.1. Wind data

The study of wind speed is important for the renewable energy problem in present time. Many considers wind as a free and environmentally source of energy. The implementation of wind farm throughout the world shows an encouraging and promising energy option. Studies of wind speed in extreme value theory were made to model windstorm losses or detect areas which can be subject to hurricanes (see Rootzén and Tajvidi (1997) and Simiu and Heckert (1996)).

The wind data in the package **extremefit** comes from Airport of Brest (France) and represents the average wind speed per day from 1976 to 2005. The data set is included in the package **extremefit** and can be loaded by the following code.

```
R> library(extremefit)
R> data("dataWind")
R> attach(dataWind)
```

The commands below illustrate the function **hill.adapt** on the wind data set and computes a monthly estimation of the survival probabilities $1 - \widehat{F}_{t,h,\tau}(x)$ for a given $x = 100$ km/h with the function **predict.hill.adapt**.

```
R> pred <- NULL
R> for(m in 1:12){
+  indices <- which(Month == m)
+  X <- Speed[indices]*60*60/1000
+  H <- hill.adapt(X)
+  pred[m] <- predict(H, newdata = 100, type = "survival")$p
+}
R> plot(pred, ylab = "Estimated survival probability", xlab = "Month")
```

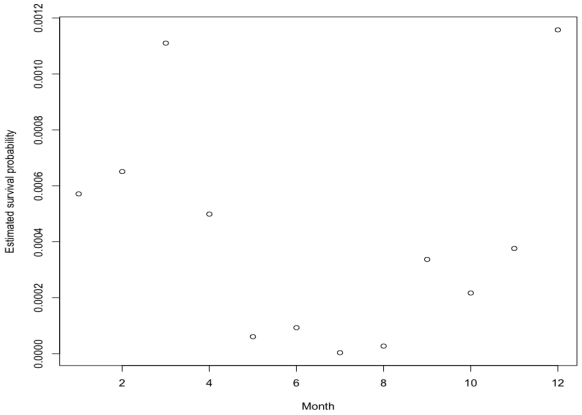

Figure 6: Plot of the estimated survival probability $1 - \widehat{F}_{t,h,\tau}(x)$ at $x = 100$ km/h.

5.2. Sea shores water quality

The study of the pollution in the aquatic environment is an important problem to have it protected. Humans tend to pollute the environment through their activities and the water quality survey is necessary. The bivalve's activity is investigated by modeling the valve movements using high frequency valvometry. The electronic equipment is described in Tran, Ciret, Ciutat, Durrieu, and Massabuau (2003) and modified by Chambon, Legeay, Durrieu, Gonzalez, Ciret, and Massabuau (2007). More information can be found in http://molluscan-eye.epoc.u-bordeaux1.fr/.

High-frequency data (10 Hz) are produced by noninvasive valvometric techniques and the study of the bivalve's behavior in their natural habitat leads to the proposal of several statistical models (Sow, Durrieu, and Briollais (2011), Schmitt, De Rosa, Durrieu, Sow, Ciret, Tran, and Massabuau (2011), Jou and Liao (2006), Coudret, Durrieu, and Saracco (2015), Azaïs, Coudret, and Durrieu (2014), Durrieu *et al.* (2015) and Durrieu, Pham, Foltête, Maxime, Grama, Le Tilly, Duval, Tricot, Naceur, and Sire (2016)). It is observed that in the presence of a pollutant, the activity of the bivalves is modified and consequently they can be used as bioindicators to detect perturbations in aquatic systems (pollutions, global warming). A group of oysters *Crassostrea gigas* of the same age are installed around the world but we concentrate on the Locmariaquer site (GPS coordinates 47°34 N, 2°56 W) in France. The oysters are placed in a traditional oyster bag. In the package **extremefit**, we provide a sample of the measurements for one oyster over one day. The data can be accessed by

```
R> library(extremefit)
R> data("dataOyster")
```

The description of the data can be found with the R command `help(dataOyster)`. The following code covers the velocities and the time covariate and also displays the data.

```
R> Velocity <- dataOyster$data[, 3]
R> time <- dataOyster$data[, 1]
R> plot(time, Velocity, type = "l", xlab = "time (hour)",
+           ylab = "Velocity (mm/s)")
```

We observe in Figure 7 that the velocity is equal to zero in two periods of time. To facilitate the study of these data, we have included a time grid where the intervals with null velocities are removed. The grid of time can be accessed by `dataOyster$Tgrid`. We shift the data to be positive.

```
R> #grid for which the velocities are different from 0
R> new.Tgrid <- dataOyster$Tgrid

R> #We shift the data to be positive
R> X <- Velocity + (-min(Velocity))
```

The bandwidth parameter is selected by cross-validation method ($h_{cv} = 0.2981812$) using **bandwidth.CV** but we select it manually in the following command due to long computation time.

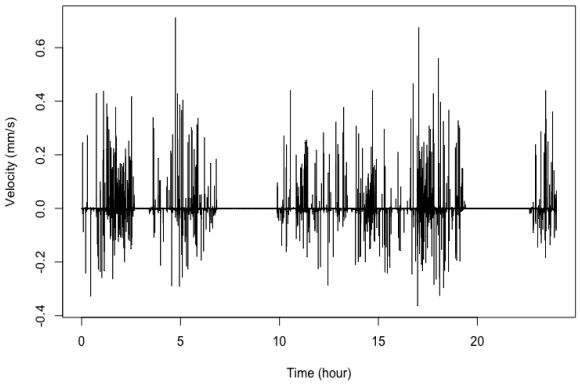

Figure 7: Plot of the velocity of the valve closing and opening over one day.

```
R> hcv <- 0.2981812
R> TS.Oyster <- hill.ts(X ,t = time, new.Tgrid, h = hcv,
+                       TruncGauss.kernel, CritVal = 3.4)
```

The estimations of the extreme quantile of order 0.999 and the probabilities of rare events are computed as described in Section 3.2. The critical value of the sequential test is $D = 3.4$ when considering a truncated Gaussian kernel, see Table (1). A global study on a set of 16 oysters on a 6 months period is given in Durrieu *et al.* (2015).

```
R> pred.quant.Oyster <- predict(TS.Oyster, newdata = 0.999, type = "quantile")
R> plot(time, Velocity, type = "l", ylim = c(-0.5, 1),
+     xlab ="Time (hour)", ylab = "Velocity (mm/s)")
R> quant0.999 <- rep(0, length(seq(0, 24, 0.05)))
R> quant0.999[match(new.Tgrid, seq(0, 24, 0.05))] <-
+    as.numeric(pred.quant.Oyster$y) -
+    (-min(Velocity))
R> lines(seq(0, 24, 0.05), quant0.999, col = "red")
```

In Durrieu *et al.* (2015) and Durrieu *et al.* (2016), we observe that valvometry using extreme value theory allows in real-time *in situ* analysis of the bivalves behavior and appears as an effective early warning tool in ecological risk assessment and marine environment monitoring.

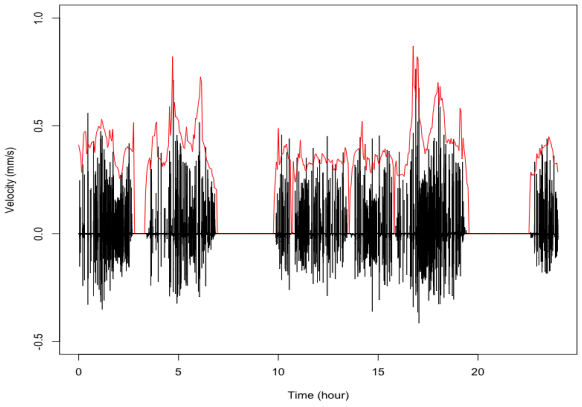

Figure 8: Plot of the estimated 0.999-quantile (red line) and the velocities of valve closing (black lines).

5.3. Electric consumption

The study of the electric consumption is an important challenge due to the expansion of the human population that increases the need of electricity. Multiple models have been used to forecast the electric consumption, as regression and time series models (Bianco, Manca, and Nardini (2009) and Ranjan and Jain (1999), Harris and Liu (1993) and Bercu and Proïa (2013)). Durand, Bozzi, Celeux, and Derquenne (2004) used hidden Markov model to forecast the electric consumption. A research project conducted in France (Lorient, GPS coordinates 47°45 N, 3°22 W) concerns the measurements of electric consumption using Linky, a smart communicating electric meter (`http://www.enedis.fr/linky-communicating-meter`). Installed in end-consumer's habitations and linked to a supervision center, this meter is in constant interaction with the network. The Linky electric meter allows a measurement of the electric consumption every 10 minutes.

To prevent from major power outages, the SOLENN project (`http://www.smartgrid-solenn.fr/en/`) is testing an alternative to load shedding. Data of electric consumption are collected on selected habitations to study the effect of a decrease on the maximal power limit. For example, an habitation with a maximal electric power contract of 9 kiloVolt ampere is decreased to 6 kiloVolt ampere. This experiment enables to study the consumption of the habitation with the application of an electric constraint related to the need of the network. For instance, after an incident such as a power outage on the electric network, the objective is to limit the number of habitations without electricity. If during the time period where the electric constraint is applied, the electric consumption of the habitation exceeds the restricted maximal power, the breaker cuts off and the habitation has no more electricity. The consumer can, at that time, close the circuit breaker and gets the electricity back. In any cases, at the end of the electric constraint, the network manager can close the breaker using the Linky electric meter which is connected to the network. The control of the cut off breakers is crucial to prevent a dissatisfaction from the customers and to detect which habitations are at risk.

The extreme value modeling approach described in Section 3 was carried out on the electric consumption data for one habitation from the 24th December 2015 to the 29th June

2016. This data are accessible on the **extremefit** package and Figure 9 displays the electric consumption of one habitation. This habitation has a maximal power contract of 9 kVA.

```
R> data("LoadCurve")
R> Date <- as.POSIXct(LoadCurve$data$Time*86400, origin = "1970-01-01")
R> plot(Date, LoadCurve$data$Value/1000, type = "l",
+     ylab = "Electric consumption (kVA)", xlab = "Date")
```

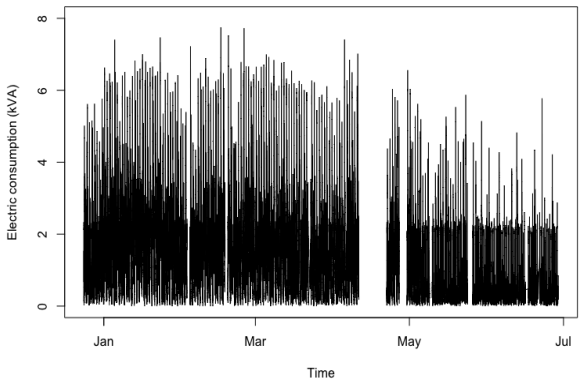

Figure 9: Electric consumption of one customer from the 24th December 2015 to the 29th June 2016.

We consider the following grid of time T_{grid}:

```
R>  Tgrid <-  seq(min(LoadCurve$data$Time), max(LoadCurve$data$Time),
+             length = 400)
```

We observe in April 2016 missing values in Figure 9 due to a technical problem. We modify the grid of time by removing the intervals of T_{grid} with no data.

```
R> new.Tgrid <-LoadCurve$Tgrid
```

We choose the truncated Gaussian kernel and the associated critical value of the goodness-of-fit test is $D = 3.4$ (see Table 1). The bandwidth parameter is selected by cross-validation method ($h_{cv} = 3.44$) using **bandwidth.CV** but we select it manually in the following command due to long computation time.

```
R> HH <- hill.ts(LoadCurve$data$Value, LoadCurve$data$Time, new.Tgrid,
+                h = 3.44, kernel = TruncGauss.kernel, CritVal = 3.4)
```

To detect the probability to cut off the breaker, we compute for each time in the grid the estimates of the probability to exceed the maximal power of 9kVA and of the extreme 0.99-quantile. Figure 10 displays the electric consumption during the period of study and the estimated quantile of order 0.99.

```
R> Quant <- rep(NA, length(Tgrid))
R> Quant[match(new.Tgrid,Tgrid)] <- as.numeric(predict(HH,
+    newdata = 0.99, type = "quantile")$y)
R> plot(Date, LoadCurve$data$Value/1000, ylim = c(0, 8),
+      type = "l", ylab = "Electric consumption (kVA)", xlab = "Time")
R> lines(as.POSIXlt((Tgrid)*86400, origin = "1970-01-01",
+                   tz = "Europe/Paris"), Quant/1000, col = "red")
```

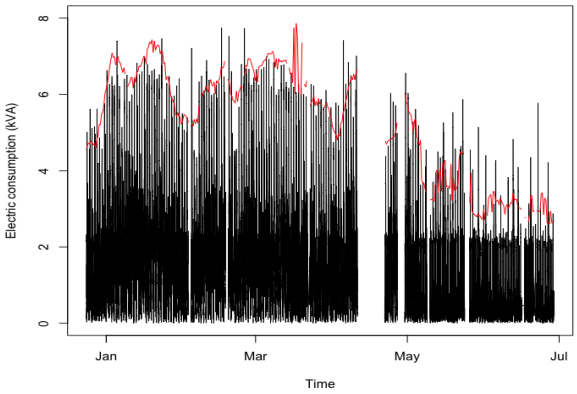

Figure 10: Plot of the estimated 0.99-quantile (red line) for each time from the 24th December 2015 to the 29th June 2016.

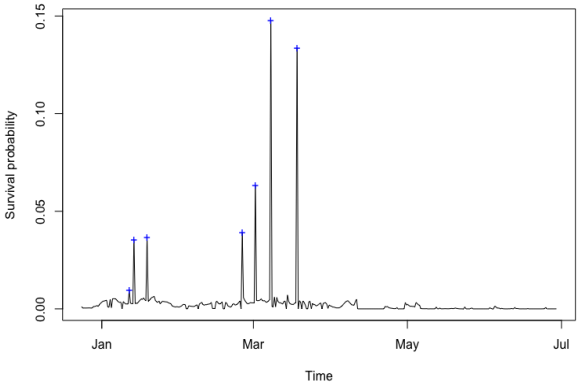

Figure 11: Plot of the estimated survival probability depending on the time and the maximal power. The plus symbol correspond to the electric constraint period.

The plus symbol appear at the times when the maximal power was decreased corresponding to the 11th, 13th, 18th of January, the 25th of February and the 1st, 7th, 18th of March in 2016, with a respectively decrease to 6.3, 4.5, 4.5, 4.5, 3.6, 2.7 and 2.7 kVA. Figure 11 displays the estimated survival probability which depends on the time and the maximal power. We can

observe that the survival probability is higher than usual during this period. Furthermore, we have the auxiliary information that this habitation cuts off its breaker on every electric constraint period except the 11th of January, 2016.

Using prediction method coupled with the method implemented in the package **extremefit**, it will be possible to detect high probability of cutting off a breaker and react accordingly.

6. Conclusion

This paper focus on the functions contained in the package **extremefit** to estimate extreme quantiles and probabilities of rare events. The package **extremefit** works also well on large data set and the performance was illustrated on simulated data and on real data sets.

The choice of the pointwise data driven threshold allows a flexible estimation of the extreme quantiles. The diffusion of the use of this method for the scientific community will improve the choice of estimation of the extreme quantiles and probability of rare events using the peak-over-threshold approach.

Acknowledgements

This work was supported by the ASPEET Grant from the Université Bretagne Sud and the Centre National de la Recherche Scientifique. Kevin Jaunatre would like to acknowledge the financial support of the SOLENN project.

References

Azaïs R, Coudret G, Durrieu G (2014). "A hidden renewal model for monitoring aquatic systems biosensors." *Environmetrics*, **25**, 189–199.

Beirlant J, Goegebeur Y, Teugels J, Segers J (2004). *Statistics of Extremes: Theory and Applications*. Wiley, Chichester.

Bercu S, Proïa F (2013). "A SARIMAX coupled modelling applied to individual load curves intraday forecasting." *Journal of Applied Statistics*, **40**(6), 1333–1348.

Bianco V, Manca O, Nardini S (2009). "Electricity consumption forecasting in Italy using linear regression models." *Energy*, **34**(9), 1413–1421.

Buishand TA, de Haan L, Zhou C (2008). "On Spatial Extremes: With Application to a Rainfall Problem." *The Annals of Applied Statistics*, **2**(2), 624–642.

Chambon C, Legeay A, Durrieu G, Gonzalez P, Ciret P, Massabuau JC (2007). "Influence of the parasite worm Polydora sp. on the behaviour of the oyster crassostrea gigas: a study of the respiratory impact and associated oxidative stress." *Marine Biology*, **152**(2), 329–338.

Coudret R, Durrieu G, Saracco J (2015). "Comparison of kernel density estimators with assumption on number of modes." *Communication in Statistics - Simulation and Computation*, **44**(1), 196–216.

Danielsson J, de Vries CG (1997). "Tail index and quantile estimation with very high frequency data." *Journal of empirical Finance*, **4**(2), 241–257.

Davison A, Smith R (1990). "Models for Exceedances over High Thresholds." *J. Roy. Statist. Soc. B*, **52**(3), 393–442.

Drees H, Kaufmann E (1998). "Selecting the optimal sample fraction in univariate extreme value estimation." *Stochastic Process. Appl.*, **75**, 149–172.

Durand JB, Bozzi L, Celeux G, Derquenne C (2004). "Analyse de courbes de consommation électrique par chaînes de Markov cachées." *Revue de statistique appliquée*, **52**(4), 71–91.

Durrieu G, Grama I, Pham Q, Tricot JM (2015). "Nonparametric adaptive estimator of extreme conditional tail probabilities quantiles." *Extremes*, **18**, 437–478.

Durrieu G, Pham QK, Foltête AS, Maxime V, Grama I, Le Tilly V, Duval H, Tricot JM, Naceur CB, Sire O (2016). "Dynamic extreme values modeling and monitoring by means of sea shores water quality biomarkers and valvometry." *Environmental Monitoring and Assessment*, **188**(7), 1–8.

El Methni J, Gardes L, Girard S, Guillou A (2012). "Estimation of extreme quantiles from heavy and light tailed distributions, Journal of Statistical Planning and Inference." *Journal of Statistical Planning and Inference*, **142**(10), 2735–2747.

Embrechts P, Resnick SI, Samorodnitsky G (1999). "Extreme value theory as a risk management tool." *North American Actuarial Journal*, **3**(2), 30–41.

Gardes L, Girard S (2010). "Conditional extremes from heavy -tailed distributions: an application to the estimation of extreme rainfall return levels." *Extremes*, **13**, 177–204.

Gencay R, Selcuk F (2004). "Extreme value theory and Value-at-Risk: Relative performance in emerging markets." *International Journal of Forecasting*, **20**(2), 287–303.

Gilleland E (2011). *extRemes: Extreme Value Analysis*. R package version 2.0-5, URL https://cran.r-project.org/web/packages/extRemes/index.html.

Gilleland E, Ribatet M, Stephenson A (2013). "A software review for extreme value analysis." *Extremes*, **16**(1), 103–119.

Grama I, Spokoiny V (2007). "Pareto approximation of the tail by local exponential modeling." *Bulletin of Academi of Sciences of Moldova*, **53**(1), 1–22.

Grama I, Spokoiny V (2008). "Statistics of extremes by oracle estimation." *Annals of Statistics*, **36**(4), 1619–1648.

Guillou A, Hall P (2001). "A diagnostic for selecting the threshold in extreme-value analysis." *J. Roy. Statist. Soc. Ser. B*, **63**, 293–305.

Hall P, Welsh AH (1985). "Adaptive Estimates of Parameters of Regular Variation." *Ann. Statist.*, **13**(1), 331–341.

Harris JL, Liu LM (1993). "Dynamic structural analysis and forecasting of residential electricity consumption." *International Journal of Forecasting*, **9**(4), 437–455.

Hofert M, Kojadinovic I, Maechler M, Yan J (2010). *copula: Multivariate Dependence with Copulas*. R package version 0.999-14, URL http://copula.r-forge.r-project.org/.

Hosking JRM (2009). *lmom: L-moments*. R package version 2.5, URL https://cran.r-project.org/web/packages/lmom/index.html.

Huisman R, Koedijk CG, Kool CJM, Palm F (2001). "Tail index estimates in small samples." *Journal of Business and Economic Statistics*, **19**(2), 208–216.

Jou LJ, Liao CM (2006). "A dynamic artificial clam (Corbicula fluminea) allows parcimony on-line measurement of waterborne metals." *Environmental Pollution*, **144**, 172–183.

Katz RW, Parlange MB, Naveau P (2002). "Statistics of extremes in hydrology." *Advances in water resources*, **25**(8), 1287–1304.

Loader C (1999). *Local regression and likelihood*. Springer.

McNeil AJ (1997). "Estimating the tails of loss severity distributions using extreme value theory." *ASTIN bulletin*, **27**(01), 117–137.

McNeil AJ (1998). "Calculating quantile risk measures for financial return series using extreme value theory." *ETH Zürich, Departement Mathematik*.

Pfaff B, McNeil A, Stephenson A (2008). *evir: Extreme Values in R*. R package version 1.7-3, URL https://cran.r-project.org/web/packages/evir/index.html.

R Development Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

Ranjan M, Jain V (1999). "Modelling of electrical energy consumption in Delhi." *Energy*, **24**(4), 351–361.

Ribatet M, Singleton R, team RC (2011). *SpatialExtremes: Modelling Spatial Extremes*. R package version 2.0-2, URL http://spatialextremes.r-forge.r-project.org/.

Rootzén H, Tajvidi N (1997). "Extreme value statistics and wind storm losses: a case study." *Scandinavian Actuarial Journal*, **1997**(1), 70–94.

Schmitt FG, De Rosa M, Durrieu G, Sow M, Ciret P, Tran D, Massabuau JC (2011). "Statistical study of bivalve high frequency microclosing behavior: scaling properties and shot noise modeling." *International Journal of Bifurcation and Chaos*, **21**(12), 3565–3576.

Sharma P, Khare M, Chakrabarti S (1999). "Application of extreme value theory for predicting violations of air quality standards for an urban road intersection." *Transportation Research Part D: Transport and Environment*, **4**(3), 201–216.

Simiu E, Heckert N (1996). "Extreme wind distribution tails: a peaks over threshold approach." *Journal of Structural Engineering*, **122**(5), 539–547.

Sornette D, Knopoff L, Kagan Y, Vanneste C (1996). "Rank-ordering statistics of extreme events: Application to the distribution of large earthquakes." *Journal of Geophysical Research: Solid Earth*, **101**(B6), 13883–13893.

Southworth H, Heffernan JE (2010). *texmex: Statistical modelling of extreme values*. R package version 2.1, URL http://code.google.com/p/texmex/.

Sow M, Durrieu G, Briollais L (2011). "Water quality assessment by means of HFNI valvometry and high-frequency data modeling." *Environmental Monitoring and Assessment*, **182**(1-4), 155–170.

Staniswalis J (1989). "The kernel estimate of a regression function in likelihood-based models." *Journal of the American Statistical Association*, **84**(405), 276–283.

Stephenson A, Ferro C (2002). *evd: Functions for extreme value distributions*. R package version 2.3-0, URL https://cran.fhcrc.org/web/packages/evd/index.html.

Stephenson A, Ribatet M (2010). *evdbayes: Bayesian Analysis in Extreme Value Theory*. R package version 1.1-1, URL https://cran.r-project.org/web/packages/evdbayes/index.html.

Tran D, Ciret P, Ciutat A, Durrieu G, Massabuau JC (2003). "Estimation of potential and limits of bivalve closure response to detect contaminants: application to cadmium." *Environmental Toxicology and Chemistry*, **22**(4), 914–920.

Wuertz D, *et al.* (2013). *fExtremes: Rmetrics - Extreme Financial Market Data*. R package version 3010.81, URL https://cran.r-project.org/web/packages/fExtremes/index.html.

Affiliation:

Gilles Durrieu, Ion Grama, Kevin Jaunatre, Jean-Marie Tricot
Laboratoire de Mathématiques de Bretagne Atlantique
Université de Bretagne Sud and UMR CNRS 6205
Campus de Tohannic, BP573, 56000 Vannes, France
Email : gilles.durrieu@univ-ubs.fr, ion.grama@univ-ubs.fr, kevin.jaunatre@univ-ubs.fr, jean-marie.tricot@univ-ubs.fr

Quang-Khoai Pham
Department of Mathematics
Forestry University of Hanoi
Hanoi, Vietnam
Email : quangkhoaihd@gmail.com