REVSTAT – Statistical Journal
Volume 17, Number 2, April 2019, 187-207

TESTING CONDITIONS AND ESTIMATING PARAMETERS IN EXTREME VALUE THEORY: APPLICATION TO ENVIRONMENTAL DATA

Authors: Helena Penalva
– Escola de Ciências Empresariais do Instituto Politécnico de Setúbal, and CEAUL, Universidade de Lisboa, Portugal
helena.penalva@esce.ips.pt

Dora Prata Gomes
– Faculdade de Ciências e Tecnologia and CMA/FCT, Universidade Nova de Lisboa, Portugal
dsrp@fct.unl.pt

M. Manuela Neves
– Instituto Superior de Agronomia, and CEAUL, Universidade de Lisboa, Portugal
manela@isa.ulisboa.pt

Sandra Nunes
– Escola de Ciências Empresariais do Instituto Politécnico de Setúbal, and CMA/FCT, Universidade Nova de Lisboa, Portugal
sandra.nunes@esce.ips.pt

Received: October 2018 Revised: January 2019 Accepted: March 2019

Abstract:

- *Extreme Value Theory* has been asserting itself as one of the most important statistical theories for the applied sciences providing a solid theoretical basis for deriving statistical models describing extreme or even rare events. The efficiency of the inference and estimation procedures depends on the tail shape of the distribution underlying the data. In this work we will present a review of tests for assessing extreme value conditions and for the choice of the extreme value domain. Motivated by two real environmental problems we will apply those tests showing the need of performing such tests for choosing the most appropriate parameter estimation methods.

Key-Words:

- *Environmental data; extreme values; heavy-tailed distributions; semi-parametric estimation; statistical testing.*

AMS Subject Classification:

- 62G32, 62E20, 62G10.

1. MOTIVATION AND INTRODUCTION

Extreme Value Theory (EVT) is concerned with the behaviour of extreme values, i.e, values occurring at the tails of a probability distribution. Society, human life, etc. tend to adapt to near-normal conditions, and these conditions tend to produce fairly minimal impacts. In contrast, unusual and extreme conditions can have a substantial impact despite, by definition, occurring in a very low proportion of times. EVT is the branch of probability and statistics dedicated to characterizing the very low or quite high values of a variable, *the tail of the distribution*. EVT had its beginnings in the early to middle part of XX century and Emil Gumbel was the pioneer in applications of statistics of extremes. In *Statistics of Extremes* [23], he presents several applications of EVT on real world problems in engineering and in meteorological phenomena. In this book appear the first applications in hydrology.

Results in EVT rely on certain assumptions. However in some situations they can be not fulfilled. So, before dealing with an application, it is important to have an *a priori* knowledge on whether the underlying distribution verifies those assumptions. On the other hand statistical inference procedures should be performed according to the most adequate domain of attraction for the underlying distribution. So, tests for extreme value conditions and for the choice of the tail must be done before the application of any inferential procedure.

The motivation for this work came from a first study in Neves *et al.* [34] and Penalva *et al.* [36] presenting a review of tests and parameter estimation procedures applied to the daily mean flow discharge rate in the hydrometric station of Fragas da Torre in the river Paiva. The data were collected from 1946/47 to 2005/2006, i.e., 60 years of data. In Penalva *et al.* [36] we drew the attention for the need of a previous analysis for assessing extreme value conditions and for the choice of the extreme value domain, in order to choose the more adequate parameter estimators. We will review briefly the analysis already performed considering the data now available during 66 years, 1946/2012 and using, for comparison, two recent classes of estimators of the tail index of the extreme value distribution, introduced in Penalva *et al.* [37] and Gomes *et al.* [21].

The procedures proposed are also applied and commented to another data set referring to burned areas of wildfires in Portugal during 33 years (1984–2016).

So, the aim of this work is to perform an univariate extreme value analysis illustrating and reviewing tests on the extreme value condition and on the statistical choice of the tail of the underlying distribution. This should be the first step in order to choose the more adequate estimators. Some recent estimators of the tail index are also compared.

The paper proceeds as follows. Section 2 contains the main results that are the basis of the theoretical background. In Section 3 the exploratory analysis of the first case-study aforementioned is performed, parametric and semi-parametric statistical approaches in EVT are briefly reviewed and first estimates of the main parameters are presented. In Section 4, statistical testing procedures for extreme value conditions and for choosing the tail are presented and applied to the data. Section 5 is dedicated to perform the study and estimation in a second case-study, showing the adequate procedure of performing the study. Finally Section 6 presents a first practical application on the effect of taking into consideration or not the choice of the tail of the underlying distribution and consequently the adequate EVI estimation. For the first case study, where estimation discrepancies were detected when the choice of the tail was made previously or not, high quantiles are estimated. A few comments on some other parameters that could be considered and the work in progress finish this section.

2. THEORETICAL BACKGROUND

Let us assume that we have a sample $(X_1, \ldots, X_n)$ of independent and identically distributed (iid) or possibly stationary, weakly dependent random variables from an unknown cumulative distribution function (cdf) F. Let us consider the notation $(X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n})$ for the sample of ascending order statistics associated to that sample.

The interest is focused on the distribution of the maxima, that is, $M_n := \max(X_1, \ldots, X_n)$, for which we have

$$\begin{aligned} \mathbb{P}\left(M_n \leq x\right) &= \mathbb{P}\left(X_1 \leq x, \ldots, X_n \leq x\right) \\ &= \mathbb{P}\left(X_1 \leq x\right) \ldots \mathbb{P}\left(X_n \leq x\right) = F^n(x). \end{aligned} \tag{2.1}$$

We often deal with the maxima, given the “kind of symmetry”, $\min\big(X_1, \ldots, X_n\big) = -\max\big(-X_1, \ldots, -X_n\big)$.

This problem has similarities to that one of determining the distribution of $S_n = \sum_{i=1}^n X_i$. Obviously S_n and possibly M_n may tend to infinity, and their distribution is a degenerate one. The *central limit theorem* gives an answer to this problem under some conditions, showing that the normal distribution is obtained as the non-degenerate limit of S_n properly normalized by $E[S_n]$ and $\sqrt{Var[S_n]}$.

As n goes to ∞, the distribution F^n in (2.1) has a trivial limit: 0, if $F(x) < 1$ and 1, if $F(x) = 1$. So the idea for M_n was the same: first subtract a $n-$dependent constant, then rescale by a $n-$dependent factor. The first question is then whether one can find two sequences, $\{a_n\} \in \mathbb{R}^+$ and $\{b_n\} \in \mathbb{R}$ and a non-trivial distribution function, G, such that $\lim_{n\to\infty} \mathbb{P}\left((M_n - b_n)/a_n \leq x\right) = G(x)$.

First results on the G distribution are due to Fréchet [17], Fisher and Tippet [12], Gumbel [22] and von Mises [40]. But were Gnedenko [19] and de

Haan [24] who gave conditions for the existence of those sequences $\{a_n\} \in \mathbb{R}^+$ and $\{b_n\} \in \mathbb{R}$ such that when $n \to \infty$ and $\forall x \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{M_n - b_n}{a_n} \le x\right) = \lim_{n\to\infty} F^n(a_n x + b_n) = \mathrm{EV}_\xi(\mathrm{x}). \tag{2.2}$$

EV_ξ is a nondegenerate distribution function, denoted as the Extreme Value cdf, given by

$$\mathrm{EV}_\xi(\mathrm{x}) = \begin{cases} \exp[-(1+\xi x)^{-1/\xi}], & 1+\xi x > 0 \ \text{ if } \xi \ne 0 \\ \exp[-\exp(-x)], & x \in \mathbb{R} \qquad\ \ \text{if } \xi = 0. \end{cases} \tag{2.3}$$

When the above limit holds we say that F is in the domain of attraction (for maxima) of EV_ξ and write $F \in \mathcal{D}_{\mathcal{M}}(\mathrm{EV}_\xi)$.

The shape parameter ξ, in (2.3), is called the *extreme value index* (EVI) and it is the primary parameter of interest in EVT analysis. The EV_ξ incorporates the three (Fisher-Tippett) types: Gumbel, with $\xi = 0$, the right tail of F is of an exponential type; Fréchet with $\xi > 0$, the right tail is heavy, of a negative polynomial type, and F has an infinite right endpoint and Weibull with $\xi < 0$, the right tail is light, and F has a finite right endpoint ($x^* < +\infty$).

These models can also incorporate location (λ) and scale ($\delta > 0$) parameters, and in this case, the EV cdf is given by,

$$\mathrm{EV}_\xi(x;\lambda,\delta) \equiv \mathrm{EV}_\xi((x-\lambda)/\delta). \tag{2.4}$$

We may then consider, when the sample size $n \longrightarrow \infty$, the approximation

$$P[M_n \le x] = F^n(x) \approx \mathrm{EV}_\xi((x-b_n)/a_n).$$

3. FIRST CASE-STUDY – A REVIEW

The source of river Paiva is in the Serra de Leomil in the North of Portugal and it is a tributary of the river Douro, with a watershed area of approximately 700 Km. The discharge rate study of this river is a matter of major importance since it is one of the main alternatives to the river Douro as source of water supply in the south of Oporto region. The data are daily mean flow discharge rate values (m^3/s) from 1 October, 1946 to 30 September, 2012 - collected from the “SNIRH: Sistema Nacional de Informação dos Recursos Hídricos”.

The descriptive study of these data revealed a tail heavier than that of the normal. Results in Table 1 are similar to those in Penalva *et al.* [36].

EVT has been developed under two frameworks. The first one is the parametric framework, that considers a class of models associated to the limiting

min	1st Qu.	Median	Mean	3rd Qu.	max
0.00	9.11	17.1	34.4	37.3	920.0

n	Skewness	Kurtosis	St Dev
11946	4.14	27.13	50.26

Table 1: Descriptive statistics for daily mean flow discharge rate values.

behaviour of the maxima, given in (2.2). The main assumption behind the parametric approach is that estimators are calculated considering the data following, approximately, an exact EV probability distribution function, defined by a number of parameters. In this approach several methodologies have been developed for estimating parameters: Block Maxima; Largest Observations; Peaks Over Threshold, to refer the most well known.

In the semi-parametric framework, the only assumption made is that the limit in (2.2) holds, i.e., that the underlying distribution verifies the extreme value condition. The EVI, ξ, that appears in (2.3), plays the central role in this framework. Under this approach several EVI-estimators have been developed. Some of the most relevant and also the most recent ones will be used here in the estimation.

As an illustration of parametric approaches to estimate EVT parameters, only the *Block Maxima* (BM) approach will be considered in this work. Other procedures can be seen in Penalva *et al.* [36].

3.1. The Block Maxima (BM) method

The so-called *Block Maxima* (BM), *Annual Maxima* or *Gumbel's* method is the first parametric approach for modelling extremes, Gumbel [23]. In this approach the $n-$sized sample is splitted into m sub-samples (usually m corresponds to the number of the observed years) of size l ($n = m \times l$) for a sufficiently large l. EV_ξ or one of the models, Gumbel, Fréchet or Weibull, with unknown $\xi \in \mathbb{R}$, $\lambda \in \mathbb{R}$ or $\delta \in \mathbb{R}^+$ are then fitted to the m maxima values of the m sub-samples.

Table 2 and Figure 1 show a very light positive asymmetry and kurtosis. It is also reasonable to consider data not correlated.

min	1st Qu.	Median	Mean	3rd Qu.	max
32.2	177.25	261.5	279.24	371.5	920.0

m	Skewness	Kurtosis	St Dev
66	0.99	2.308	157.17

Table 2: Basic descriptive statistics for the maximum values in each year.

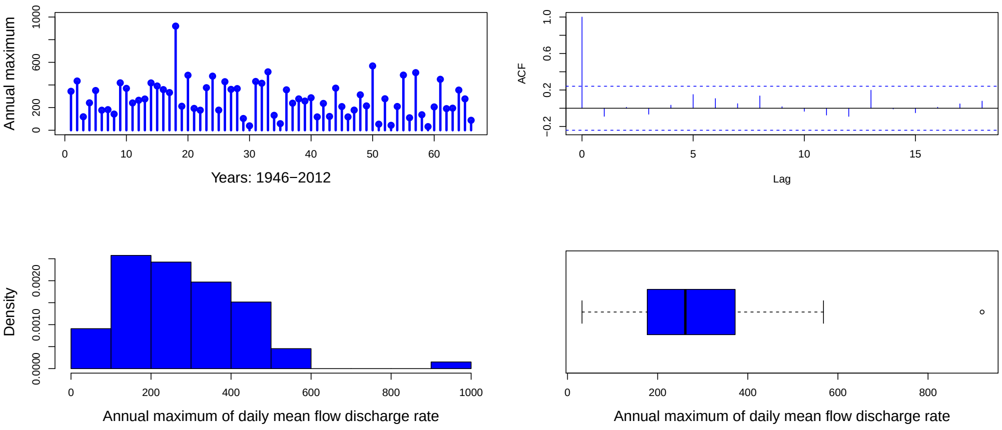

Figure 1: Plots of the maximum value in each year, the partial autocorrelation function, the histogram and the boxplot.

Maximum likelihood estimates and standard errors were easily obtained using the `evd` package in R software [38].

$\hat{\xi}$	$\hat{\lambda}$	$\hat{\delta}$
-0.03 (0.08)	207.74 (17.52)	127.11 (12.72)

Table 3: Maximum likelihood estimates and standard errors (in parenthesis).

3.2. Semi-parametric estimators

In this framework we do not need to fit a specific parametric model based on scale, shape and location parameters. We construct an EVI-estimator based on the largest k top observations, with k intermediate, i.e. such that $k = k_n \to \infty$ and $k/n \to 0$, as $n \to \infty$, assuming only that the model F underlying the data is in $\mathcal{D}_{\mathcal{M}}(\mathrm{EV}_\xi)$, in specific sub-domains of $\mathcal{D}_{\mathcal{M}}(\mathrm{EV}_\xi)$, with EV_ξ provided in (2.3).

Most estimators show a strong dependence on that value k. They usually present: a small bias and a high variance for small values of k;bias increases and variance decreases when k increases; the need of looking for an adequate value of k for which we have a minimum Mean Square Error. Thus, an intensive research has been performed trying to obtain estimators overcoming these difficulties. Currently there are several different EVI-estimators, so we decide to present and compare here a very few. Here we will illustrate the application of the following estimators: the classical Hill estimator, Hill [27], and a recent class of estimators, the *Lehmer mean-of-order-p* (L_p) estimators, Penalva *et al.* [37] and Penalva [35], both defined for $\xi > 0$. Two of the estimators developed for $\xi \in \mathbb{R}$ are here considered: the Moment estimator, Dekkers *et al.* [8] and the Mixed Moment estimator, Fraga Alves *et al.* [16].

Recently, Caeiro *et al.* [4] introduced a class of *reduced bias* EVI-estimators. This class can not only reduce the bias of the classical estimators but also do not increase the asymptotic variance of the estimators, for adequate levels of k and adequate estimation of parameters of second-order $(\beta, \rho) \in (\mathbb{R}, \mathbb{R}^-)$. These are the scale and the shape second-order parameters, controlling the rate of first-order convergence, and necessary for establishing distributional properties of the estimators. Details on second-order conditions can be found in Beirlant *et al.* [2], de Haan and Ferreira [25] and Fraga Alves *et al.* [15], among others. Those estimators are then denoted *minimum-variance reduced biased* (MVRB) EVI-estimators. We will consider two of those estimators, one based on the Hill and the other on the L_p estimators, see Gomes *et al.* [21].

Let $X_{1:n} \le X_{2:n} \le \cdots \le X_{n:n}$ be the order statistics associated to the sample $(X_1, X_2, \ldots, X_n)$.

Let us define the log-excesses as $V_{ik} := \ln X_{n-i+1:n} - \ln X_{n-k:n}$, and $\mathrm{M}_{k,n}^{(l)} := \frac{1}{k}\sum_{i=1}^{k} [V_{ik}]^l$, for $l \in \mathbb{R} \setminus \{0\}$, and $\mathrm{L}_{k,n}^{(r)} := \frac{1}{k}\sum_{i=1}^{k} \left[1 - \frac{X_{n-k:n}}{X_{n-i+1:n}}\right]^r$, for $r \ge 1$.

The aforementioned estimators have the functional definitions:

- The Hill estimator, H, defined for $\xi > 0$, as

$$(3.1) \qquad \widehat{\xi}^{\mathrm{H}}(k) \equiv \mathrm{H}(k) := \frac{1}{k}\sum_{i=1}^{k} V_{ik}, \quad k = 1, 2, \ldots, n-1.$$

- The Moment estimator, M, defined for $\xi \in \mathbb{R}$, as

$$(3.2) \qquad \widehat{\xi}_{k,n}^{M} := M_{k,n}^{(1)} + 1 - \frac{1}{2}\left(1 - \frac{(M_{k,n}^{(1)})^2}{M_{k,n}^{(2)}}\right)^{-1}, \quad k = 1, 2, \ldots, n-1.$$

- The Mixed Moment estimator, MM, defined for $\xi \in \mathbb{R}$, as

$$(3.3) \qquad \widehat{\xi}_{k,n}^{MM} := \frac{\widehat{\varphi}_{k,n} - 1}{1 + 2\min(\widehat{\varphi}_{k,n} - 1, 0)}, \quad k = 1, 2, \ldots, n-1,$$

where

$$\widehat{\varphi}_{k,n} := \frac{M_{k,n}^{(1)} - L_{k,n}^{(1)}}{\left(L_{k,n}^{(1)}\right)^2}.$$

- The class of *Lehmer mean-of-order-p* (L_p) estimators, defined for $\xi > 0$ and

$p > 0$, as
(3.4)

$$\widehat{\xi}^{\mathrm{L}}(k) \equiv \mathrm{L}_p(k) := \frac{1}{p}\,\frac{\sum\limits_{i=1}^{k} V_{ik}^{p}}{\sum\limits_{i=1}^{k} V_{ik}^{p-1}}, \quad k = 1, 2, \ldots, n-1, \quad \big[\mathrm{L}_1(k) \equiv \mathrm{H}(k)\big].$$

- The class of *corrected-Hill* (CH) EVI-estimators, defined by

  $$\text{(3.5)} \qquad \mathrm{CH}(k) := \mathrm{H}(k)\Big(1 - \hat{\beta}(n/k)^{\hat{\rho}}/(1-\hat{\rho})\Big), \quad k = 1, 2, \ldots, n-1,$$

  where $H(k)$ is the Hill estimator and $\hat{\beta}$ e $\hat{\rho}$ are consistent estimators of parameters β e ρ. The use of CH(k) enables us to eliminate the dominant component of bias of the H EVI-estimator, H(k), keeping its asymptotic variance.

- More generally than the class in (3.5), we shall now also consider the direct reduction of the dominant bias component of $\mathrm{L}_p(k)$, in (3.4), working with the RB Lehmer's EVI-estimators, Gomes *et al.* [21], defined by

  $$\text{(3.6)} \qquad \mathrm{L}_p^{\mathrm{RB}}(k) := \mathrm{L}_p(k)\Big(1 - \hat{\beta}(n/k)^{\hat{\rho}}/(1-\hat{\rho})^{p}\Big), \quad k = 1, 2, \ldots, n-1,$$

  $[\mathrm{L}_1^{\mathrm{RB}} \equiv \mathrm{CH} \ \text{ in } \ (3.5)]$

Figure 2 shows the sample paths of estimates obtained when using the aforementioned estimators.

Values of p in $\mathrm{L}_p(k)$ and $\mathrm{L}_p^{\mathrm{RB}}(k)$ were chosen using criteria given in Penalva [35].

The discrepancies observed, already noticed in Penalva *et al.* [36], regarding the results of the above EVI-estimators and also compared with the results obtained under the parametric approaches claim for tests on extreme value domain of attraction. This emphasizes the care to be taken with the choice of the estimators, because even having very nice and stable paths, if conditions of their applicability are not verified, they may not stabilize near the true value of the parameter.

4. TESTING CONDITIONS IN EVT LIMITING RESULTS

In any of the above procedures it is assumed that the underlying cdf F belongs to $\mathcal{D}_{\mathcal{M}}(EV_{\xi})$, for a appropriate value of ξ, or it is in specific sub-domains of $\mathcal{D}_{\mathcal{M}}(EV_{\xi})$. This condition is known as the *extreme value condition.*

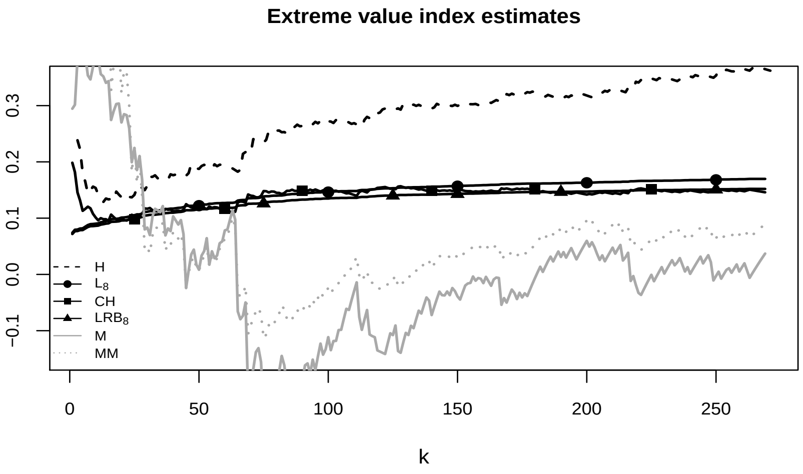

Figure 2: Sample paths of the EVI-estimates considered.

4.1. Testing the extreme value condition

It is then important, before any application, to check the assumption:

$$H_0 : F \in \mathcal{D}_{\mathcal{M}}(EV_\xi) \;\text{ for some }\; \xi \in \mathbb{R}. \tag{4.1}$$

Some tests for the hypothesis H_0 are available, such as those in Dietrich *et al.* [9], Drees *et al.* [10] and Hüsler and Li [28].

Let $X_1, X_2, \ldots, X_n$ be iid random variables with cdf F and suppose that some additional second order conditions hold then, for $\eta > 0$, Dietrich *et al.* [9] introduced the test statistic written as

$$E_n := k \int_0^1 \left(\frac{\log X_{n-\lfloor kt \rfloor,n} - \log X_{n-k,n}}{\widehat{\xi}_+} - \frac{t^{-\widehat{\xi}_-} - 1}{\widehat{\xi}_-} \left(1 - \widehat{\xi}_-\right) \right)^2 t^\eta dt, \tag{4.2}$$

where k is again an intermediate sequence such that $k = k_n \to \infty$, $k/n \to 0$ and $k^{1/2}A(n/k) \to 0$ as $n \to \infty$ and A is related to the second order condition already referred to and $\widehat{\xi}_+$ and $\widehat{\xi}_-$ are the moment estimators, Dekkers *et al.* [8], of $\xi_+ := \max(0, \xi)$ and $\xi_- := \min(0, \xi)$.

Hüsler and Li [28] present an algorithm for testing H_0 using the test statistic E_n in (4.2). They have carried out an extensive simulation study with guidelines for obtaining the value of η and have provided quite accuracy tables for the quantiles $\chi_{1-\alpha}$ of the variable limiting of E_n, see Hüsler and Li [28] for details. Values of E_n are compared with values of $\chi_{1-\alpha}$: if $E_n > \chi_{1-\alpha}$ hypothesis H_0 is rejected with a type I error α. Otherwise there is no reason to reject H_0.

For our data, the application of the test based on (4.2), provided values of the test statistic smaller than the corresponding asymptotic $0.95-$quantile for a large range of $k-$values. So, since the sample path of test statistic is almost always outside the rejection region, except for a small range of k, we find no evidence to reject the null hypothesis, see Figure 3.

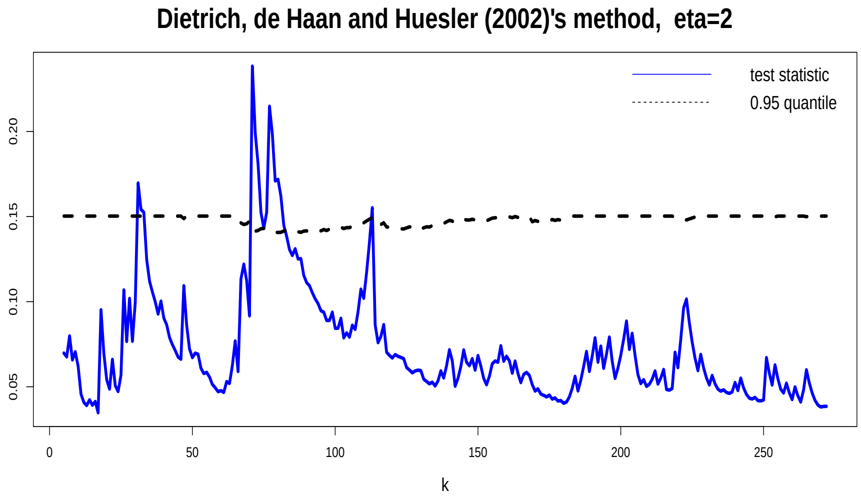

Figure 3: Plot of the sample paths for the E-test, based on the test statistic in (4.2) and the corresponding quantile. Available sample size $n = 11946$.

See also Hüsler and Li [28], Neves and Fraga Alves [32] and Penalva *et al.* [36] for a description of other tests.

4.2. Statistical choice of extreme domains of attraction

Once the hypothesis $H_0 : F \in \mathcal{D}_{\mathcal{M}}(EV_\xi)$ is not rejected, it is of major importance to decide for the type of the tail, i.e., the natural hypothesis testing are now:

$$(4.3) \qquad H_0 : F \in \mathcal{D}_{\mathcal{M}}(EV_0) \quad vs \quad H_1 : F \in \mathcal{D}_{\mathcal{M}}(EV_\xi)_{\xi \neq 0},$$

or against the one-sided alternatives

$$F \in \mathcal{D}_{\mathcal{M}}(EV_\xi)_{\xi<0} \quad \text{or} \quad F \in \mathcal{D}_{\mathcal{M}}(EV_\xi)_{\xi>0}.$$

This is known as the *statistical choice of extreme domains of attraction.*

Under the semiparametric framework, several tests have been proposed in literature, among which we can mention: Galambos [18], Castillo *et al.* [5];

Hasofer and Wang [26]; Falk [11]; Correia and Neves [7], that considered the Hasofer and Wang statistic and presented a slight modification. An extensive simulation study has been performed in Fraga Alves and Gomes [13], Marohn [29, 30], Fraga Alves [14] and Segers and Teugels [39]. Castillo *et al.* [5] considered tests to distinguish between polynomial and exponential tails, based on properties of the *coefficient of variation* (CV).

Neves and Fraga Alves [32, 33] studied the following tests statistics, that will be here applied.

The **Ratio-test**:

$$R_n^*(k) := \frac{X_{n:n} - X_{n-k:n}}{\frac{1}{k}\sum_{i=1}^{k}\left(X_{n-i+1:n} - X_{n-k:n}\right)} - \log k \xrightarrow[n\to\infty]{d} \mathrm{EV}_0. \tag{4.4}$$

The **Gt-test**:

$$G_n(k) := \frac{\frac{1}{k}\sum_{i=1}\left(X_{n-i+1:n} - X_{n-k:n}\right)^2}{\left(\frac{1}{k}\sum_{i=1}^{k} X_{n-i+1:n} - X_{n-k:n}\right)^2}, \tag{4.5}$$

and

$$G_n^*(k) = \sqrt{k/4}\,\left(G_n(k) - 2\right) \xrightarrow[n\to\infty]{d} N(0,1).$$

The **HW-test**:

$$W_n(k) := \frac{1}{k}\left[1 - \frac{G_n(k) - 2}{1 + \left(G_n(k) - 2\right)}\right], \tag{4.6}$$

and

$$W_n^*(k) = \sqrt{k/4}\,\left(kW_n(k) - 1\right) \xrightarrow[n\to\infty]{d} N(0,1).$$

For the two-sided tests R^*, G^* or W^*, the null hypothesis is rejected if $R^*(G^*)(W^*) < \chi_{\alpha/2}$ or $R^*(G^*)(W^*) > \chi_{1-\alpha/2}$, where χ_p is the p probability quantile of the corresponding limiting distribution.

For the one-sided tests, the null hypothesis is rejected in favour of either unilateral alternatives, for example, for R_n^*,

$$H_1^l : F \in D_M(EV_\xi)_{\xi<0} \quad \text{or} \quad H_1^r : F \in D_M(EV_\xi)_{\xi>0},$$

if

$$R_n^*(k) < \chi_\alpha \quad \text{or} \quad R_n^*(k) > \chi_{1-\alpha}.$$

Figure 4 illustrates the application of those tests.

These tests suggest the non rejection of the null hypothesis, leading us to consider that the underlying distribution of the data are in the domain of attraction of the Gumbel distribution.

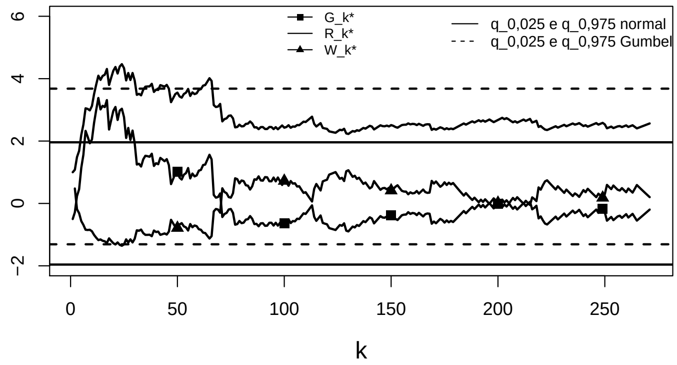

Figure 4: Sample paths of the statistics R_n^*, with the associated quantiles $\chi_{0.025}$ and $\chi_{0.975}$ for the standard Gumbel distribution in dashed lines, and the G_n^* and W_n^* sample paths statistics, with the associated quantiles of the standard normal distribution in solid lines.

Such as we have already pointed out in Penalva *et al.* [36], with fewer years of data, we think that this explains the discrepancy observed in Figure 2, where were plotted sample paths of very well behaved EVI-estimators, but not adequate to the tail of the data under study. We claim again for the need of performing at first the tests described and illustrated briefly in this Section.

5. SECOND CASE-STUDY – THE ANALYSIS

The second set of data analysed in this work, and also studied in Gomes *et al.* [20] based on a shorter period of time, consists of the burned area (ha), in Portugal, related to each of the wildfires occurred in a period from 1984 to 2016, exceeding 100 ha, making a total of 6507 observations. The data analysed here do not seem to have a significant temporal structure. This new data set is used to illustrate what we have just commented.

The main results of a graphical and descriptive analysis are shown in Table 4 and in Figure 5. Tables and graphics provide evidence on the heaviness of the right tail. Notice that similar conclusions were obtained by Beirlant *et al.* [1], for data analysis of burned area of wildfires exceeding 100 ha, recorded in Portugal from 1990 till 2003 ($n = 2627$).

See in Figure 6 the application of the test to the extreme value condition, based on (4.2). We find no evidence to reject the null hypothesis, i.e., $F \in \mathcal{D}_{\mathcal{M}}(EV_\xi)$.

min	1st Qu.	Median	Mean	3rd Qu.	max
100	138.55	215.81	485.35	427.51	58012.75

n	Skewness	Kurtosis	St Dev
6507	19.01	568.90	1407.58

Table 4: Descriptive statistics for burned area of wildfires exceeding 100ha.

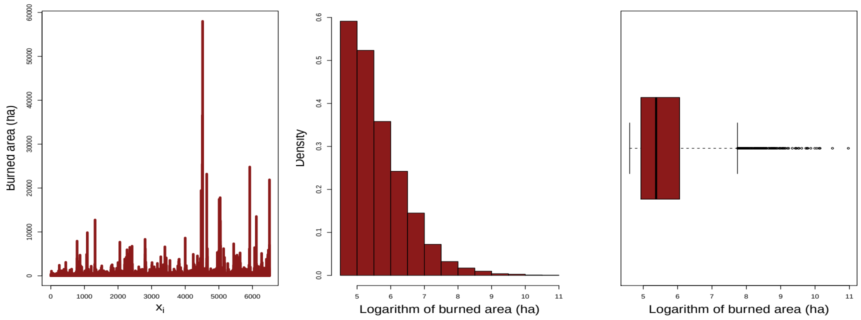

Figure 5: Plot of burned areas, histogram and boxplot, for wildfires, exceeding 100 ha.

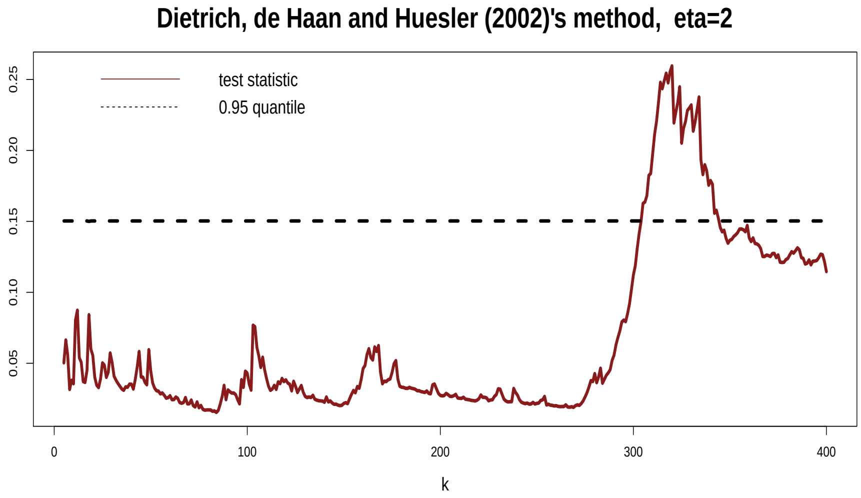

Figure 6: Plot of the sample paths for the E-test, based on (4.2) statistics, with the corresponding quantile. Available sample size $n = 6507$.

The tests to the statistical choice of the tail, such as was described and presented in Subsection 4.2, produced now the plots presented in Figure 7. Those tests suggest the rejection of the null hypothesis, leading us to consider that the underlying distribution of the data is in the domain of attraction of the Fréchet distribution.

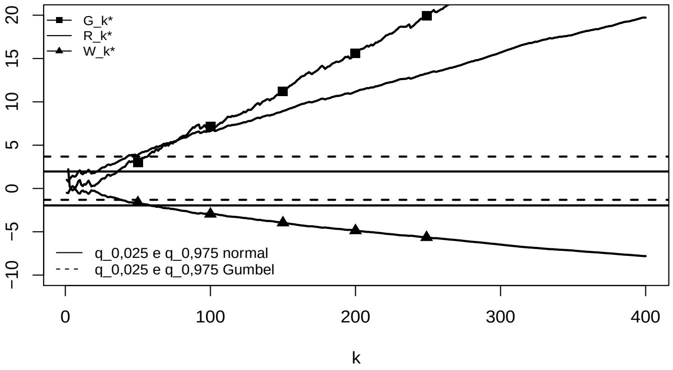

Figure 7: Sample paths of the statistics R_n^*, with the associated quantiles $\chi_{0.025}$ and $\chi_{0.975}$ for the standard Gumbel distribution in dashed lines, and the G_n^* and W_n^* sample paths statistics, with the associated quantiles of the standard normal distribution in solid lines.

Here we will consider again, in the BM methodology, blocks as the years of observations, $m = 33$. Figure 8 and Table 5 were obtained for the burned area of wildfires exceeding 100 ha.

min	1st Qu.	Median	Mean	3rd Qu.	max
641.33	2860.10	6235.83	8956.80	8652.43	58012.75
	m	Skewness	Kurtosis	St Dev	
	33	2.90	9.85	10889.31	

Table 5: Basic descriptive statistics for maximum values in each year.

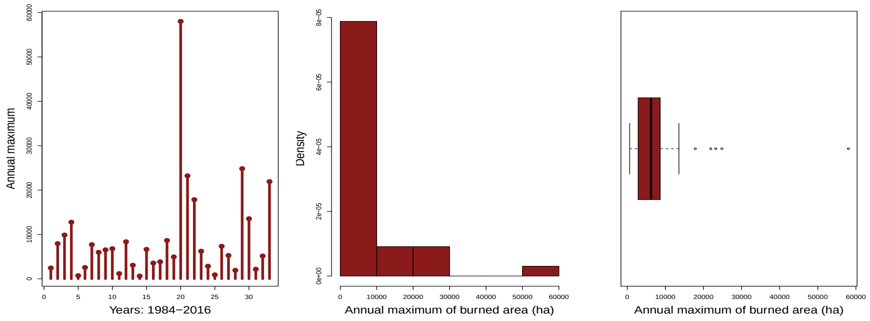

Figure 8: Maximum value of burned areas in each year, histogram and boxplot.

Below are given the estimates of the main parameters.

$\hat{\xi}$	$\hat{\lambda}$	$\hat{\delta}$
0.52 (0.21)	4007.95 (754.45)	3599.50 (727.93)

Table 6: Maximum likelihood estimates (standard errors in parenthesis).

The ξ estimate corroborates the first idea pointing that the data present clearly a tail heavier than that one of the first case-study.

Figure 9 shows the sample paths of estimates obtained using the aforementioned estimators. Values of p in $\mathrm{L}_p(k)$ and $\mathrm{L}_p^{\mathrm{RB}}(k)$ were also chosen using criteria given in Penalva [35]. A quick analysis of the sample paths of the EVI-estimates allow us to consider as $\widehat{\xi}$ a value between 0.55 and 0.65, which is also in agreement with a heavy tail detected for the underlying cdf F and with the result obtained under the parametric approach.

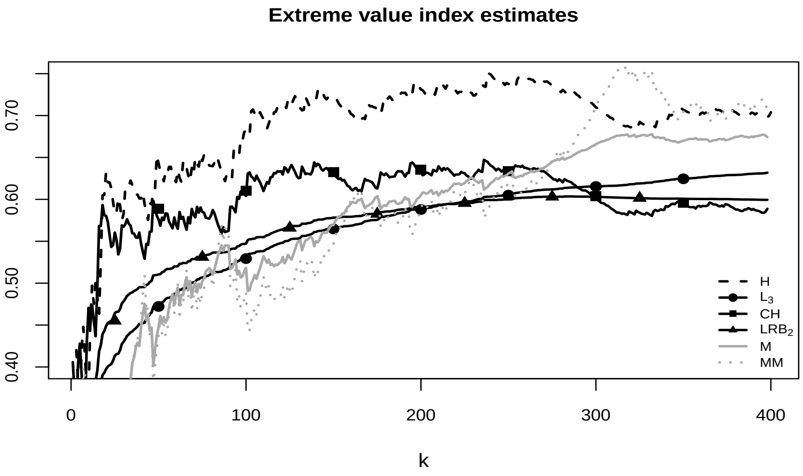

Figure 9: Sample paths of the EVI-estimates considered.

6. FIRST COMMENTS ON PRACTICAL EFFECTS OF MISSING THOSE TESTS. A FEW COMMENTS

We showed, with this work based on the two case-studies, that the realization of tests on the extreme value conditions and on the statistical choice of the tail of the underlying distribution are with no doubt the first step to properly apply the several estimation approaches and to choose the more adequate estimators.

A first illustration of the practical effects in the estimation of other important parameters when the choice of the tail is performed or not *a priori*, is presented. It is well known how an accurate EVI estimation is important because it dominates the tail behaviour of a distribution. However in several situations, such as risk management or catastrophic situations, where human lives can be in danger, in addition to modelling the tails, other parameters are of the major importance to be estimated, such as extreme quantiles, return levels or return periods of the distribution of the process at risk. For the first case study, high quantiles were estimated.

While it is true that EVI determines the asymptotic behaviour of the tail and the quantiles of a distribution, other parameters (for example, scale and location) are no less important for an accurate estimation of quantiles, see Matthys and Beirlant [31] and Caeiro and Gomes [3], among others.

In the first example studied, Section 3., and in the parametric approach, a negative value, although very close to zero, was obtained for $\widehat{\xi}$. Now considering the location and scale parameters estimates and by inverting the EV_ξ cdf in (2.3), for $\xi \neq 0$, the extreme quantiles, for very small values of p, can be easily estimated as

$$\widehat{\chi}_{1-p} := \widehat{\lambda} - \frac{\widehat{\delta}}{\widehat{\xi}}\left[1 - (-\ln(1-p))^{-\widehat{\xi}}\right]. \tag{6.1}$$

For example, for $p = 0.01, 0.001, 0.0001$, the corresponding quantile estimates are $\widehat{\chi}_{0.99} = 753.9114$; $\widehat{\chi}_{0.999} = 1000.7254$ and $\widehat{\chi}_{0.9999} = 1230.6420$.

In the semi-parametric framework, and using the estimates displayed in Figure 2 that show a more stable sample path (and also the Hill estimates as reference), as usually is done, high quantile estimates, also for the previous values of p were calculated.

It was used the moment estimator described in Matthys and Beirlant [31], subsection 2.3, defined as:

$$\widehat{\chi}^{\hat{\xi}}_{1-p,k+1} := X_{n-k:n}\, \hat{a}^{\hat{\xi}}_{n,k+1}\, \frac{{c_n}^{\hat{\xi}} - 1}{\hat{\xi}}; \qquad c_n := \frac{k}{np} \quad \text{for} \quad k < n \tag{6.2}$$

with

$$\hat{a}^{\hat{\xi}}_{n,k+1} = \frac{X_{n-k:n}\; H}{\rho_1(\hat{\xi})}, \qquad \rho_1(\xi) = \begin{cases} 1 & \text{for}\, \xi \geq 0 \\ 1/(1-\xi) & \text{for}\, \xi < 0. \end{cases}$$

where $\hat{\xi}$ is a consistent estimator of ξ. Here the H, L_8, CH and LRB_8 estimates, displayed in Figure 2, were used in (6.2).

Figure 10 shows the paths of $\widehat{\chi}_{0.99}(k)$, $\widehat{\chi}_{0.999}(k)$ and $\widehat{\chi}_{0.9999}(k)$.

However, if we have first performed the statistical test in (4.3), we were led not to reject the null hypothesis so we will consider $\xi = 0$. In this case the

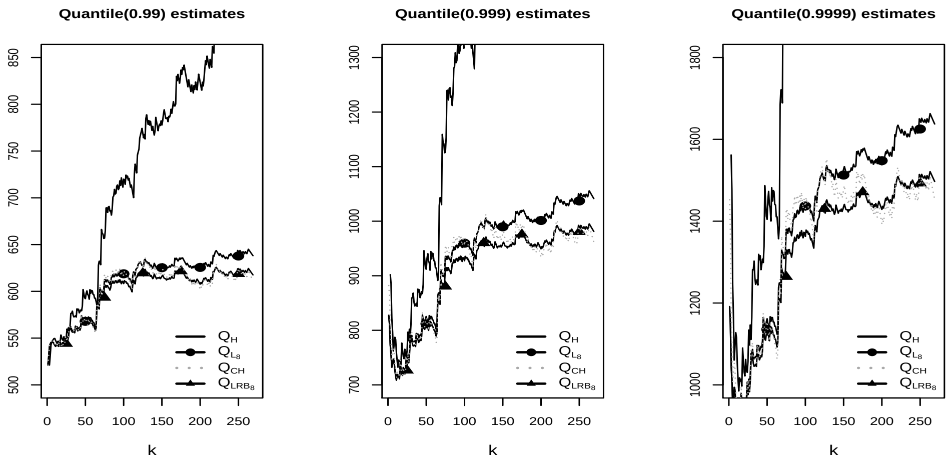

Figure 10: Sample paths of the quantiles estimates.

extreme quantiles can be estimated under the approach aforementioned, based on the inversion of the EV$_\xi$ cdf in (2.3), for $\xi = 0$, i.e.

$$\widehat{\chi}_{1-p}(k) := \widehat{\lambda} - \widehat{\delta} \ln\left(-\ln(1-p)\right), \tag{6.3}$$

and for the previous values of p we will obtain $\widehat{\chi}_{0.99} = 788.3877$; $\widehat{\chi}_{0.999} = 1079.6836$ and $\widehat{\chi}_{0.9999} = 1370.4655$.

We see that the quantiles estimates show large discrepancies among the procedures used. It is then advisable to perform a careful choice of the tail and also of the EVI-estimators in which the quantile estimates are based. This is out of scope of this article and an important topic for future research.

The next challenge is modelling and estimating clusters of extreme values since they are linked with incidences and durations of catastrophic phenomena. Here, an important parameter comes into play, the extremal index θ, that characterizes the degree of local dependence in the extremes of a stationary sequence. It needs to be adequately estimated, not only by itself but because its influence on other relevant parameters, such as a high quantile. Ignoring θ may lead to an underestimation of marginal quantile of F and an overestimation of quantiles of the EV.

ACKNOWLEDGMENTS

The authors are grateful to the two anonymous referees for their careful reviews and helpful suggestions and comments, which have highly improved the final version of this article.

This work has been supported by **FCT**—Fundação para a Ciência e a Tecnologia, Portugal, through the projects UID/MAT/00006/2013 (CEAUL), UID/MAT/00006/2019 (CEAUL) and UID/MAT/0297/2013 (CMA/UNL).

We also thank our colleagues José Miguel Cardoso Pereira and Ana Sá for providing us with the data on burned areas.

REFERENCES

[1] Beirlant, J.; Fraga Alves, M.I. and Gomes, M.I. (2016). Tail fitting for truncated and non-truncated Pareto-type distributions, *Extremes*, **19**, 429–462.

[2] Beirlant, J.; Goegebeur, Y.; Segers, J. and Teugels, J. (2004). *Statistics of Extremes. Theory and Applications.* Wiley.

[3] Caeiro, F. and Gomes, M.I. (2008). Minimum-variance reduced-bias tail index and high quantile estimation. *REVSTAT - Statistical Journal*, **6**, 1–20.

[4] Caeiro, F.; Gomes, M.I. and Pestana, D.D. (2005). Direct reduction of bias of the classical Hill estimator, *REVSTAT - Statistical Journal*, **3**, 111–136.

[5] Castillo, J. del; Daoudi, J. and Lockhart, R. (2014). Methods to distinguish between polynomial and exponential tails. *Scandinavian Journal of Statistics*, **41**, 382–393.

[6] Castillo, E.; Galambos, J. and Sarabia, J.M. (1989). *The selection of the domain of attraction of an extreme value distribution from a set of data.*. In "Extreme value theory (Oberwolfach, 1987) – Lecture Notes in Statistics" (J. Hüsler and R.-D. Reiss, Eds.), Springer, Berlin-Heidelberg, **51**, 181–190.

[7] Correia, A.L. and Neves, M. (1996). *Escolha estatística em modelos extremais–testes de ajustamento.* In "Bom Senso e Sensibilidade" (J. Branco, P. Gomes and J. Prata, Eds.), Actas do III Congresso Anual da Sociedade Portuguesa de Estatística, Edições Salamandra, 223-236.

[8] Dekkers, A.L.M.; Einmahl, J.H.J. and de Haan, L. (1989). A moment estimator for the index of an extreme-value distribution, *Annals of Statistics*, **17**, 4, 1833–1855.

[9] Dietrich, D.; de Haan, L. and Hüsler, J. (2002). Testing extreme value conditions, *Extremes*, **5**, 1, 71–85.

[10] Drees, H.; de Haan, L. and Li, D. (2006). Approximations to the tail empirical distribution function with application to testing extreme value conditions, *Journal of Statistical Planning and Inference*, **136**, 3498–3538.

[11] Falk, M. (1995). On testing the extreme value index via the POT-method, *Annals of Statistics*, **23**, 2013–2035.

[12] Fisher, R.A.and Tippett, L.H.C. (1928). Limiting forms of the frequency distributions of the largest or smallest member of a sample, *Proceedings of the Cambridge Philosophical Society*, **24**, 180–190.

[13] Fraga Alves, M.I. and Gomes, M.I. (1996). Statistical choice of extreme value domains of attraction – a comparative analysis. *Communications in Statistics – Theory and Methods*, **25**, 4, 789–811.

[14] Fraga Alves, M.I (1999). Asymptotic distribution of Gumbel statistic in a semi-parametric approach, *Portugaliae Mathematica*, **56**, 3, 282–298.

[15] Fraga Alves, M.I., Gomes, M.I., de Haan, L. and Neves, C. (2007). A note on second order conditions in extreme value theory: linking general and heavy tails conditions. *REVSTAT - Statistical Journal*, **5**, 3, 285–305.

[16] Fraga Alves, M.I.; Gomes, M.I.; de Haan, L. and Neves, C. (2009). Mixed moment estimator and location invariant alternatives, *Extremes*, **12**, 149–185.

[17] Fréchet, M. (1927). Sur la loi de probabilité de l'écart maximum, *Annales de la Société Polonaise de Mathématique (Cracovie)*, **6**, 93–116.

[18] Galambos, J. (1982). A statistical test for extreme value distributions. In "Nonparametric Statistical Inference" (B.V. Gnedenko *et al.*, ed.), North Holland, Amsterdam, 221–230.

[19] Gnedenko, B. V. (1943). Sur la distribution limite d'une série aléatoire, *Annals of Mathematics*, **44**, 423–453.

[20] Gomes, M.I.; Figueiredo, F. and Neves, M.M. (2012). Adaptive estimation of heavy right tails: resampling-based methods in action, *Extremes*, **15**, 463–489.

[21] Gomes, M.I.; Penalva, H., Caeiro, F. and Neves, M.M. (2016). Nonreduced versus reduced-bias estimators of the extreme value index-efficiency and robustness. In "COMPSTAT 2016 22^{nd} International Conference on Computational Statistics" (A. Colubi, A. Blanco and C. Gatu, Eds), 279–290.

[22] Gumbel, E.J. (1935). Les valeurs extrêmes des distributions statistiques, *Annales de l'institut Henri Poincaré*, **5**, 2, 115–158.

[23] Gumbel, E.J. (1958, 2004). *Statistics of Extremes*, Columbia University Press, New York.

[24] de Haan, L. (1970). *On Regular Variation and its Applications to the Weak Convergence of Sample Extremes*, Mathematical Centre Tract 32, Amesterdam, Dordrecht: D. Reidel.

[25] de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction.* Springer Science+Business Media, LLC, New York.

[26] Hasofer, A. and Wang, J.Z. (1992). A test for extreme value domain of attraction. *Journal of the American Statistical Association*, **87**, 171–177.

[27] Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution, *Annals of Statistics*, **3**, 1163–1174.

[28] Hüsler, J. and Li, D. (2006). On testing extreme value conditions, *Extremes*, **9**, 69–86.

[29] Marohn, F. (1998a). An adaptive efficient test for Gumbel domain of attraction. *Scandinavian Journal of Statistics*, **25**, 311–324.

[30] Marohn, F. (1998b). Testing the Gumbel hypothesis via the POT-method. *Extremes*, **1**, 2, 191–213.

[31] Matthys, G. and Beirlant, J. (2003). Estimating the extreme value index and high quantiles with exponential regression models, *Statistica Sinica*, **13**, 853–880.

[32] Neves, C. and Fraga Alves, M.I. (2007). Semi-parametric approach to the Hasofer-Wang and Greenwood statistics in extremes, *Test*, **16**, 297–313.

[33] Neves, C. and Fraga Alves, M.I. (2008). Testing extreme value conditions – an overview and recent approaches, *REVSTAT - Statistical Journal*, **6**, 1, 83–100.

[34] Neves, M.M.; Penalva, H. and Nunes, S. (2015). *Extreme value analysis of river levels in a hydrometric station in the North of Portugal.* In "Current Topics on Risk Analysis: ICRA6 and RISK 2015 Conference proceedings" (M. Guillén, A. Juan, H. Ramalhinho, I. Serra and C. Serrat, Edts.), 533–538.

[35] Penalva, H. (2017). *Contributos Computacionais e Metodológicos na Estimação do Índice de Valores Extremos.* Tese de Doutoramento, ISA - Universidade de Lisboa, Portugal. http://hdl.handle.net/10400.5/14946

[36] Penalva, H.; Nunes, S. and Neves, M. (2016). Extreme Value Analysis – a brief overview with an application to flow discharge rate data in a hydrometric station in the north of Portugal. *REVSTAT - Statistical Journal*, **14**, 2, 193-215.

[37] Penalva, H.; Caeiro, F.; Gomes, M.I. and Neves, M. (2016). *An Efficient Naive Generalization of the Hill Estimator – Discrepancy between Asymptotic and Finite Sample Behaviour*. Notas e Comunicações CEAUL 02/2016.

[38] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

[39] Segers, J. and Teugels, J. (2000). Testing the Gumbel hyphotesis by Galton's ratio, *Extremes*, **3**, 3, 291–303.

[40] von Mises, R. (1936). La distribution de la plus grande de n valeurs., *American Mathematical Society*, Reprinted in Selected Papers Volumen II, Providence, R.I. (1954), 271–294.