Modeling Peaks Over A Threshold Using R (Case Wadis Soubella, Algeria)

Fares Belagoune[1,a)],Djamel Boutoutaou[2, b)] and Mehdi Bellout [3]

[1] *Magister in Hydraulic sciences, University of Ouargla, Ouargla 30000, Algeria*.
[2] *Laboratoire d'exploitation et de Valorisation des ressources Naturelles en zones arides. Université d'Ouargla, Ouargla 30000, Algeria.*
[3] *Magister in Hydraulic sciences, University of Ouargla, Ouargla 30000, Algeria.*

[a)]*Corresponding author: faresbelagoune@yahoo.fr*
[b)] *boutoutaoudjamel@yahoo.fr*

Abstract. Floods are complex, natural hazards that, to varying degree, affect some parts of the world every year. The objective of this study is to Modeling Peaks Over a Threshold of El-Ham Basin River Using R. The free software environment for statistical computing and graphics has been developed and it is maintained by statistical programmers, with the support of an increasing community of users with many different backgrounds, which allows access to both well-established and experimental techniques. In This work R and some of its packages are presented powerful tools to explore and extract patterns from raw information, to pre-process input data of hydrological models, and post-processing its results. The Generalized Pareto Distribution (GPD) is the limiting distribution of normalized excesses over a threshold, as the threshold approaches the endpoint of the variable. The POT package contains useful tools to perform statistical analysis for peaks over a threshold using the GPD approximation.

INTRODUCTION

The study on floods in Algeria established by the National Agency of Water Resources (ANRH shows that the country is confronted with the phenomenon of very destructive floods and floods especially in arid and semi-arid regions Flooding is a submersion (fast or slow) likely to affect large areas of natural and urban, it corresponds to the overflow of water during a flood. A flood is a rapid and temporary flow of a river. It is described by three parameters: height, speed and current velocity. Floods occur when soil and vegetation cannot absorb any runoff water and cause an elevation of the bed of the stream. Most often, it does not overflow, but the water runs sometimes in amounts that cannot be transported in the beds of rivers, or retained in natural or artificial lagoons. The river overflows and then produced a flood. Flooding of rivers in these areas is less known. They are characterized by their sudden duration (rain showers, thunderstorm).The duration of the flood is of the order of minutes to hours. The human and material damage caused by these floods were still high [2]. In This work R [1] and some of its packages are presented powerful tools to explore and extract patterns from raw information, to pre-process input data of hydrological models, and post-processing its results. The Generalized Pareto Distribution (GPD) is the limiting distribution of normalized excesses over a threshold, as the threshold approaches the endpoint of the variable [3]. The POT package contains useful tools to perform statistical analysis for peaks over a threshold using the GPD approximation. There is many packages devoted to the extreme value theory (evd, ismev, evir); however, the POT package is specialized in peaks over threshold analysis. Moreover, this is currently the only one which proposes many estimators for the GPD. A user's guide (as a package vignette) and two demos are also included in the package [4].

DESCRIPTION OF THE STUDIED AREA

The watershed Hodna an area of 26,000 km² is the fifth largest basin of Algeria (Figure 1), is located 150 km as the crow flies south of the Mediterranean coast (Gulf of Bejaia). The altitude of the summits of the mountains of Hodna decreasing from east to west is between 1900 and 1000 m, while in the south a few peaks located in the Saharan Atlas reach 1200 m. The situation Hodna basin between two sets of mountains to the north and south basin organizes around a closed almost flat at 400 m above sea basin, and receives the flow of surface water in the region. In the center of the bowl, Chott El Hodna has an area of 1150 km ². The catchment area of the Oued El-Ham is located northwest of Hodna he occupies all all of this party basin, the basin is localized geographically between 35 ° 15 'and 36 ° 15' North latitude and between 3 ° and 4 ° 15 'East longitude. It drains an area of 5605

km² (with a perimeter of 360 km) to the hydrometric station of the Rocad-Sud located at the outlet of the basin (Figure 2).

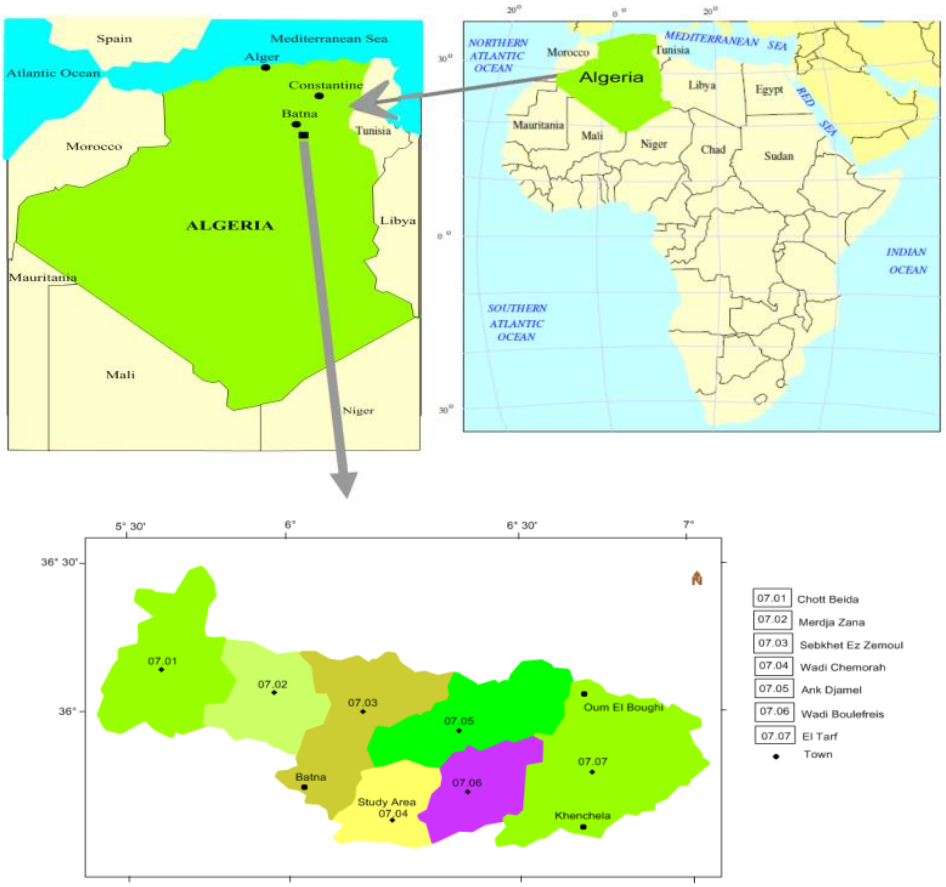

FIGURE 1. Map showing the watershed Hodna

The climate of the region is semi-arid, characterized by winter rains and summer drought. Interannual average precipitation over the whole basin of Wadis El-Ham is 185 mm, with a high interannual variability (Coefficient of Variation interannual CV = 0.40). Average maximum temperatures in the basin range from 24 ° to 27 ° C plain and 19 ° to 21 ° C in the highlands. The same for the average minimum temperatures, they vary from 9 ° to 12 ° C in the plains, and from 19 ° to 21 ° C in the highlands. The annual thermal gradient as a function of altitude is 0.75 ° C per 100 m [5].

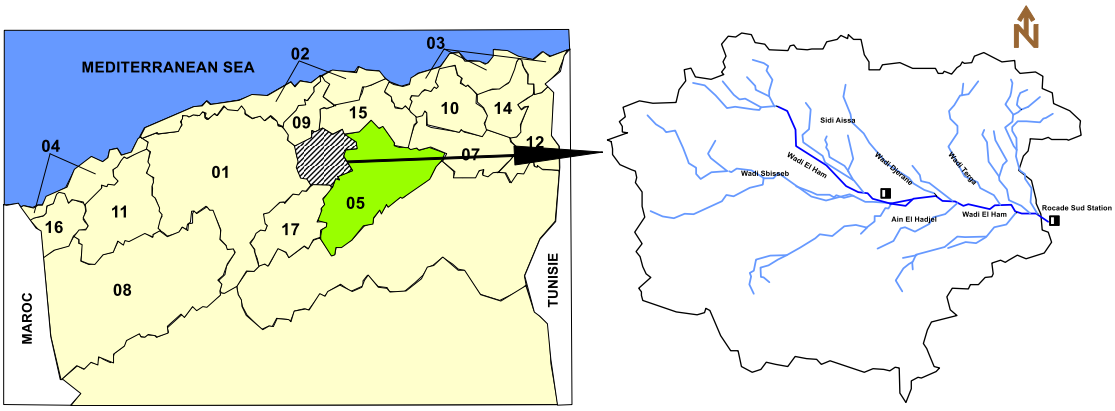

FIGURE 2. Location and Drainage watershed of Wadis El-ham

MATERIALS AND METHODS

The Pot Package

The POT package is an add-on package for the R statistical software. The main goal of this package is to develop tools to perform statistical analyses of Peaks over a Threshold (POT). The package can be downloaded from CRAN (The Comprehensive R Archive Network).

The Univariate Case

Even if this package is only related to peaks over a threshold, a classical introduction to the EVT (The Extreme Value Theory) must deal with "block maxima". Let $X_1,\ldots,X_n$ be a series of independent and identically distributed random variables with common distribution function F. Let Mn = max($X_1,\ldots,X_n$). Suppose there exists normalizing constants $a_n > 0$ and b_n such that:

$$\Pr\left[\frac{M_n - b_n}{a_n} \leq y\right] = F^n(a_n y + b_n) \to G(y), \quad n \to +\infty \tag{1}$$

For all y ∈ R, where G is a non-degenerate distribution function. According to the Extremal Types Theorem (Fisher and Tippett, 1928), G must be either Fréchet, Gumbel or negative Weibull. Jenkinson (1955) noted that these three distributions can be merged into a single parametric family: the Generalized Extreme Value (GEV) distribution. The GEV has a distribution function defined by:

$$G(y) = \exp\left[-\left(\xi\frac{y-\mu}{\sigma}\right)\right]_+^{-1/\xi} \tag{2}$$

Where (μ,σ,ξ) are the location, scale and shape parameters respectively, σ > 0 and $z^+ = \max(z,0)$. The Fréchet case is obtained when ξ > 0, the negative Weibull when ξ < 0 while the Gumbel case is defined by continuity when ξ → 0; From this result, Pickands (1975) showed that the limiting distribution of normalized excesses of a threshold. μ as the threshold approaches the endpoint μ end of the variable of interest is the Generalized Pareto Distribution (GPD). That is, if X is a random variable, then:

$$\Pr[X \leq y | X > \mu] \to H(y), \qquad \mu \to \mu_{end} \tag{3}$$

With:

$$H(y) = 1 - \left(1 + \xi\frac{y-\mu}{\sigma}\right)_+^{-1/\xi} \tag{4}$$

Where (μ,σ,ξ) are the location, scale and shape parameters respectively, σ > 0 and $z^+ = \max(z; 0)$. Note that the Exponential distribution is obtained by continuity as ξ → 0. In practice, these two asymptotical results motivated modeling block maxima with a GEV, while peaks over threshold with a GPD.

Sample Collection and Analysis

The study is based on measurements the instantaneous flow performed by ANRH. For the watershed of Wadis El-Ham, it has three hydrometric stations, two stations (Ain N'ssissa and Ced Fages) controlling tributaries and the third installed at the outlet control throughout the basin (Figure 3). In this study, it was limited to data from the hydrometric station of the outlet called Ain El-Hadjel (X = 35 ° 40 '26" N, Y = 3° 52' 54" E, Z = 609m).

RESULTS AND DISCUSSION

Threshold Selection

The location for the GPD or equivalently the threshold is a particular parameter as must often it is not estimated as the other ones. All methods to define a suitable threshold use the asymptotic approximation defined by equation (3). In other words, we select a threshold for which the asymptotic distribution H in equation (4) is a good approximation. The POT package has several tools to define a reasonable threshold. For this purpose, the user must use tcplot, mrlplot, lmomplot, exiplot and diplot functions. The main goal of threshold selection is to selects enough events to reduce the variance; but not too much as we could select events coming from the central part of the distribution1 and induce bias.

Threshold Choice plot: tcplot

Let $X \sim GP(\mu_0,\sigma_0,\xi_0)$. Let μ_1 be a another threshold as $\mu_1 > \mu_0$. The random variable $X|X > \mu_1$ is also GPD with updated parameters $\sigma_1 = \sigma_0 + \xi_0 (\mu_1 - \mu_0)$ and $\xi_1 = \xi_0$. Let

$$\sigma_* = \sigma_1 - \xi_1 \mu_1 \tag{5}$$

With this new parameterization, σ_* is independent of μ_1. Thus, estimates of σ_* and ξ_1 are constant for all $\mu_1 > \mu_0$ if μ_0 is a suitable threshold for the asymptotic approximation. Threshold choice plots represent the points defined by:

$$\{(\mu_1, \sigma_*): \mu_1 \leq x_{max}\} \quad \text{and} \quad \{(\mu_1, \xi_1): \mu_1 \leq x_{max}\}$$

Where $x_{\max}$ is the maximum of the observations *x*.
Results of the tcplot function are display in Figure 3. We can see clearly that a threshold around 0. 89 is a reasonable choice. However, in practice decision are not so clear-cut as for this synthetic example.

Mean Residual Life Plot: mrlplot

mean residual life plot is based on the theoretical mean of the GPD. Let *X* be a random variable distributed as GPD (μ,σ,ξ). Then, theoretically we have:

$$\mathbb{E}[X] = \mu + \frac{\sigma}{1-\xi}, \qquad for\ \xi < 1 \tag{6}$$

When $\xi \geq 1$, the theoretical mean is infinite.

In practice, if *X* represents excess over a threshold μ_0, and if the approximation by a GPD is good enough, we have:

$$\mathbb{E}[X - \mu_0 \,|X > \mu_0] = \frac{\sigma_{\mu_0}}{1-\xi} \tag{7}$$

For all new threshold μ_1 such as $\mu_1 > \mu_0$, excesses above the new threshold are also approximate by a GPD with updated parameters. Thus,

$$\mathbb{E}[X - \mu_1 \,|X > \mu_1] = \frac{\sigma_{\mu_1}}{1-\xi} = \frac{\sigma_{\mu_0} + \xi_{\mu_1}}{1-\xi} \tag{8}$$

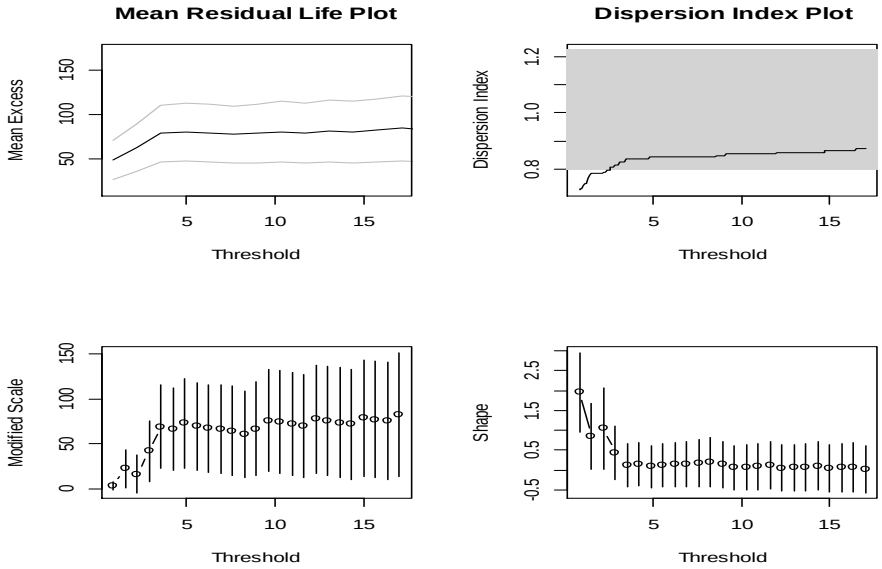

FIGURE 3. Tools for the threshold selection

Dispersion Index Plot: diplot

The Dispersion Index plot is particularly useful when dealing with time series. The EVT states that excesses over a threshold can be approximated by a GPD. However, the EVT also states that the occurrences of these excesses must be represented by a Poisson process. Let X be a random variable distributed as a Poisson distribution with parameter λ . That is:

$$\Pr[X = k] = e^{-\lambda}\frac{\lambda^k}{k!}, \qquad k \in \mathbb{N} \tag{9}$$

Thus, we have $\mathbb{E}[X] = \mathrm{Var}[X]$. Cunnane (1979) introduced a Dispersion Index statistic defined by: Where s^2 is the intensity of the Poisson process and λ the mean number of events in a block - most often this is a year. Moreover, a confidence interval can be computed by using a χ^2 test:

$$I_\alpha = \left[\frac{\chi^2_{\frac{(1-\alpha)}{2},M-1}}{M-1}, \frac{\chi^2_{1-\frac{(1-\alpha)}{2},M-1}}{M-1}\right] \tag{10}$$

Where $\Pr[DI \in I_\alpha] = \alpha$

Fitting the GPD

The univariate case

The main function to fit the GPD is called fitgpd. This is a generic function which can fit the GPD according several estimators. There are currently 17 estimators available: method of moments moments, maximum likelihood mle, biased and unbiased probability weighted moments pwmb, pwmu, mean powerdensity divergence mdpd, median med, pickands' pickands, maximum penalized likelihood mple and maximum goodness-of-fit mgf estimators. For the mgf estimator, the user has to select which goodness-of-fit statistics must be used. These statistics are the Kolmogorov-Smirnov, Cramer von Mises, Anderson Darling and modified Anderson Darling. See the html help page of the fitgpd function to see all of them. Details for these estimators can be found in (Coles, 2001), (Hosking and Wallis, 1987), (Juàrez and Schucany, 2004), (Peng and Welsh, 2001) and (Pickands, 1975). The MLE is a particular case as it is the only one which allows varying threshold. Moreover, two types of standard errors are available: "expected" or "observed" information of Fisher. The option obs.fish specifies if we want observed (obs.fish = TRUE) or expected (obs.fish = FALSE). Here is the scale and shape parameter estimates of the GPD for the 7 estimators implemented.

	scale	shape
mom	2.059277	0.1440857
mle	1.862201	0.2365360
pwmu	1.821022	0.2431135
pwmb	1.831041	0.2389492
pick	1.491969	0.5178477
med	1.556960	0.4591639
mdpd	1.794216	0.2836382

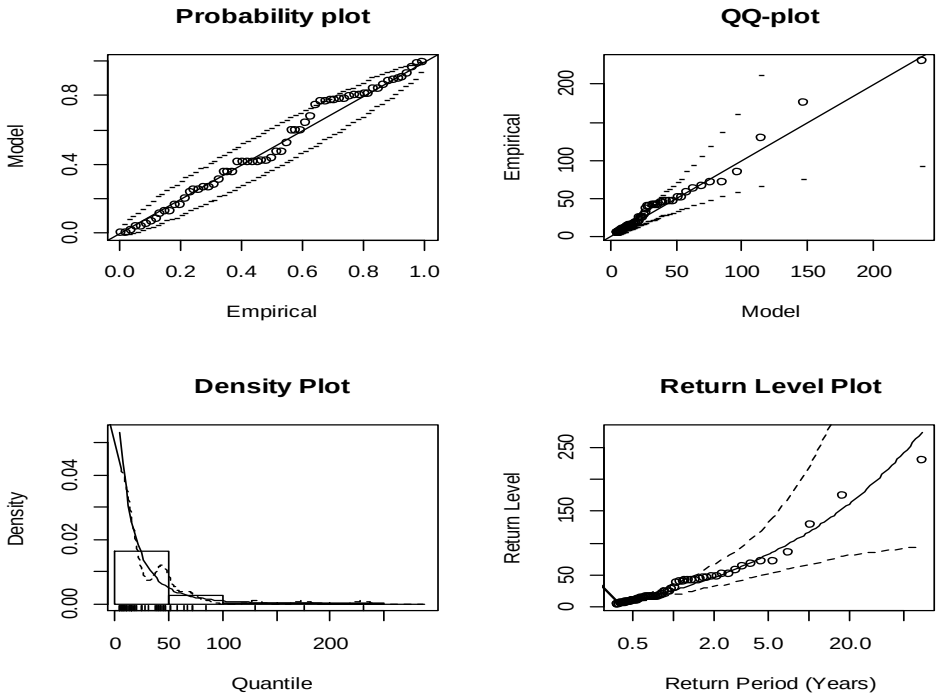

FIGURE 4. Graphic tools for model diagnostic

CONCLUSION

In this article we have proposed an efficient way to perform canonical correlation analysis in R. The functions provided in the POT package. The POT package can also: Simulate and compute density, quantile and distribution functions for the GPD. Fit the GPD with a varying threshold using MLE. Perform analysis of variance for two nested models.

REFERENCES

1. R®. Development Core Team (2012). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
2. Belagoune.F . Etude des crues des cours d'eaux en milieu semi aride. Thesis. University of Ouargla, Algeria,2011.
3. Pickands.J. Statistical inference using extreme order statistics. Annals of Statistics 1975; 3:119–131.
4. Ribatet.M. Cemagref Unité de Recherche HH Lyon, France,2007.
5. Hasbaia et al.Variabilité de l'érosion hydrique.Variabilite de l'érosion hydrique dans le bassin du Hodna: cas du sous-bassin versant de l'oued elham,Algeria,2012.