Metodološki zvezki, Vol. 10, No. 1, 2013, 17-29

Topics in Data Analysis Using R in Extreme Value Theory

Helena Penalva[1], Manuela Neves[2] and Sandra Nunes[3]

Abstract

The Statistical Extreme Value Theory has grown gradually from the beginning of the 20^{th} century. Its unquestionable importance in applications was definitely recognized after Gumbel's book in 1958, *Statistics of Extremes*. Nowadays there is a wide number of applied sciences where extreme value statistics are largely used. So, accurately modeling extreme events has become more and more important and the analysis requires tools that must be simple to use but also should consider complex statistical models in order to produce valid inferences. To deal with accurate, friendly, free and open-source software is of great value for practitioners and researchers. This paper presents a review of the main steps for initializing a data analysis of extreme values in R environment. Some well documented packages are briefly described and two data sets will be considered for illustrating the use of some functions.

1 Introduction and Motivation

In extreme value theory we need to deal with events that are more extreme than any that have already been observed. The question is then how to make inference beyond the sample data. Clearly statistical inference can be deduced only from those observations which are extreme in some sense. Then we need to use techniques in such a way that it is possible to make statistical inference about rare events, using only a limited amount of data! So, in analysis of extreme values, assumptions on the tail of the data underlying distribution need to be considered.

Historically there have been two main application areas of extreme value theory: the environmental area with the study of sea levels, wind speeds, river flow, etc., and the reliability and structural safety area. Nowadays extreme value theory has emerged as one of the most important statistical areas in several applied sciences, such as insurance, risk assessment, telecommunications, geology and seismic risk, biology and public health.

The first book that gave a great promotion to extreme value statistics, showing this theory as a tool for modeling the extremal behavior of physical processes was *Statistics of Extremes*, Gumbel (1958). Many applications are presented and studied and this book

[1] Escola Superior de Ciências Empresariais, Instituto Politénico de Setúbal, Portugal; helena.penalva@esce.ips.pt

[2] Instituto Superior de Agronomia, Universidade Técnica de Lisboa, Portugal; manela@isa.utl.pt

[3] Escola Superior de Ciências Empresariais, Instituto Politénico de Setúbal, Portugal; sandra.nunes@esce.ips.pt

is still relevant today. Some years later, in 1984, *Statistical Extremes and Applications* a book edited by Tiago de Oliveira, was considered "... a complete perspective of the field, also with a series of promising directions of research and some recent results". More recently we can refer to books with large number of applications such as Coles (2001), Embrechts et al. (2003), Beirlant et al. (2004), Castillo et al. (2005) and Reiss and Thomas (2007), among others.

Statistical inference about extreme events deals with in the estimation of the probability of occurrence of extreme events. There are a few parameters whose estimation is of major importance. The extreme value index, which is directly related with the heaviness of the tail of the underlying distribution, is the basis of all parameters of extreme events like, for example, the high quantile of probability $1-p$, with p small, usually denoted as the *return level* associated to the *return period* $1/p$ or the right endpoint of an underlying model.

In all areas of application it is of major importance to use adequate and accurate statistical methods. The objective of this paper is to present the initial topics of a review of some functions and packages included in R (R Development Core Team, 2012) environment, for the analysis of extreme values. Those functions will be applied in some data analyses.

Section 2 introduces notations and gives a brief background of models that form the basis for the theory of statistical extremes is given. Section 3 presents a description of the main packages used in extreme value analysis available in R. Section 4 is devoted to illustrate some analyses through R, using two data sets.

2 Background on extreme value analysis

In Extreme Value Theory (EVT) , opposite to the Central Limit Theory , we are not interested in modeling the central values but the tail of the underlying distribution.

Let $X_1, \cdots, X_n$ be independent and identically distributed (i.i.d.) random variables with distribution function (d.f.) F with mean value μ and finite variance, σ^2. The central limit theory is concerned with the limit behavior of the partial sums, $\sum_{i=1}^n X_i$ and, in its simplest form, states that

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow[n\to\infty]{} N(0,1). \tag{2.1}$$

In statistical applications the central limit theorem leads to the approximation of the sample mean $\overline{X_n}$, for large n, as

$$\overline{X_n} \sim \mathcal{N}(\mu, \sigma/\sqrt{n}).$$

Theory of sample extremes is concerned with the limiting behavior of the maximum $M_n = \max\left(X_1, \cdots, X_n\right)$ or the minimum $m_n = \min\left(X_1, \cdots, X_n\right)$ of a sample $(X_1, \cdots, X_n)$, as $n \to \infty$.

Figure 1 shows how the normal distribution fits reasonably well to the sample average of standard uniform random variables but hardly to the sample maximum.

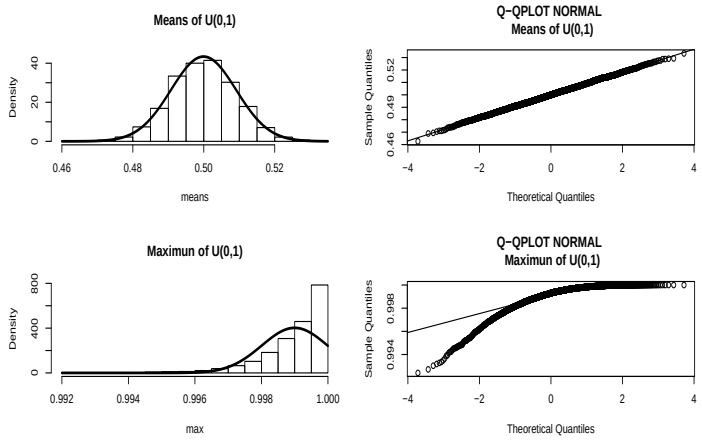

Figure 1: Normal fit and Q-Q plots of the sample average and maximum of random samples ($n = 1000$) from an uniform distribution.

Let us consider the sample $(X_1, \cdots, X_n) \overset{i.i.d.}{\sim} F$ and suppose we want to derive the distribution of the maximum M_n. We know that,

$$P[M_n \leq x] = P[X_1 \leq x, \ldots, X_n \leq x] = P[X_1 \leq x] \ldots P[X_n \leq x] = F^n(x).$$

The d.f. F^n is of little help in practice since F is itself unknown and also the misspecification of F can lead to large errors in the distribution of the maximum.

The asymptotic theory of sample extremes distribution has been developed under analogous arguments to those considered in the central limit theory.

First results date back to Fréchet (1927), who obtained the asymptotic distribution of the maximum. Fisher and Tippet (1928) and von Mises (1936) presented the first studies on the *extremal limit problem*. However Gnedenko (1943) was the first who gave conditions for the existence of sequences $\{a_n\} \in \mathbb{R}^+$ and $\{b_n\} \in \mathbb{R}$ such that,

$$\lim_{n\to\infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = \lim_{n\to\infty} F^n(a_n x + b_n) = G(x) \quad \forall x \in \mathbb{R}, \tag{2.2}$$

where $G(x)$ is a nondegenerate distribution function.

This function, called Extreme Value d.f. , is usually denoted by EV_γ and is given by

$$EV_\gamma(x) = \begin{cases} \exp[-(1+\gamma x)^{-1/\gamma}], & 1+\gamma x > 0,\ if\ \gamma \neq 0\ ; \\ \exp[-\exp(-x)], & x \in \mathbb{R},\ if\ \gamma = 0. \end{cases} \tag{2.3}$$

We say that F is in the domain of attraction (for maxima) of EV_γ and write $F \in \mathcal{D}_\mathcal{M}(EV_\gamma)$.

The EV_γ incorporates the three (Fisher-Tippett) types:

- Gumbel: $\Lambda(z) = \exp(-\exp(-z)) = EV_0(z) \quad z \in \mathbb{R},\ \gamma = 0.$
  It is the limit for exponential tailed distributions.

- Fréchet: $\Phi_\gamma(z) = \exp(-z^{-1/\gamma}) = EV_\gamma(\frac{z-1}{\gamma}) \quad z > 0,\ \gamma > 0.$
  It is the limit for heavy tailed distributions.

- Weibull: $\Psi_\gamma(z) = \exp(-(-z)^{1/\gamma}) = EV_\gamma(\frac{-z-1}{\gamma}) \quad z < 0,\ \gamma < 0.$
  It is the limit for light tailed distributions.

If interest is in the minimum, it is easy to convert what is said above considering that

$$\min\big(X_1, \cdots, X_n\big) = -\max\big(-X_1, \cdots, -X_n\big).$$

The EV_γ d.f. given in (2.3) can also incorporate location, μ, and scale, $\sigma > 0$, parameters.

The shape parameter γ is called extreme value index and measures the heaviness of the right-tail, $\overline{F} := 1 - F$.

- If $\gamma = 0$, the right tail is of exponential type. The *right endpoint* of F, $x^* \equiv x^F := \sup\{x : F(x) < 1\}$, can then be either finite or infinite. It is the domain of attraction for maxima for many common distributions such as normal, lognormal, exponential and gamma distributions.

- If $\gamma > 0$, the right tail is heavy, it is of a negative polynomial type and F has an infinite *right endpoint*. Moments of order r are infinite if $r \geq 1/\gamma$. Pareto and Cauchy are examples of distributions in the domain of attraction for maxima of the Fréchet d.f.

- If $\gamma < 0$, the right tail is light and F has a finite *right endpoint*, $x^* = \mu - \frac{\sigma}{\gamma} < +\infty$. As an example of a distribution belonging to the domain of attraction for maxima of the Weibull d.f. we can refer to the uniform distribution.

Figure 2 shows plots of density functions for the Gumbel, Weibull and Fréchet (left) and a zoom of the plot for comparing with the standard normal density (right).

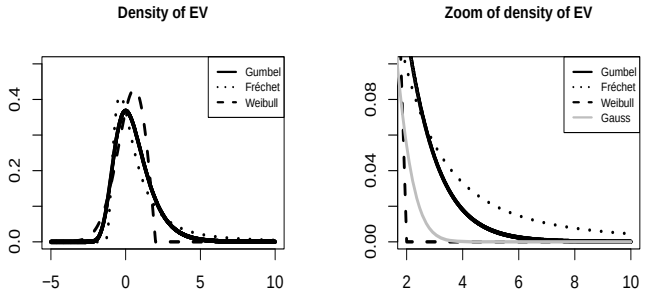

Figure 2: Density functions for the Gumbel, Weibull and Fréchet (left) and a zoom of those functions compared with the standard normal density (right).

The extreme value index γ is the basis of parameters of extreme events that we need to estimate in several situations. Let us define some of them, for an underlying distribution F:

- *high quantile* of probability $1-p$, with p small, (*return level*).
  Here the interest relies in extreme quantiles, situated in the border or even beyond the range of the available data, then the return level is defined for very small values of p, say $p < 1/n$, as $\chi_{1-p} := \inf\{x : F(x) \geq 1-p\} =: F^{\leftarrow}(1-p)$, where $F^{\leftarrow}$ stands for the left-continuous inverse function of F;

- the *probability of exceedance* of a high level $x \equiv x_H$, $p_{x_H} := P(X > x_H) = 1 - F(x_H)$;

- the *return period* of a high level x_H, $r_{x_H} := 1/(1-F(x_H))$, in any i.i.d. scheme.

The limit distribution family, EV_γ, that appeared in (2.3), seems to present some difficulties due to the normalizing constants, $\{a_n\}$ and $\{b_n\}$ being unknown. However that limit can be interpreted, for sufficiently large n, as

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) \approx EV_\gamma(x),$$

or equivalently

$$P\left(M_n \leq x\right) \approx EV_\gamma\left(\frac{x - b_n}{a_n}\right),$$

that is also a distribution function of the same family, so we have not to worry about the knowledge of those constants.

3 The R Environment for the analysis of extremes

The software R (R Development Core Team, 2012) is a language and an open source environment for statistical computing, which allows the manipulation and data analysis, numerical computation and graphical production. From the homepage http://www.r-project.org/, R can be downloaded and detailed information can be obtained. A great advantage of R is that, besides being freely available without propriety licensing requirements, several statistical techniques were incorporated in its base and a large number of functions has been developed and made available by users as contributed packages.

A package consists of a collection of functions, examples and documentation and all source code is published. We can then see the exact algorithms and code being used. A collection of manuals are accessible from http://cran.r-project.org/.

R software contains packages with several functions for modeling extreme data. Gilleland et al. (2012) give an excellent software review for extreme value analysis. They describe and compare packages available in R with other software. For other packages in R or for other software see Gilleland et al. (2012).

In this paper we will consider `evd` and `ismev`, two packages prepared for extreme value analysis. Let us expose the main functionalities of these packages.

- The `evd` (extreme value distributions) package (Stephenson, 2002) was originally developed for distributions of extreme values, but has been extended to include functions for statistical modeling using maximum likelihood estimation for univariate and bivariate maxima models. It allows graphical model diagnostics for all

fitted models; profile likelihood and profile likelihood confidence intervals for any parameter of interest, including return levels. Cluster identification and extremal index estimation is also available. It is also possible to use non-parametric estimation. A user's guide document is available.

- The `ismev` package (Stephenson, 2012) is an R version of S functions, originally written by Heffernan and ported into R by Alec Stephenson. It was written to support univariate extreme value modeling including the computations carried out in Coles (2001). This package allows to perform univariate maximum likelihood estimates of extreme value distribution for the block maxima and for threshold model approaches (generalized Pareto distribution and Poisson point process models); the diagnosis of the quality of the fitted distributions and to select an appropriate threshold for the threshold models. It also allows the incorporation of non-stationary. This package has also a user's guide document.

Some of the main functions that are included in these two packages will be used in the following section. Some other packages can be mentioned, such as `evir`, `fExtremes`, `evdbayes`, `copula`, `POT`, etc., but will not be considered in this work. Some of them even show some overlapping functionalities, see Gilleland *et. al* (2012) for more details.

4 Modeling extreme data with R - two case studies

To illustrate the first steps for performing a data analysis of extreme values using some functions of the two packages mentioned in the previous section, we used two data sets available in the package `evd`: 'lisbon' and 'sask'. These two data sets have been studied in Tiago de Oliveira (1997) and here is illustrated how those results can be obtained using R environment.

To apply some of the main functions in each package, we have decided to study 'lisbon' data through `evd` package and 'sask' data using `ismev` package.

4.1 Example 1: Maximum wind speed data in Lisbon

These data correspond to maximum wind speeds, in kilometers per hour, at Lisbon, between 1941 and 1970. The data are available in R package `evd` and can be obtained typing `data(lisbon)`.

To obtain the main descriptive statistics, we can load the package `fBasics` and use `basicStats(lisbon)`. Table 1 shows those measures. There is a slight positive asymmetry and the kurtosis indicates possibly a tail lighter than the normal one, see the histogram of the data in Figure 3 (left).

Table 1: Basic descriptive statistics for Lisbon data.

1st Qu.	Median	Mean	3rd Qu.	St. Dev.	Skewness	Kurtosis
90.25	100.0	101.3	110.25	13.90	0.30	-0.40

From Figure 3 (right) it seems reasonable to assume that the data are stationary over the period under observation. As data refer to the maximum at each year, at least weak dependence can be assumed, what enables to use results presented in Section 2, see Leadbetter et al. (1983). Then we will try to fit an EV_γ to this data set.

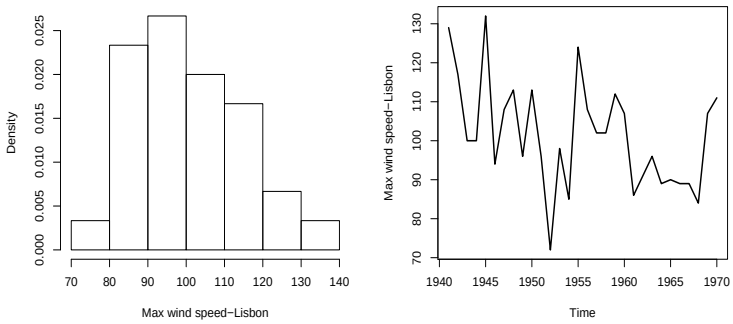

Figure 3: Maximum wind speed in Lisbon over the period 1941-1970. Histogram (left) and the time series (right)

Using `fgev(lisbon)` we get the maximum-likelihood fitting for the Extreme Value distribution. The estimates for the location, scale and shape, with the standard errors in parenthesis, are

$$\hat{\mu} = 96.03(2.62) \qquad \hat{\sigma} = 12.85(1.83) \qquad \hat{\gamma} = -0.20(0.13)$$

Function `confint` gives, by default, 95% Wald confidence intervals for all the parameters. Then typing `confint(fgev(lisbon))`, we get the 95% confidence intervals $(90.90\ ;\ 101.16)$, $(9.26\ ;\ 16.45)$ and $(-0.45\ ;\ 0.05)$, for μ, σ and γ, respectively.

Although the maximum likelihood estimate for γ is negative, what would correspond to a bounded distribution, the confidence interval for γ includes zero, so leads us to not reject the null hypothesis, $\gamma = 0$. The Gumbel distribution is then a possible candidate to model maximum wind speed data in Lisbon, what was also investigated in Tiago de Oliveira (1977).

Greater accuracy for the confidence intervals is usually attained by the profile likelihood. Plots for the profile log-likelihood for all parameters are easily obtained by

```
par(mfrow = c(1,3))
plot(profile(fgev(lisbon)),ci = c(0.95, 0.99))
```

In this case we considered 95% and 99% confidence intervals for each parameter, see Figure 4. The confidence interval limits can be obtained through the functions

`confint(profile(fgev(lisbon)), level=0.95)` or
`confint(profile(fgev(lisbon)), level=0.99)`.

Diagnostic plots for assessing the accuracy of the EV model can be performed using `plot(fgev(lisbon))` and are shown in Figure 5. Both probability plot and quantile plot show a reasonable EV fit, the plots are almost linear, so they do not cause doubts

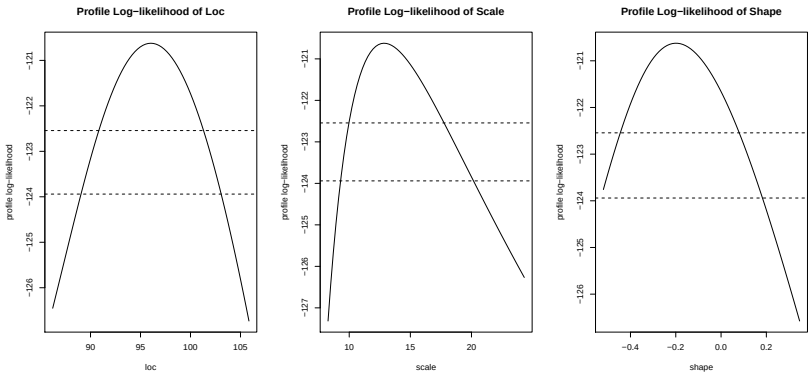

Figure 4: Profile log-likelihood for the three parameters in Lisbon data.

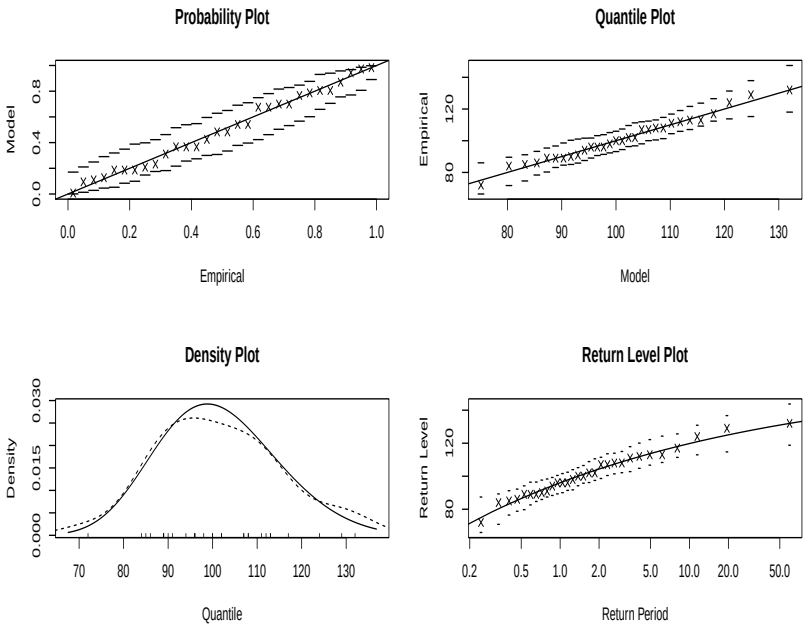

Figure 5: Diagnostic plots for EV fit to the Lisbon data.

on the adequacy of EV fitting. The return level plot is not linear and shows a slight convexity (actually our estimated value is negative!). However as the confidence interval for γ contained zero, we considered Gumbel distribution as possibly a more adequate model. Typing `fgev(lisbon,shape=0)`, we get the maximum-likelihood fitting for the Gumbel distribution. The parameters estimates, with standard errors in parenthesis, are

$$\hat{\mu} = 94.71(2.41) \qquad \hat{\sigma} = 12.49(1.68)$$

Using `plot(fgev(lisbon,shape=0))`, the diagnostic plots in Figure 6 are obtained showing that Gumbel distribution seems more appropriate for these data.

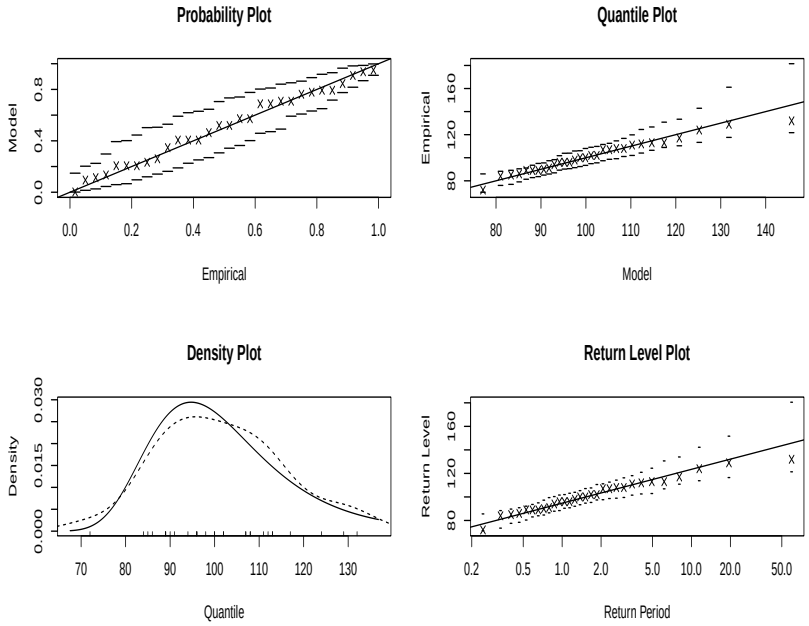

Figure 6: Diagnostic plots for Gumbel fit to the Lisbon data.

To confirm that suspicion, we carried out an analysis of deviance between the two models, using

```
M1<-fgev(lisbon); M2<-fgev(lisbon,shape=0); anova(M1,M2).
```

As we can see in Table 2, the results obtained show no significant differences between both models, so Gumbel model, with two parameters, is a good choice for modeling these data.

Table 2: Deviance analysis

Model	Mdf	Deviance	df	chisq	P-value
M1	3	241.25			
M2	2	243.32	1	2.0742	0.1498

4.2 Example 2: Flood discharges of the North Saskachevan River at Edmonton

These data that have been dealt with by Tiago de Oliveira (1977) are still available in R package `evd` and can be obtained typing `data(sask)`. These data correspond to maximum annual flood discharges, in units of 1000 cubic feet per second, of the North Saskachevan River at Edmonton, over a period of 47 years. But, as it is said in the

description "... unfortunately, the data are ordered from largest to smallest", so it is not possible to plot the corresponding time series.

Basic descriptive statistics, see Table 3, show a strong positive asymmetry and a very high kurtosis what indicates possibly a heavy right tail. Mean and median values give some support to this supposition.

Table 3: Basic descriptive statistics for flood discharges of the North Saskachevan River.

1st Qu.	Median	Mean	3rd Qu.	St. Dev.	Skewness	Kurtosis
30.33	40.4	51.5	61.33	32.38	2.00	4.62

As we mentioned before, for analyzing this data we used the package `ismev`. The maximum-likelihood fitting for the EV d.f. was obtained through the function `gev.fit(sask)`, giving the following estimates for the location, scale and shape parameters, with the standard errors in parenthesis:

$$\hat{\mu} = 35.07(2.44) \qquad \hat{\sigma} = 14.29(2.23) \qquad \hat{\gamma} = 0.43(0.16)$$

We have a positive estimate of γ and the 95% Wald confidence interval for γ is $(0.12\,;\,0.75)$, clearly indicating a heavy right tail for the distribution. The 95% and 99% confidence intervals for the profile log-likelihood for γ can be obtained running the commands

```
gev.profxi( gev.fit(sask), 0.1, 0.9) and
gev.profxi( gev.fit(sask), 0.05, 0.98, conf=0.99).
```

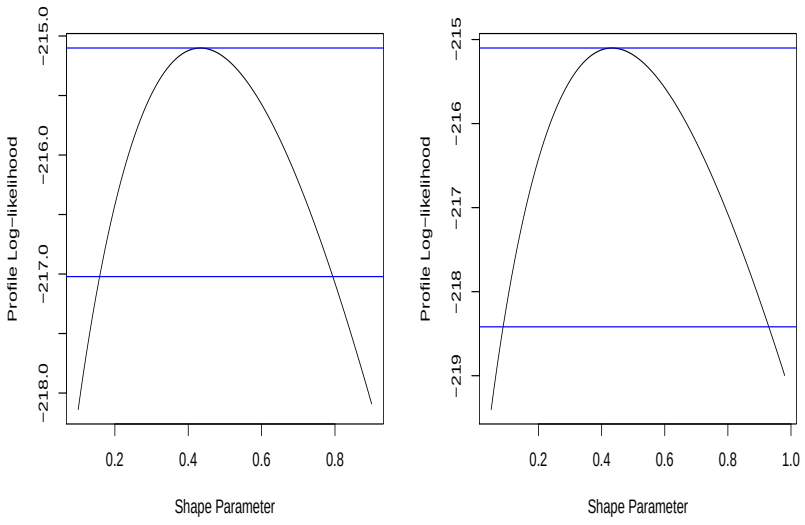

Figure 7: Profile log-likelihoods for shape parameter, using 95% and 99% confidence respectively

Profile log-likelihood, see Figure 7, gives a positive estimate for γ. Diagnostic plots for assessing the accuracy of the EV model for these data are shown in Figure 8 and can be obtained by the function `gev.diag(gev.fit(sask))`.

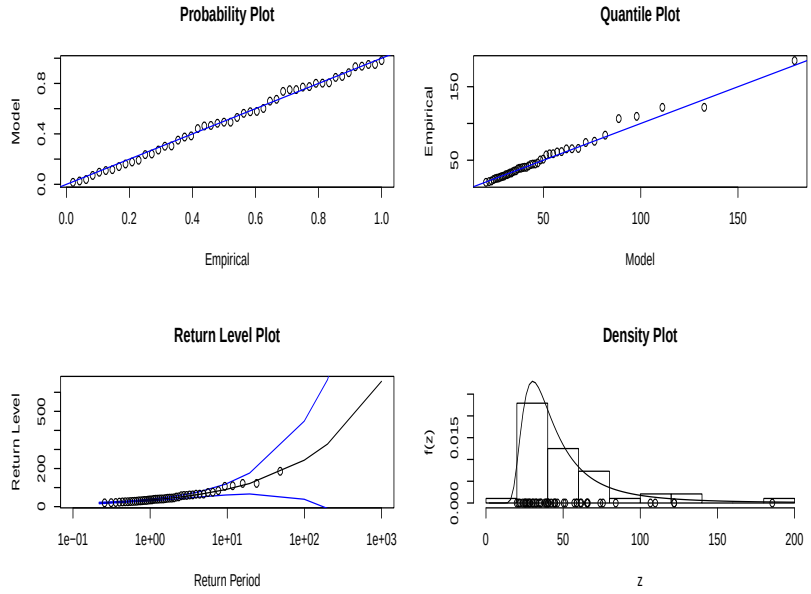

Figure 8: Diagnostic plots for EV fit to the flood discharges of the North Saskachevan River.

These plots show the Fréchet model as being adequate for modeling these data.
With the package `ismev`, the profile log-likelihood for m-year return levels, see Figure 9, can be obtained running the commands

`gev.prof( gev.fit(sask), m=10, 60, 200)` for 10-year return level and
`gev.prof( gev.prof( gev.fit(sask), m=50, 80, 550)` for 50-year return level.

5 Concluding remarks

In this work we intended to show the facilities and functionalities of R software to be used by practitioners in the diverse areas of application. The objective was to show how to begin a data analysis in extreme value theory in R environment. A brief idea of the theory behind the analyses performed is presented. Here, only the first approach in modeling extreme values, the block-maxima EV model, was considered. Other parametric methodologies, such as, the $r-$ largest observations, the POT method, the PWM method and the more recent semi-parametric approaches could be considered, but are out of scope of the idea of this paper. Some work is now in progress for using those methodologies in R and with several data sets.

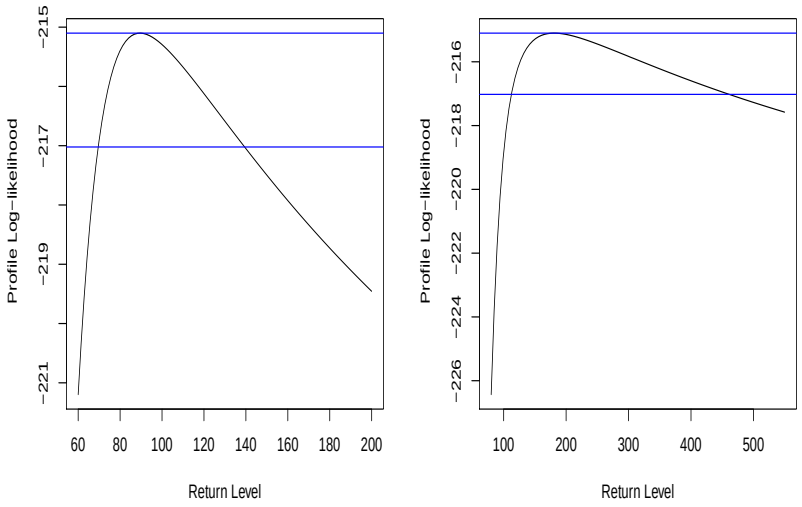

Figure 9: Profile log-likelihoods for 10 and 50 years return levels, respectively

Acknowledgments

Research of the second author was partially supported by National Funds through **FCT**—Fundação para a Ciência e a Tecnologia, projects PEst-OE/MAT/UI0006/2011 and EXTREMA, PTDC/FEDER. Research of the third author was partially supported by National Funds through **FCT**—Fundação para a Ciência e a Tecnologia, projects PEst-OE/MAT/UI0297/2011 (CMA/FCT/UNL).
The authors would like to thank the Editor and two referees whose critical but constructive remarks and useful suggestions have greatly improved the contents of this paper.

References

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

[2] Castillo, E., Hadi, A. S., Balakrishnan, N. and Sarabia, J. M., (2005). *Extreme Value and Related Models in Engineering and Science Applications.* New York: John Wiley & Sons.

[3] Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag, London.

[4] Embrechts, P., Klüppelberg, C. and Mikosch, T. (2003). *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag, London.

[5] Fisher, R. A. and Tippett, L. H. C. (1928): On the estimation of the frequency distributions of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, **24**, 180-190.

[6] Fréchet, M. (1927): Sur la loi de probabilité de l'écart maximum. *Ann. Soc. Polon. Math. (Cracovie)*, **6**, 93-116.

[7] Gilleland, E., Ribatet, M. and Stephenson, A. G. (2012): A software review for extreme value analysis. Springer (Springerlink.com). Published online:20 July 2012)DOI 10.1007/s10687-012-0155-0).

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

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

[10] Leadbetter, M., Lindgren, G. and Rootzén, H. (1983). *Extremes and related properties of random sequences and series.* Springer-Verlag, New York.

[11] R Development Core Team (2012), R: A Language and Environment for Statistical Computing, R. Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/. ISBN 3-900051-07-0.

[12] Reiss, R. D. and Thomas, M. (2007). *Statistical Analysis of Extreme Values: With Applications to Insurance, Finance, Hydrology and Other Fields*. Third Edition, Springer Verlag.

[13] Stephenson, A. G. (2002): evd: Extreme Value Distributions. *R News, URL http://CRAN.R-project.org/doc/Rnews/*, **2(2)**, 31-32.

[14] Stephenson, A. G. (2012): ismev: An Introduction to Statistical Modeling of Extreme Values. Original S functions written by Janet E. Heffernan with R port and R documentation. *R package version 1.38, URL http://CRAN.R-project.org/package=ismev*.

[15] Tiago de Oliveira, J. (ed.) (1984): *Statistical Extremes and Applications*. D. Reidel, Dordrecht, Holland.

[16] Tiago de Oliveira, J. (1997): *Statistical Analysis of Extremes*. Pendor.

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