Visualizing and modeling extreme data in R environment: a practical approach

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

[1]Escola de Ciências Empresariais do Instituto Politécnico de Setúbal, Portugal, helena.penalva@esce.ips.pt
[2]Escola Superior de Ciências Empresariais do Instituto Politécnico de Setúbal e CMA/FCT/UNL, Portugal, sandra.nunes@esce.ips.pt
[3]Instituto Superior de Agronomia e CEAUL, Universidade de Lisboa, Portugal, manela@isa.utl.pt

Research partially supported by National Funds through FCT, Foundation for Science and Tecnhology projects PEst-OE/MAT/UI0297/2011 (CMA/FCT/UNL) and PEst-OE/MAT/UI0006/2011 (CEAUL) and EXTREMA, PTDC/MAT/101736/2008.

Introduction

- Extreme value theory has been applied in various sciences, such as, biology and public health insurance, geology and seismic risk, risk assessment and telecommunications.
- Extreme value theory deals with events that are more extreme than any that have already been observed. The question is: how to make inference beyond the sample data? Statistical inference can be deduced only from those observations which are extreme in some sense.
- It is of great value for researchers to deal with accurate, friendly, free and open-source software.
- The objective of this work is to review some of the available R packages and functions therein by performing an exploratory data analysis, revisiting graphical techniques and also modeling extreme value data.

Background on extreme value analysis: Extreme Value d.f.

- The first results on asymptotic theory of samples extremes were due to Fréchet (1927), Fisher and Tippet(1928) and von Mises (1936). However was Gnedenko (1943) who gave conditions for the existence of sequences $\{a_n \in \mathbb{R}^+\}$ e $\{b_n \in \mathbb{R}\}$ such that,

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

where G is a nondegenerate distribution function. This function, called Extreme Value d.f. , is given by

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

The shape parameter, γ, is called extreme value index. The EV_γ d.f. can also incorporate location, $\mu\in\mathbb{R}$, and scale, $\sigma>0$, parameters. The EV_γ is the general form for the following distribution for maxima:

- Gumbel: $\Lambda(z)=\exp(-\exp(-z))=EV_0(z)$ $\quad z\in\mathbb{R}, \gamma=0$ - the right tail is of an exponential type.
- Fréchet: $\Phi_\gamma(z)=\exp(-z^{-1/\gamma})=EV_\gamma(\frac{z-1}{\gamma})$ $\quad z>0, \gamma>0$ - the right tail is heavy.
- Weibull: $\Psi_\gamma(z)=\exp(-(-z)^{1/\gamma})=EV_\gamma(\frac{-z-1}{\gamma})$ $\quad z<0, \gamma<0$ - the right tail is light.

Background on extreme value analysis: GP d.f.

- Similar theory holds for excess over a high threshold u. The correspondent stochastic behavior is approximately supported by the generalized Pareto (GP) distribution family, if block maxima have approximating distribution G.

$$H(y)=1-(1+\gamma y/\tilde{\sigma})^{-1/\gamma}, \{y: y>0 \text{ and } (1+\gamma y/\tilde{\sigma})>0\}$$

- The GP distribution of the excesses over u has two parameters: $\tilde{\sigma}$ - scale and γ - shape. The parameter γ is equal to that of the corresponding EV distribution and $\tilde{\sigma}=\sigma+\gamma(u-\mu)$.

Background on extreme value analysis: other parameters

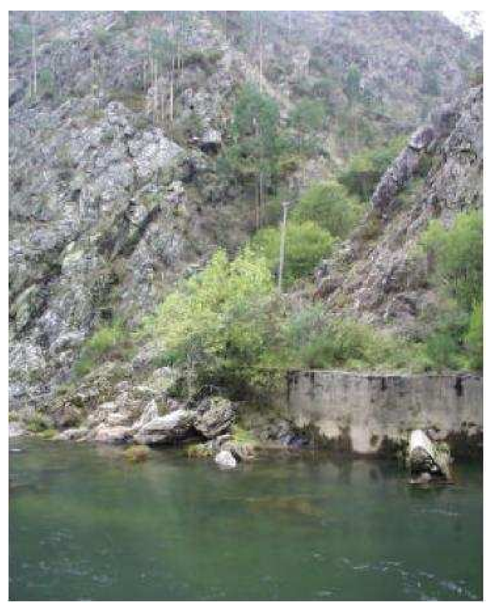
- The extreme value index, γ, is the basis of all parameters of extreme events, such as: high quantile of probability $1-p$, with p small, (return level); the probability of exceedance of a high level; the return period of a high level and the right endpoint of an underlying model.

The software R for the analysis of extremes

- The software R is an open source language for statistical computing, which allows the manipulation and data analysis, numerical computation and graphical production. A great advantage of R is to allow that several statistical techniques can be developed and implemented by users and made available as additional packages.
- Some packages for analysis of extremes values: evd; ismev; evir; fExtremes; POT; evdbayes; copula; SpatialExtremes

The Data

- We choose a set of data to illustrate the use of two parametric methodologies: the "block maxima" methodology based on EV distribution and the POT (Peaks Over a Threshold) methodology based on GP distribution.
- The data refers do daily mean river levels from hydrometric station at Fragas da Torre, during the years from 1946/47 to 1995/96. The 50 years observed correspond exactly to the period between October 1^{st} of 1946 to September 30^{th} of 1996.
- The source of Paiva river is in the Serra de Leomil, in the north of Portugal, an its hydrographic basin has an area of approximately 700 Km^2.
- The flow study of this river is a matter of major importance since it is one of the main alternatives to the Douro river as source of water supply in the south of Oporto region. In particular there is a project to build a dam precisely at Fragas da Torre.
- At the figure on the right we can see a picture of the hydrometric station at Fragas da Torre.



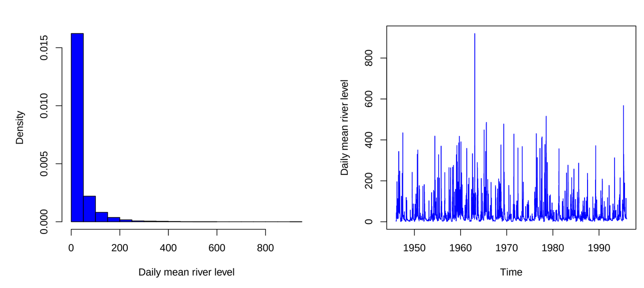
Source: IDAD, 2002

Exploratory Analysis

- Gomes(1993) chose a sub-sample constituted only by the months between November and February for which stationarity can be accepted. However, as the interest is to deal with extreme values, we saw that for most of the years, the months March and April presented very high values, so we decided to include them in this analysis. Stationarity can also be accepted for this sub-sample.
- To obtain the main descriptive statistics we can load the package *fBasics* and run the following commands:

- *>library(fBasics)* → load a package fBasics
- *>data(fraga)* → obtain the data
- *>basicStats(fraga)* → obtain some descriptive statistics
- *>hist(fraga)* → plot the histogram
- *>fraga2=ts(fraga, start=1946, frequency=181)* → define the data as times series data
- *>plot(fraga2)* → plot the times series data
- The histogram and the positive asymmetry indicate a tail much heavier that the normal one. From chronogram it seems reasonable to assume that the data are stationary.

Skewness	Kurtosis
4.12	27.60



- To verify the sub-sample stationarity we used the Augmented Dickey-Fuller Test wherein H1 represents the hypothesis of stationarity. Running the command:
  - *>adf.test(fraga, alternative = "stationary")*
- A p-value smaller than 0,01 was obtained therefore we can admit the stationarity of the data.

Block Maxima - Fitting EV Model

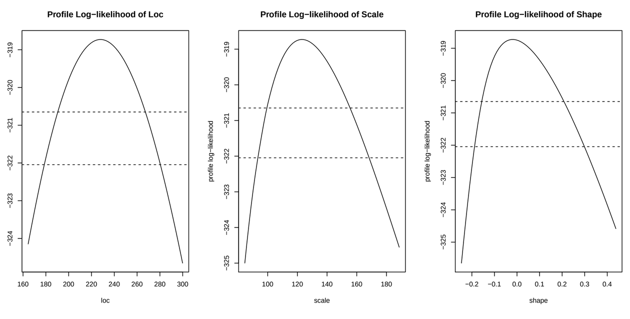
- The maximum likelihood fitting for the EV distribution, the Wald confidence intervals and the profile log-likelihood intervals for all the parameters can be obtained using the following functions:

- *>library(evir)* → using package evir
- *>gev(fraga, block=181)* → maximum likelihood estimates for the parameters, their standard deviations and the maximum of each block, among others
- *>library(evd)* → using package evd
- *>fgev(gev$data)*→ fit the EV distribution to the block maxima data
- *>confint(fgvev(gev$data), level=0.95)* → 95% Wald confidence interval for all the parameters
- *>plot(profile(fgvev(gev$data)), ci=c(0.95, 0.99))* → plot the profile log-likelihood for all the parameters
- *>confint(profile(fgvev(gev$data)),level=0.95)* → 95% confidence interval attained by the profile log-likelihood for all the parameters

$\hat{\mu}=227.35(19.34)$ $\quad\hat{\sigma}=122.73(13.87)$ $\quad\hat{\gamma}=-0.02(0.09)$

Wald confidence intervals

μ	σ	γ
(189.44;265.26)	(95.54;149.91)	(-0.2;0,16)



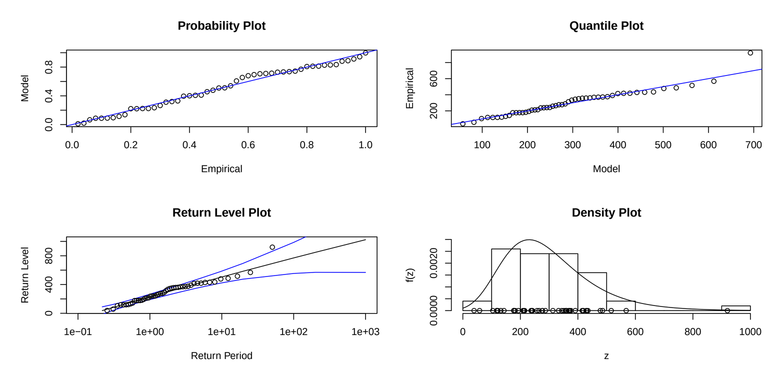
Profile log-likelihood intervals

μ	σ	γ
(190.53;267.22)	(99.55;155.46)	(-0.16;0,21)

- Notice that the confidence intervals for γ includes zero, so leads to not reject the null hypothesis, $\gamma=0$. The Gumbel distribution is then a possible candidate to model maximum river levels data.
  Greater accuracy for the confidence intervals is usual attained by the profile log-likelihood. In this case the results are similar.

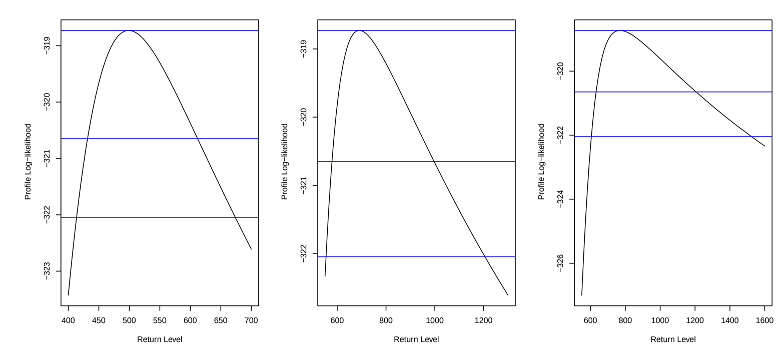
Block Maxima - Diagnostic Plots for EV Model

- *>library(mgcv); >library(ismev)* → using package ismev
- *>gev.diag(gev.fit(gev$data))* → diagnostic plots for assessing the accuracy of the EV model
- Both probability plot and quantile plot show a reasonable EV fit. The return level plot is linear and the density estimate is satisfactory modeling the histogram.



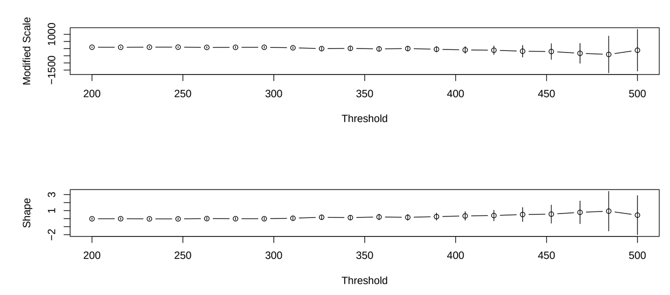
Block Maxima - Profile Log-likelihoods for Return Levels

- *>library(ismev)*
- *>gev.prof(gev.fit(gev$data), m=10, conf=c(0.95, 0.99), 400, 700)*
- *>gev.prof(gev.fit(gev$data), m=50, conf=c(0.95, 0.99), 550, 1300)*
- *>gev.prof(gev.fit(gev$data), m=100, conf=c(0.95, 0.99), 550, 1600)*



POT Methodology - To choose a threshold

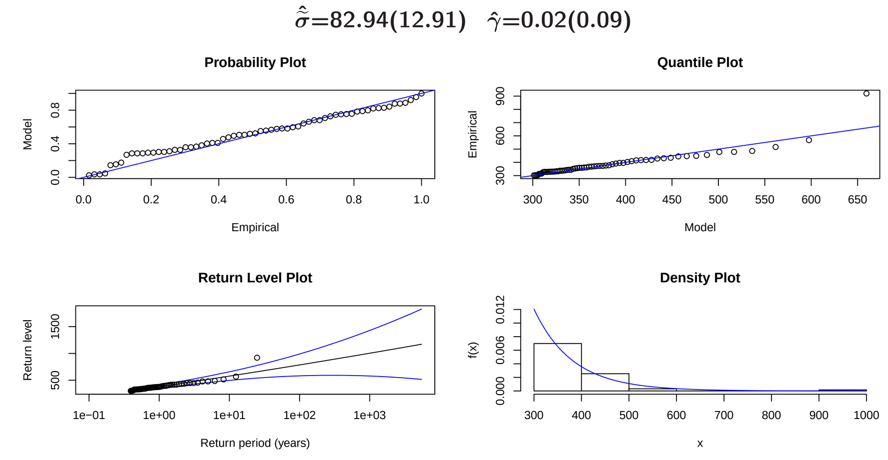
- *>library(POT)* → using package POT
- *>mrl.plot(fraga)* → mean residual life plot
- *>library(ismev)* → using package ismev
- *>gpd.fitrange(fraga, 200, 500, nint=20, show=TRUE)* → fitting the GP model over a range of thresholds
- Both graphs suggest a threshold around 300.



POT Methodology - Fitting GP Model and Diagnostic

- *>library(ismev)*
- *>fit300=gpd.fit(fraga, 300)* → maximum likelihood fitting for the GP model
- *>gpd.diag(fit300)* → diagnostic plots
- Both probability plot and quantile plot show a reasonable GP fit. The return level plot is linear and the density estimate is satisfactory modeling the histogram.

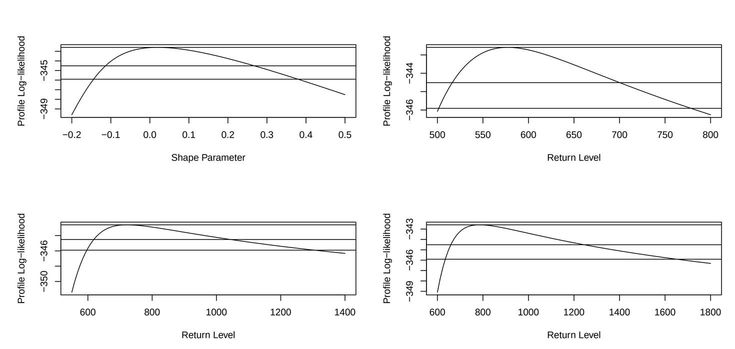
$\hat{\tilde{\sigma}}=82.94(12.91)$ $\quad\hat{\gamma}=0.02(0.09)$



POT Methodology - Profile for γ and Return Levels

- The profile log-likelihood for the shape parameter and for the return levels can be obtained using the following functions.

- *>library(ismev)*
- *>gpd.profxi(gpd.fit(fraga, 300), conf=c(0.95,0.99), -0.2, 0.5)* → profile log-likelihood plot for the shape parameter
- *>gdp.prof(gpd.fit(fraga, 300), m=10, conf=c(0.95,0.99), 500, 800)* → return level for 10 years
- *>gdp.prof(gpd.fit(fraga, 300), m=50, conf=c(0.95,0.99), 550, 1400)* → return level for 50 years
- *>gdp.prof(gpd.fit(fraga, 300), m=100, conf=c(0.95,0.99), 600, 1800)* → return level for 100 years



References

- 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.
- Fréchet, M. (1927): Sur la loi de probabilité de l'écart maximum. *Ann. Soc. Polon. Math. (Cracovie)*, **6**, 93-116.
- Gomes, M.I. (1993). On the estimation of parameters of rare events in environmental time series. Statistics for the Environment, 226-241.
- Gnedenko, B. V.(1943): Sur la distribution limite d'une série aléatoire *Annals of Mathematics*, **44**, 423-453.
- IDAD - Instituto do Ambiente e Desenvolvimento (2002): Captação no rio Paiva e Adução até 'ETA de Lever. Volume VI- Resumo Não Téncico. IMA-1202-01/27.
- 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.
- Pfaff, B. and McNeil, A. (2012). evir: Extreme Values in R. R package version 1.7-3, URL http://CRAN.R-project.org/package=evir.
- Stephenson, A. G. (2002). evd: Extreme Value Distributions. R News, URL http://CRAN.R-project.org/doc/Rnews/, 2(2), 31-32.
- 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.
- 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.

EVT2013 - Extremes in Vimeiro Today, Vimeiro, September2013