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

Helena Penalva1, Sandra Nunes2, Manuela Neves3

1Escola de Ciências Empresariais do Instituto Politécnico de Setúbal, Portugal, helena.penalva@descs.ips.pt
2Escola Superior de Ciências do Ambiente e Desenvolvimento, CIMA/FACT/UNL, Portugal, sandra.nunes@descs.ips.pt
3Instituto Superior de Agronomia e CEUL, Universidade de Lisboa, Portugal, manuela.neves@descs.ips.pt

Research partially supported by National Funds through FCT, Foundation for Science and Technology projects PEst-OE/MAT/UI0297/2011 (CIMA/FACT/UNL) and PEst-OE/MAT/UI0298/2011 (CEUL) and EXTREMA, P2/DIC/MAT/10378/2008.

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 based only from those observations which are extremes in a 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 themes by performing an exploratory data analysis, reviewing graphical techniques and also modeling extreme value data.

Background on extreme value analysis - Extreme Value D.f.

- The first results on an asymptotic theory of sample extremes were due to Fréchet (1927), Fisher and Tippett(1928) and von Mises (1928). However, see Vorobev (1948) who gives condition for the existence of sequences $K(n)/n$ such that:

$$U_n \sim (-n)^{1/\gamma}F_n$$

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

$$F(x;\gamma) = \exp\left\{-\left(\frac{x}{\gamma}\right)^\gamma\right\}$$

The shape parameter, $\gamma$, is called extreme value index. The $\gamma$ can also incorporate location, shape and scale, $\gamma$ parameter. The $\gamma$ is in general form for the following distribution for values:

- Gumbel: $\gamma = 0$, the right tail is of an exponential type.
- Fréchet: $\gamma < 0$, the right tail is of a power type.
- Weibull: $\gamma > 0$, the right tail is heavy.

Background on extreme value analysis: GP d.f.

- In a similar way for exceedance over a high threshold, $\alpha$, the so-called generalized Pareto (GP) distribution family, $F(x;\alpha,\gamma)$, where $F$ has a non-degenerate distribution function $F$, is given by

$$F(x;\alpha,\gamma) = \begin{cases} \frac{1}{\alpha} \left(1 + \frac{\gamma}{\alpha} x\right)^{-1/\gamma}, & \text{if } \gamma \neq 0, \\
1 - e^{-\frac{x}{\alpha}}, & \text{if } \gamma = 0. \end{cases}$$

The $\alpha$ or distribution of the exceedances over $\alpha$ has two parameters: $\gamma$, scale $\alpha$ and $\alpha$, shape $\gamma$. The parameter $\gamma$ is equal to that of the corresponding $\gamma$ distribution and $\alpha = \gamma^{-\gamma}$.

Background on extreme value analysis: other parameters

- The extreme value index $\gamma$ is in the basis of all parameters of extreme events, such as, high quantile of probability $\gamma$ with a small return level; the probability excess of a high level, the return period of a high level and the right-skewness 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 analysis, numerical computation and graphical production. A great advantage of R is that it allows several statistical techniques to be developed and implemented by users and made available as additional packages.
- Some packages for analysis of extreme values are: actuar, evd, POT, which include, Copula, SpatialExtremes, POT, etc.

The Data

- We choose a set of data to illustrate the use of two parametric methods: the “Gumbel maxima” methodology based on the exceedance data and the POT (Peaks Over a Threshold) methodology based on an exceedance.
- The data refers to daily max easy river levels from hydrometric stations at Fragas da Torre, during the years from 1946/47 to 1995/96. The 50 years observed corresponded exactly to the period between October 2nd of 1949 to September 30th of 1999.
- The source of Paiva river is the Serra de Lousã, in the north of Portugal, as its hydrographic basin has an area of approximately 700 km².
- The flow study of this river is a matter of major importance since it is one of the main alternatives to the Sinos river as sources of water supply in the south of Oporto region. In particular there is an interest in include a dams specifically at Fragas da Torre.
- At the figure on the right we can see a picture of the hydrometric station at Fragas da Torre.

Exploratory Analysis

- Gomes(1993) chose a sub-sample constituted only by the months between November and February for which stationarity can be presented very high values, so we decided to include them in this analysis. Stationarity can also be accepted for this sub-sample.
- The maximum likelihood fitting for the extremes values: evd; ismev; evir; fExtremes; POT; evdbayes; copula; SpatialExtremes; POT Methodology - To choose a threshold

- POT Methodology - Fitting of EV Model

- POT Methodology - Profile log-likelihoods for Return Levels

- POT Methodology - Fit the model using POT

- POT Methodology - Model and Diagnostic

- POT Methodology - Profile for $\gamma$ and Return Levels

References