Introduction to statistical modeling of extreme values

Facultad de Ciencias
Universidad Zaragoza

Universidad Zaragoza
1542

Inés Ortega Palacios

Trabajo de fin de grado en Matemáticas
Universidad de Zaragoza

Directores del trabajo: Jesús Abaurrea y Ana C. Cebrián
28 de junio de 2017

Prologue

The general aim of this work is to give a description of extreme value models. These models will be used to analyse the occurrence of floods in the Ebro river in Zaragoza. Different approaches have been considered in order to express the distribution of these extremes.

The first studies in extreme value theory were written in the first half of the 20th century. An article written by L.H.C. Tippett and R.A. Fisher in 1928 and the extended theory that B.V. Gnedenko wrote in 1948 based on their work were the first results in this area. In the second half of the same century, E.J. Gumbel captured the first statistical applications in his book *Statistics of extremes* (1958) [9] and described the Gumbel distribution to model rare physical phenomena. Other remarkable publications used in this work are *Extremes and related properties of random sequences and processes* by M.R. Leadbetter [13] and *An introduction to statistical modeling of extreme values* by S. Coles [3]. The specific objective of this work is the study of the distributions of extremes in order to apply them in a practical case.

The classical extreme value models define extremes as maximum (or minimum) values per unit of time. The class of distributions described in the first chapter is a 3-parameter distribution called the generalized extreme value (GEV) distribution.

The "Excess over threshold" method suggests that an extreme value is an observation which exceeds a concrete threshold. The choice of this threshold can be challenging depending on the data set. The second chapter shows the limit distribution which is fitted to these extreme values, which is called the generalized Pareto distribution.

The occurrence of excesses over a fixed threshold follows a Poisson process. A characterization of a Poisson process and its relation with both GEV and Pareto distributions are given in the third chapter.

An important consideration is the condition for which the extreme distributions can be fitted to dependent sequences of observations. A cluster process is considered to represent the occurrences.

In order to predict floods in the Ebro river in Zaragoza, estimations of distributions of both models and return levels are calculated in the last chapter. The statistical programming language R is the main tool used in this work in order to estimate these extreme value models and their return levels for this data set.

Resumen

La teoría de valores extremos es una rama de la estadística que centra su interés en el comportamiento de los valores más altos (o más bajos) de la variable a estudiar.

Los primeros resultados de la teoría de valores extremos datan de la primera mitad del siglo XX. Los artículos expuestos por L. Tippett y R.A. Fisher en 1928 fueron pioneros en el área y en 1958, E.J. Gumbel plasmó esta teoría en su libro *Statistics of extremes* [9] en el que incluía la distribución de valores extremos que hoy en día lleva su nombre. En la actualidad, la teoría de valores extremos es aplicada en muchos campos, como la hidrología, el análisis de riesgos en finanzas o la geología.

El primer enfoque que se considera en este trabajo al describir valores extremos es el análisis de máximos. Sea una sucesión de variables aleatorias independientes $X_1, \ldots, X_n$ con la misma función de distribución F. Teóricamente, la distribución de $M_n = \max(X_1, \ldots, X_n)$ se puede obtener de la forma

$$\wp(M_n \leq z) = \wp(X_1 \leq z) \times \cdots \times \wp(X_n \leq z) = (F(z))^n$$

Usualmente, la distribución F es desconocida. El teorema de Fisher-Tippet-Gnedenko, desarrollado por los dos primeros en 1928 y probado por el tercero en 1943, permite obtener una aproximación de la distribución de máximos, M_n. Dadas dos sucesiones de constantes $\{a_n > 0\}$ y $\{b_n\}$ tales que $\text{lím}_{n \to \infty} \wp((M_n - b_n)/a_n \leq z) = G(z)$ con $G(z)$ función de distribución no degenerada, entonces G es de la forma

$$G(z) = exp\left(-\left(1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right)^{-1/\xi}\right)$$

con $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$, $-\infty < \mu < \infty$, $\sigma > 0$ y $-\infty < \xi < \infty$.

Esta distribución es conocida como "Distribución de Valores Extremos Generalizada (VEG)". Dado un conjunto de datos en un caso práctico, se puede ajustar la VEG a la serie de máximos del conjunto, generalmente anuales, y así obtener estimaciones de los parámetros de la distribución.

El segundo enfoque considerado en el trabajo es la descripción de valores extremos de un conjunto de datos como aquellas observaciones que superen un umbral fijo. Este método se denomina "Método de excesos sobre umbral". Las dos distribuciones que se deben ajustar en este caso son la distribución del número de ocurrencias sobre el umbral en un periodo de tiempo y la distribución de los excesos sobre umbral.

Dada una variable aleatoria X con función de distribución desconocida F, la distribución de excesos sobre un umbral u se describe como $F_u(x) = \wp\{X - u \leq x | X > u\}$ con $x \geq 0$.

Utilizando los resultados de la teoría del análisis de máximos, F^n puede ajustarse por una VEG y así se puede obtener la distribución asintótica de los excesos sobre el umbral:

$$H(y) = 1 - \left(1 + \frac{\xi y}{\sigma_u}\right)^{-1/\xi}$$

definida en $\{y : y > 0,\ (1 + \xi y/\sigma_u > 0)\}$ con $\sigma_u = \sigma + \xi(u - \mu)$. Esta distribución se denomina "Pareto generalizada".

Antes de ajustar una distribución Pareto a un conjunto de datos se debe fijar un umbral. Su elección puede ser complicada, ya que si se elige un umbral demasiado alto, el número de excesos no sería suficiente para ajustar la distribución y un umbral demasiado bajo proporcionaría un gran número de excesos de los cuales no todos serían valores extremos reales. Utilizando una gráfica de vida media residual, que relaciona el valor del umbral con la media de los valores de los excesos para ese umbral, se obtiene un procedimiento para seleccionar u en los puntos donde la gráfica es lineal.

En este trabajo se estudia cómo se pueden hacer estimaciones del número de ocurrencias y comprobar que estos resultados guardan una estrecha relación con la estimación de los parámetros en la distribución de VEG y la distribución Pareto generalizada.

Al considerar la aplicación de la teoría de valores extremos a casos prácticos, se debe tener en cuenta que ésta supone en todo momento la independencia de las variables a estudio y en la realidad la independencia no es habitual. En el tercer capítulo se explican las condiciones que debe cumplir un conjunto de datos para que se puedan ajustar las distribuciones de extremos. Es suficiente con ver que si la distribución tiene un carácter estacionario y no tiene dependencia a largo plazo, muchos de los resultados para series independientes pueden aplicarse en este caso.

En el cuarto capítulo se considera el conjunto de datos de niveles de agua diarios, medidos en metros, del Río Ebro a su paso por Zaragoza desde 1961 a 2016. Por sus condiciones físicas, el nivel del río presenta dependencia a corto plazo. Por lo tanto, es correcto aplicar en este caso la teoría de valores extremos expuesta en el trabajo para modelizar la ocurrencia de riadas en Zaragoza. Utilizando el lenguaje de programación estadística R, se ajustan una distribución VEG y una distribución Pareto generalizada de modo que se obtienen probabilidades de niveles de retorno de la altura del agua.

Contents

Chapter 1

Classical extreme value theory and models

The statistical behaviour of the distribution of maxima of independent random variables with a common distribution function is the main objective of this chapter. Thus, basic ideas of extreme value theory and its models will be presented.

The central result is the Fisher-Tippett-Gnedenko Theorem, which describes the form of the limit distribution of the normalised maxima. The three models obtained in this result form the Generalized Extreme Value family of distributions which will be explained in detail in Section 1.1.2. Another consideration is the characterisation of the model for minima and the analytical solution that can be found for return levels (Section 1.1.3).

The Generalized Extreme Value parameters and return levels will be estimated using the maximum likelihood estimation method.

1.1. Asymptotic Models

1.1.1. Model Formulation

Let $X_1, \cdots, X_n$ be independent identically distributed random variables and let

$$M_n = \max(X_1, \cdots, X_n) \tag{1.1}$$

Let F be the distribution function of $X_1, \cdots, X_n$. The distribution function of M_n can be calculated exactly for all values of n using the independence of the random variables as follows:

$$\wp(M_n \leq z) = \wp(X_1 \leq z, \cdots, X_n \leq z) = \wp(X_1 \leq z) \times \cdots \times \wp(X_n \leq z) = (F(z))^n$$

Although this may seem useful, there is a problem when calculating this if the distribution function F is unknown.

When F is estimated, it is clear that there can be little discrepancies due to the lack of data. In this case, some calculations are needed to obtain an estimation of F^n, but the n^{th}-power of a little error leads to a problem with huge discrepancies. Thus, in order to get the distribution function of maxima, F^n, another way to solve the problem is to assume that F is unknown and try to estimate F^n using observed extreme data.

Now, let the behaviour of F^n when $n \to \infty$ be considered. As F is a distribution function, let z_+ be the upper-end point of F, that is, $z_+ = \sup\{x \in \mathbb{R} : F(x) < 1\}$. Then the following equality is immediately obtained

$$\forall z < z_+,\ \wp(M_n \leq z) = F^n(z) \to 0 \text{ when } n \to \infty$$

and in the case $z_+ < \infty$, for any $z \geq z_+$

$$\wp(M_n \leq z) = F^n(z) = 1$$

Thus, the distribution of M_n degenerates to a point mass on z_+.

The difficulty of calculating M_n can be partly saved by giving a linear normalization of M_n. With an appropriate choice of two sequences of constants, $(a_n > 0)$ and (b_n),

$$M_n^* = \frac{M_n - b_n}{a_n}$$

and this allows to study the behaviour of M_n^* instead of M_n, which will be useful in order to explain the general extreme value theory.

1.1.2. Extremal Types Theorem and its generalization

Some results, based on [13, Section 1.4], need to be considered before giving an accurate definition of the Extremal Types theorem.

Definition 1.1. *A distribution function G is said to be max-stable if, for every $n \in \mathbb{N}, n > 1$, there are constants $a_n > 0$ and b_n such that*

$$G^n(a_n z + b_n) = G(z)$$

From this concept and in order to continue with the results, several definitions of the domain of attraction can be given. Quoting [14],

Definition 1.2. *Let $X_1, \ldots, X_n$ be mutually independent random variables with common distribution function $F(z)$ and M_n the maximum of these random variables. Suppose there exist a pair of sequences $(a_n > 0)$ and (b_n) and a distribution function $G(z)$ such that*

$$\lim_{n\to\infty} \wp\left\{\frac{M_n - b_n}{a_n} \le z\right\} = \lim_{n\to\infty} F^n(a_n z + b_n) = G(z)$$

for all z at which $G(z)$ is continuous. Then $F(z)$ lies in the domain of attraction of G , $F \in D(G)$.

Theorem 1.3. *(i) A non-degenerate distribution function G is max-stable if and only if there is a sequence $\{F_n\}$ of distribution functions and constants $a_n > 0$ and b_n such that*

$$F_n(a_{nk}^{-1} z + b_{nk}) \to G^{1/k}(z) \tag{1.2}$$

as $n \to \infty$ for each $k = 1, 2, \cdots$
(ii) In particular, if G is non-degenerate, $D(G)$ is non-empty if and only if G is max-stable. Then also $G \in D(G)$

The following theorem contains the most important result about extreme value distribution functions. Fisher and Tippett started the research in 1928, and later, Gnedenko formalized it in 1948.

Theorem 1.4 (Fisher - Tippett - Gnedenko theorem)**.** *If there exist sequences of constants $(a_n > 0)$ and (b_n) such that*

$$\lim_{n\to\infty} \wp\left(\frac{M_n - b_n}{a_n} \le z\right) = G(z)$$

where G is a non-degenerate distribution function, then G belongs to one of the following families
I: $G(z) = \exp\left(-\exp\left(-\left(\frac{z-b}{a}\right)\right)\right)$, $-\infty < z < \infty$ (Gumbel distribution)

$$\text{II: } G(z) = \begin{cases} 0 & z \le b \\ \exp\left(-\left(\frac{z-b}{a}\right)^{-\alpha}\right) & z > b \end{cases} \quad \text{(Fréchet distribution)}$$

$$\text{III: } G(z) = \begin{cases} \exp\left(-\left(-\left(\frac{z-b}{a}\right)^{\alpha}\right)\right) & z < b \\ 1 & z \ge b \end{cases} \quad \text{(Weibull distribution)}$$

for a scale parameter $a > 0$, a location parameter b and a shape parameter $\alpha > 0$.

This theorem shows that the only possible limiting distribution for M_n^* given M_n and sequences $(a_n > 0)$ and (b_n), is one of these three types. In some way, this theorem gives an extreme value analogue of the central limit theorem, as M_n^* is the normalized distribution of M_n.

By reformulating the three models in Theorem 1.4, they can be combined into a single family of models as follows:

Theorem 1.5. *If there exist sequences of constants $(a_n > 0)$ and (b_n) such that*

$$\lim_{n\to\infty} \wp\left(\frac{M_n - b_n}{a_n} \leq z\right) = G(z)$$

for a non-degenerate distribution function, G, then G is a member of the family

$$G(z) = exp\left(-\left(1+\xi\left(\frac{z-\mu}{\sigma}\right)\right)^{-1/\xi}\right) \tag{1.3}$$

defined on the set: $\{z : 1+\xi(z-\mu)/\sigma > 0\}$ with location parameter $-\infty < \mu < \infty$, scale parameter $\sigma > 0$ and shape parameter $-\infty < \xi < \infty$. From now on, this family will be called the generalized extreme value (GEV) family of distributions.

Note 1.6. *The model (1.3) has a location parameter, μ; a scale parameter, σ; and a shape parameter, ξ. Observe that given the three distributions in Theorem 1.4, type II corresponds to the case $\xi > 0$; type III to the case $\xi < 0$ and the Gumbel distribution to the GEV family's subset with $\xi = 0$*

The notation in Theorem 1.4 can be simplified. Before giving a new version of the theorem, the following equivalence relation is defined.

Definition 1.7. *Two distributions F and F^* are said to be of the same type if there exist constants a and b such that $F^*(az+b) = F(z)$ for all z.*

Theorem 1.8. *Every max-stable distribution is of extreme value type and equal to $G(az+b)$ for some $a > 0$ and b where for*

I: $G(z) = \exp(-\exp(-z))$, $-\infty < z < \infty$

II: $G(z) = \begin{cases} 0 & z \leq 0 \\ \exp(-z^{-\alpha}) & z > 0 \end{cases}$

III: $G(z) = \begin{cases} \exp\left(-(-z)^{\alpha}\right) & z < 0 \\ 1 & z \geq 0 \end{cases}$

Conversely, each distribution of extreme value type is max-stable.

With all these results a brief proof of the theorem can be given.

Theorem 1.9. *Let M_n be as defined in (1.1), then for some constants $a_n > 0$ and b_n it satisfies*

$$\wp\left(\frac{M_n - b_n}{a_n} \leq z\right) \to G(z) \tag{1.4}$$

for some non-degenerate G if and only if G is one of the three extreme value type distributions defined in Theorem 1.8.

Proof. If (1.4) is true, then Theorem 1.3 shows that G has to be max-stable, and as it is shown in Theorem 1.8 is of the extreme value type. Conversely, if G is an extreme value function type, in particular, it is max-stable, and Theorem 1.3 shows that $G \in D(G)$ and the result holds. □

1.1.3. Return levels

Quantiles can be obtained by finding the inverse of (1.3).

In common terminology, z_p is defined as the return level associated to the return period $1/p$. The probability of the occurrence of z_p is p and once every $1/p$ years the annual maximum is expected to be greater than z_p.

The return level z_p is exceeded by the annual maximum in a particular year with probability p, that is, z_p is expected to be exceeded by the annual maximum once every $1/p$ years.

Given some data series, they can be grouped in packages of n observations, for a large value n, and generate a series of block maxima. Often they are chosen to correspond to a time period, for example one year. Using this information, this series can be fitted to a GEV distribution.

In order to estimate the return level, the general equation of the GEV distribution leads to,

$$1-p=\exp\left(-\left(1+\xi\left(\frac{z_p-\mu}{\sigma}\right)\right)^{-1/\xi}\right) \Rightarrow 1+\xi\left(\frac{z_p-\mu}{\sigma}\right)=\left[-\log(1-p)\right]^{-\xi}$$

and therefore

$$z_p=\begin{cases} \mu-\frac{\sigma}{\xi}\left[1-\{-\log(1-p)\}^{-\xi}\right] & \xi\neq 0 \\ \\ \mu-\sigma\log\{-\log(1-p)\} & \xi=0 \end{cases} \tag{1.5}$$

Note 1.10. *When plotting z_p against $-\log(1-p)$ on a logarithmic scale three cases be can distinguished depending on the value of ξ. Let $y_p=-\log(1-p)$. Then, z_p can be plotted against $x_p=log(y_p)$ so,*

- *If $\xi=0$, the following expression holds,*

$$z_p=\mu-\sigma\log\{-\log(1-p)\}=\mu-\sigma\log(y_p)=\mu-\sigma x_p$$

Thus, the plot is linear.

- *If $\xi\neq 0$, z_p satisfies*

$$\mu-\frac{\sigma}{\xi}\left[1-\{-\log(1-p)\}^{-\xi}\right]=\mu-\frac{\sigma}{\xi}\left[1-y_p^{-\xi}\right]=\mu-\frac{\sigma}{\xi}\left[1-\exp(-\xi x_p)\right]$$

  - *If $\xi<0$, the plot is convex and it has its asymptotic limit when $p\to 0$ at $\mu-\frac{\sigma}{\xi}$.*
  - *If $\xi>0$, the plot is concave and has not finite bound.*

This graph is called the return level plot.

1.1.4. General distribution for minima

As some applications may need a model of minima instead of maxima, some transformations can be applied to the previous results in order to give an expression of the distribution of $\tilde{M}_n=min\{X_1,\dots,X_n\}$, where the X_i are independent random variables with a common distribution function.

Let $Y_i=-X_i$ for $i=1,\dots,n$. Each large value of Y_i corresponds to the small value of X_i so if $\tilde{M}_n=min\{X_1,\dots,X_n\}$ and $M_n=max\{Y_1,\dots,Y_n\}$ it is easy to see that $\tilde{M}_n=-M_n$ and for a large n,

$$\wp\{\tilde{M}_n\leq z\}=\wp\{-M_n\leq z\}=\wp\{M_n\geq -z\}=1-\wp\{M_n\leq -z\}\approx$$

$$1-exp\left(-\left(1+\xi\left(\frac{-z-\mu}{\sigma}\right)\right)^{-1/\xi}\right)=1-exp\left(-\left(1-\xi\left(\frac{z-\tilde{\mu}}{\sigma}\right)\right)^{-1/\xi}\right)$$

on $\{z:1-\xi(z-\tilde{\mu})/\sigma>0\}$ where $\tilde{\mu}=-\mu$. This distribution is the GEV distribution for minima. Therefore, in a similar way to Theorem 1.5, there is a theorem for the distribution of minima ([3, Theorem 3.3]).

Theorem 1.11. *If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that*

$$\wp\left(\frac{\tilde{M}_n - b_n}{a_n} \leq z\right) \rightarrow \tilde{G}(z)$$

as $n \rightarrow \infty$ for a non-degenerate distribution function, $\tilde{G}$, then $\tilde{G}$ is a member of the GEV family of distributions for minima where $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$.

1.2. Inference for the GEV Distribution

1.2.1. Maximum likelihood estimation

Let $Z_1, \dots, Z_n$ be independent random variables with the GEV distribution function and suppose that $\xi \neq 0$. Knowing that the general distribution function is defined as in Theorem 1.5, the density function of a random variable Z with parameters (μ, σ, ξ) can be obtained as follows,

$$f(\mu, \sigma, \xi | z) = \frac{d}{dz} G(z) = \frac{1}{\sigma}\left(1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right)^{-1-1/\xi} - \left(1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right)^{-1/\xi}$$

Thus, the log-likelihood of $Z_1, \dots, Z_n$ can be calculated using the definition,

$$l\,(\mu, \sigma, \xi | Z_1, \dots, Z_n) = \log\left(\prod_{i=1}^{n} f(\mu, \sigma, \xi | Z_i)\right) = \sum_{i=1}^{n} \log f(\mu, \sigma, \xi | Z_i) =$$

$$= \sum_{i=1}^{n} \log\left[\frac{1}{\sigma}\left(1 + \xi\left(\frac{Z_i - \mu}{\sigma}\right)\right)^{-1-1/\xi} - \left(1 + \xi\left(\frac{Z_i - \mu}{\sigma}\right)\right)^{-1/\xi}\right] =$$

$$= -m \log \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{n} \log\left[1 + \xi\left(\frac{Z_i - \mu}{\sigma}\right)\right] - \sum_{i=1}^{n}\left[1 + \xi\left(\frac{Z_i - \mu}{\sigma}\right)\right]^{-1/\xi} \tag{1.6}$$

where the parameters (μ, σ, ξ) satisfy

$$1 + \xi\left(\frac{Z_i - \mu}{\sigma}\right) > 0, \; for\, i = 1, \dots, n$$

In the case $\xi = 0$, the Gumbel limit of the GEV distribution shall be used obtaining

$$l(\mu, \sigma | Z_1, \dots, Z_n) = \sum_{i=1}^{n} \log\left(\frac{1}{\sigma} \exp\left\{-\left(\frac{Z_i - \mu}{\sigma}\right)\right\} \exp\left\{-\exp\left\{-\left(\frac{Z_i - \mu}{\sigma}\right)\right\}\right\}\right) =$$

$$= -m \log \sigma - \sum_{i=1}^{n}\left(\frac{Z_i - \mu}{\sigma}\right) - \sum_{i=1}^{n} \exp\left\{-\left(\frac{Z_i - \mu}{\sigma}\right)\right\} \tag{1.7}$$

There is no analytical solution for the maximization of (1.6) and (1.7), but a numerical solution can be obtained by using standard numerical algorithms.

1.2.2. Inference for return levels

Once the maximum likelihood estimators $(\tilde{\mu}, \tilde{\sigma}, \tilde{\xi})$ of the GEV distribution are calculated, they can be substituted in order to estimate the maximum likelihood estimator of z_p for $0 < p < 1$ ($1/p$ return level). Using (1.5),

$$\tilde{z}_p = \begin{cases} \tilde{\mu} - \frac{\tilde{\sigma}}{\tilde{\xi}}\left[1 - y_p^{-\tilde{\xi}}\right] & \tilde{\xi} \neq 0 \\ \\ \tilde{\mu} - \tilde{\sigma} \log y_p & \tilde{\xi} = 0 \end{cases}$$

where $y_p = -\log(1-p)$. An approximation of the variance of $\tilde{z}_p$ can be found by using the Delta method.

Theorem 1.12 (Delta method)**.** *Suppose that $\mathbb{X} = (X_1, \ldots, X_k)$ is a random vector satisfying*

$$\sqrt{n}\,(\mathbb{X} - \mu) \xrightarrow{d} N(0, \Sigma)$$

where Σ is the covariance matrix. Let $h : \mathbb{R}^k \to \mathbb{R}$ be a differentiable function and let

$$\nabla h(\mathbb{X}) = \begin{pmatrix} \frac{\partial h}{\partial X_1} \\ \vdots \\ \frac{\partial h}{\partial X_k} \end{pmatrix} \tag{1.8}$$

then

$$\sqrt{n}\,[h(\mathbb{X}) - h(\mu)] \xrightarrow{d} N\left(0, \nabla h(\mu)^T \cdot \Sigma \cdot \nabla h(\mu)\right)$$

Therefore, by applying the Delta method, an approximation of $Var(h(\mathbb{X}))$ can be found by using the covariance matrix Σ.

Hence, an approximation of the variance of $\tilde{z}_p$ is given by

$$Var\,(\tilde{z}_p) = \nabla z_p^T \cdot V \cdot \nabla z_p^T$$

where V is the variance-covariance matrix of $(\tilde{\mu}, \tilde{\sigma}, \tilde{\xi})$ and

$$\nabla z_p^T = \left[\frac{\partial z_p}{\partial \mu}, \frac{\partial z_p}{\partial \sigma}, \frac{\partial z_p}{\partial \xi}\right] = \left[1, -\xi^{-1}\left(1 - y_p^{-\xi}\right), \sigma\xi^{-2}\left(1 - y_p^{-\xi}\right) - \sigma\xi^{-1} y_p^{-\xi} \log y_p\right]$$

evaluated at $(\tilde{\mu}, \tilde{\sigma}, \tilde{\xi})$.

Chapter 2

Threshold Models

Threshold models are based on the study of the limit distribution of exceedances over a fixed threshold: the generalized Pareto distribution is the distribution which models the behaviour of these variables.

An important decision in order to obtain a good extreme value behaviour of excesses is the choice of the appropriate threshold. In this chapter, two intuitive methods to find a valid threshold will be explained, as well as an analytical solution to the return levels for the Pareto distribution. Estimators of parameters and return levels will be obtained using the maximum likelihood estimation method.

The form of the probability and quantile plots of the model against density and return level will be shown, as these plots are useful to evaluate the quality of a fitted generalized Pareto distribution.

2.1. Asymptotic Model Characterization

Let $X_1, X_2, \cdots$ be a sequence of independent random variables with common marginal distribution function, F. It can be intuitive to say that X_i is an extreme event if it exceeds some fixed threshold, u. A theoretical description for this behaviour can be given by the following conditional probability.

Definition 2.1 (Excess distribution function)**.** *Let X be a random variable with distribution function F. For a fixed u,*

$$F_u(x) = \wp\{X - u \leq x | X > u\},\ x \geq 0 \tag{2.1}$$

is the excess distribution function of the random variable X over the threshold u.

Using this definition we have the following equality,

$$\wp\{X > u + x | X > u\} = \frac{1 - F(u+x)}{1 - F(u)} \tag{2.2}$$

If the distribution function F was known there would be no problem in calculating that probability and obtaining a formula for the distribution. As this is usually not the case, under the same conditions where the GEV distribution function can be used as an approximation to the distribution function for maxima of long sequences, an explicit expression of (2.1) can be obtained by substituting the distribution given in Theorem 1.5 for the distribution function for maxima, F^n.

2.1.1. The Generalized Pareto Distribution: Description and justification

The main Pareto distribution result is given in the following theorem.

Theorem 2.2 (Pickands (1975), Balkema and de Haan (1974))**.** *Let $X_1, X_2, \cdots$ be a sequence of independent random variables with common distribution function F, and let*

$$M_n = max\{X_1, \cdots, X_n\}.$$

Denote an arbitrary element in the sequence as X, and suppose that F satisfies Theorem 1.5 , then for large enough u, the distribution function of $(X-u)$, conditional on $X>u$, is approximately

$$H(y) = 1-\left(1+\frac{\xi y}{\sigma_u}\right)^{-1/\xi} \tag{2.3}$$

defined on $\{y : y>0,\ (1+\xi y/\sigma_u > 0)\}$ *where* $\sigma_u = \sigma + \xi(u-\mu)$.

This distribution is called the generalized Pareto distribution (GPD) where σ_u is the scale parameter and ξ is the shape parameter.

Proof. Assuming that Theorem 1.5 is true, for large n,

$$F^n(z) \approx exp\left\{-\left(1+\xi\left(\frac{z-\mu}{\sigma}\right)\right)^{-1/\xi}\right\}$$

for some parameters $\mu, \sigma > 0$ and ξ. Taking a logarithmic scale,

$$n\log F(z) \approx -\left(1+\xi\left(\frac{z-\mu}{\sigma}\right)\right)^{-1/\xi} \tag{2.4}$$

Note that for large enough n, a logarithmic expression can be approximated by its Taylor expansion. That implies:

$$\log(1+z) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} z^n = z - \frac{z^2}{2} + \frac{z^3}{3} - \frac{z^4}{4} + \cdots$$

which means that $\log F(z) \approx -(1-F(z))$. Replacing the approximation in (2.4),

$$n(1-F(u)) \approx \left(1+\xi\left(\frac{u-\mu}{\sigma}\right)\right)^{-1/\xi} \implies 1-F(u) \approx \frac{1}{n}\left(1+\xi\left(\frac{u-\mu}{\sigma}\right)\right)^{-1/\xi} \tag{2.5}$$

Equivalently, for $y>0$,

$$1-F(u+y) \approx \frac{1}{n}\left(1+\xi\left(\frac{u+y-\mu}{\sigma}\right)\right)^{-1/\xi}$$

Hence,

$$\wp\{X > u+y \mid X>u\} \approx \frac{n^{-1}\left[1+\xi\left(u+y-\mu\right)/\sigma\right]^{-1/\xi}}{n^{-1}\left[1+\xi\left(u-\mu\right)/\sigma\right]^{-1/\xi}} =$$

$$= \left[1+\frac{\xi\left(u+y-\mu\right)/\sigma}{1+\xi\left(u-\mu\right)/\sigma}\right]^{-1/\xi} = \left[1+\frac{\xi y}{\sigma_u}\right]^{-1/\xi} \tag{2.6}$$

where $\sigma_u = \sigma + \xi\left(u-\mu\right)$ □

Note 2.3. *If $\xi < 0$, the distribution of excesses has an upper bound for $u - \sigma_u/\xi$ but if $\xi > 0$, the distribution has not an upper limit. For $\xi = 0$, a distribution approximation can be obtained by taking the limit $\xi \to 0$. Then,*

$$H(y) = 1 - exp\left(-\frac{y}{\sigma_u}\right) \tag{2.7}$$

2.1.2. Examples

Given a distribution function, F, the Pareto distribution can be calculated as follows:

Example 2.4. *For $X \sim U(0,1)$, the distribution function $F(z) = z$, where $z \in [0,1]$. So,*

$$\wp\{X > u+y \mid X > u\} = \frac{1-F(u+y)}{1-F(u)} = \frac{1-(u+y)}{1-u} = 1 - \frac{y}{1-u}$$

for $0 \leq y \leq 1-u$. This is a generalized Pareto distribution with $\sigma_u = 1-u$ and $\xi = -1$

Example 2.5. *Given $X \sim \exp(1)$, the distribution function is of the form $F(x) = 1-e^{-x}$ for $x > 0$. Using the definition of Pareto distribution,*

$$\wp\{X > u+y \mid X > u\} = \frac{1-F(u+y)}{1-F(u)} = \frac{e^{-(u+y)}}{e^{-u}} = e^{-y}$$

for $y > 0$. Thus, it corresponds to the generalized Pareto distribution with $\xi = 0$ and $\sigma_u = 1$

The GEV distribution describes the limit distribution of normalised maxima while the GPD considers the limit distribution of excesses over thresholds using the GEV distribution as the approximation of the distribution of the maxima. Actually, the value of ξ is common across the two models and the value of σ is threshold-dependent, except in the particular case where $\xi = 0$.

2.2. Modeling Threshold Excesses

2.2.1. Methods of Threshold Selection

Given a sequence of independent and identically distributed variables, $X_1, \cdots, X_n$ and a high threshold, u, the excesses of these variables are $\{x_i : x_i > u\}$. Grouping these values as $x_{(1)} \leq \cdots \leq x_{(k)}$ (all the values in the data set that are greater than the fixed value of the threshold u), the threshold excesses are defined by: $y_j = x_{(j)} - u$ for $j = 1, \cdots, k$. Applying Theorem 2.2, these excesses correspond to independent observations of a random variable whose distribution function can be approximated by a member of the generalized Pareto family if the considered threshold is extreme enough.

One problem in order to use this approach is the choice of the threshold. Choosing a too high threshold leads to small samples and the model would not be very accurate due to the lack of data. If the chosen threshold was too low, a very little number of observations would be real extreme values.

Two different methods can be considered in order to select that threshold.

- The first method consists on the study of the mean residual life plot, which focuses on the mean of the generalized Pareto distribution. If Y is a random variable with generalized Pareto distribution with parameters σ_u and ξ, it holds

  $$E(Y) = \begin{cases} \frac{\sigma_u}{1-\xi} & \xi < 1 \\ \infty & \xi \geq 1 \end{cases}$$

  This method is based on the behaviour of the distribution of the excesses when the value of the threshold u changes.

  Let σ_{u_0} be the scale parameter of the generalized Pareto distribution which corresponds to the excess of the threshold u_0 and let $\xi < 1$. Given $Y_1, \cdots, Y_n$ a series of variables such that the generalized Pareto distribution is a valid distribution for the excesses over a threshold u_0 and taking Y an arbitrary element of the series,

  $$E(Y - u_0 \mid Y > u_0) = \frac{\sigma_{u_0}}{1-\xi}$$

If the distribution is valid for u_0, it is valid for any $u > u_0$. Hence, for $u > u_0$

$$E(Y-u \mid Y>u) = \frac{\sigma_u}{1-\xi} = \frac{\sigma_{u_0} + \xi(u-u_0)}{1-\xi}$$

as $\sigma_u = \sigma + \xi(u-\mu)$. It is actually a linear function of u. According to this, the function $E(Y-u_0 \mid Y > u_0)$ is expected to change linearly with u for values of u for which the Pareto distribution is appropriate.

Let $x_{(1)}, \cdots, x_{(k)}$ be the n_u observations that exceed a threshold u, and x_{max} the largest of the X_i. The mean residual life plot is defined as the locus of points of the form

$$\left\{ \left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{(i)} - u) : u < x_{max} \right) \right\} \tag{2.8}$$

Thus, under a right choice of threshold u for which the generalized Pareto distribution approximates to the distribution of excesses, $x_{(1)}, \cdots, x_{(k)}$, the mean residual plot must be linear.

- The second method is based on the estimation of the model using a sequence of values for the threshold u, taking into account the linear relation of u and σ_u and knowing that ξ should be constant with respect to u.

  If the generalized Pareto distribution is valid for a threshold u_0, then according to Theorem 2.2 excesses for a higher threshold also follow a Pareto distribution. Let σ_u be the value of the scale parameter of a Pareto distribution for a threshold $u > u_0$. Then following (2.3), $\sigma_u = \sigma_{u_0} + \xi\,(u-u_0)$, so the scale parameter changes if $\xi \neq 0$. This inconvenience is avoided by reparameterizing

  $$\sigma^* = \sigma_u - \xi u = \sigma_{u_0} - \xi u_0$$

  and therefore σ^* and ξ should be constant for any $u > u_0$ above u_0 if u_0 is a valid threshold.

  The estimations of these parameters will not be exactly constant due to sampling variability, but approximately if u_0 is a valid threshold. Therefore, confidence intervals for both quantities, σ^* and ξ can be plotted against u. As they both have to be constant as u changes, u_0 should be the lowest value of u for which the estimates remain near-constant.

  The confidence interval for $\hat{\xi}$ can be obtained from the variance-covariance matrix V (see Section 2.2.2) and confidence intervals for $\hat{\sigma}^*$ con be obtained using the delta method,

  $$Var(\sigma^*) \approx \nabla\sigma^{*T} V \nabla\sigma^*$$

  where

  $$\nabla\sigma^{*T} = \left[\frac{\partial\sigma^*}{\partial\sigma_u}, \frac{\sigma^*}{\partial\xi} \right] = [1, -u]$$

2.2.2. Parameter and return level estimation

Method of Maximum likelihood

Once a valid threshold is determined, a useful method to estimate the parameters of the generalized Pareto distribution is maximum likelihood.

Let $x_{(1)}, \cdots, x_{(k)}$ be the k excesses of a valid threshold u. For $\xi \neq 0$, using (2.3), if H is the distribution function, the likelihood function is

$$L(\sigma,\xi) = \prod_{i=1}^{k} h(x_i) = \prod_{i=1}^{k} H'(x_i) = \prod_{i=1}^{k} \frac{1}{\xi}\left(1 + \frac{\xi x_i}{\sigma}\right)^{-1-1/\xi} \frac{\xi}{\sigma} = \prod_{i=1}^{k} \sigma^{-1}\left(1 + \frac{\xi x_i}{\sigma}\right)^{-1-1/\xi}$$

Hence, taking logarithms,

$$l(\sigma,\xi) = \sum_{i=1}^{k} \log\sigma^{-1} + \sum_{i=1}^{k} \log\left(1+\frac{\xi x_i}{\sigma}\right)^{-1-1/\xi} = -k\log\sigma - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{k}\log\left(1+\frac{\xi x_i}{\sigma}\right) \tag{2.9}$$

with $1+\frac{\xi x_i}{\sigma} > 0$ for $i = 1,\cdots,k$. If not, $l(\sigma,\xi) = -\infty$.

Similarly, using (2.7), the log-likelihood function for σ for $\xi = 0$ is

$$l(\sigma) = -k\log\sigma - \sigma^{-1}\sum_{i=1}^{k} x_i \tag{2.10}$$

It is not possible to make an analytical maximization of the log-likelihood in (2.9). Numerical methods are needed, ensuring that the techniques are evaluated under the valid parameter space and paying special attention in order to avoid instabilities for $\xi \approx 0$.

Return levels

Suppose that a generalized Pareto distribution, X, with parameters σ and ξ is valid for a threshold u. That is, assuming $\xi > 0$ and $x > u$,

$$\wp\{X > x \mid X > u\} = \left[1+\xi\left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi}$$

Calling $\wp\{X > u\} = \zeta_u$, by definition of conditional probability,

$$\wp\{X > x\} = \zeta_u\left[1+\xi\left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi}$$

Hence, using the same idea of return level given in Section 1.1.3, the level x_m that exceeds on average once every m time periods (usually years) is the solution of

$$\zeta_u\left[1+\xi\left(\frac{x_m-u}{\sigma}\right)\right]^{-1/\xi} = \frac{1}{m}$$

and solving this equation, for a sufficiently large m,

$$x_m = u + \frac{\sigma}{\xi}\left[(m\zeta_u)^{\xi} - 1\right] \tag{2.11}$$

Equivalently, if $\xi = 0$ and using (2.7) for m sufficiently large, it leads to

$$x_m = u + \sigma\log(m\zeta_u) \tag{2.12}$$

Note 2.6. *From* (2.11) *and* (2.12)*, the plot of x_m against m on a logarithmic scale leads to the same cases obtained in Note 1.10: if $\xi = 0$ linearity, if $\xi > 0$ convexity and if $\xi < 0$ concavity.*

It is common to give return levels on an annual scale. If n_y is the number of observations per year and the N-year return level is the level expected once every N years then $m = N \times n_y$. Hence, using (2.11)*, the N-year return level is defined by*

$$z_N = u + \frac{\sigma}{\xi}\left[(Nn_y\zeta_u)^{\xi} - 1\right]$$

The estimation of return levels requires the previous estimation of ζ_u, σ, ξ. Maximum likelihood estimations of the parameters obtained in (2.9) and (2.10) can be used as estimations of ξ and σ.

A natural estimator of ζ_u,

$$\hat{\zeta}_u = \frac{k}{n},$$

which represents the sample proportion of points that exceed the threshold u, can be used in order to obtain an estimation. The number of excesses of u has a binomial distribution $Bin(n, \zeta_u)$ so the natural estimator is the maximum likelihood estimator. The variance of $\hat{\zeta}_u$ can be obtained from the properties of the binomial distribution as follows,

$$Var(\hat{\zeta}_u) \approx \hat{\zeta}_u \left(1 - \hat{\zeta}_u\right)/n$$

Therefore, the variance-covariance matrix of $(\hat{\zeta}_u, \hat{\sigma}, \hat{\xi})$ is approximately

$$V = \begin{bmatrix} \hat{\zeta}_u \left(1 - \hat{\zeta}_u\right)/n & 0 & 0 \\ 0 & v_{1,1} & v_{1,2} \\ 0 & v_{2,1} & v_{2,2} \end{bmatrix} \tag{2.13}$$

where $\hat{\sigma}$ and $\hat{\xi}$ are the maximum likelihood estimations of σ and ξ and $v_{i,j}$ denotes the (i,j)th term of the variance-covariance matrix of $\hat{\sigma}$ and $\hat{\xi}$. By the multivariate delta method, $Var(x_m) \approx \nabla {x_m}^T V \nabla x_m$, where

$$\nabla {x_m}^T = \left[\frac{\partial x_m}{\partial \zeta_u}, \frac{\partial x_m}{\partial \sigma}, \frac{\partial x_m}{\partial \xi}\right] =$$

$$= \left[\sigma m^{\xi} {\zeta_u}^{\xi-1}, \xi^{-1}\left\{(m\zeta_u)^{\xi} - 1\right\}, -\sigma\xi^{-2}\left\{(m\zeta_u)^{\xi} - 1\right\} + \sigma\xi^{-1}(m\zeta_u)^{\xi}\log(m\zeta_u)\right],$$

evaluated at $(\hat{\zeta}_u, \hat{\sigma}, \hat{\xi})$.

2.2.3. Model checking

In order to verify the GPD behaviour of the sample, plotting quantiles and probabilities against density and return level functions can be very useful.

Let u be a threshold and let $x_{(1)}, \dots, x_{(k)}$ be threshold excesses as obtained in Section 2.2.1. Thus, $x_{(1)} \leq \dots \leq x_{(k)}$. The probability plot consists of the points

$$\left\{\left(\frac{i}{k+1}, H(x_{(i)})\right); i = 1, \dots, k\right\}$$

where H is the Pareto distribution function with parameters $(\hat{\sigma}, \hat{\xi})$ given by (2.3). If $\xi = 0$ then the Pareto distribution obtained in (2.7) should be considered.

Now, let

$$H^{-1}(y) = u + \frac{\hat{\sigma}}{\hat{\xi}}\left[y^{-\hat{\xi}} - 1\right]$$

The quantile plot is given by

$$\left\{\left(H^{-1}\left(\frac{i}{k+1}\right), x_{(i)}\right); i = 1, \dots, k\right\}$$

According to [3], if the GPD is a good model for the excesses over the threshold u, then both plots should be linear.

Chapter 3

Point processes

The method of excesses over threshold (EOT) is based on the fact that excesses over a fixed value u have a generalized Pareto distribution and the occurrence of these excesses is a Poisson process.

This chapter will show that the point process of extreme values is closely related to both the GEV family of distributions and the GPD. Methods for the study of extremes on dependent sets of data are also given in this chapter in order to apply these results in a practical case.

3.1. Point processes of exceedances

3.1.1. Poisson process and some definitions

An intuitive way to describe a point process N is defining it as a random distribution of points X_i in space. For a group of points (X_i) and a set $A \in \mathbb{R}$, $N(A)$ can be described as a measure that counts the number of points X_i in A.

Definition 3.1. *A Poisson process in $\mathbb{R}^+$ with parameter $\lambda(t)$ is a point process which verifies that for any t,*

$$\wp(N(t,t+\delta)=1|H_t)=\lambda(t)+o(\delta)$$

$$\wp(N(t,t+\delta)>1|H_t)=o(\delta)$$

where H_t is the process behaviour until time t and $N(t_1,t_2)$ is the number of points in $(t_1,t_2]$. If $\lambda(t)$ is a constant, the process is said to be homogeneous (HPP), and non-homogeneous (NHPP) otherwise.

The verification of a Poisson process might not be intuitive but using the previous definition there are some characterizations of the process that can be easier to check [1, 12].

- Let $N(A)$ be the random variable that denotes the number of points that occur at an arbitrary period of time A. In a HPP with disjoint sets $A_1, A_2, \ldots$, the random variables $N(A_1), N(A_2), \ldots$ are independent and have Poisson distribution $A_i \sim Poisson(\lambda|A_i|)$ where $|A_i|$ is the length of A_i for $i = 1, 2, \ldots$.

- Arrival times at a point process are defined as time from one to another consecutive event, that is, $T_{r1} = T_1, T_{r2} = T_2 - T_1, \ldots$. In the case of a HPP, arrival times are independent random variables with distribution $Exp(\lambda)$.

Given any period of time A, the intensity measure of the process is defined as

$$\Lambda(A) = E\{N(A)\} = \int_A \lambda(t)dt$$

which gives the expected number of points in the set A (see [2]). The intensity function $\lambda(t)$ is deduced from this definition as the derivative of $\Lambda(A)$, that is, $\lambda(t) = \Lambda'(t)$

3.1.2. Convergence law of point processes

Let $X_1, X_2, \dots$ be independent and identically distributed random variables which satisfy Theorem 1.4 and let a bidimensional scaling point process be defined as follows

$$N_n = \left\{ \left(\frac{i}{n+1}, \frac{X_i - b_n}{a_n} \right), i = 1, \dots, n \right\} \tag{3.1}$$

where a_n and b_n normalize the behaviour of random variables.

Let $A = [0,1] \times (u, \infty)$ be a region of $\mathbb{R}^2$ for some value u. The probability of each point of N_n dropping in A is given by

$$p = \wp\{(X_i - b_n)/a_n > u\} \approx \frac{1}{n} \left[1 + \xi \left(\frac{u - \mu}{\sigma} \right) \right]^{-1/\xi}$$

As X_i are mutually independent, the distribution $N_n(A)$ is binomial with $N_n(A) \approx \text{Bin}(n,p)$.

For a large enough n, $\text{Bin}(n,p) \approx \text{Poi}(np)$. Thus, $N_n(A) \to N(A)$ where $N(A) \sim \text{Poi}(\Lambda(A))$ and

$$\Lambda(A) = \left[1 + \xi \left(\frac{u - \mu}{\sigma} \right) \right]^{-1/\xi}$$

This result is formalized in the following theorem.

Theorem 3.2. *Let $X_1, X_2, \dots$ be independent and identically distributed random variables such that there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ that satisfy*

$$\lim_{n \to \infty} \wp\{(M_n - b_n)/a_n \leq z\} = G(z)$$

where $G(z)$ is a member of the GEV family of distributions, and let z_- and z_+ be the lower and upper end points of G and N_n a bidimensional scaling point process defined as in (3.1). *Then N_n converges on regions in $(0,1) \times [u, \infty)$ for any $u > z_-$ to a Poisson process N with intensity measure on an arbitrary region $A = [t_1, t_2] \times [z, z_+)$ given by*

$$\Lambda(z) = (t_2 - t_1) \left[1 + \xi \left(\frac{z - \mu}{\sigma} \right) \right]^{-1/\xi} \tag{3.2}$$

From the theorem (see [3, Theorem 7.1]), it is easy to see that the intensity function of this process is given by

$$\lambda(z) = \left[1 + \xi \left(\frac{z - \mu}{\sigma} \right) \right]^{-1/\xi}$$

3.2. Relation with other extreme value models

This section shows how the GEV distribution for maxima and the Pareto distribution can also be obtained in terms of Poisson processes.

3.2.1. Relation with the GEV family of distributions

Let M_n be the maximum of $X_1, \dots, X_n$ as usual and

$$N_n = \left\{ \left(\frac{i}{n+1}, \frac{X_i - b_n}{a_n} \right) for\ i = 1, \dots, n \right\}$$

Taking $A_z = \{(0,1) \times (z, \infty)\}$, the following events verify

$$\left\{ \frac{M_n - b_n}{a_n} \leq z \right\} \approx \{N_n(A_z) = 0\}$$

and therefore,

$$\wp\left\{\frac{M_n - b_n}{a_n} \leq z\right\} = \wp\{N_n(A_z) = 0\} \to \wp\{N(A_z) = 0\}$$

and since N follows a Poisson distribution,

$$\wp\{N(A_z) = 0\} = \exp\{-\Lambda(A_z)\} = \exp\left[-\left(1+\xi\frac{z-\mu}{\sigma}\right)^{-1/\xi}\right]$$

so the limit of the normalized distribution of maxima is the GEV family of distributions.

3.2.2. Relation with generalized Pareto family of distributions

Let $\Lambda(A_z) = \Lambda_1([t_1,t_2]) \times \Lambda_2([z,\infty))$ be a factorization of the intensity measure defined in (3.2), $\Lambda(A_z)$, such that

$$\Lambda_1([t_1,t_2]) = (t_2 - t_1) \;\; and \;\; \Lambda_2([z,\infty)) = \left(1+\xi\frac{u-\mu}{\sigma}\right)^{-1/\xi}$$

Then,

$$\wp\{(X_i - b_n)/a_n > z|\,(X_i - b_n)/a_n > u\} = \frac{\Lambda_2([z,\infty))}{\Lambda_2([u,\infty))} =$$

$$= \left[\frac{1+\xi(z-\mu)/\sigma}{1+\xi(u-\mu)/\sigma}\right]^{-1/\xi} = \left[1+\frac{\xi(z-u)/\sigma}{1+\xi(u-\mu)/\sigma}\right]^{-1/\xi} = \left[1+\xi\left(\frac{z-u}{\sigma_u}\right)\right]^{-1/\xi}$$

where $\sigma_u = \sigma + \xi(u-\mu)$, which leads to the generalized Pareto distribution.

3.3. Estimations of the Poisson process

3.3.1. Estimation of parameters

Data modelling needs the estimation of the process given a set of observed points $x_1,\ldots,x_n \in \mathscr{A}$. Let a non-homogeneous one-dimensional processes be considered and let the intensity function be of the form $\lambda(\cdot,\theta)$, where θ is an unknown vector. Thus, the only parameters which have to be estimated are the componente of the vector θ. Likelihood estimation can be used writing probabilities as functions of θ (see [6]).

Let $I_i = [x_i, x_i+\delta_i]$ for $i = 1,\ldots,n$ be small intervals that represent the observations' neighbourhoods and $\mathscr{I} = \mathscr{A} \setminus \bigcup_{i=1}^{n} I_i$. Using the Poisson process definition,

$$\wp\{N(I_i) = 1\} = exp\{-\Lambda(I_i;\theta)\}\Lambda(I_i;\theta)$$

where

$$\Lambda(I_i;\theta) = \int_{x_i}^{x_i+\delta_i} \lambda(u)du \approx \lambda(x_i)\delta_i$$

and substituting,

$$\wp\{N(I_i) = 1\} \approx exp\{-\lambda(x_i)\delta_i\}\lambda(x_i)\delta_i \approx \lambda(x_i)\delta_i$$

for small δ_i. Therefore the following equality holds,

$$L(\theta;x_1,\ldots,x_n) = \wp\{N(\mathscr{I} = 0), N(I_1) = 1,\ldots,N(I_n) = 1\} =$$

$$= \wp\{N(\mathscr{I} = 0)\}\prod_{i=1}^{n}\wp\{N(I_i) = 1\} \approx \exp\{-\Lambda(\mathscr{A};\theta)\}\prod_{i=1}^{n}\lambda(x_i)\delta_i$$

Maximization of this likelihood often requires numerical methods.

Estimations of parameters (μ,σ,ξ) can be obtained from the GEV and Pareto estimation methods as the connection between parameters from Poisson processes and these distributions has been observed.

3.3.2. Estimation of return levels

Estimations of return levels in stationary processes are easy to obtain. For non-stationary sequences, estimations can be obtained as well but they depend on the model.

Let z_m be the m-year return level as usual. In the case of the stationary point process, the following result holds

$$1-\frac{1}{m}=\wp\{max(X_1,\dots,X_n)\leq z_m\}\approx\prod_{i=1}^{n}p_i=\prod_{i=1}^{n}p=p^n$$

where

$$p=\begin{cases}1-n^{-1}\left[1+\xi(z_m-\mu)/\sigma\right]^{-1/\xi} & \text{if } [1+\xi(z_m-\mu)/\sigma]>0\\ 1 & \text{otherwise}\end{cases}$$

and (μ,σ,ξ) are the parameters of the Poisson process. Making calculations and substituting,

$$1-\frac{1}{m}=\left[1-\frac{1}{n}\left[1+\xi\frac{z_m-\mu}{\sigma}\right]^{-1/\xi}\right]^n\Rightarrow n-n\left(1-\frac{1}{m}\right)^{1/n}=\left[1+\xi\frac{z_m-\mu}{\sigma}\right]^{-1/\xi}$$

Therefore,

$$z_m=\mu+\frac{\sigma}{\xi}\left[\left[n-n\left(1-\frac{1}{m}\right)^{1/n}\right]^{-\xi}-1\right]$$

3.4. Extremes on dependent sequences

Every extreme value model that has been described was obtained supposing that events were independent. Nevertheless, this assumption is not usually true. In this section, some results will be considered in order to describe dependence conditions for which extreme value theorems are still valid.

Definition 3.3. *A series $X_1,X_2,\dots$ is said to be stationary if for any finite set $t_1<\dots<t_n$ and $h\in\mathbb{Z}$,*

$$(X_{t_1},\dots,X_{t_n})\overset{d}{=}(X_{t_1+h},\dots,X_{t_n+h})$$

Definition 3.4. *A stationary series satisfies the $D(u_n)$ condition if $\forall i_1,\dots,i_k,j_1,\dots,j_l$ with $j_1-i_k>h$ we have*

$$|\wp\{X_{i_1}\leq u_n,\dots,X_{i_k}\leq u_n,X_{j_1}\leq u_n,\dots,X_{j_l}\leq u_n\}$$
$$-\wp\{X_{i_1}\leq u_n,\dots,X_{i_k}\leq u_n\}\wp\{X_{j_1}\leq u_n,\dots,X_{j_l}\leq u_n\}|\leq\alpha(n,h) \qquad (3.3)$$

where $\alpha(n,h_n)\rightarrow 0$ for sequence (h_n) such that $\lim_{n\to\infty}h_n/n=0$.

This condition is weaker than the independence of events and it is verified in cases where observations are mostly independent when there is a sufficiently large temporal distance between them. Using this, the following theorem holds for a stationary sequence.

Theorem 3.5. *Let $X_1,X_2,\dots$ be a stationary sequence and define $M_n=max\{X_1,\dots,X_n\}$. Then if the $D(u_n)$ condition is satisfied for $u_n=a_nz+b_n$, $\forall z\in\mathbb{R}$, for sequences of constants $\{a_n>0\}$ and $\{b_n\}$ such that*

$$\lim_{n\to\infty}\wp\left\{\frac{M_n-b_n}{a_n}\leq z\right\}=G(z)$$

where G is a non-degenerate distribution function, then G is a member of the GEV family of distributions.

Therefore, the $D(u_n)$ condition can be understood as a property that guarantees the weakness of the dependence of data so that it does not affect the distribution of maxima.

The main assumption in order to fit a Pareto distribution is that threshold excesses series must be independent (see Theorem 2.2). In the case of stationary sequences, excesses can only be treated as if they were independent events if Definition 3.4 is satisfied. However, the behaviour of the extreme distribution of neighbouring excesses has not yet been considered.

When the variables are not independent, threshold excesses are expected to appear in clusters, which implies that one excess can be easily followed by another. Thus, the log-likelihood method cannot be applied due to the dependence of observations.

In order to find a set of approximately independent excesses, the most common method is declustering (see [3]), which consists on a filtering of the dependent excesses in order to obtain a set of exceedances that are approximately independent.

The first step is defining clusters of excesses using an empirical rule. The maximum excess in each cluster is taken and assuming cluster maxima to be independent, a generalized Pareto distribution can be fitted to these values.

Chapter 4

Practical example: extreme distributions of water levels in the Ebro river in Zaragoza

This chapter will show a practical example of how to fit extreme value distributions to a data set. The data that has been used are daily mean water levels measured in metres from 1961 to 2016 in the Ebro river in Zaragoza (Spain). The data series can be found on the *Sistema Automático de Información Hidrológica* (SAIH) official webpage [4]. The date of each observation is also available in the SAIH file.

The behaviour of water level over time will be studied in order to fit an appropriate distribution to the observations. A GEV distribution and a Pareto distribution can be fitted to this set of data and an analysis of return levels can be easily made once the parameters of the distributions are estimated.

The data will be analysed using the system for statistical computation and graphics, R. The R language (see [15]) is worldwide used by mathematicians for the development of statistical software and data analysis. The analysis of extremes is distributed in many packages with a lot of different applications (see [5]). The packages which will be used in this analysis are the Rcmdr, the extRemes and the ismev ([7, 8, 10]). The complete R code used in this chapter can be found in Appendix A.

4.1. Data and exploratory analysis

The application of the extreme value models requires the data to be independent and identically distributed. However, it has been seen in Section 3.4 that weaker conditions can be considered. If the series is stationary and it does not show long term dependence, the extreme value theorems are valid . Thus, the first hypothesis which has to be analysed is the stationarity of the data series and the dependence of water level values.

Because of their physical characteristics, the observations of the height of the water in the river are clearly dependent but they do not show a long time dependence, that is, the water level one day is related to the level the following day but it is independent to the height of the water six months later. The two main reasons for which there can be changes in parameters of the distributions are the existence of trend and stationarity behaviour (see [11]).

Figure 4.1 (left) represents a scatter plot of water level observations where each point is the mean level of the water in metres per month and year. With this plot it is easy to see that the points representing the 1960s are shifted and the minimum values in those years are lower than in the rest of the decades. Thus, events are not identically distributed and this can affect the estimations of parameters. Therefore the analysis shall be focused on data from 1970 to 2016.

In order to study the existence of a seasonal pattern of the series, a box plot representing the variability of data per month can be given in order to understand the behaviour at each time of the year. In

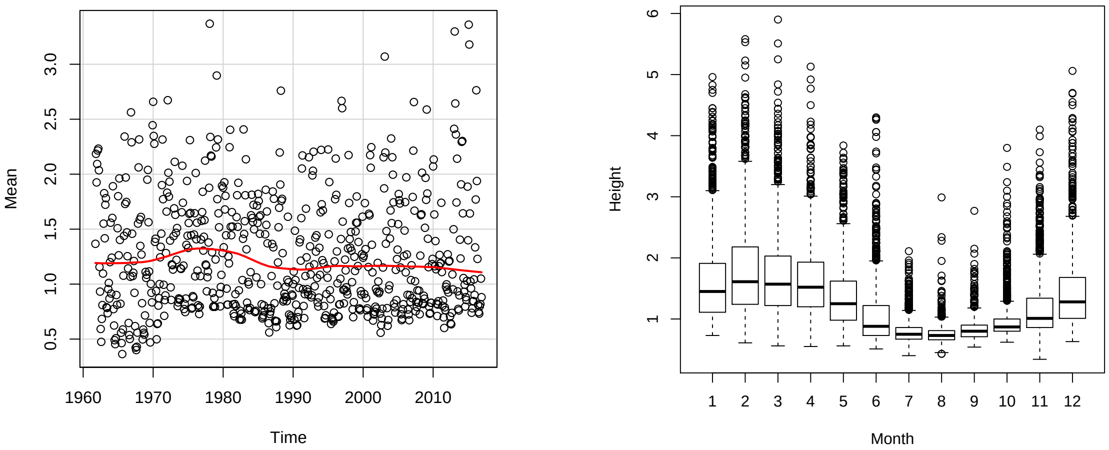

Figure 4.1: Scatter plot of data from 1961 to 2016 (left) and box plot of water levels per month of the year (right)

Figure 4.1 (right), months with approximately identically distribution can be checked.

The plot suggests that water levels present seasonal behaviour along the year. December and January through April seem to be approximately identically distributed. Thus, these months are suitable for the analysis. From this set of data, the extreme distributions will correspond in fact to the study of annual maximums.

4.2. Extreme behaviour and extreme distributions of data

4.2.1. GEV approximation

From the data in December and January through April, the set of annual maxima are represented in a scatter plot. It is easy to see that in Figure 4.2, the points appear as a point cloud with no pattern. This confirms the hypothesis of identical distribution of the annual maximum set of points and therefore a generalized extreme value distribution can be fitted to this set.

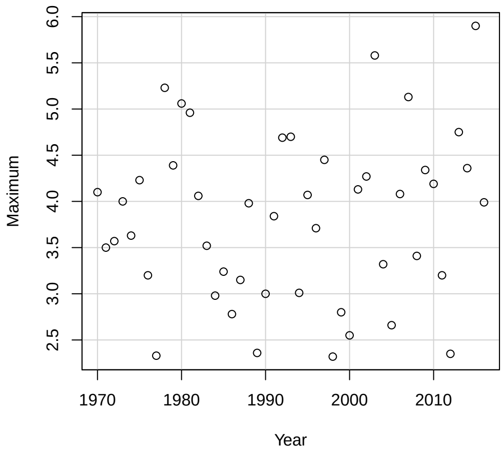

Figure 4.2: Scatter plot of maximum water level per year

From the set of data given in Figure 4.2, a distribution of the following form can be obtained using `fevd` (extRemes R Package [8]),

$$G(z) = exp\left(-\left(1+\xi\left(\frac{z-\mu}{\sigma}\right)\right)^{-1/\xi}\right)$$

where the location parameter $\hat{\mu} = 3.48$, the scale parameter $\hat{\sigma} = 0.85$ and the shape parameter $\hat{\xi} = -0.23$. As $\hat{\xi} < 0$, the GEV distribution corresponds to a Weibull distribution according to Note 1.6.

Once the GEV distribution is estimated, return levels can easily be obtained. The direct application of Section 1.1.3, the choice of a range of values and the function `return.level` (extRemes R Package) gives,

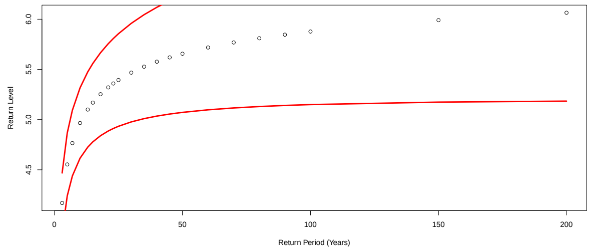

Figure 4.3: Return levels and their 95% confidence intervals for the GEV distribution

$1/p$	5	10	25	50	100	200
95% lower CI	4.24	4.61	4.93	5.07	5.14	5.18
z_p	4.55	4.97	5.39	5.66	5.88	6.06
95% upper CI	4.87	5.31	5.86	6.24	6.60	6.95

According to Figure 4.3, the maximum is expected to be greater than 4.5 metres once every 5 years and the probability of getting a level of 4.55 metres is $1/5 = 0.2$. In a similar way, taking $1/p = 25$, the maximum exceeds 5.4 metres once every 25 years and the river has a probability of reaching the height 5.4 metres of $p = 0.4$.

4.2.2. GPD approximation

The first step to fit a generalized Pareto distribution is the selection of the threshold. It is an important step since if the threshold is too high there might be a lack of data while if it is too low the GPD approximation will not be valid as there might be some excesses which are not real extreme values.

Figure 4.4 shows the mean residual life plot with 95% of confidence interval of the high level of water obtained using `mrl.plot` (ismev R Package [10]). It plots a threshold u against the mean of the exceedances of the threshold, for a range of thresholds.

The plot is linear from $u \approx 3$ until $u \approx 4$. Applying Section 2.2.1, if the threshold $u = 3$ is valid, any threshold $u > 3$ is valid as well. Thus, the mean residual life plot should be linear after $u = 3$. Due to a lack of exceedances over thresholds $u > 4$, the estimation of the mean residual life plot is unreliable although it is in the confidence interval. Because of this, the mean residual life plot is not linear after $u = 4$.

Thus, $u = 3$ is considered. As data presents a short time dependence, the estimation requires the use of the declustering method for EOT explained in Section 3.4. The function `decluster` (extRemes R Package [8]) applies a declustering method to the threshold excesses which consists in providing the maximum water level of each cluster of exceedances, where each cluster is a set of months given by water levels from December of a year and from January to April of the following year.

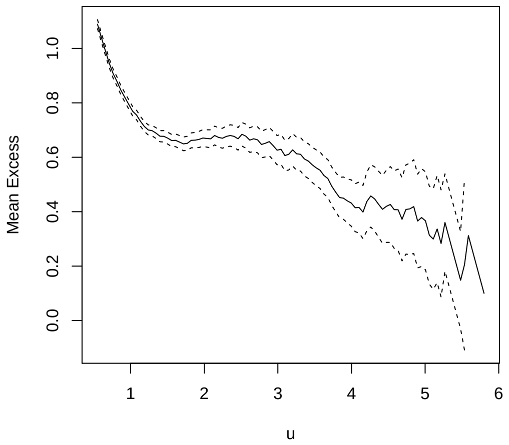

Figure 4.4: Mean residual life plot of data with 95% of confidence interval

The number of excesses over the threshold $u = 3$ is 384, but the number of events taken into account in order to fit a GPD is equal to 104. Then, the GPD obtained by using `fevd` (extRemes R Package [8]) is given by

$$H(x) = 1 - \left(1 + \frac{\hat{\xi}x}{\hat{\sigma}}\right)^{-1/\hat{\xi}}$$

where the scale parameter $\hat{\sigma} = 0.79$ and the shape parameter $\hat{\xi} = -0.14$. Applying Note 2.3, as $\hat{\xi} < 0$, the distribution has an upper bound at $u - \hat{\sigma}/\hat{\xi} = 8.64$ metres and the theoretical mean of excesses is $\bar{x} = \hat{\sigma}/(1-\hat{\xi}) = 0.69$ metres. Return levels of the excesses are obtained applying Section 2.2.2. For a given range of values and using `return.level` (extRemes R Package),

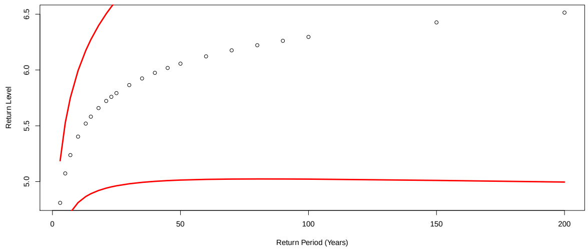

Figure 4.5: Return levels and their 95% confidence intervals for the GP distribution

$1/p$	5	10	25	50	100	200
95% lower CI	4.62	4.81	4.96	5.01	5.03	4.99
z_p	5.07	5.40	5.79	6.05	6.29	6.50
95% upper CI	5.53	5.99	6.62	7.10	7.57	8.03

Looking at Figure 4.5 and the table, the probability of the occurrence of one particular water level is the inverse of the return period and the return level is expected to be exceeded once every $1/p$ years.

The values of the estimated return levels are slightly different as they are calculated using different sets of data, but the values of z_p obtained in each of the methods are in the confidence intervals of the other method respectively, so they are assumed to be good estimations.

Appendix A

R Code

A.1. Data and exploratory analysis Code

```
#Month and year variable (Time)
> Datos$TIEMPO <- with(Datos, ANIO+((MES.NUMERO-1)/12))
# Create data and mean per month variable
> Mediames <- as.data.frame(tapply(Datos$TIEMPO,list(Datos$TIEMPO),min,na.rm=TRUE))
> Mediames$Mean <- tapply(Datos$ALTURA,list(Datos$TIEMPO),mean,na.rm=TRUE)
#Change of name "TIEMPO"
> names(Mediames)[1]<- "Time"

#Scatter plot of mean water level per month and year
> scatterplot(Mean~Time, reg.line=FALSE, smooth=TRUE, spread=FALSE, boxplots=FALSE,
    span=0.3, ellipse=FALSE, levels=c(.5, .9), data=Mediames)

#Calculate subset of data as the years required are from 1970 to 2016
> Datos70 <- subset(Datos, subset= ANIO>1969)
> Datos70 <- within(Datos70, {mesnumerofactor <- as.factor(MES.NUMERO)})
#Change of name of variable "MES"
> names(Datos70)[7]<-"Month"
#Change of name "ALTURA"
> names(Datos)[5]<-"Height"

#Boxplot of Height of water per month of the year
> Boxplot(Height~Month,data=Datos70,id.method="none")
```

A.2. GEV approximation Code

```
#Subset of data of months January to April and December
> Datosfin<-subset(Datos70,subset=MES.NUMERO<5|MES.NUMERO==12)

#Calculate set of data of maximum water level per year
> MaxGEV <- as.data.frame(tapply(Datosfin$ANIO,list(Datosfin$ANIO),min,na.rm=TRUE))
> MaxGEV$Maximum <- tapply(Datosfin$Height,list(Datosfin$ ANIO),max,na.rm=TRUE)
> max<-as.vector(MaxGEV$Maximum)
#Change name of variable "ANIO"
> names(MaxGEV)[1]<-"Year"

#Scatter plot of the distribution of maxima
> scatterplot(Maximum~Year, reg.line=FALSE, smooth=TRUE, spread=FALSE,
    boxplots=FALSE, span=0.3, ellipse=FALSE, levels=c(.5, .9), data=MaxGEV)
```

```
#Fit a GEV distribution to the set of maxima
>fit1<-fevd(max)
#Calculate return levels given certain return periods and 95% confidence intervals
> vec<-c(3,5,7,10,13,15,18,21,23,25,30,35,40,45,50,60,70,80,90,100,150,200)
> z<-return.level(fit1,return.period=vec,alpha=0.05,do.ci=TRUE)

#Plot of return levels and their 95% confidence intervals for GEV
#Vector of lower 95% CI
> return1<-c(z[1],z[2],z[3],z[4],z[5],z[6],z[7],z[8],z[9],z[10],z[11],z[12],z[13],
z[14],z[15],z[16],z[17],z[18],z[19],z[20],z[21],z[22])
#Vector of return levels
> return2<-c(z[23],z[24],z[25],z[26],z[27],z[28],z[29],z[30],z[31],z[32],z[33],z[34],
z[35],z[36],z[37],z[38],z[39],z[40],z[41],z[42],z[43],z[44])
#Vector of upper 95% CI
> return3<-c(z[45],z[46],z[47],z[48],z[49],z[50],z[51],z[52],z[53],z[54],z[55],z[56],
z[57],z[58],z[59],z[60],z[61],z[62],z[63],z[64],z[65],z[66])

> plot(vec,return2,xlab="Return Period (Years)",ylab="Return Level")
> lines(vec,return1,type="l", col="Red",lwd=3)
> lines(vec,return3,type="l", col="Red",lwd=3)
```

A.3. GPD approximation Code

```
#Vector of height of water
> alt<-as.vector(Datosfin$Height)
#Mean residual life plot with 95% confidence interval
> mrl.plot(alt,conf=0.95)

#Vector of groups for declustering
> bis<-c(152,151,151,151)
> vect<-c(rep(1,120),rep(2,151),rep(3:46,rep(bis,11)),rep(47,152),rep(48,31))

#Decluster of excesses with threshold u=3
> altura<-Datosfin$Height
> dec<-decluster(altura,threshold=3,method="intervals",groups=vect)

#Fit a generalized Pareto distribution
> fin<-fevd(dec, threshold=3, type="GP",time.units="days",period.basis="year")

#Calculate return levels given certain return periods and 95% confidence intervals
> vec<-c(3,5,7,10,13,15,18,21,23,25,30,35,40,45,50,60,70,80,90,100,150,200)
> z<-return.level(fin,return.period=vec,alpha=0.05,do.ci=TRUE)

#Plot of return levels and their 95% confidence intervals for GPD
#Vector of lower 95% CI
> return1<-c(z[1],z[2],z[3],z[4],z[5],z[6],z[7],z[8],z[9],z[10],z[11],z[12],z[13],
z[14],z[15],z[16],z[17],z[18],z[19],z[20],z[21],z[22])
#Vector of return levels
> return2<-c(z[23],z[24],z[25],z[26],z[27],z[28],z[29],z[30],z[31],z[32],z[33],z[34],
z[35],z[36],z[37],z[38],z[39],z[40],z[41],z[42],z[43],z[44])
#Vector of upper 95% CI
> return3<-c(z[45],z[46],z[47],z[48],z[49],z[50],z[51],z[52],z[53],z[54],z[55],z[56],
z[57],z[58],z[59],z[60],z[61],z[62],z[63],z[64],z[65],z[66])

> plot(vec,return2,xlab="Return Period (Years)",ylab="Return Level")
> lines(vec,return1,type="l", col="Red",lwd=3)
> lines(vec,return3,type="l", col="Red",lwd=3)
```

Bibliography

[1] A.C. CEBRIÁN GUAJARDO *Análisis, modelización y predicción de episodios de sequía*, Departamento de métodos estadísticos, Universidad de Zaragoza, 1999.

[2] S.G. COLES & R.S.J. SPARKS, *Extreme value methods for modelling historical series of large volcanic magnitudes*, Statistics in volcanology. Geol Soc, London, 2006

[3] S. COLES, J. BAWA, L. TRENNER & P DORAZIO, *An introduction to statistical modeling of extreme values*, Springer, Vol 208, 2001.

[4] CONFEDERACIÓN HIDROGRÁFICA DEL EBRO, *Datos de altura media diaria en metros para estación de aforo 9011*, 1961 to 2016. Data available at `http://www.saihebro.com/saihebro/index.php`.

[5] C. DUTANG & K. JAUNATRE, *CRAN Task View: Extreme Value Analysis*, 2017 `https://CRAN.R-project.org/view=ExtremeValue`.

[6] P. EMBRECHTS, C. KLÜPPELBERG & T. MIKOSCH, *Modelling extremal events: for insurance and finance*, Springer Science & Business Media, Vol 33, 2013.

[7] J.FOX & M.BOUCHET-VALAT, *Rcmdr: R Commander*, R package version 2.3-2, 2017, `http://socserv.socsci.mcmaster.ca/ifox/Misc/Rcmdr/`

[8] E. GILLELAND, *Package 'extRemes'*, 2016, `https://cran.r-project.org/web/packages/extRemes/extRemes.pdf`.

[9] E.J. GUMBEL, *Statistics of extremes*, New York: Columbia University Press, 1958.

[10] J.E. HEFFERNAN, A.G. STEPHENSON & E. GILLELAND, *Ismev: an introduction to statistical modeling of extreme values*, R package version, Vol 1, 2012, `https://cran.r-project.org/web/packages/ismev/ismev.pdf`.

[11] R.J. HYNDMAN & G. ATHANASOPOULOS, *Forecasting: principles and practice*, O Texts, 2014.

[12] N. LASKIN, *Communications in Nonlinear Science and Numerical Simulation*, Elvesier, Vol 8, 3, 2003, 201–213.

[13] M.R. LEADBETTER, G. LINDGREN & H. ROOTZÉN, *Extremes and related properties of random sequences and processes*, Springer Science & Business Media, 2012.

[14] J. PICKANDS III, *Statistical inference using extreme order statistics*, The Annals of Statistics, 1995, 119–131.

[15] R CORE TEAM, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017, `https://www.R-project.org`.

Index