Politecnico di Torino

Faculty of Engineering

Master degree course in Mathematical Engineering

Extreme Value Theory

With application to huge rainfalls

Master degree Thesis

Author:
Francesca Garibaldi

Supervisor:
Franco Pellerey

Academic Year 2018/2019

Ai miei genitori

Grazie per tutti i sacrifici che avete fatto pur di permettermi di frequentare l'università.Grazie per esservi sempre fidati di me ed aver creduto nelle mie capacità anche quando io non ci riuscivo. Ma soprattutto grazie per aver accettato ogni mio risultato, bello o brutto che fosse, senza mai dubitare del mio impegno.

Al Professor Pellerey

Grazie per tutta la disponibilità e la pazienza dimostratemi in questi mesi. Grazie per essere sempre riuscito a dedicarmi anche solo un istante del suo tempo, persin quando era oberato di lavoro.

A Luca

Grazie per tutte le ore passate al telefono a tranquillizzarmi prima di un esame o a confortarmi quando temevo che non ce la avrei fatta. Grazie per tutte le attenzioni e l'affetto che sei sempre riuscito a dimotrarmi nonostante la lontananza. Ma soprattutto grazie per aver condiviso con me, nel bene e nel male, tutti i momenti più importanti di questi ultimi due anni.

Alla mia piccola Cami

A te che ti sei presa cura di tutto mentre ero via, senza mai chiedere aiuto pur di non disturbare. A te che sei sempre pronta ad accogliermi con un sorriso e un abbraccio quando torno a casa, anche se non ci sentiamo quasi mai. Voglio solo che tu sappia che farò di tutto perchè anche tu possa vivere un'esperienza come la mia, perché sei il mio orgoglio e te la meriti.

Alle mie amiche

Siete state delle ottime compagne non solo di studio e di progetti, ma anche di cibo e divertimento. Avervi conosciute è stata un'enorme fortuna per me perché la vostra presenza ha sicuramente controbuito a rendere questi anni indimenticabili.

Alla mie coinquiline, passate e presenti.

A voi che avete conosciuto ogni lato del mio carattere, ma che nonostante tutto avete sempre trovato il modo di ricordarmi quanto valgo. Vi ringrazio per tutti i momenti passati insieme in quel di Via Monginevro 28.

Alla persona che meno di tutte ha creduto che ce la potessi fare e che ha sempre preteso più di chiunque altro. A me stessa. Forse è giunto il momento di ricredersi.

Abstract

The analysis of extreme events, such as floods, earthquakes or sever windstorms, has grown of importance over the years, since forecasting the occurrence of such intense phenomena is useful for both risk management and mitigation.

In this work, the statistical methods deriving from Extreme value theory are used to analyze rainfall measurements collected in the province of Genoa, as in the last fifty years the area has been hit by a series of flash floods that have caused considerable damages.

After having exposed the classical approach of Extreme value analysis, some generalizations are presented. Such characterizations of extremes have been developed with the aim to include additional information when modelling rare events and are known as: the r Largest Order Statistics approach and Threshold exceedances model.

As regards the application case, inference for distribution parameters is based on maximum likelihood estimation, while the goodness of fit is assessed through diagnostic plots.

Contents

List of Figures **iii**

1 Introduction **1**

2 Extreme value theory **2**
2.1 Univariate models 2
2.1.1 Asymptotic Model and Extremal Types Theorem 2
2.1.2 The Generalized Extreme Value Distribution 3
2.1.3 Notes on the proof of Extremal Types Theorem 6
2.2 Bivariate models 7
2.2.1 Asymptotic characterization 7
2.2.2 Model examples for Componentwise Maxima 9
2.2.3 Extremal dependence 10

3 Model generalizations **12**
3.1 The r Largest Order Statistic Model 12
3.2 The Threshold Exceedances Model 14
3.2.1 The Generalized Pareto Distribution 14
3.2.2 Notes on the justification for the Generalized Pareto Model 15
3.2.3 Return levels 16

4 Inference and Model assessment **17**
4.1 Profile likelihood and Likelihood ratio test 17
4.2 Diagnostic Plots 18
4.3 Threshold selection 20
4.3.1 Mean residual life plot 20
4.3.2 Alternative technique 21

5 Real data application **22**
5.1 Introduction 22
5.2 Daily rainfalls data 23
5.3 Univariate analysis 24
5.3.1 Block maxima approach 24
5.3.2 r Largest Order Statistic approach 30
5.3.3 Peaks over threshold approach 33

5.4 Bivariate analysis 37
5.4.1 Componentwise Block maxima approach 37
5.5 Comparing models on different time periods 40

6 Conclusion **46**

A Code **47**

Bibliography **57**

List of Figures

2.1 Density functions for Gumbel ($\mu = 0, \sigma = 1$), Fréchet ($\mu = 0, \sigma = 1, \xi = 4$) and Weibull($\mu = 0, \sigma = 1, \xi = 2$) distributions. 4

5.1 Polcevera's basin 23
5.2 Annual maximum rainfalls recorded at Isoverde station 24
5.3 Profile likelihood for ξ with 95% normal approximated (blue) and profile (black) confidence intervals 25
5.4 Diagnostic plots for GEV fit 26
5.5 Profile likelihood for 10-year return level with 95% normal approximated (blue) and profile confident intervals (black) 27
5.6 Profile likelihood for 50-year return level with 95% normal approximated (blue) and profile confident intervals (black) 28
5.7 Standard error pattern for location parameter 30
5.8 Standard error pattern for scale parameter 30
5.9 Standard error pattern for shape parameter 31
5.10 Diagnostic plots for order statistic model with r = 3 32
5.11 Mean residual life plot for daily rainfall data 33
5.12 Parameter estimates against threshold values for daily rainfall data 34
5.13 Excesses over threshold $u = 100$ 34
5.14 Profile likelihood for the shape parameter in the GPD model . . 35
5.15 Diagnostic plots for GPD model 36
5.16 Profile likelihood for 100-year return level in GPD model 36
5.17 Annual maximum rainfalls recorded at Isoverde and Mignanego stations 37
5.18 Empirical estimates for $\chi(u)$ and $\overline{\chi(u)}$ with approximate 95% confidence intervals 38
5.19 Componentwise maxima transformed to have uniform marginals distributions 39
5.20 Annual maxima recorded between 1960 and 1985 40
5.21 Annual maxima recorded between 1986 and 2014 40
5.22 Profile likelihood for ξ with 95% confident interval 41
5.23 Diagnostic plot for Gumbel fit 42
5.24 Comparison between model densities 44
5.25 Comparison between cumulative distribution functions 44

Chapter 1

Introduction

The adjective "extremes" refers to the complete class of phenomena whose occurrence is rare and which can lead, precisely because of their unusual character, to disastrous environmental, economic and human impacts.

The probabilistic theory dealing with these events is called Extreme Value Theory and aims to develop mathematical methods and models able to describe and to predict the occurrence of such rare phenomena.

Initially, the analysis of extremes was introduced in hydrology and meteorology in order to study flood levels and natural disasters; however, in recent years, the domain of application of this statistical methodology has managed to include other disciplines and applied sciences such as finance, traffic prediction, insurance and structural engineering.

The first studies on extreme values date back to the first part of the twentieth century, when Tippet and Fisher (1928) stated an asymptotic argument which represents the cornerstone of extreme value theory: the Extremal Types Theorem. In the subsequent years, the asymptotic theory was extended and codified by Gnedenko (1948) and Gumbel(1958), while the characterization of extremes as observations exceeding an high threshold is due to Pickands (1970).

The present work is articulated in four chapters. Firstly, Chapter 2 contains an overview of the classical theory of extremes: at the beginning, the univariate case is discussed, with reference to the way of modeling block maxima data; after that the basic theory of bivariate extremes is presented, with mention to asymptotic dependence. Two model generalizations are then described in Chapter 3. They mainly rely on different characterizations of extremes that allow a greater number of observations to be used when fitting extreme value distributions.
Inference for model parameters, together with goodness of fit procedures are described in Chapter 4.
Finally, in Chapter 5, all previously mentioned techniques are applied to a data set consisting of rainfall measurements collected in the province of Genoa, with the aim of analyzing the behavior of the intense meteorological phenomena that have affected the area over the last fifty years.

Chapter 2

Extreme value theory

This chapter concerns the classical approach of extreme value theory, which will be also referred as EVT. In the first part the study will focus on univariate sequences of independent and identically distributed random variables, while in the second part the bivariate case will be discussed.

2.1 Univariate models

2.1.1 Asymptotic Model and Extremal Types Theorem

The classical approach for studying extremes is mainly based on an asymptotic result which can be considered as an analog of the central limit theorem. Before exposing it, some model specifications are needed.
Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables. The main purpose is that of studying the behaviour of the sample maximum

$$M_n = \max(X_1, X_2, \ldots, X_n)$$

Supposing that the variables have common distribution function F, the exact distribution of M_n can be found as follows:

$$\begin{aligned} \mathbb{P}\{M_n \leq z\} &= \mathbb{P}\{X_1 \leq z, X_2 \leq z, \ldots, X_n \leq z\} \\ &= \mathbb{P}\{X_1 \leq z\} \cdot \mathbb{P}\{X_2 \leq z\} \cdot \cdots \cdot \mathbb{P}\{X_n \leq z\} \\ &= F^n(z). \end{aligned} \tag{2.1}$$

This formulation cannot be used in practice, since it requires to know the distribution F. Furthermore, the substitution of empirical estimates of F into 2.1 cannot be considered, since small discrepancies in the approximation of F, can generate significant errors when estimating F^n.
An alternative approach suggests to study the asymptotic behaviour of F^n, trying to find suitable families of models that can be estimated using extreme observations. Defining z^+ the smallest value of z for which $F(z) = 1$, it follows that

$$\lim_{z \to \infty} F^n(z) = 0 \qquad \forall z < z_+,$$

so, the limit distribution degenerates to a point of mass. This problem can be overcome by re-normalizing M_n through linear transformation:

$$M_n^* = \frac{M_n - b_n}{a_n}, \tag{2.2}$$

where $\{a_n > 0\}$ and $\{b_n\}$ are suitable sequences of constants such that the location and scale of M_n^* are stabilized as n grows.
Now it is possible to enunciate the fundamental theorem of EVT ([1]).

Extremal types theorem

Theorem 2.1.1. *If there exist sequences of constants* $\{a_n > 0\}$ *and* $\{b_n\}$ *such that*

$$\mathbb{P}\left\{\frac{M_n - b_n}{a_n} \leq z\right\} \to G(z) \quad as\, n \to \infty \tag{2.3}$$

where G is a non-degenerate distribution function, then G belongs to one of the following families :

I. $$G(z) = \exp\left\{-\exp\left[-\left(\tfrac{z-b}{a}\right)\right]\right\}, \quad -\infty < z < \infty;$$

II. $$G(z) = \begin{cases} 0, & z \leq b, \\ \exp\{-(\tfrac{z-b}{a})^{-\alpha}\}, & z > b; \end{cases}$$

III. $$G(z) = \begin{cases} \exp\{-[-(\tfrac{z-b}{a})^{-\alpha}]\}, & z < b, \\ 1, & z \geq b; \end{cases}$$

for parameters $a > 0$, b *and, in the case of families II, III,* $\alpha > 0$.

Theorem 2.1.1, stated by Fisher and Tippet in 1928, exposes the entire range of limit distributions for M_n^*.
Families I, II and III are referred collectively as **extreme value distributions**, and are widely know as *Gumbel, Fréchet* and *Weibull* respectively. All three families are characterized by parameters b and a, which are called **location** and **scale** parameter, respectively. In addiction, the Fréchet and Weibull distributions have a **shape** parameter α. The main aspect of Theorem 2.1.1 is that it provides an extreme value analog of the central limit theorem: it suggests that the above mentioned types of extreme value distributions are the only possible limits for the distribution of M_n^* independently of F.

2.1.2 The Generalized Extreme Value Distribution

The three families mentioned in Theorem 2.1.1 differ according to the types of tail behavior of the distribution function F. This can be pointed out by simply studying the behavior of G near its upper-end point z_+. In particular, it results that for the Fréchet and Gumbel distributions $z_+ = +\infty$, while, for the Weibull,

z_+ is finite. Moreover, the different rates of decay in the tail of F make sure that the density of G has exponential and polynomial decay for the Gumbel and Fréchet distributions respectively. As a results, the three classes lead to quite different representations of the extreme value behavior.

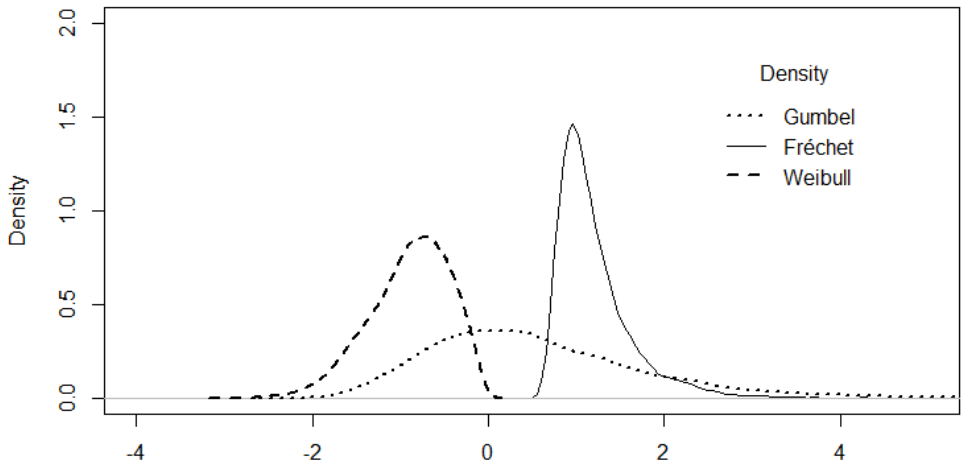

Figure 2.1: Density functions for Gumbel ($\mu = 0, \sigma = 1$), Fréchet ($\mu = 0, \sigma = 1, \xi = 4$) and Weibull($\mu = 0, \sigma = 1, \xi = 2$) distributions.

Unfortunately, the formulation of G in Theorem 2.1.1, which separates between the three families, is not so helpful in practice, since it requires to know which class to adopt before estimating its parameters. However, it is possible to reformulate the models in Theorem 2.1.1, combining the Gumbel, Fréchet and Weibull distributions into a single family of models, characterized by distribution functions of the form:

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, \tag{2.4}$$

defined on the set $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$, where the parameters satisfy the conditions : $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$.
The model in Eq 2.4 is called **generalized extreme value distribution** (GEV) and has three parameters :

- a location parameter μ;
- a scale parameter σ;
- a shape parameter ξ .

The Fréchet and the Weibull families correspond to the cases $\xi > 0$ and $\xi < 0$, respectively. The Gumbel distribution ($\xi = 0$), instead, is obtained by taking the limit of 2.4 as $\xi \to 0$ and has the following form:

$$G(z) = \exp\left[-\exp\left\{-\left(\frac{z-\mu}{\sigma}\right)\right\}\right], \quad -\infty < z < \infty \tag{2.5}$$

Thus, Theorem 2.1.1 can be restated as follows ([1]):

Theorem 2.1.2. *If there exist sequences of constants* $\{a_n > 0\}$ *and* $\{b_n\}$ *such that*

$$\mathbb{P}\left\{\frac{M_n - b_n}{a_n} \leq z\right\} \to G(z) \quad as\, n \to \infty \tag{2.6}$$

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

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, \tag{2.7}$$

defined on the set $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$*, where the parameters satisfy the conditions :* $-\infty < \mu < \infty$*,* $\sigma > 0$ *and* $-\infty < \xi < \infty$*.*

This unification simplifies statistical implementations: in fact no a priori hypothesis about which class to adopt is still required, but the most suitable family can be chosen through inference on the parameter ξ.
The limit in 2.6 suggests that, for large values of n, maxima of long sequences can be modelled through GEV distribution. In particular, it must be observed that, since

$$\mathbb{P}\left\{\frac{M_n - b_n}{a_n} \leq z\right\} \approx G(z) \tag{2.8}$$

for large n , then

$$\begin{aligned}\mathbb{P}\{M_n \leq z\} &\approx G\left(\frac{z-b_n}{a_n}\right) \\ &= G^*(z),\end{aligned} \tag{2.9}$$

where G^* is said to be a distribution of the **same type**[1] of G and so also belongs to the GEV family.
The previous arguments suggest a strategy for modelling extremes of independent series of observations $X_1, X_2, \ldots$, widely known as **block maxima approach**. As the name says, data are divided into sequences of length n, which must be a large value. Then the block maxima $M_{n,1}, \ldots, M_{n,m}$ are extracted and modeled through GEV distribution. In most applications blocks tend to

[1]The distributions F and F^* are of the same type if there are constants $a > 0$ and b such that $F^*(ax + b) = F(x)\ \forall x$.

correspond to one year of observations, so the considered block maxima are simply annual maxima.
By inverting Eq.2.4, estimates of extreme quantiles are obtained:

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}[1 - \{-\log(1-p)\}^{-\xi}], & \text{for} \quad \xi \neq 0, \\ \mu - \sigma \log(-\log(1-p)), & \text{for} \quad \xi = 0, \end{cases} \tag{2.10}$$

where $G(z_p) = 1 - p$. The value z_p is called **return level** associated with the **return period** $1/p$, since it is expected to be exceeded on average once every $1/p$ years. In other words, z_p is the value exceeded by the annual maximum with probability p.

2.1.3 Notes on the proof of Extremal Types Theorem

Since the formal proof of the extremal types theorem is not in the scope of this work, only some ideas of the justification are presented in this section. First, it is necessary to report the following definition ([1])

Definition 2.1.3. A distribution G is said to be **max-stable** if, for every $n = 2, 3, \ldots$, there are constants $\alpha_n > 0$ and β_n such that

$$G^n(\alpha_n z + \beta_n) = G(z).$$

So, if $X_1, X_2, ...X_n$ are i.i.d G and G is max stable, then taking the sample maxima $M_n = \max(X_1, X_2, ...X_n)$ leads to an identical distribution, except for a change in scale and location.
The following theorem connects the previous argument with the concept of extreme value distribution ([1]).

Theorem 2.1.4. *A distribution is max-stable if, and only if, it is a generalized extreme value distribution.*

The main idea is to consider M_{nk}, which can be seen as the maximum over a series on $n \times k$ observations, or as the maximum of k maxima, each of them being the greatest of n measurements. Let then suppose that

$$\mathbb{P}\left\{\frac{M_n - b_n}{a_n} \leq z\right\} \approx G(z) \tag{2.11}$$

for large n. Hence, for any integer k, the following relation is satisfied

$$\mathbb{P}\left\{\frac{M_{nk} - b_{nk}}{a_{nk}} \leq z\right\} \approx G(z). \tag{2.12}$$

But, since M_{nk} can be seen as the maximum of k variables, all having the same distributions as M_n, it results that

$$\mathbb{P}\left\{\frac{M_{nk} - b_{nk}}{a_{nk}} \leq z\right\} = \mathbb{P}\left\{\frac{M_n - b_n}{a_n} \leq z\right\}^k. \tag{2.13}$$

The combination of 2.12 and 2.13, gives

$$\mathbb{P}\{M_{nk} \leq z\} \approx G\left(\frac{z - b_{nk}}{a_{nk}}\right) \tag{2.14}$$

and

$$\mathbb{P}\{M_{nk} \leq z\} \approx G^k\left(\frac{z - b_n}{a_n}\right). \tag{2.15}$$

As a result, since G and G^k are identical except for re-normalization coefficients, G is max stable and, for Theorem 2.1.4, it belongs to the GEV family ([1]).

2.2 Bivariate models

2.2.1 Asymptotic characterization

Let $(X_1, Y_1), (X_2, Y_2), \ldots$ be a sequence of vectors which are independent versions of a random vector having distribution function $F(x, y)$. The classical approach prescribes to study the behavior of extremes through some limit distribution for block maxima. In this perspective, the following definition is required ([1]):

Definition 2.2.1. Let $M_{x,n} = \max_{i=1,\ldots,n}\{X_i\}$ and $M_{y,n} = \max_{i=1,\ldots,n}\{Y_i\}$. Then, the vector

$$\mathbf{M}_n = (M_{x,n}, M_{y,n})$$

is called **vector of componentwise block maxima.**

Note that the index i, for which X_i is the maximum over its sequence, it's not necessarily the same as that of the Y_i sequence. Therefore, the vector $\mathbf{M}_n$ may not correspond to a bivariate observation of the original series.
As in the univariate case, the main interest is that of studying the behavior of $\mathbf{M}_n$ as $n \to \infty$. By looking at $\{X_i\}$ and$\{Y_i\}$ separately, it is possible to extend to them the results achieved in the previous section, since they correspond to i.i.d sequences of random variables. In particular, the analysis can be simplified by assuming a known marginal distribution for both the X_i and Y_i. The most common representation suggests to use a standard Fréchet distribution, which corresponds to a GEV with $\mu = 0, \sigma = 1$ and $\xi = 1$:

$$F(z) = \exp\left(-\frac{1}{z}\right), \quad z > 0.$$

In order to obtain non degenerate marginal limits, it is better to consider the following standardized version for $\mathbf{M}_n$,

$$\mathbf{M}_n^* = \left(\max_{i=1,\ldots,n}\{X_i\}/n, \max_{i=1,\ldots,n}\{Y_i\}/n\right). \tag{2.16}$$

Now it is possible to state the bivariate analog of Theorem 2.1.1 ([1]).

Theorem 2.2.2. *Let $\boldsymbol{M}_n^* = (M_{x,n}^*, M_{y,n}^*)$ be defined by 2.16, where the (X_i, Y_i) are independent vectors with standard Fréchet marginal distributions. Then if*

$$\mathbb{P}\{M_{x,n}^* \leq x, M_{y,n}^* \leq y\} \xrightarrow{d} G(x,y),$$

where G is a non-degenerate distribution function, G has the form

$$G(x,y) = \exp\{-V(x,y)\}, \quad x > 0, y > 0 \tag{2.17}$$

where

$$V(x,y) = 2\int_0^1 \max\left(\frac{w}{x}, \frac{1-w}{y}\right) dH(w), \tag{2.18}$$

and H is a distribution function in $[0,1]$ satisfying the mean constraint

$$\int_0^1 w dH(w) = 1/2. \tag{2.19}$$

Theorem 2.2.2 states that the standardized vector of componentwise block maxima converges in distribution to a member of the family introduced in 2.17, which is known as the class of **bivariate extreme value distributions.** In particular, the elements of this family are generated by functions H, satisfying 2.19, which may not be differentiable.
The function V is called **exponent measure** and satisfies the following property

$$V(a^{-1}x, a^{-1}y) = aV(x,y) \quad \forall a > 0.$$

More precisely, V is said to be **homogeneous of order** -1. Applying this property to 2.17 implies that

$$G^n(x,y) = G(n^{-1}x, n^{-1}y),$$

so, the distribution G has a bivariate version of the property of max-stability defined in the previous section.
It is possible to extend the distribution in 2.17, in order to have the complete class of bivariate limits for arbitrary GEV margins. This is simply achieved by generalizing the marginal distributions, i.e. by defining

$$\tilde{x} = \left[1 + \xi_x\left(\frac{x-\mu_x}{\sigma_x}\right)\right]^{\frac{1}{\xi_x}} \quad \text{and} \quad \tilde{y} = \left[1 + \xi_y\left(\frac{y-\mu_y}{\sigma_y}\right)\right]^{\frac{1}{\xi_y}}.$$

As a result, the marginal distributions of

$$G(x,y) = \exp\{-V(\tilde{x}, \tilde{y})\}$$

are GEV with parameters (μ_x, σ_x, ξ_x) and (μ_y, σ_y, ξ_y), provided that $[1 + \xi_x(x - \mu_x)/\sigma_x] > 0$, $[1 + \xi_y(y - \mu_y)/\sigma_y] > 0$, and that the function V satisfies 2.18 for some choice of H ([1]).

2.2.2 Model examples for Componentwise Maxima

According to Theorem 2.2.2, each function $H(x)$ on $[0,1]$ satisfying the constraint in 2.19 can generate a valid member of the class of bivariate limit distributions. However, it is not so easy to find parametric families whose mean is constant and for which the integral in 2.18 can be computed. The general approach prescribes to work with sub families for H, and hence G, that are easily tractable and from which the entire class of limit distributions can be approximated.
One standard model is represented by the **logistic family**:

$$G(x,y) = \exp\left\{-\left(x^{-1/\alpha} + y^{-1/\alpha}\right)^{\alpha}\right\}, \quad x > 0, y > 0, \tag{2.20}$$

for some parameter $\alpha \in (0,1)$. It can be shown that in this case H is differentiable and has density function

$$h(w) = \frac{1}{2}(\alpha^{-1} - 1)\{w(1-w)\}^{-1-1/\alpha}\{w^{-1/\alpha} + (1-w)^{-1/\alpha}\}^{\alpha-2}$$

on $0 < w < 1$. Note that $h(w)$ is symmetric about $w = \frac{1}{2}$ and this implies that variables x and y are exchangeable in 2.20.
The α parameter is known as **logistic dependence parameter**, since it determines the level of dependence between the two processes. In fact, as $\alpha \to 1$

$$G(x,y) \to \exp\{-(x^{-1} + y^{-1})\},$$

corresponding to independent variables; while, decreasing values for α, lead to dependence. More precisely, perfect dependence is achieved as limit for $\alpha \to 0$, in which case

$$G(x,y) \to \exp\{-\max(x^{-1}, y^{-1})\}.$$

The previous model can be generalized in the asymmetric case, leading to the **bilogistic model**, for which

$$G(x,y) = \exp\left\{(1-w)\,u^{1-\alpha} + w\,(1-u)^{1-\beta}\right\}$$

and the distribution function H has density

$$h(w) = \frac{1}{2}(1-\alpha)(1-w)^{-1}w^{-2}(1-u)u^{1-\alpha}\{\alpha(1-u) + \beta u\}^{-1}$$

on $0 < w < 1$. Parameters α and β are such that $0 < \alpha < 1$ and $0 < \beta < 1$, and u is the solution of

$$(1-\alpha)(1-w)(1-u)^{\beta} - (1-\beta)wu^{\alpha} = 0.$$

More precisely, the quantity $\alpha - \beta$ determines the level of asymmetry in the dependence structure and hence, the special case with $\alpha = \beta$ corresponds to the logistic model ([1]).

2.2.3 Extremal dependence

The concept of dependence is crucial when studying combination of processes, even at high levels.
Classical methods for bivariate extremes are limited to the case in which extremes events are dependent, since, in all the other cases, their application can lead to over-estimate the probability of extreme joint events.
Therefore, it is advisable to quantify the strength of dependence between processes tails before modelling them via traditional methodologies.
The following section presents two measures that are useful for quantifying the extremal dependence for generic bivariate random vectors.

Measures of Extremal Dependence

Let consider a generic random vector (X, Y), then it is possible to show that there is a unique function $C(\cdot, \cdot)$, with domain $[0, 1] \times [0, 1]$, such that

$$F(x, y) = C\{F_X(x), F_Y(y)\}$$

where F_X and F_Y are the marginal distribution functions of X and Y.
Function C is called **copula** and describes the relationship between X and Y, independently from the marginal distributions. In other words C can be seen as the joint distribution of variables, after transformation to $U = F_X(X)$ and $V = F_Y(Y)$, having uniform standard margins ([2]).
As an example, the bivariate logistic model, defined in the previous section, has the following parametric copula

$$C_\alpha(u, v) = \exp\left\{-\left[(-\log u)^{1/\alpha} + (-\log v)^{1/\alpha}\right]^\alpha\right\},$$

on $[0, 1] \times [0, 1]$ where, as previously said, the value of α quantifies the extent of dependence between the two variables .
It can be useful to summarize the information contained in the copula through some one-dimensional function or parameter, simplifying both dependence interpretation and inference.
For this purpose let define

$$\chi = \lim_{u \to 1} \mathbb{P}\{F_Y(Y) > u | F_X(X) > u\} = \lim_{u \to 1} \mathbb{P}\{V > u | U > u\},$$

which measures the tendency of one variable to be large conditional on the other variable being large. The same measure can be also obtained as limit of the following asymptotically equivalent function

$$\begin{aligned} \chi(u) &= 2 - \frac{\log \mathbb{P}\{F_X(X) < u, F_Y(Y) < u\}}{\log \mathbb{P}\{F_X(X) < u\}} \\ &= 2 - \frac{\log \mathbb{P}\{F_X(X) < u, F_Y(Y) < u\}}{\log u} \end{aligned} \tag{2.21}$$

for $0 < u < 1$.

In particular, the subsequent properties are satisfied ([1]):

- $0 \leq \chi \leq 1$;
- if $\chi = 0$, variables X and Y are said to be **asymptotically independent**;
- for bivariate extreme value distribution, $\chi = 2 - V(1,1)$. As a consequence, for the logistic model, $\chi = 2 - 2^{\alpha}$, as $V_{\alpha}(1,1) = 2^{\alpha}$;
- within the class of asymptotically dependent variables, the value of χ grows as the level of dependence increases at extreme levels.

According to the previous properties, χ represents a measure of the strength of dependence when dealing with asymptotically dependent variables. However, in the case of asymptotically independent distributions, it is unable to provide any type of information, being $\chi = 0$ within this class. In order to overcome this deficiency it is possible to define a second measure. For $0 < u < 1$, let

$$\begin{aligned} \overline{\chi}(u) &= \frac{2\log \mathbb{P}\{F_X(X) > u\}}{\log \mathbb{P}\{F_X(X) > u, F_Y(Y) > u\}} - 1 \\ &= \frac{2\log(1-u)}{\log \mathbb{P}\{F_X(X) > u, F_Y(Y) > u\}} - 1 \end{aligned} \tag{2.22}$$

and

$$\overline{\chi} = \lim_{u \to 1} \overline{\chi}(u).$$

The previous measure satisfies the following properties ([1]):

- $-1 \leq \overline{\chi} \leq 1$;
- $\overline{\chi} = 1$, for **asymptotically dependent variables**;
- for independent variables $\overline{\chi} = 0$;
- within the class of asymptotically independent variables, the value of $\overline{\chi}$ grows as the level of dependence increases at extreme levels.

Resuming, the above mentioned measures, taken together, are useful to quantify the extremal dependence for any random vector (X, Y). In fact, if $\overline{\chi} = 1$, the variables are asymptotically independent and χ can be use to measure the extent of extremal dependence. On the contrary, $\chi = 0$ implies that X and Y are asymptotically dependent and, this time, the value of $\overline{\chi}$ must be used to quantify the strength of dependence at high levels([1]).
Estimates for $\chi(u)$ and $\overline{\chi}(u)$ can be obtained using empirical observations. More precisely, such estimates can be plotted as the value of u changes, in order to study their behavior as $u \to 1$. The resulting graph is also known as **chiplot** and provides an informal mean, useful to make inference on extremal dependence.

Chapter 3

Model generalizations

Extreme observations are scarce by definition and this lack of data often leads to models characterized by high variance. Considering only block maxima, when just few years of observations are available, can cause a great waste of data, since there could be more than one extreme measurement in a single block.
The desperate need of additional information has resulted in the search of different characterizations, that allow the use of additional observations when modeling extreme values. The two main generalizations are:

- the r Largest Order Statistic Model
- the Threshold Exceedances Model.

Both characterizations are presented in the following sections.

3.1 The r Largest Order Statistic Model

This first model generalizes the block maxima approach by considering, for each block, not only the maximum, but the set of the r largest order statistics, for small values of r.
Just as in the previous chapter the main aim is that of describing the asymptotic behavior of i.i.d sequences of random variables X_i.
Let define

$$M_n^{(k)} = k\text{th largest order statistic of} \quad \{X_1, \ldots, X_n\}$$

then, Theorem 2.1.1 can be generalized as follows ([1]).

Theorem 3.1.1. *If there exist sequences of constants* $\{a_n > 0\}$ *and* $\{b_n\}$ *such that*

$$\mathbb{P}\left\{\frac{M_n - b_n}{a_n} \leq z\right\} \to G(z) \quad as \quad n \to \infty \tag{3.1}$$

for some non degenerate distribution function G*, so that* G *is the GEV distribution function given by 2.4, then, for fixed* k*,*

$$\mathbb{P}\left\{\frac{M_n^{(k)} - b_n}{a_n} \leq z\right\} \to G_k(z) \quad as \quad n \to \infty$$

on $\{z : 1 + \xi(z - \mu)/\sigma > 0\}$, *where*

$$G_k(z) = \exp\{-\tau(z)\} \sum_{s=0}^{k-1} \frac{\tau(z)^s}{s!} \tag{3.2}$$

with

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

Theorem 3.1.1 shows that the normalized kth largest order statistic in a block has limit distribution 3.2, where $G_k(z)$ has the same parameters of the GEV distribution of the block maxima.
However, this result is not so helpful in practice, since in most applications, for some value of r, the whole set of largest order statistics is available for each block. So, rather than estimating the behaviour of a single component, it would be preferable to approximate the limiting joint distribution of the entire vector

$$\mathbf{M}_n^{(r)} = (M_n^{(1)}, \ldots, M_n^{(r)}).$$

In the following theorem, the joint density function of the limit distribution is presented ([1]).

Theorem 3.1.2. *If there exist sequences of constants* $\{a_n > 0\}$ *and* $\{b_n\}$ *such that*

$$\mathbb{P}\left\{\frac{M_n - b_n}{a_n} \leq z\right\} \to G(z) \quad as \quad n \to \infty \tag{3.3}$$

for some non degenerate distribution function G,then, for fixed r, the limiting distribution as $n \to \infty$ *of*

$$\tilde{\boldsymbol{M}}_n^{(r)} = \left(\frac{M_n^{(1)} - b_n}{a_n}, \ldots, \frac{M_n^{(r)} - b_n}{a_n}\right)$$

falls within the family having joint probability density function

$$f(z^{(1)}, \ldots, z^{(r)}) = \exp\left\{-\left[1 + \xi\left(\frac{z^{(r)} - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \times \prod_{k=1}^{r} \sigma^{-1}\left[1 + \xi\left(\frac{z^{(k)} - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1}, \tag{3.4}$$

where $-\infty < \mu < \infty$, $\sigma > 0$ *and* $-\infty < \xi < \infty$; $z^{(r)} \leq z^{(r-1)} \leq \cdots \leq z^{(1)}$; *and* $z^{(k)} : 1 + \xi(z^{(k)} - \mu)/\sigma > 0$ *for* $k = 1, \ldots, r$.

The r largest order statistic model certainly represents an improvement, in terms of waste of data, with respect to the block maxima approach. However, in most applications, data are not available in this form and it can happen that the number of extreme observations in each block is different. For this reason it would be better to avoid the use of blocking methods, especially when the entire series of measurements is available.
The approach described in the following section moves in this direction.

3.2 The Threshold Exceedances Model

As previously said, when the data at disposal consist of entire time series of measurements, it would be better to exploit all information.
In this perspective, let $X_1, X_2, \dots$ be a sequence of random variables i.i.d F, then all values of the X_i exceeding some high threshold u can be read as extreme observations.
Now, let X be an arbitrary term of the i.i.d sequence, the following conditional probability can be used to describe the behavior of extreme events

$$\mathbb{P}\{X > u + y | X > u\} = \frac{1 - F(u+y)}{1 - F(u)}, \quad y > 0. \tag{3.5}$$

Unfortunately, in real applications, the CDF F is unknown and this makes it impossible to directly apply Eq. 3.5. Therefore, it would be desirable to find approximations of the parent distribution of excesses, which are valid for fairly large threshold values.

3.2.1 The Generalized Pareto Distribution

In this context, the following theorem contains an important result ([1]).

Theorem 3.2.1. *Let $X_1, X_2, \dots$ be a sequence of independent random variables with common distribution F, and let*

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

Denote an arbitrary term in the X_i sequence by X, and suppose that F satisfies 2.1.2, so that for large n,

$$\mathbb{P}\{M_n \leq z\} \approx G(z),$$

where

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

for some $\mu, \sigma > 0$ and ξ. Then , for large enough u, the distribution function of $(X - u)$, conditional on $X > u$, is approximately

$$H(y) = 1 - \left(1 + \frac{\xi y}{\hat{\sigma}} \right)^{-\frac{1}{\xi}} \tag{3.6}$$

defined on $\{y : y > 0 \quad and \quad (1 + \xi y/\hat{\sigma}) > 0\}$, where

$$\hat{\sigma} = \sigma + \xi(u - \mu). \tag{3.7}$$

Theorem 3.2.1 suggests that, if the block maxima can be approximated through GEV distribution, then the excesses over a threshold u have limit distribution within the family mentioned in Eq. 3.6, which is known as **generalized**

Pareto family.
It is important to note that the parameters of $H(y)$ in 3.6 are completely determined by those of the relative GEV distribution of block maxima. In particular the parameter ξ is the same for the two models and its value is decisive for the behavior of the Pareto distribution.
More precisely, if $\xi < 0$, the distribution of excesses has an upper bound of $u - \hat{\sigma}/\xi$, while for $\xi \geq 0$ it is unbounded. Once again the case $\xi = 0$ is equivalent to considering the limit of 3.6 as $\xi \to 0$, which leads to

$$H(y) = 1 - \exp\left(-\frac{y}{\hat{\sigma}}\right), \quad y > 0,$$

which corresponds to an exponential distribution with parameter $1/\hat{\sigma}$ ([1]).

3.2.2 Notes on the justification for the Generalized Pareto Model

In this brief section a simplified proof of Theorem 3.2.1 is presented.
Let consider a random variable X with distribution function F. Theorem 2.1.1 states that

$$F^n(z) \approx \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \tag{3.8}$$

for large n, and for some parameters $\mu, \sigma > 0$ and ξ. Thus, by taking logarithmic transformation,

$$n \log F(z) \approx -\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}. \tag{3.9}$$

For large values of z, a Taylor expansion implies that

$$\log F(z) \approx -\{1 - F(z)\}, \tag{3.10}$$

so, by substituting 3.9 into 3.10, it follows that, for large values of u,

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

Similarly, for $y > 0$,

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

Therefore,

$$\begin{aligned}
\mathbb{P}\{X > u + y | X > u\} &\approx \frac{n^{-1}[1 + \xi(u+y-\mu)/\sigma]^{-1/\xi}}{n^{-1}[1 + \xi(u-\mu)/\sigma]^{-1/\xi}} \\
&= \left[1 + \frac{\xi(u+y-\mu)/\sigma}{1 + \xi(u-\mu)/\sigma}\right]^{-1/\xi} \\
&= \left[1 + \frac{\xi y}{\hat{\sigma}}\right]^{-1/\xi},
\end{aligned} \tag{3.11}$$

where

$$\hat{\sigma} = \sigma + \xi(u - \mu),$$

as required ([1]).

3.2.3 Return levels

Also in this case it is possible to derive return levels from the estimated extreme value model. Let assume that the family of distributions in Eq. 3.6 is appropriate for modeling the exceedances over a threshold u for a random variable X, that is

$$\mathbb{P}\{X > z | X > u\} = \left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}.$$

for $z > u$. As a consequence

$$\begin{aligned}\mathbb{P}\{X > z\} &= \mathbb{P}\{X > u\}\mathbb{P}\{X > z | X > u\} \\ &= \mathbb{P}\{X > u\}\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}.\end{aligned}$$

Now, let denote with $p_u = \mathbb{P}\{X > u\}$, it follows that the level z_m exceeded on average once every m observations corresponds to the solution of the equation

$$p_u\left[1 + \xi\left(\frac{z_m - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}} = \frac{1}{m}. \tag{3.12}$$

By solving Eq 3.12, it results that

$$z_m = u + \frac{\sigma}{\xi}[(mp_u)^{\xi} - 1],$$

which is valid for $z_m > u$. By definition, z_m is the **m-observation return level**, however, since in most application the focus goes on the N-year return level, it would be more convenient to consider

$$z_N = u + \frac{\sigma}{\xi}[(Nn_y p_u)^{\xi} - 1],$$

for which $m = N \times n_y$ is the number of observations in N years, each of which consisting of n_y measurements ([1]).

Chapter 4

Inference and Model assessment

The previous chapter provides an outline of the main theoretical results of EVT. In the following sections some specifications about inference and model assessment are presented, since they will be exploited in the real data analysis.

4.1 Profile likelihood and Likelihood ratio test

All the models presented in the previous chapter can be used to fit extreme value distributions to data. In order to make inference on the value of parameters, maximum likelihood procedures are usually used. In particular, let recall that, if $x_1, x_2, \ldots, x_n$ are independent observations of a random variable X having parametric distribution function F, indexed by $\theta \in \mathbb{R}^d$, the related **likelihood function** is

$$L(\theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

where $f(x_i, \theta)$ is the probability density function of X. The value of θ for which $L(\theta)$ is maximized is called **maximum likelihood estimator** and, under appropriate regularity conditions, it is possible to show that it has limit distribution within the class of multivariate normal distributions. This last asymptotic argument can be exploited to compute confidence intervals for the parameters; however, more accurate results can be obtained through an alternative technique based on the so called **profile likelihood.**
Let $l(\theta)$ be the logarithmic transformation of $L(\theta)$, also called **log-likelihood**,

$$l(\theta) = \log L(\theta) = \prod_{i=1}^{n} \log f(x_i, \theta).$$

The function can be also written in the form $l(\theta_i, \theta_{-i})$, where θ_{-i} indicates the set of all components of θ excluding θ_i. Hence, the **profile log-likelihood** for

an arbitrary θ_i can be defined as

$$l_p(\theta_i) = \max_{\theta_{-i}} l(\theta_i, \theta_{-i}).$$

In other words the profile log-likelihood for θ_i is obtained by maximizing the log likelihood with respect to all components of θ, except θ_i.
For this type of likelihood function is possible to derive an asymptotic result very useful for both parameter inference and model selection ([1]).

Theorem 4.1.1. *Let $x_1, x_2, \ldots, x_n$ be independent realizations from a parametric distribution function F, and let $\hat{\theta}$ denote the maximum likelihood estimator of the d-dimensional model parameter $\theta = (\theta^{(1)}, \theta^{(2)})$, where $\theta^{(1)}$ is a k-dimensional subset of θ. Then, under suitable regularity conditions, for large n*

$$D_p(\theta^{(1)}) = 2\{l(\hat{\theta}) - l_p(\theta^{(1)})\} \sim \chi^2_k. \tag{4.1}$$

Function D_p in Eq.4.1 is called **deviance function** and, as it measures the amount of uncertainty affecting the maximum likelihood estimator, it can be used to derive alternative confidence intervals for parameters.
As an example, the region $C_\alpha = \{\theta_i : D_p(\theta_i) \leq c_\alpha\}$ corresponds to the $(1-\alpha)$ confidence interval for θ_i , where c_α is the $(1-\alpha)$ quantile of the χ^2_1 distribution.

Another important application of Theorem 4.1.1 concerns model selection.
In particular, let assume that $\mathcal{M}_1$ is a model with parameter vector $\theta = (\theta^{(1)}, \theta^{(2)}) \in \mathbb{R}^d$ and let $\mathcal{M}_0$ be a nested model of $\mathcal{M}_1$, i.e its parameters correspond to a $(d\text{-}k)$-dimensional subset of θ, e.g $\theta^{(2)}$. Then, let $l_1(\mathcal{M}_1)$ and $l_0(\mathcal{M}_0)$ be the maximized log-likelihood for $\mathcal{M}_1$ and $\mathcal{M}_0$, respectively, and define the statistic

$$D = 2\{l_1(\mathcal{M}_1) - l_0(\mathcal{M}_0)\}.$$

For Theorem 4.1.1, a $(1-\alpha)$ confidence region for the true value of $\theta^{(1)}$ is represented by $C_\alpha = \{\theta^{(1)} : D_p(\theta^{(1)}) \leq c_\alpha\}$, where c_α is the $(1-\alpha)$ quantile of the χ^2_k distribution. As a consequence, if 0 belongs to C_α, or equivalently if $D < c_\alpha$, then $\mathcal{M}_0$ represents a plausible reduction of $\mathcal{M}_1$ ([1]).
The just described test is known as **likelihood ratio test** and, as already pointed out, provides a useful tool for model selection when dealing with nested models.

4.2 Diagnostic Plots

After a model is estimated, it is necessary to assess its validity in terms of goodness of fit. Since, in practice, it's not possible to find others sources of data against which to compare the model, the only alternative for checking its accuracy is to evaluate how much it is in agreement with the observations used to extrapolate it. For this reason goodness of fit techniques are mainly based on the comparison between model based and empirical estimates of the distribution function. Two of the most common graphical procedures in this sense are the **probability plot** and the **quantile plot**.

Probability plot

Let consider a series $x_1, x_2, \ldots, x_n$ of independent realizations from a distribution function F, and let $\hat{F}$ indicate an estimate obtained through some extrapolation procedure. If $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ denote the order statistic of the sample (i.e $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$), then the **empirical distribution function** of F can be defined as

$$\tilde{F} = \frac{i}{n+1} \quad \text{for} \quad x_{(i)} \leq x \leq x_{(i+1)}.$$

The probability plot allows to check the accuracy of the estimated model by comparing $\hat{F}(x_{(i)})$ and $\tilde{F}(x_{(i)})$ for $i = 1, \ldots, n$. More precisely, it is defined as the set of points:

$$\left\{ \left(\hat{F}(x_{(i)}), \frac{i}{n+1} \right) : i = 1, \ldots, n \right\}.$$

If $\hat{F}$ provides a good approximation of the real distribution, the points in the probability plot should lie close to the bisector. As a consequence, all departures from linearity must be considered a symptom of lack of fit ([1]).

Quantile plot

Given $\hat{F}$ and the series of order statistics as above, the quantile plot consists in the set of points

$$\left\{ \left(\hat{F}^{-1}\left(\frac{i}{n+1} \right), x_{(i)} \right) : i = 1, \ldots, n \right\},$$

where $\hat{F}^{-1}(\frac{i}{n+1})$ and $x_{(i)}$ provide model based and empirical estimates for the $\frac{i}{n+1}$ quantile of the distribution function F. It contains the same information as the probability plot, just presented on a different scale. However, since it tends to be more sensitive to lack of fit, especially in tail of data, it should be preferred for validity checks ([1]).

In the context of Extreme value models, in addiction to the just describe techniques, it is possible to define another graphical procedure for model assessment, called **return level plot**.
In the following rows the graph will be described in the case of the annual maximum distribution, however, note that it is possible to obtain equivalent results for the Generalized Pareto model.
Let recall the definition presented in Section 2.1.2. The return level associated with the return period $1/p$ is

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}[1 - \{-\log(1-p)\}^{-\xi}], & \text{for} \quad \xi \neq 0, \\ \mu - \sigma \log(-\log(1-p)), & \text{for} \quad \xi = 0, \end{cases} \tag{4.2}$$

where $G(z_p) = 1 - p$. By imposing $y_p = -\log(1-p)$, Eq.4.2 can be re-written as:

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}[1 - y_p^{-\xi}], & \text{for} \quad \xi \neq 0, \\ \mu - \sigma \log(y_p), & \text{for} \quad \xi = 0. \end{cases} \tag{4.3}$$

The curve obtained by plotting z_p against $\log y_p$, corresponds to the so called **return level plot** and has the following characteristics ([1]):

- for $\xi < 0$, it is convex and has asymptotic limit at $\mu - \sigma/\xi$, as $p \to 0$;
- for $\xi > 0$, it is concave with no finite bound ;
- finally, it is linear for $\xi = 0$.

The described graph allows for model interpretation and validation. In particular, the choice of plotting probabilities on a logarithmic scale helps to highlight the effect of extrapolation in the tail, making it possible to display return level even for long return period. Moreover, the addition of empirical estimates for the return level function in the plot allows validity checks. If the fitted GEV distribution is adequate for the data, the empirical estimates are expected to be in agreement with the model based curve, while, any disagreement must be interpreted as a sign of scarce adequacy.

4.3 Threshold selection

Theorem 3.2.1 in Section 3.2 is based on the characterization of extreme values as observations exceeding an high threshold u. In particular, for large enough u, such exceedances can be approximated through a member of the Generalized Pareto family. In this context the choice of a suitable threshold is required before applying estimation procedures and such selection imply a balance between bias and variance. In fact, too high a threshold will lead to a model characterized by large variance, since only few observations will be used to fit it; while, too small value for u will produce bias, as the asymptotic assumption behind the approximation may no longer be satisfied.
Two methods are available for threshold selection.

4.3.1 Mean residual life plot

This first methodology is based on the mean of the generalized Pareto distribution. If X is distributed according to a generalized Pareto with parameters σ and ξ, then for $\xi < 1$

$$\mathbb{E}(Y) = \frac{\sigma}{1-\xi},$$

otherwise the mean is infinite.
So, let u_0 be the threshold value for which the excesses generated by the series $X_1, X_2, \ldots, X_n$ can be approximated through a member of the generalized Pareto family. Then, for each $u > u_0$

$$\mathbb{E}(X - u | X > u) = \frac{\sigma_u}{1-\xi},$$

where σ_u is the scale parameter corresponding to the excesses over the threshold u. In particular, by virtue of Eq. 3.7 in Theorem 3.2.1, $\sigma_u = \sigma_{u_0} + \xi(u - u_0)$,

where σ_{u_0} represents the scale parameter corresponding to the threshold u_0.
As a consequence, the mean of excesses is expected to depend linearly from u above a value $u = u_0$ for which the assumptions of Theorem 3.2.1 are satisfied.
The previous arguments suggest a technique for selecting the more suitable threshold, which consists in plotting the sample mean of excesses against the value of u, looking for a point from which the graph is approximately linear in u. Such plot is known as **mean residual life plot** and can be defined as the set of point

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

where $x_{(1)}, \ldots, x_{(n_u)}$ are the n_u measurements exceeding u, and x_{max} is the largest of the X_i ([1]).

4.3.2 Alternative technique

In general, the mean residual life plot results difficult to interpret. For this reason a complementary technique have been developed, in order to assess the hypothesis derived from the previous graph.
The procedure mainly consists in estimating the generalized Pareto model for a range of thresholds, with the aim of studying the behavior of parameters. In fact, as already discussed in the previous paragraph, if u_0 is the value at which the generalized Pareto provides a reasonable approximation for the distribution of excesses, then, for any threshold $u > u_0$, the scale parameter can be expressed in the form

$$\sigma_u = \sigma_{u_0} + \xi(u - u_0).$$

Now, let consider the following constant reparameterization of σ_u

$$\sigma^* = \sigma_u - \xi u,$$

it follows that, if u_0 represents a suitable threshold, then both σ^* and ξ are expected to be approximately stable above u_0 ([1]).
In the light of this argument the procedure prescribes to plot the estimates for the generalized Pareto parameters as u changes, looking for the lowest value $u = u_0$ for which they remain stable.

Chapter 5

Real data application

5.1 Introduction

The history of the Genoa area is deeply linked to that of the torrents that cross it, like the Polcevera, the Bisagno and other minor watercourses.
These waterways have long been associated with disastrous geo-hydrological events that injured the city with serious consequences both in terms of victims, damage and destruction.
From the flood of 1970, that caused 44 victims and over 2000 displaced, Genoa has continued to be hit and to record damage and victims.
Such extremes phenomena, well known to the national news, have been repeated with different intensities until today.
From the available data in the Genoa area, it emerges that many of the historical events that caused substantial damages, including deaths and missing people, were due to rains of short duration and high intensity, very characteristic of the Ligurian territory.
Despite the progress of modern weather forecasts, such intense phenomena are still difficult to predict accurately.
In Liguria, as in other areas of the country, intense weather events produce rainfall that easily exceeds 70 mm in an hour, 120 mm in 2 hours, and 200 mm in 6 hours. Precipitation falls in small-sized hydrographic basins, leading to flash floods, which in turn produce a considerable transport of sediments that can undermine rivers estuary.
Modelling such meteorological phenomena can be useful to predict future extremes events and manage their risks.
In this chapter the main results of the Extreme Value Theory are applied to rainfall observations collected in the province of Genoa. Firstly, some approaches derived from the univariate theory are exploited to model the measurements of a single station. Then, the bivariate analysis is applied to a data set consisting of couple of observations from two different rain gauges.

5.2 Daily rainfalls data

The data at disposal are extracted from the Ligurian weather and climate database in the form of daily aggregate rainfall measurements.
Only meteorological stations located in areas of great interest, as the torrents basins, are considered in the choice of the sample. Moreover, since hydrological annals are often characterized by big amounts of missing values and meteorological stations can be active in different periods, the following requirements are needed, in order to ensure the consistency of data:

- the measurements must cover a period of at least 50 years;
- the maximum number of missing values for each year must be at most 60 days;
- there must be at least two stations active in the same period.

The resulting dataset is composed by the observations collected at two rain gauges located in the Polcevera's basin, which were active from 1960 to 1998 and from 2003 to 2014.

Figure 5.1: Polcevera's basin

5.3 Univariate analysis

In this section the observations recorded at the Isoverde station are considered for the analysis . Apart from the four years with no measurements (1999-2002), during which the rain gauge was disabled, the dataset contains 43 missing values. All days with no observations are deleted from the data set, since their percentage is negligible and the elimination wont introduce bias in the model.

5.3.1 Block maxima approach

In order to fit the GEV distribution using the block maxima approach, the series of annual maximum rainfalls is considered.

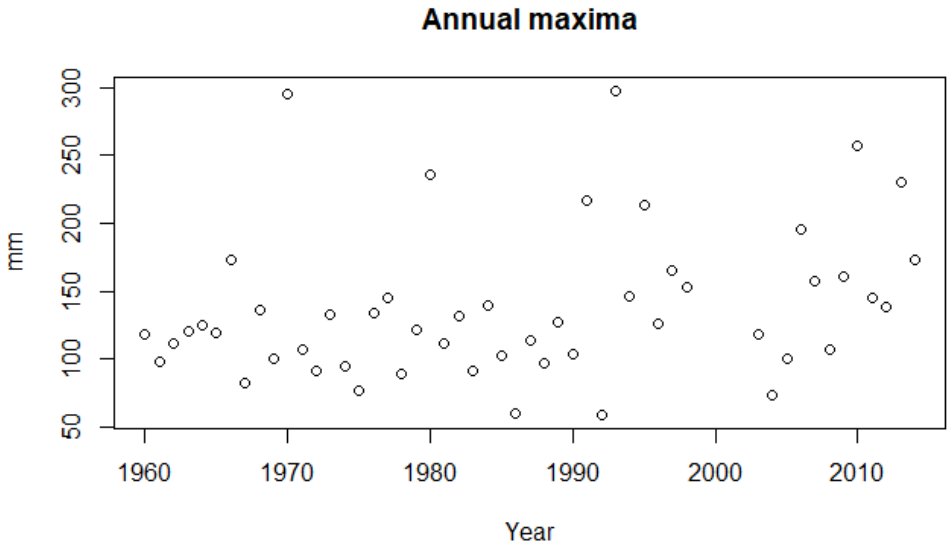

Figure 5.2: Annual maximum rainfalls recorded at Isoverde station

In Figure 5.2 it is possible to identify different extreme observations, probably related to the floods that hit the area from 1970 to 2014. Moreover, since there is no strong evidence of change in the pattern variation the data can be assumed to be independent observations drawn out from the GEV distribution.

Model fitting

Maximization of the GEV log-likelihood is performed using the function `fevd` of the package `extRemes`. The method function `ci`, instead, allows to find confidence intervals for parameters and return levels.

```
GEV3<-extRemes::fevd(maxima, type='GEV')
normcint<-ci(GEV3, alpha=0.05, type="parameter")
```

The resulting estimates and approximate 95% confidence intervals for the three parameters are:

Parameter	95% lower CI	Estimate	95% upper CI
μ	100.6075	112.0366	123.4658
σ	28.5475	37.1118	45.6760
ξ	-0.1060	0.0983	0.3027

The value of $\hat{\xi}$ is positive, but the 95% confidence interval extends also below zero, so the statistical evidence for an unbounded distribution is not so strong. In this situation the profile log-likelihood can be use to obtain more accurate intervals.

```
cint<-ci(GEV3, alpha=0.05, type="parameter" which.par=3, method='proflik',
xrange=c(-0.13, 0.4))
```

Parameter	95% lower CI	Estimate	95% upper
ξ	-0.0847	0.0983	0.3308

The confidence interval obtained from the profile likelihood is slightly translated with respect to the previous calculation, adding no useful information for the inference on the shape parameter, especially because 0 still lies in the interval. (see Figure 5.3).

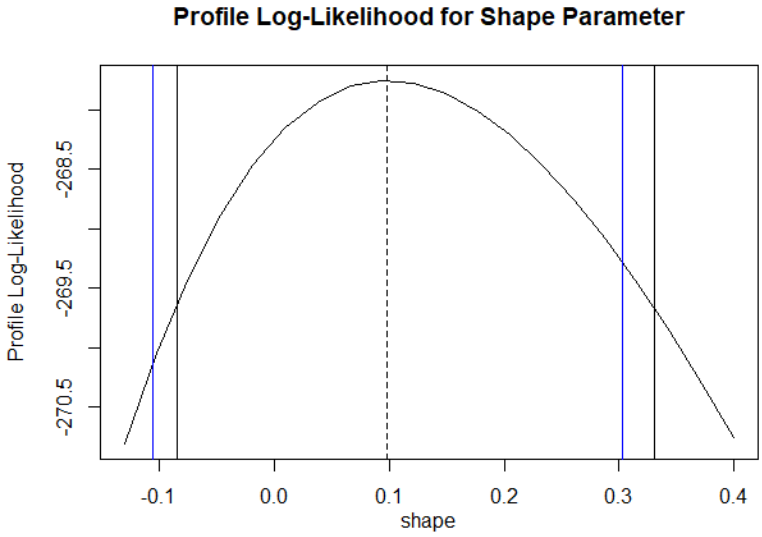

Figure 5.3: Profile likelihood for ξ with 95% normal approximated (blue) and profile (black) confidence intervals

Diagnostic plots

Some diagnostic plots can be used to assess the goodness of the GEV model.

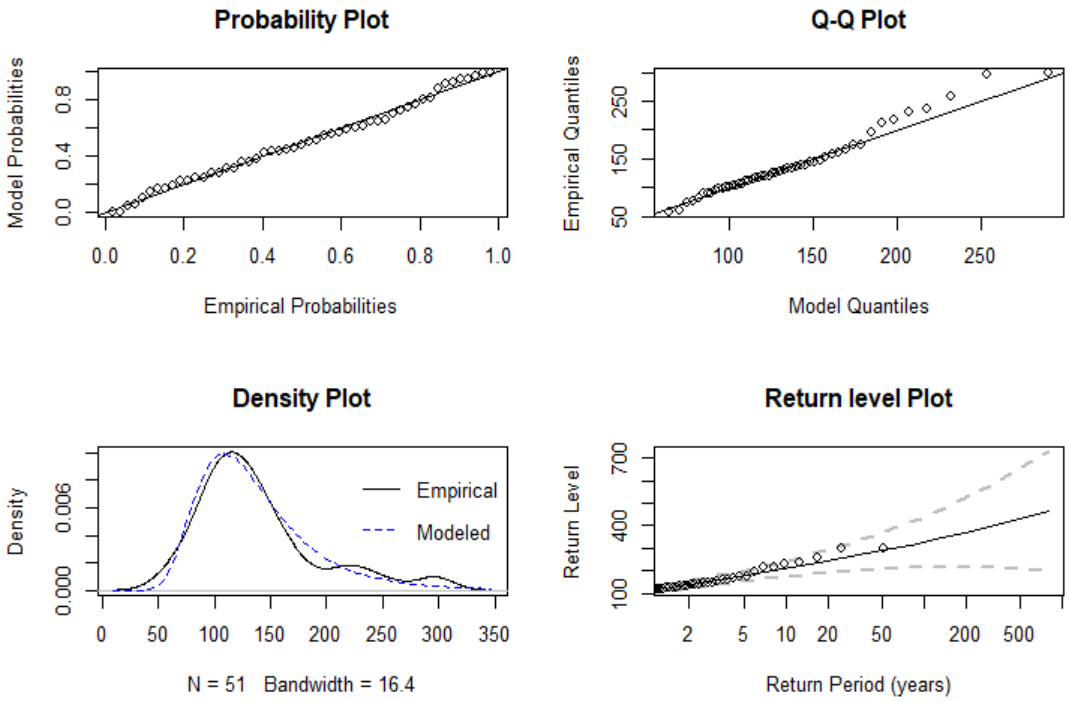

Figure 5.4: Diagnostic plots for GEV fit

The points in the probability plot are near-linear, and the estimated density curve seems consistent with empirical one, suggesting that the fitted GEV provides a quite good approximation of the data. However, the q-q plot presents some departures from linearity in correspondence of the tails, which are probably due to the increasing level of uncertainty that characterizes model extrapolation at high levels. The empirical estimates in the return level plot lie very close to the model based line, which results to be almost linear, since the approximation for the shape parameter is near to zero. However, even if the return level estimates seem convincing, the increasing confidence bands for large return periods indicate, once again, the uncertainty that affects the model at high levels.
Anyway, alternatives methodologies, like r-largest order statistics and peaks over threshold approaches, can be exploited in the attempt to improve model accuracy.

Return levels

Estimates and 95% confidence intervals for the return levels are obtained using the function `return.level`.

```
ret.lev<-return.level(GEV3, return.period=c(10,20,30,50,100), do.ci=TRUE)
```

	95% lower CI	Estimate	95% upper CI
10-year return level	173.4266	205.5154	237.6042
20-year return level	191.5596	240.0557	288.5545
30-year return level	199.9981	261.0624	322.1268
50-year return level	208.2666	288.5415	368.8417
100-year return level	214.9789	327.9238	440.8687

The previous results suggest that the level $z \approx 205.5$ mm is expected to be exceeded on average once every 10 years. However, caution is necessary when dealing with return-level estimates, since inferences can be poor, especially for long return periods. This can be deduced from the return level plot in Figure 5.4 : the confidence bands (dashed lines) tend to become wider as the return period increases.
Once again greater accuracy for confidence intervals can be obtained considering the profile likelihood.

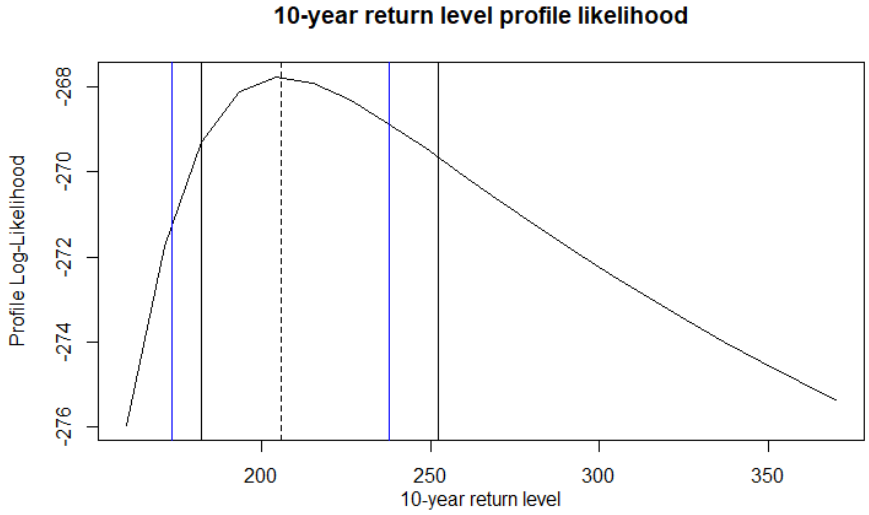

Figure 5.5: Profile likelihood for 10-year return level with 95% normal approximated (blue) and profile confident intervals (black)

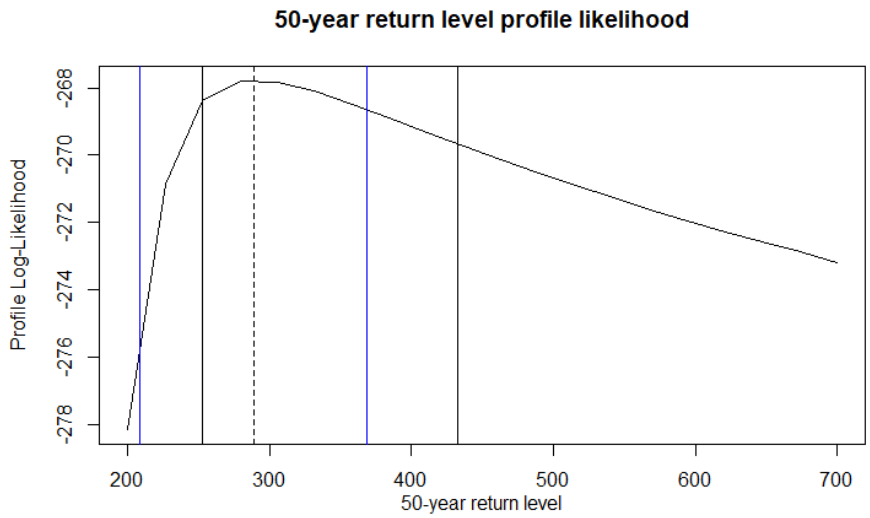

Figure 5.6: Profile likelihood for 50-year return level with 95% normal approximated (blue) and profile confident intervals (black)

From the previous figures it is possible to observe that, as the return level grows, the confidence intervals obtained from the profile log-likelihood differ more and more from those computed through normal approximation. In particular, the great gap between upper bounds in Figure 5.6 is due to the fact that the normal approximation is not able to reflect the uncertainty that affects the model.

Fitting Gumbel distribution

Since zero is contained in the confidence interval for ξ and the diagnostic plots in Figure 5.4 arise some doubts about the goodness of the GEV distribution, the suitability of modeling the data using the Gumbel family can be assessed.

```
GUMBEL<-extRemes::fevd(maxima, type="Gumbel")
```

The maximum likelihood estimates for the location and scale parameters with 95% confidence intervals are:

Parameter	95% lower CI	Estimate	95% upper CI
μ	102.9351	114.0063	125.0775
σ	30.0251	38.4878	46.9055

For what concerns the return levels, the following estimates and intervals are obtained:

	95% lower CI	Estimate	95% upper CI
10-year return level	175.8900	200.6181	225.3462
20-year return level	197.9750	228.3227	258.6705
30-year return level	210.6054	244.2605	277.9156
50-year return level	226.3341	264.1835	302.0229
100-year return level	247.5116	291.0561	334.6006

The main difference between the two models is related to the accuracy of estimation: in fact both parameters and return levels have narrower confidence intervals for the Gumbel family. Since an increase in model precision would be appreciated, statistical criteria and tests can be exploited in order to compare the goodness of fit for the two models.
The AIC and BIC criteria, for example, are useful when comparing maximum likelihood estimates on the same data. They provide a measure of statistical quality, penalizing model complexity, and are defined as :

$$AIC = 2k - 2\ln(L)$$

$$BIC = K\ln(n) - 2\ln(L)$$

where k is the number of parameters of the model and L is maximum value of the likelihood function. Let note that the model to be selected is that with the lowest value for both estimators.
In the considered example the compute values are:

	AIC	BIC
GEV	541.5011	547.2966
Gumbel	540.4965	544.3601

Both measures result smaller for the Gumbel family, suggesting that this distribution would be preferable for the data. However, it is better to perform a likelihood ratio test to assess the reduction to the Gumbel family.
The command `lr.test(GUMBEL, GEV3, alpha = 0.05)` returns the following values:

D	C_α	α	p-value
0.99537	3.8415	0.05	0.3184

The previous hypothesis is confirmed by the likelihood ratio test : the D statistic is smaller than the $(1-\alpha)$ quantile of the Chi-square distribution, so is not possible to reject the null-hypothesis of reduction to the Gumbel family.

5.3.2 r Largest Order Statistic approach

In general the r largest order statistic approach can be used to get improved accuracy, since with respect to the block maxima analysis it allows to include more information.
In order to choose how many statistics to consider, the model is fitted for different values of the r parameter and the behaviour of the standard error is analyzed. The function `rlarg.fit` in the `ismev` package implements the maximum-likelihood fitting for the order statistic model.

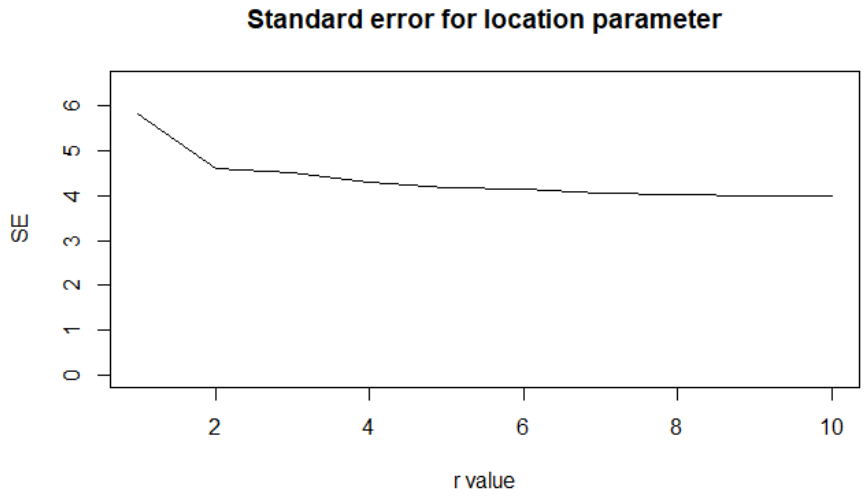

Figure 5.7: Standard error pattern for location parameter

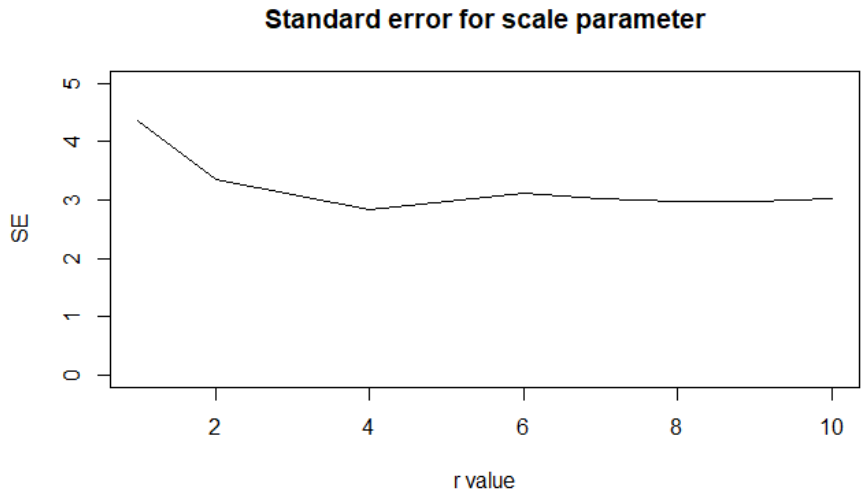

Figure 5.8: Standard error pattern for scale parameter

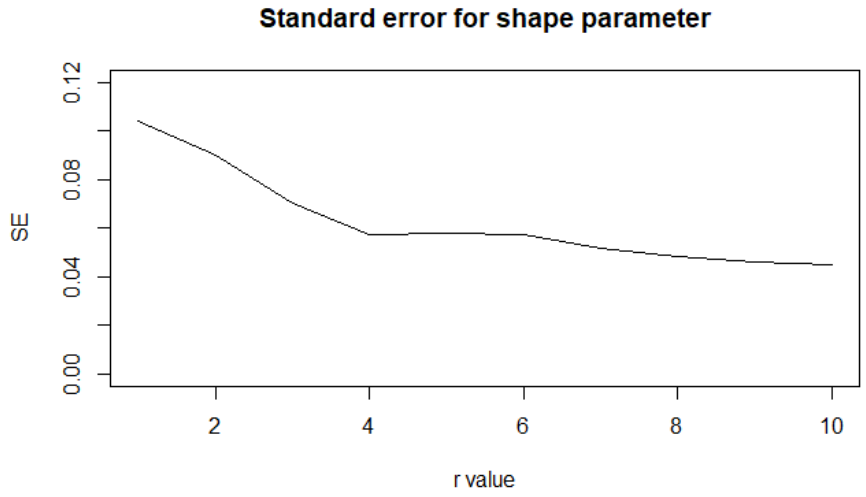

Figure 5.9: Standard error pattern for shape parameter

The standard error always decreases as the value of r grows, indicating an increase in model precision. However, all three graphs present an elbow between 2 and 4, suggesting that there would not be such significant improvement in considering a greater number of order statistics.
The following table shows the values for $\hat{\mu}, \hat{\sigma}$ and $\hat{\xi}$, obtained by fitting the model with different numbers of ordered statistics.

r	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$
1	112.0314	37.1128	0.0985
3	113.7729	36.5578	0.0793
5	115.1741	35.7699	0.0876
7	114.6938	35.7034	0.1021
9	114.7851	35.4204	0.1031

The estimates for the three parameters remain approximately stable as r grows. This lend support to the approximation of data using the r-largest order statistic approach, in fact, if the approximation results valid for a particular value of r, the estimates are expected to be stable when using fewer order statistics.
Anyway, the selected value for r may not be too large, as the asymptotic arguments supporting the model can be violated, leading to a biased approximation. Therefore, the model is fitted using r=3 order statistics and the obtained maximum likelihood estimates are:

$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$
113.7729	36.5578	0.0793

The approximations for the three parameters are very similar to those obtained through block maxima approach. This means that model extrapolation on the

basis of a greater number of observations has lead to similar conclusions. Probably the order statistics subsequent to the annual maxima are not able to provide enough information, useful for modeling extreme events.

Diagnostic plots

In order to assess the goodness of fit some diagnostic plot are realized.

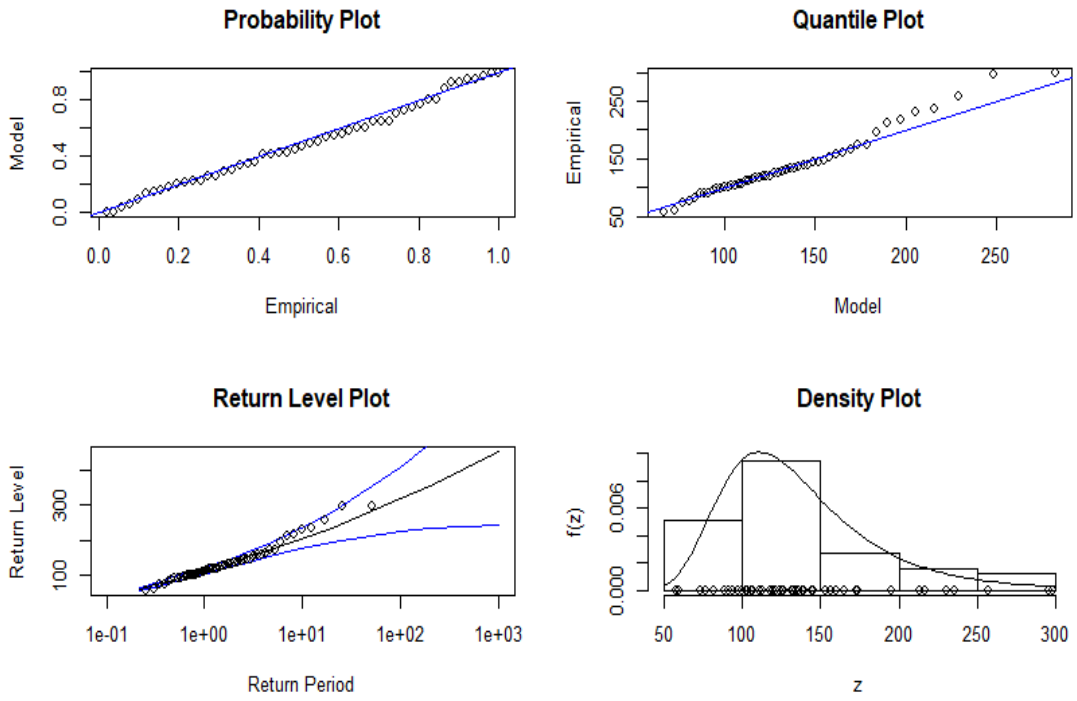

Figure 5.10: Diagnostic plots for order statistic model with r = 3

The graphs in Figure 5.10 are very similar to those obtained for the GEV model in the previous section, which is not surprising, as the estimated parameters for the two distributions are very similar.
The probability plot seems convincing, while the quantile plot presents some departures from linearity in correspondence of the extremes. Model accuracy tends to decrease when extrapolating at high levels and this is deductible from the increasing width of confidence bands in the return level plot. Therefore, caution is required when trying to make inferences using this approximation, since, for example, estimates for return levels can be poor, especially for long return periods.

5.3.3 Peaks over threshold approach

Threshold choice

To support the choice of the threshold both techniques discussed in Section 4.3 are used.
Firstly, the function `mrlplot`, in package `extRemes`, is employed to produce a mean residual life plot, including 95% confidence intervals.

```
> extRemes::mrlplot(rain_mm)
> title('Mean residual life plot')
```

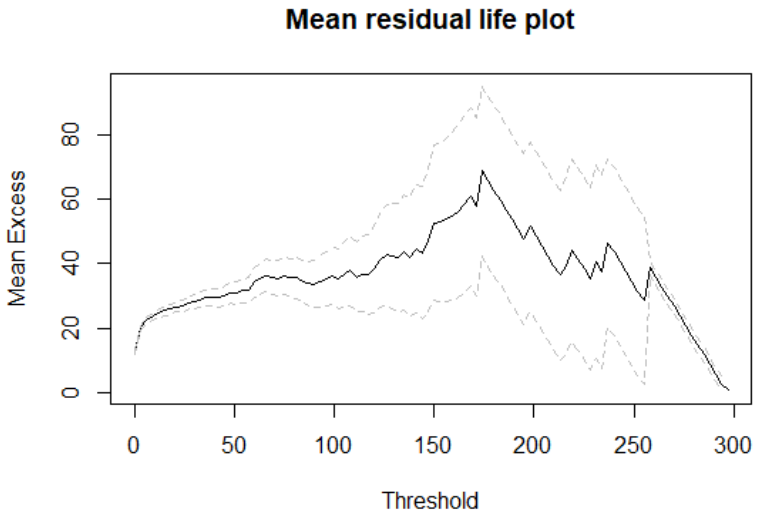

Figure 5.11: Mean residual life plot for daily rainfall data

As previously told, the plot represents the sample mean of excesses when the value of the threshold u changes, and it is expected to be approximately linear in u, above a certain threshold u_0 for which the excesses can be modeled by the Generalized Pareto distribution.
The graph in Figure 5.11 seems approximately linear between $u = 100$ and $u \approx 170$ and after that it declines steeply. A certain stability in the trend seems to be reached after $u \approx 170$. Unfortunately, only seven observations excess this threshold, not enough to fit an asymptotic model. It would be probably better to chose $u = 100$.
Since the interpretation of the mean residual life plot is not so easy, the complementary technique explained in Section 4.3 can be used to confirm the earlier hypothesis. The function `gpd.fitrange` in package `ismev` fits the GPD model for different values of the threshold u and realizes graphs of the parameter estimates together with confidence intervals.

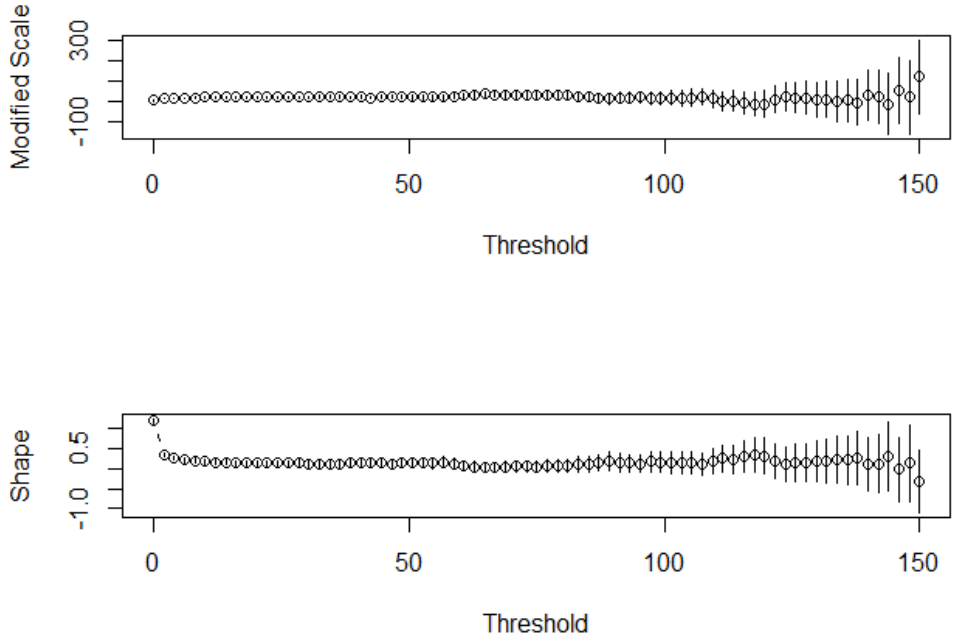

Figure 5.12: Parameter estimates against threshold values for daily rainfall data

The model-based technique suggests to look for stability of the Generalized Pareto parameters. In Figure 5.12 the estimates of ξ and σ_u become unstable for $u > 100$ and this supports the arguments derived from the mean residual life plot. Taking $u = 100$, the excesses data set is composed by 80 observations, good compromise between bias and variance.

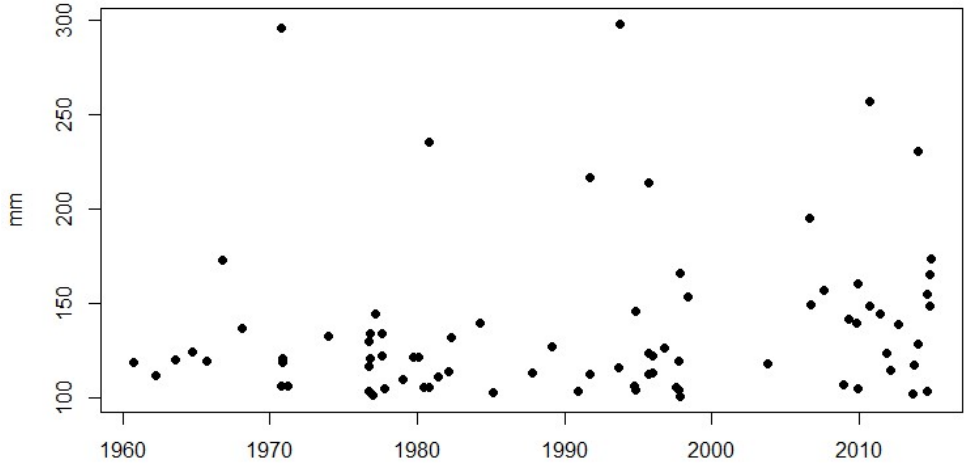

Figure 5.13: Excesses over threshold $u = 100$

Model fitting

Now it is possible to fit the Generalized Pareto Distribution, once again using the function `fevd` of package `extRemes`.

```
> GPD3<-extRemes::fevd(rain_mm, threshold=100, type="GP")
```

The maximum likelihood estimates together with the 95% confidence intervals are:

Parameter	95% lower CI	Estimate	95% upper CI
$\hat{\sigma}$	21.0394	31.5098	41.9802
$\hat{\xi}$	-0.1162	0.1352	0.3867

The value of $\hat{\xi}$ suggests an unbounded distribution ($\xi > 0$), however conclusions cannot be drawn, since 0 lies in the interval. Once again confidence intervals derived form the profile likelihood can be used to provide more accurate range for the parameter values.

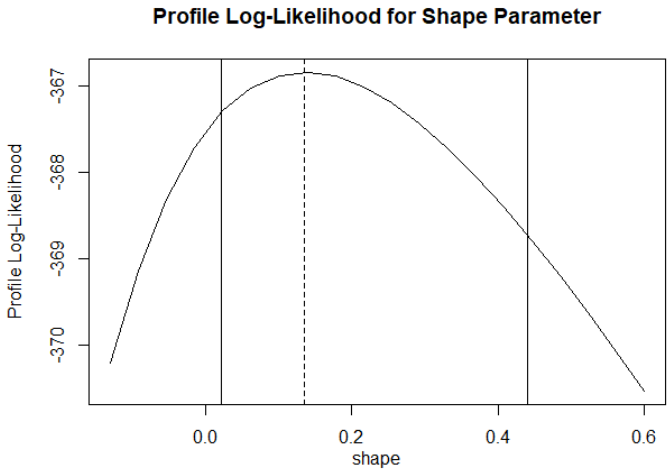

Figure 5.14: Profile likelihood for the shape parameter in the GPD model

The new confidence interval represented in Figure 5.14 turn out to be exclusively above zero, supporting the earlier hypothesis for an unbounded distribution.

Diagnostic plots

Figure 5.15 shows the diagnostic plots for the fitted GPD model.
Even if, in the probability plot, points lie almost perfectly on the bisector, the departures from linearity in the q-q plot give rise to some doubts about the

reliability of the model. Moreover, the return level plot presents extremely wide confidence bands as the return period increases.

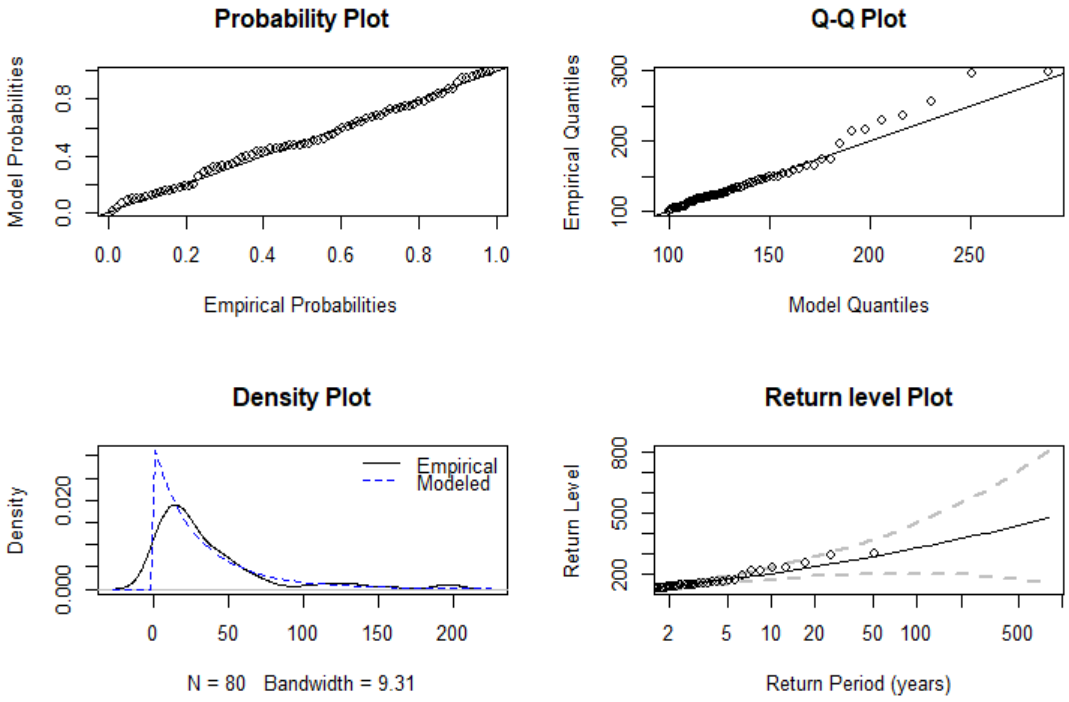

Figure 5.15: Diagnostic plots for GPD model

The uncertainty that characterizes the large values of the model can be also derived by computing the profile likelihood for high return levels.

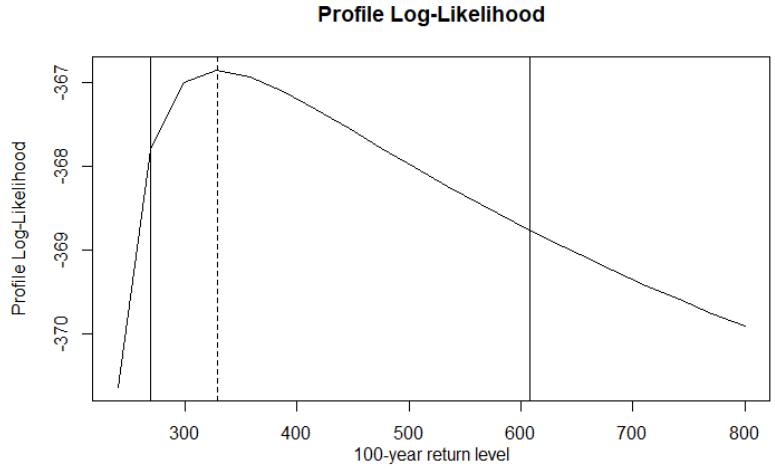

Figure 5.16: Profile likelihood for 100-year return level in GPD model

Figure 5.16 shows the profile likelihood for the 100-year return level. The curve is greatly asymmetric and the upper bound of 608.05 in the confidence interval is very high, providing evidence of the great amount of uncertainty that affects the model. Note that the normal approximation for the 100-year return level confidence interval, $[204.03, 453.52]$, is not able to reflect the previous arguments, that is why it is better to use the profile likelihood whenever an accurate measure of uncertainty is needed.

5.4 Bivariate analysis

In this section, the results for bivariate extremes, discussed in the Chapter 2, are applied to data sets consisting of componentwise rainfalls observations. Before model fitting, empirical estimates of the dependence functions $\chi(u)$ and $\overline{\chi(u)}$ are constructed and their behaviour is analyzed. After that the componentwise block maxima is exploit to fit asymptotic distributions.

5.4.1 Componentwise Block maxima approach

Asymptotic dependence

Figure 5.17 shows the annual maximum rainfalls recorded at Isoverde and Mignanego stations during the observation period (1960-1998 and 2003-2014). There is evidence for large values at one location to correspond to the maxima of the other.

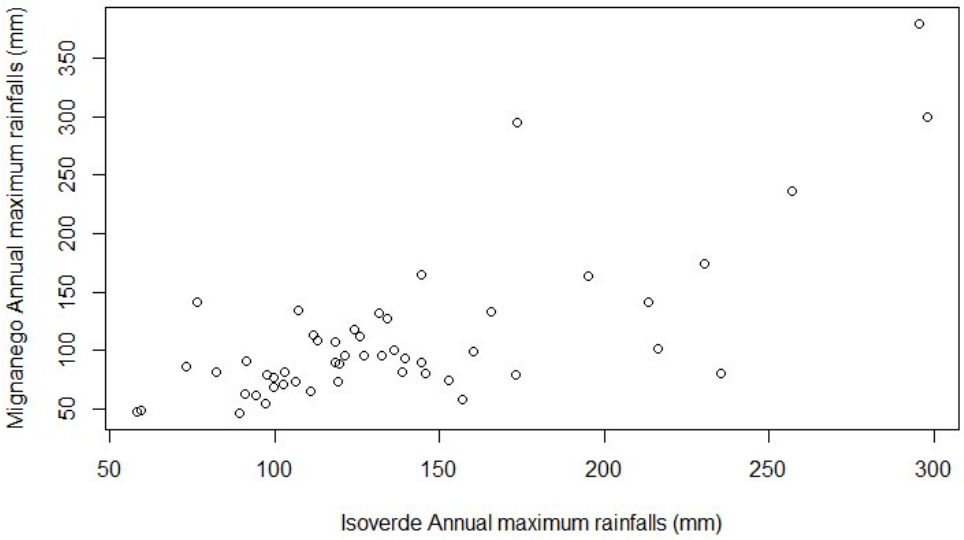

Figure 5.17: Annual maximum rainfalls recorded at Isoverde and Mignanego stations

The empirical estimates for $\chi(u)$ and $\overline{\chi(u)}$ shown in Figure 5.18 are obtained using the function `chiplot` in package `evd`.

The value $\overline{\chi} = 1$ as limit of the $\overline{\chi(u)}$ is plausible, indicating that the distributions of extremes are asymptotically dependent. In this contest the value of χ can be used as measure of dependence strength. According to the solid line in the plot $\chi(u) > 0\ \forall u$ and as $u \to 1$ also $\chi(u) \to 1$. However the wide confidence bands and the unstable behaviour of the graph don't allow to draw reliable conclusions about the value of the measure. This condition is typical of componentwise block maxima : sometimes data are insufficient or too sparse and this makes it difficult to assess the validity of model choice.

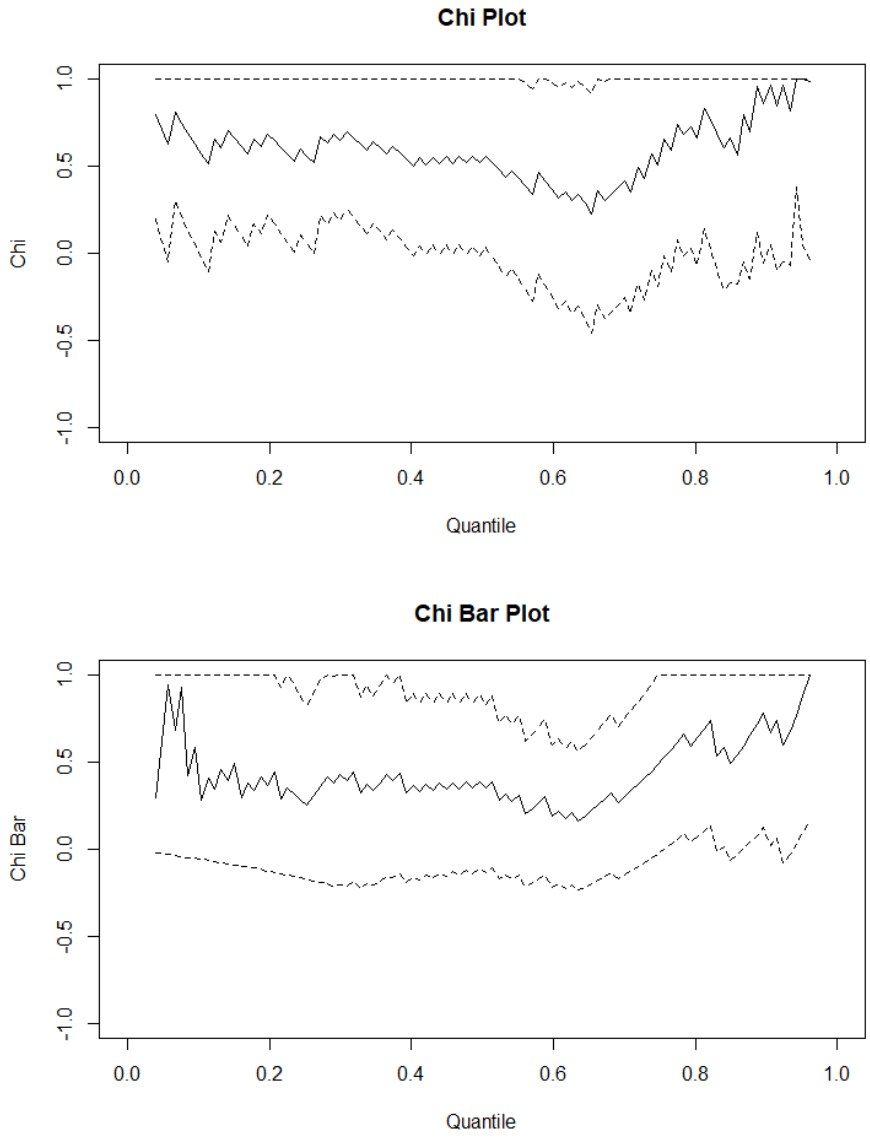

Figure 5.18: Empirical estimates for $\chi(u)$ and $\overline{\chi(u)}$ with approximate 95% confidence intervals

Model fitting

In the light of the considerations arising from the `chiplot`, a bivariate extreme value distribution would like to be fitted. Let recall the expression for the logistic family seen in Section 2.2.2.

$$G(x,y) = \exp\left\{-\left(x^{-1/\alpha} + y^{-1/\alpha}\right)^{\alpha}\right\}, \quad x > 0, y > 0. \tag{5.1}$$

The previous model is chosen, due to the quite symmetrical distribution of the plotted points in Figure 5.19, which have been re-scaled in order to have uniform marginal distributions.

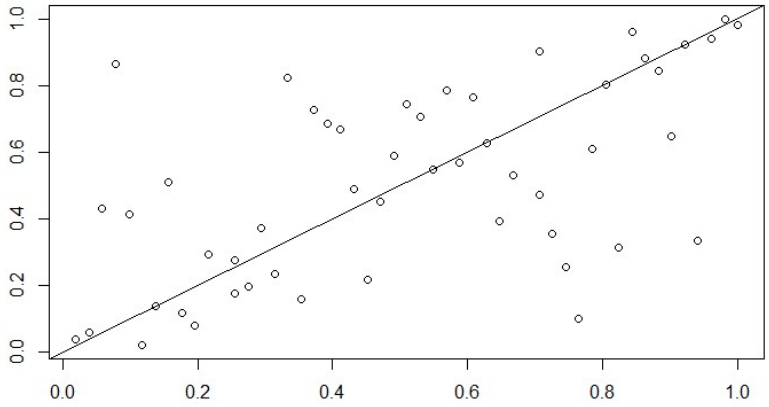

Figure 5.19: Componentwise maxima transformed to have uniform marginals distributions

The maximum likelihood estimation is performed using the function `fbvevd` in package `evd` and produces the following results:

	Isoverde			Mignanego			
	μ_x	σ_x	ξ_x	μ_y	σ_y	ξ_y	α
MLE	112.1576	36.8001	0.09778	81.2629	29.0419	0.3208	0.5259

The previous outcomes are obtained through a two-stage estimation procedure: first the two series are separately modeled using the GEV distribution, then the bivariate model is produced by maximum likelihood estimation, after transformation to standard Fréchet margins.
The estimate for the logistic parameter $\hat{\alpha} = 0.5259$ corresponds to a quite significant level of dependence and this is consistent with the previous empirical analysis. In fact, by replacing the value of $\hat{\alpha}$ in the formula $\chi = 2 - 2^{\alpha}$ the estimate $\hat{\chi} = 0.56013$ is obtained, very close to the mean value assumed in the `chiplot`, providing support for model validity.

5.5 Comparing models on different time periods

Let consider again the series of annual maxima analyzed in Section 5.3.1. This time the observations are equally divided into two groups and the related scatterplots are presented in following pictures.

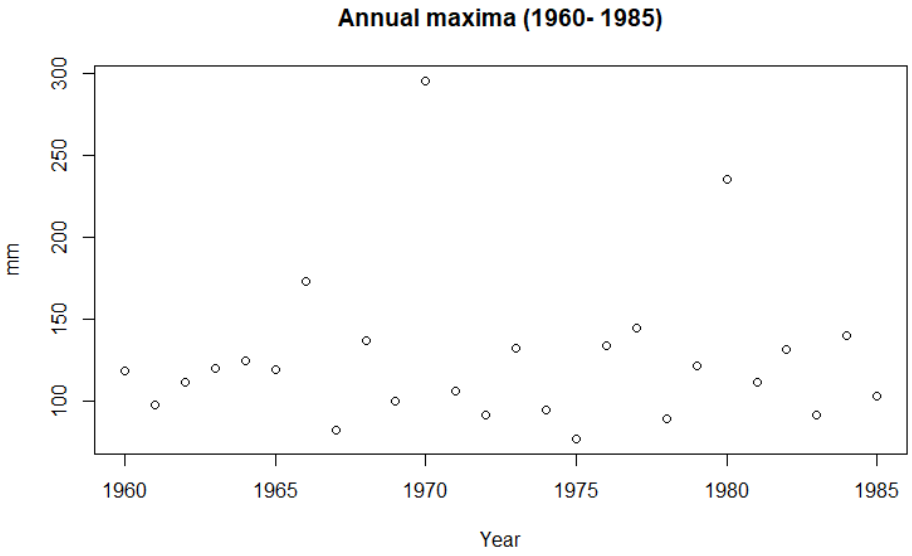

Figure 5.20: Annual maxima recorded between 1960 and 1985

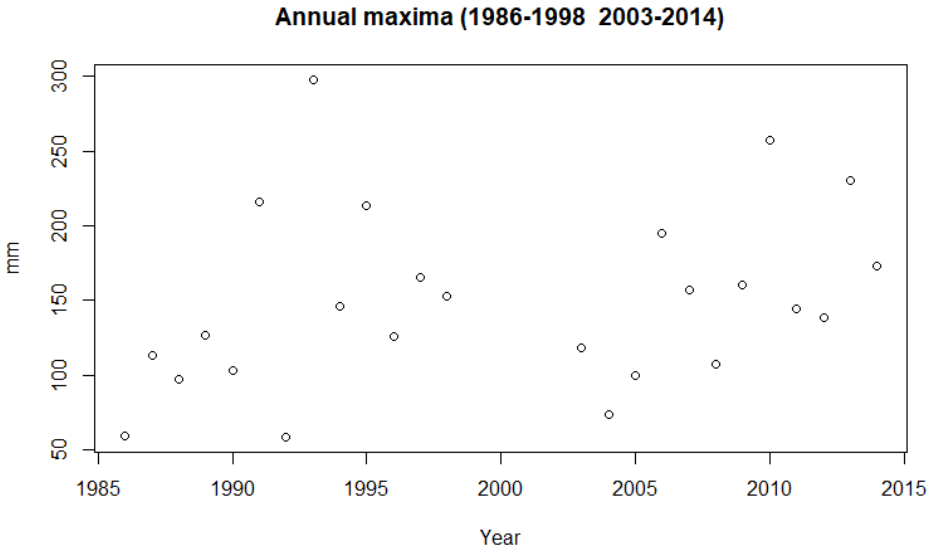

Figure 5.21: Annual maxima recorded between 1986 and 2014

From the previous pictures it is possible to deduce that the behaviour of extremes rainfalls has changed throughout the years; in particular the amount of rainfall observed during the intense events seems to have increased between the first and the second period.
In the light of these considerations, the block maxima approach is applied to each sub-sample and the fitted GEV are compared in order to point out the evolution in the extremal behaviour.
Firstly, the years going from 1960 to 1985 are considered. Maximum likelihood estimation on this data provides the following results:

Parameter	95% lower CI	Estimate	95% upper CI
$\hat{\mu}_1$	94.8936	104.9569	115.0201
$\hat{\sigma}_1$	14.8844	23.1435	31.4025
$\hat{\xi}_1$	-0.0604	0.2589	0.5782

The positive estimate for the shape parameter and the confidence interval, which lies almost exclusively in the positive domain, suggest an unbounded distribution. This hypothesis is reinforced by the profile likelihood confidence interval for ξ_1, whose lower-end limit lies above zero, as shown in Figure 5.22.

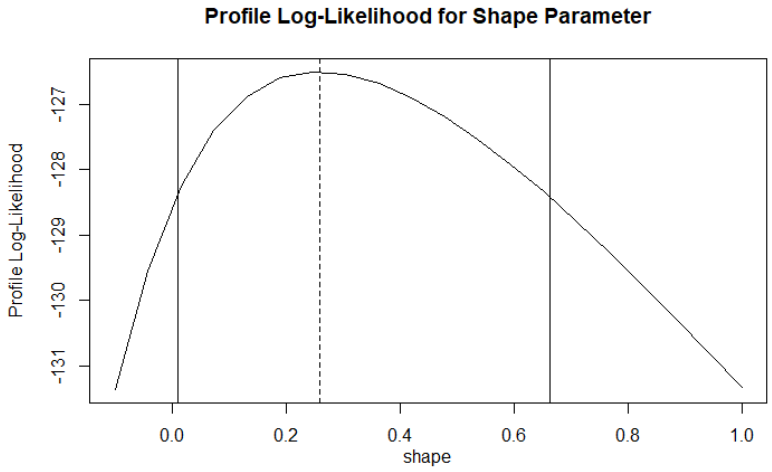

Figure 5.22: Profile likelihood for ξ with 95% confident interval

However, a likelihood ratio test can be performed to substantiate the earlier arguments. Considering a level of significance $\alpha = 0.05$, the statistic D = 4.1156 exceeds the Chi-square critical value C_α= 3.8415, leading to reject the null hypothesis of reduction to the Gumbel family, in favour of the GEV distribution.

The analysis is then repeated on the observations collected in the period going from 1986 to 1999 and from 2003 to 2014. The maximum likelihood estimates for the three parameters are shown in the following table.

Parameter	95% lower CI	Estimate	95% upper CI
$\hat{\mu}_2$	100.9297	122.9018	144.8739
$\hat{\sigma}_2$	33.2917	49.1178	64.9438
$\hat{\xi}_2$	-0.3751	-0.0586	0.2579

The values of $\hat{\mu}_2$ and $\hat{\sigma}_2$ are a little greater than those obtained for the previous time period, while the estimate for ξ_2 results negative, indicating a bounded distribution. However, since 0 lies inside the confidence interval, the evidence for the previous hypothesis is not strong. In order to make inference on the shape parameter, the Gumbel distribution is fitted to the data and the nested models are compared using the AIC/BIC criteria and the likelihood ratio test.

	AIC	BIC
GEV	278.2964	281.9530
Gumbel	276.4335	278.8713

D	C_α	α	p-value
0.13712	3.8415	0.05	0.7112

Both methods suggest to prefer the Gumbel family: the AIC and BIC criteria are smaller for the Gumbel distribution and the high p-value in the test makes it impossible to reject the null hypothesis of reduction to the nested model.
The validity of the fitted distribution is confirmed by the diagnostic plots.

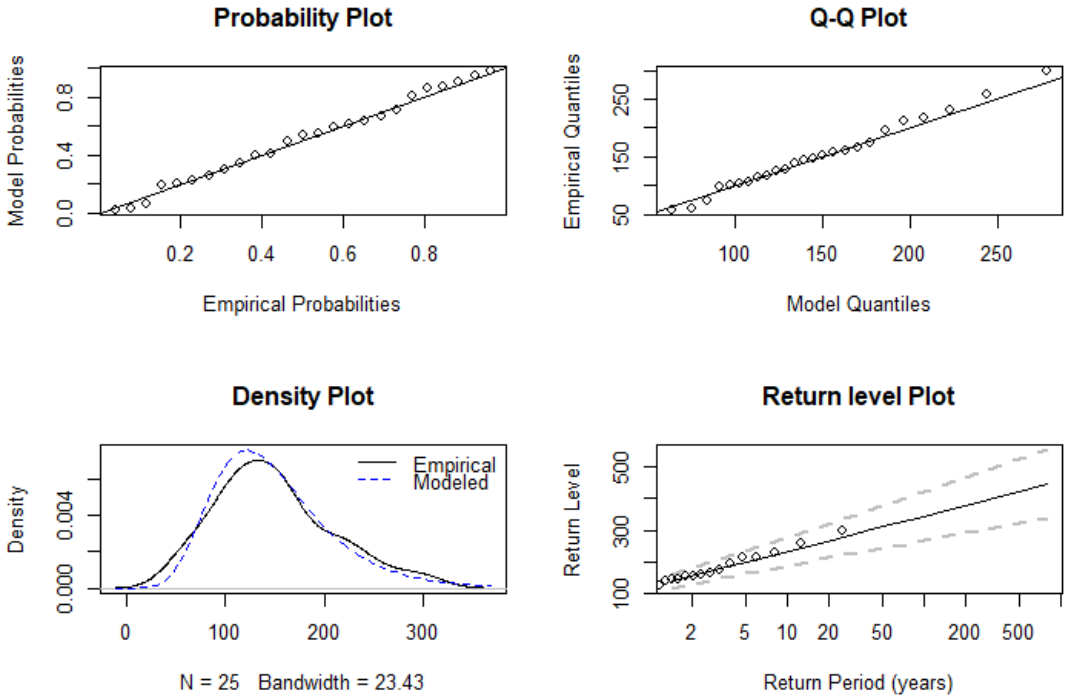

Figure 5.23: Diagnostic plot for Gumbel fit

In the probability and quantile plots, points lie close to the unit diagonal, indicating that the estimated model provides a good approximation of data, even in correspondence of the distribution tails. The reliability of the fitted distribution is also suggested by the return level plot : in fact, the return level estimates are in agreement with the modeled curve and the confidence bands remain narrow, even for long return periods.
The previous analysis have highlighted considerable differences in terms of extremal distributions between the two time periods. However, the most interesting difference arises from the computation of return levels.
Let consider, for each period, the most reliable model (i.e. GEV($\hat{\mu}_1, \hat{\sigma}_1, \hat{\xi}_1$) and Gumbel($\hat{\mu}_2, \hat{\sigma}_2$)) the following estimates are obtained:

First period (1960-1985)

	95% lower CI	Estimate	95% upper CI
10-year return level	134.9524	175.6437	216.3350
20-year return level	140.3010	208.4394	276.5779
30-year return level	139.8138	230.2682	320.7225
50-year return level	134.3987	261.0618	387.7248
100-year return level	116.1613	309.7075	503.2536

Second period (1986-2013)

	95% lower CI	Estimate	95% upper CI
10-year return level	186.7455	230.6722	274.5988
20-year return level	211.8109	265.5249	319.2388
30-year return level	226.0985	285.5748	345.0510
50-year return level	243.8698	310.6381	377.4065
100-year return level	267.7303	344.4441	421.1579

The return level derived from the GEV distribution fitted on the period going from 1960 to 1985 tend to be lower than those obtained by extrapolation on the subsequent years. This can be traced to the fact that, as already pointed out, the rainfalls observed in the first period were less abundant than those in the second one.
However, it is important to notice that, as a consequence of the changing in extreme rainfalls behavior, the return level estimates are different depending on whether the model is fitted using the entire set of annual maxima or just a sub-sample. In particular, the return levels presented in Section 4.3.1 tend to be halfway between those computed on separated periods. As an example, let consider the 10-year return level : it results to be $z \approx 205.5$, when the GEV is estimated using all annual maxima, while it becomes $z \approx 230.7$ if only the years from 1986 to 2013 are considered.
Therefore, fitting the model on the overall period, without taking into account the variation in the weather conditions, could lead to and underestimation of return levels.

All the the previous arguments suggest that the behaviour of intense rainfalls has evolved over the years. However, it would be interesting to understand if

the two modeled distributions could be considered statistically different or not. Figures 5.24 and 5.25 show the comparisons between modeled densities and probability distributions. The density curve, corresponding to the period going from 1986 to 2014, is smoother than the other one and a little bit translated. These differences are due to the location and scale parameters that, as pointed out in the previous analysis, are bigger for the second sub-sample.
In Figure 5.25, it is possible to notice that the two CDFs are not overlapping and in correspondence of the value $x = 100$ their order is reversed.

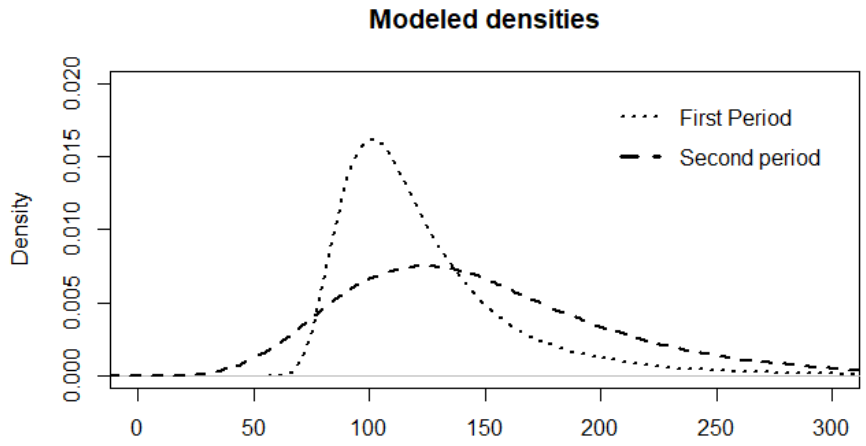

Figure 5.24: Comparison between model densities

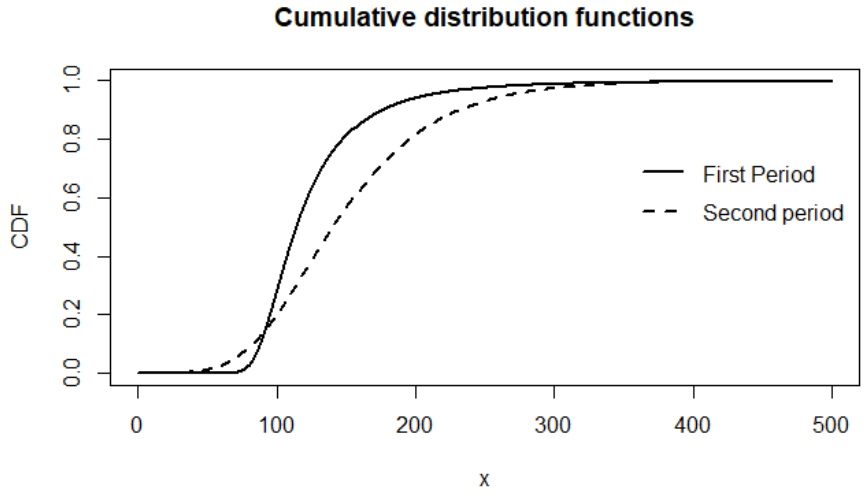

Figure 5.25: Comparison between cumulative distribution functions

In order to assess if the difference between the two distributions is statistically significant, a two-sample Kolmogorv Smirnov test can be performed, using the function `ks.test` in package `stats`. The two alternatives `"greater"` and `"less"` corresponds to the null hypothesis that the true distribution function of the first sample is respectively not greater than, or not less than the distribution of the second one.

```
> stats::ks.test(maxima1,maxima2, alternative="greater")
```

D^+	p-value
0.36769	0.03187

```
> stats::ks.test(maxima1,maxima2, alternative="less")
```

D^-	p-value
0.12	0.6928

Hence, considering a significance level $\alpha = 0.1$ it is not possible to reject the hypothesis that the distribution of the data in the first time period is stochastically bigger than the other one.

Chapter 6

Conclusion

The main purpose of this thesis has been that of exploiting the statistical procedures deriving from extreme value theory in order to model the behavior of extreme rainfalls. In particular, it was decided to study the observations collected in the province of Genoa, since in the last fifty years the area has been hit by intense meteorological phenomena.
The candidate meteorological stations for the study had to meet some specific requirements in terms of missing values and length of the activity period, in order to ensure data consistency.
Firstly, the series of annual maxima has been analyzed through block maxima approach and the Gumbel distribution has resulted to be the more suitable model in terms of accuracy for both return levels and parameter estimates.
After that, since the entire series of measurements was available, the others characterizations of extremes have been exploited in order to achieve greater precision. However, the extrapolation based on a greater number of observations has led to very similar conclusions, adding no further improvement to the analysis.
The study carried out on the pairs of observations collected from two different rain gauges has pointed out an extremal dependence between processes, suggesting that a bivariate extreme value distribution, such as the logistic model, would be suitable for the those data.
Anyway, the most interesting results derive from the analysis realized by comparing the annual maxima from two different time periods. It is possible to affirm that the behaviour of intense rainfalls has evolved over the years. In particular, since 1986, the annual maxima have increased, leading to even more frequent and intense precipitations. Such meteorological changes can result in misleading estimates, especially for what concerns return levels, in fact the failure to take into account the variations in weather conditions has led, in this case, to underestimated values.
Extreme value analysis certainly represents the more suitable technique able to predict the occurrence of extreme events. However, as it tries to make inference outside of the range of available data, a critic view is always required when applying such methodology to real life examples.

Appendix A

Code

The following code has been used to produce all the figures and results presented in the previous chapters.

```
#LIBRARIES
library(readxl)
library(evd)
library(evir)
library(extRemes)
library(lmomco)
library(truncgof)
library(POT)
library(ismev)
library(stats)

#Plotting densities for Extreme value distributions
n=10000
gumbel<-rgumbel(n, loc=0, scale=1)
frechet<- rfrechet(n, loc=0, scale=1, shape=4)
weibull<- rrweibull(n, loc=0, scale=1, shape=2)
plot(density(frechet),main="",xlab="",xlim=c(-4,5),ylim=c
    (0,2))
lines(density(gumbel),lwd=2,lty=3)
lines(density(weibull),lwd=2,lty=2)
legend(2.5,1.8,legend=c("Gumbel","Frechet","Weibull"),lty=c
    (3,1,2),lwd=c(2,1,2),title="Density",y.intersp=1.5,bty="n
    ")

# BLOCK MAXIMA APPROACH
#Data visualization
plot(annual_maxima, xlab="Year")
title("Annual maxima")

#Creating a numeric array containing annual maximum
    observations
maxima<-annual_maxima[,2]
maxima=as.matrix(maxima)
```

```
summary(maxima)

#Histogram and density for annual maximum observations
hist(maxima, main=('Frequency distribution'))
par(mfrow=c(1,1))
plot(density(maxima,adjust=2),main=('Density distribution'))

#Fitting GEV to block maxima
GEV3<-extRemes::fevd(maxima, type='GEV')
summary(GEV3)

#95% confidence intervals for GEV parameters
norm_cint<-ci(GEV3, alpha=0.05, type="parameter")
norm_cint

#Profile likelihood curve for shape parameter with
   confidence intervals
prof_cint<-ci(GEV3,alpha=0.05, type="parameter",which.par=3,
    method='proflik',xrange=c(-0.13, 0.4))
prof_cint
par(mfrow=c(1,1))
profliker(GEV3, type="parameter",which.par=3,xrange=c(-0.13,
    0.4),main="Profile Log-Likelihood for Shape Parameter")
abline(v=prof_cint[1])
abline(v=prof_cint[3])
abline(v=norm_cint[3,1], col="blue")
abline(v=norm_cint[3,3], col="blue")

#Diagnostic plots
par(mfrow=c(2,2))
plot(GEV3, type=c("probprob"), main="Probability Plot")
plot(GEV3, type=c("qq"), main="Q-Q Plot")
plot(GEV3, type=c("density"), main="Density Plot")
plot(GEV3, type=c("rl"), main="Return level Plot")

#Return levels with 95% confident intervals
ret.lev<-return.level(GEV3,return.period=c(10,20,30,50,100),
   do.ci=TRUE)
ret.lev

#Profile likelihood for the 10-year return-level with
   confidence intervals
par(mfrow=c(1,1))
profliker(GEV3, type="return.level", return.period=10,xrange
   =c(160,370), main="10-year return level profile
   likelihood")
ci_rl<-ci(GEV3, alpha=0.05, type="return.level",return.
   period=10,method="proflik", xrange=c(160,370))
abline(v=ret.lev[1,1], col='blue')
abline(v=ret.lev[1,3], col='blue')
abline(v=ci_rl[1])
```

```
abline(v=ci_rl[3])

#Profile likelihood for the 50-year return-level with
   confidence intervals
par(mfrow=c(1,1))
profliker(GEV3, type="return.level", return.period=50,xrange
   =c(200,700), main="50-year return level profile
   likelihood")
ci_rl<-ci(GEV3, alpha=0.05, type="return.level",return.
   period=50,method="proflik", xrange=c(200,700))
abline(v=ret.lev[4,1], col='blue')
abline(v=ret.lev[4,3], col='blue')
abline(v=ci_rl[1])
abline(v=ci_rl[3])

#Fitting Gumbel distribution
GUMBEL<-extRemes::fevd(maxima, type="Gumbel")
summary(GUMBEL)

#95% confidence intervals for Gumbel parameters
norm_cint2<-ci(GUMBEL, alpha=0.05, type="parameter")
norm_cint2

#Return levels with 95% confidence intervals
ret.lev2<-return.level(GUMBEL, return.period=c
   (10,20,30,50,100),do.ci=TRUE)
ret.lev2

#Diagnostic plots
par(mfrow=c(2,2))
plot(GUMBEL, type=c("probprob"), main="Probability Plot")
plot(GUMBEL, type=c("qq"), main="Q-Q Plot")
plot(GUMBEL, type=c("density"), main="Density Plot")
plot(GUMBEL, type=c("rl"), main="Return level Plot")

#Likelihood ratio test
lr.test(GUMBEL, GEV3, alpha = 0.05)

# r-LARGEST ORDER STATISTIC APPROACH
#Data matrix creation
#Each row is a vector of decreasing order, containing the
   largest order statistics for each year
year<-annual_maxima$Anno
A=matrix(0, nrow=length(year), ncol=366)
j=1
for(i in c(1:length(year))){
  tmp=data_rain$mm[data_rain$Anno==year[i]]
  tmp=sort(tmp, decreasing = TRUE)
  tmp=t(tmp)
  A[j, 1:length(tmp)]=tmp
  j=j+1
```

```
}

#Fitting the model for different values of r
k=10
MLE=matrix(0, nrow=k, ncol=3)
SE=matrix(0, nrow=k,ncol=3)
NLLH=array(0, dim=k)
for(i in c(1:k)){
  RLOS=rlarg.fit(A, r = i, show=FALSE)
  MLE[i,1:3]=RLOS$mle
  SE[i,1:3]=RLOS$se
  NLLH[i]=RLOS$nllh
}

#Plotting SE pattern for the three parameters
plot(SE[,1], type='l',ylab='SE', xlab='r value',main="
   Standard error for location parameter", ylim=c(0,6.5))

plot(SE[,2], type='l',ylab='SE', xlab='r value',main="
   Standard error for scale parameter", ylim=c(0,5))

plot(SE[,3], type='l',ylab='SE', xlab='r value',main="
   Standard error for shape parameter", ylim=c(0,0.12))

#Plotting MLE pattern for the three parameters
plot(MLE[,1], type='l',ylab='MLE', xlab='r value',main="
   Maximum likelihood estimation for location parameter")

plot(MLE[,2], type='l',ylab='MLE', xlab='r value',main="
   Maximum likelihood estimation for scale parameter")

plot(MLE[,3], type='l',ylab='MLE', xlab='r value',main="
   Maximum likelihood estimation for shape parameter")


#Fitting the model for the selected value of r
RLOS_BEST=rlarg.fit(A, r=3, show=FALSE)
RLOS_BEST$mle
RLOS_BEST$se

#Diagnostic plots
rlarg.diag(RLOS_BEST)

# PEACKS OVER THRESHOLD APPROACH
#Dataset of daily rainfall data
rain_series<-data_rain[1:dim(data_rain)[1] , 1:2]
plot(rain_series[1:3650,])
rain_mm<-data_rain$mm
rain_mm<-as.matrix(rain_mm)
```

```
#THRESHOLD SELECTION
#1- Mean residual life plot
extRemes::mrlplot(rain_mm)
abline(v=174)
abline(v=100)
title('Mean residual life plot')

#Number of excesses
soglia=100
count=0;
for(i in c(1:length(rain_mm))){
  if (rain_mm[i]>=soglia){
    count=count+1
  }
}

#2 - Fitting the GPD Model Over a Range of Thresholds
gpd.fitrange(rain_mm, 0, 150, 75)

#Plotting excesses
excess<-rain_series[which(rain_mm>=soglia),]
ind=which(rain_mm>=soglia)
n=365*30
plot(rain_series[1:n,])
points(excess[which(ind<n),],pch=16, col='blue')
plot(excess, pch=16, col='black')

#Empirical density of excesses
par(mfrow=c(1,1))
plot(density(rain_mm[ind], adjust=2), main=('Density
   distribution'))

#Fitting GPD model on excess dataset:
GPD3<-extRemes::fevd(rain_mm, threshold=100, type="GP")
summary(GPD3)

#95% confidence intervals for GPD parameters
norm_cint3<-ci(GPD3, alpha=0.05, type="parameter")
norm_cint3

#Profile likelihood for shape parameter with confidence
   intervals
prof_cint2<-ci(GPD3, alpha=0.05, type="parameter", which.par
   =2, method='proflik', xrange=c(-0.1, 0.6))
par(mfrow=c(1,1))
profliker(GPD3, type="parameter", which.par=2,xrange=c
   (-0.13, 0.6),main="Profile Log-Likelihood for Shape
   Parameter")
abline(v=prof_cint2[1])
abline(v=prof_cint2[3])
abline(v=norm_cint3[2,1], col="blue")
```

```
abline(v=norm_cint3[2,3], col="blue")

#Diagnostic plots
par(mfrow=c(2,2))
plot(GPD3, type=c("probprob"), main="Probability Plot")
plot(GPD3, type=c("qq"), main="Q-Q Plot")
plot(GPD3, type=c("density"), main="Density Plot")
plot(GPD3, type=c("rl"), main="Return level Plot")

#Return levels with 95% confidence intervals
ret.lev3<-return.level(GPD3, return.period=c
   (10,20,30,50,100),do.ci=TRUE)
ret.lev3

#Profile likelihood for 100-year return level with confident
    intervals
prof_cint3<-ci(GPD3, alpha=0.05, type="return.level", return
   .period=100, method="proflik",xrange=c(240,800),verbose=
   TRUE)
profliker(GPD3, type="return.level", return.period=100, main
   ='Profile Log-Likelihood', xrange=c(240,800))
abline(v=prof_cint3[1])
abline(v=prof_cint3[3])

# ANALYSIS ON SEPARATED PERIODS
# BLOCK MAXIMA APPROACH
#Creating two dataset from the database maxima
maxima1<-maxima[1:26]
maxima2<-maxima[27:length(maxima)]
#Plotting two series of maxima
par(mfrow=c(1,1))
plot(annual_maxima[1:26,], xlab="Year")
title("Annual maxima (1960- 1985)")
plot(annual_maxima[27:length(maxima),],xlab="Year")
title("Annual maxima (1986-1998  2003-2014)")

#1960 - 1985 - Fitting GEV distribution
GEV_m1<-extRemes::fevd(maxima1, type='GEV')
summary(GEV_m1)

#95% confidence intervals for GEV parameters
cint_norm1<-ci(GEV_m1, alpha=0.05, type="parameter")
cint_norm1

#Profile likelihood for shape parameter with confidence
   intervals
cint_m1<-ci(GEV_m1, alpha=0.05, type="parameter",which.par
   =3, method='proflik', xrange=c(-0.1, 1))
cint_m1
par(mfrow=c(1,1))
```

```
profliker(GEV_m1, type="parameter", which.par=3, xrange=c
   (-0.1, 1),main="Profile Log-Likelihood for Shape
   Parameter")
abline(v=cint_m1[1])
abline(v=cint_m1[3])
abline(v=cint_norm1[3,1], col="blue")
abline(v=cint_norm1[3,3], col="blue")

#Diagnostic plots
par(mfrow=c(2,2))
plot(GEV_m1, type=c("probprob"), main="Probability Plot")
plot(GEV_m1, type=c("qq"), main="Q-Q Plot")
plot(GEV_m1, type=c("density"), main="Density Plot" )
plot(GEV_m1, type=c("rl"), main="Return level Plot")

#Return levels with 95% confidence intervals
ret.lev.m1<-return.level(GEV_m1, return.period=c
   (10,20,30,50,100),do.ci=TRUE)
ret.lev.m1

#Fitting Gumbel distribution
GUMBEL_m1<-extRemes::fevd(maxima1, type="Gumbel")
summary(GUMBEL_m1)

#Diagnostic plots
par(mfrow=c(2,2))
plot(GUMBEL_m1, type=c("probprob"), main="Probability Plot")
plot(GUMBEL_m1, type=c("qq"), main="Q-Q Plot")
plot(GUMBEL_m1, type=c("density"), main="Density Plot")
plot(GUMBEL_m1, type=c("rl"), main="Return level Plot")

#Likelihood ratio test
lr.test(GUMBEL_m1, GEV_m1, alpha = 0.05)

#1986 - 2014 - Fitting GEV distribution
GEV_m2<-extRemes::fevd(maxima2, type='GEV')
summary(GEV_m2)

#95% confidence intervals for GEV parameters
cint_norm2<-ci(GEV_m2, alpha=0.05, type="parameter")
cint_norm2

#Profile likelihood curve for shape parameter with
   confidence intervals
cint_m2<-ci(GEV_m2, alpha=0.05, type="parameter",which.par
   =3, method='proflik',xrange=c(-0.28, 0.35))
cint_m2
par(mfrow=c(1,1))
profliker(GEV_m2, type="parameter", which.par=3,xrange=c
   (-0.28, 0.35),main="Profile Log-Likelihood for Shape
   Parameter")
```

```
abline(v=cint_m2[1])
abline(v=cint_m2[3])
abline(v=cint_norm2[3,1], col="blue")
abline(v=cint_norm2[3,3], col="blue")

#Diagnostic plots
par(mfrow=c(2,2))
plot(GEV_m2, type=c("probprob"), main="Probability Plot")
plot(GEV_m2, type=c("qq"), main="Q-Q Plot")
plot(GEV_m2, type=c("density"), main="Density Plot" )
plot(GEV_m2, type=c("rl"), main="Return level Plot")

#Return levels with 95% confidence intervals
ret.lev.m2<-return.level(GEV_m2,return.period=c
   (10,20,30,50,100),do.ci=TRUE)
ret.lev.m2

#Fitting Gumbel distribution
GUMBEL_m2<-extRemes::fevd(maxima2, type="Gumbel")
summary(GUMBEL_m2)

#Likelihood ratio test
lr.test(GUMBEL_m2, GEV_m2, alpha = 0.05)

#Diagnostic plots
par(mfrow=c(2,2))
plot(GUMBEL_m2, type=c("probprob"), main="Probability Plot")
plot(GUMBEL_m2, type=c("qq"), main="Q-Q Plot")
plot(GUMBEL_m2, type=c("density"), main="Density Plot" )
plot(GUMBEL_m2, type=c("rl"), main="Return level Plot")

#Return levels with 95% confidence intervals
ret.lev.m2g<-return.level(GUMBEL_m2,return.period=c
   (10,20,30,50,100),do.ci=TRUE)
ret.lev.m2g

#Plotting empirical densities
plot(density(maxima1,adjust=2),main="",xlab="")
lines(density(maxima2, adjust=2),lwd=2,lty=3)
legend(200,0.01,legend=c("First Period","Second period "),
   lty=c(1,3),lwd=c(1,2),y.intersp=1.5,bty="n")

#Extracting model parameters
# GEV (1960-1985)
loc1=as.numeric(GEV_m1$results$par[1])
scale1=as.numeric(GEV_m1$results$par[2])
shape1=as.numeric(GEV_m1$results$par[3])
# GUMBEL (1986-2013)
loc2=as.numeric(GUMBEL_m2$results$par[1])
scale2=as.numeric(GUMBEL_m2$results$par[2])
```

```
#Plotting modeled densities
n=100000
gumbel=revd(n, loc2, scale2, 0)
gev=revd(n, loc1, scale1, shape1)
plot(density(gev), xlim=c(0,300), ylim=c(0,0.02), xlab="",
   lwd=2,lty=3, main="Modeled densities")
lines(density(gumbel),lwd=2,lty=2)
legend(200,0.02,legend=c("First Period","Second period "),
   lty=c(3,2),lwd=c(2,3),y.intersp=1.5,bty="n")

#Plotting CDF
seq=c(0:500)
d1=pevd(seq, loc1, scale1, shape1, type="GEV")
d2=pevd(seq, loc2, scale2, 0, type="Gumbel")
plot(seq, d1, type="l", xlab="x",ylab= "CDF", main="
   Cumulative distribution functions", lwd=2, lty=1)
lines(d2, lwd=2, lty=2)
legend(350,0.8,legend=c("First Period","Second period "),lty
   =c(1,2),lwd=c(2,2),y.intersp=1.5,bty="n")

#Two sample Kolmogorov-Smirnov test
stats::ks.test(maxima1, maxima2, alternative="greater")
stats::ks.test(maxima1, maxima2, alternative="less")

#BIVARIATE ANALYSIS
#Creating a matrix containing bivariate rainfall
   observations
bi_max<-as.matrix(bivariate_maxima[,2:3])

#EXTREMAL DEPENDENCE
#Bivariate plot
plot(bi_max[,1], bi_max[,2],
     xlab="Isoverde Annual maximum rainfalls (mm)",
     ylab="Mignanego Annual maximum rainfalls (mm)")

#Plotting estimates of the dependence measures chi and chi-
   bar
par(mfrow=c(1,2))
chi<-chiplot(bi_max)

#Bivariate plot - Uniform scale
library(tiger)
b1<-to.uniform(bi_max[,1])
b2<-to.uniform(bi_max[,2])
par(mfrow=c(1,1))
plot(b1,b2)
abline(0,1)

#Maximum-likelihood Fitting for Bivariate Extreme Value
   Distributions
BEVD<-fbvevd(bi_max,model="log")
```

```
BEVD$estimate # dep = logistic dependence parameter
BEVD$std.err
BEVD$dep.summary #Maximum likelihood estimate for chi
```

Bibliography

[1] Stuart G. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer, London, (2001).

[2] Stuart G. Coles, Janet Heffernan, and Jonathan A. Tawn. Dependence Measures for Extreme Value Analyses . *Extremes*, 2:339–365, December 1999.

[3] Jan Beirlant, Yuri Goegebeur, Johan Segers, and Jozef Teugels. *Statistics of Extremes : Theory and Applications*. Wiley, New York, (2004).

[4] Juan Juan Cai, Anne-Laure Fougères, and Cécile Mercadier. Environmental data: multivariate Extreme Value Theory in practice . *Journal de la Société Française de Statistique*, 154(2), 2013.

[5] Stuart G. Coles and Jonathan A. Tawn. Modelling extreme multivariate events . *Journal of the Royal Statistical Society*, 53(2):377–392, 1991.

[6] Miguel de Carvalho and Alexandra Ramos. Bivariate Extreme Statistics, II. *REVSTAT- Statistical Journal*, 10(1):83–107, March 2012.

[7] Laurence de Haan and Ana Ferreira. *Extreme Value Theory: An Introduction*. Springer, New York, (2006).

[8] Eric Gilleland and Richard W. Katz. extRemes 2.0: An Extreme Value Analysis Package in R. *Journal of Statistical Software*, 72, 2016.

[9] Emil Julius Gumbel. *Statistics of Extremes*. Dover Publications, Mineola, New York, (2004).

[10] Amir Khorrami. Extreme Value Analysis of Severe Rainfalls in the Piemonte and Valle d'Aosta Regions. Master's thesis, March 2014.

[11] Samuel Kotz and Saralees Nadarajah. *Extreme Value Distributions: Theory and Applications*. Imperial College Press, London, (2000).

[12] G. MacDonald, Steven Koonin, Herbert Levine, and H. Abarbanel. Statistics of Extremes Events with Applications to Climate, January 1992.

[13] Rolf-Dieter Reiss and Michael Thomas. *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields*. Birkhäuser Basel, Berlin, (2007).

[14] Jonathan A. Tawn. Bivariate extreme value theory: models and estimation. *Biometrika*, 75:165–195, 1988.

[15] J. Pickands. Statistical inference using extreme order statistics. *Annals of Statistics*, 3:119–131, 1975.

[16] Eric Gilleland, Mathieu Ribatet, and Alec G. Stephenson. A software review for extreme value analysis. *Extremes*, 16:103–119, 2013.

[17] Wikipedia contributors. Extreme value theory — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Extreme_value_theory&oldid=902554360`, 2019. [Online; accessed 2-July-2019].

[18] Christophe Dutang and Kevin Jaunatre. CRAN Task View: Extreme Value Analysis. `https://CRAN.R-project.org/view=ExtremeValue`.

[19] Arpal. Ligurian weather and climate database. `http://www.banchedati.ambienteinliguria.it/index.php/aria/meteo`.