Università degli Studi di Cagliari

DOTTORATO DI RICERCA
Dipartimento di Ingegneria del Territorio
Ciclo XXIII

TITOLO TESI

Extreme Events in Hydrology:

an approach using Exploratory Statistics and the Generalized Pareto Distribution. Performances and properties of the GPD estimators with outliers and rounded-off datasets.

Settore disciplinare di afferenza: ICAR/02 Costruzioni idrauliche e marittime e idrologia

Presentata da: Dott. Michelangelo Puliga
Coordinatore Dottorato: Prof. Ing. Giorgio Querzoli
Tutor: Prof. Ing. Roberto Deidda

Esame finale anno accademico 2009 - 2010

"Mediocristan is where we must endure the tyranny of the collective, the routine, the obvious and the predicted. Extremistan is where we are subjected to the tyranny of the singular, the accidentally, the unseen and the unpredicted. "

Nicholas Nassim Taleb.
The Black Swan: The Impact of the Highly Improbable
(April 2007)

Abstract

Two large databases of daily cumulated rainfall are checked with the tools of the Exploratory statistics. The analysis allows to discover not common artefacts in the first database (rounding-off of data with different rounding-off rules) and several errors in the other one. The best statistical model to fit data is selected using the L-Moments ratio diagram as a tool to explore the accommodation of each dataset to other alternative models. This tool suggests the Generalized Pareto Distribution as the best statistical model for this data, but the application of this distribution requires an estimate of the optimal threshold for each dataset. A detailed analysis of the present techniques for the optimal threshold selection is performed and a new approach based on quantile sums is proposed. Furthermore the performances of the GPD parameters estimators are checked for robustness against spurious rounded-off data and severe outliers.

Acknowledgements

I first thank prof. Roberto Deidda for the opportunity of this doctoral research, for the constant help in learning a critical approach toward statistics and science. Never trust too much models, data and techniques without an extensive check of each object. I want to thank Dr. Valerio Lucarini (University of Reading) for an introductory part on climate models and data assessment,and for his precious creative and open mind approach to science. Finally I thank my girlfriend Silvia for her support during the work like a bright star in a deep sky.

Contents

List of Figures

List of Tables

Abbreviations

pdf	probability density function
thrprc	percentage of discarded data at the optimal threshold
CDF	Cumulative distribution function
MAD	Median absolute deviation
AD	Anderson Darling distribution
CM	Cramér Von Mises distribution
GEV	Generalized Extreme Value distribution
GPD	Generalized Pareto distribution
GLD	Generalized Logistic distribution
PWM	Probability weighted moments
LM	L-Moments
MOMENTS	Moments estimation for the GPD
MLE	Maximum Likelihood estimation for the GPD
MPLE	Penalized Maximum Likelihood estimation for the GPD

Symbols

x, X	random variable
m	empirical mean
s	standard deviation
q	quantile
u	threshold
$q_{1/4}$	first quartile
$q_{3/4}$	third quartile
$F(x), G(x)$	cumulative distribution function
$f(x)$	probability density function
$\Gamma(x)$	Gamma function
$N(\sigma, \mu)$	Normal distribution
$L(x; \theta$	Likelihood function

μ	location parameter or theoretical mean for the Normal distribution
σ^2	variance of the Normal distribution
α	scale parameter
ξ	shape parameter
η	estimator efficiency

Introduction and Outline of the Work

Extreme Events are an important research topic in many scientific domains: from Finance to Meteorology, from Geophysics to Engineering wherever an event largely deviates from the average size the theory of Extreme Values (EVT) must be introduced. Examples of extreme events in Geophysics are the floods, the wind gusts, the size of the waves, the earthquakes; in Finance the Stock Exchange crashes, the failure of Big companies, and the awesome Credit crunch of the last years. All these phenomena are characterized by a variability across many orders of magnitude.

It is common to observe rainfalls of few mm a day but sometimes the daily cumulated rainfalls are of hundreds of mm: in this case it is probable that the river floods the plane. Conversely other quantities like the weight of an animal population, its lifetime or the measures of position of the planets respect to the Sun have a limited variability. All these phenomena are described using a bell shape curve: the Normal distribution. This mathematical function states that the probability of large deviation is so small that can be considered null. In fact a deviation of 3 times the value of the standard deviation has a probability of only 0.1% while at the same level of probability a distribution of Extreme value manifests events that are even 30 times or more greater than the median. In the book "The Black Swan" the philosopher and economist Nicholas Nassim Taleb calls the region of Extreme Events as the Extremistan while the region where the Normal distribution applies and the events are mild is the Mediocristan (with a little negative sense). An Extreme Event for Taleb is a Black Swan: an unexpected bird (swans are white usually) discovered in Australia few centuries ago. The presence of unexpected values, the severity of the statistics and the impossibility to predict the extreme values with a good confidence level are the key concepts of the science of the extremes. We can deal with the high level of incertitude only developing robust tools. It is well known that hydraulic operas like dams, and bridges must be conceived to resist not to the ordinary water flow regimes but to the furious discharge of a flood. This concept would apply to the statistics too, the classical estimators of the Mediocristan region like the

mean, the standard deviation show their limits when data have severe outliers,or when data are corrupted. Furthermore the variance - a dispersion index conceived specially for symmetric distribution, - working with the distribution of extremes that usually are highly skewed has little sense. We see that a more correct measure is obtained with the quantile levels but only with large confidence values.

Before introducing the mathematical models for Extreme Events we need to focus our attention that the extreme events are not always cause of extreme damages. Instead extreme damages could result even from mild events if the infrastructure level of a country is not good. Then the study of the Extremes must be separated from the evaluation of the risks. The correlation of the two quantities risk and extremes is present and usually is positive but the level of risk is not a trivial linear function of the extreme size such as doubling the extreme size means to double the risk and the consequent damages.

Finally the Extreme events are not only big but also rare. The concept of rarity is better understood introducing the return time (or return period), that is the averaged time between two events of the same size. It is common to use the return period as an indicator of the probability of big events: regions with severe extreme have lower return period for big events.

The Extreme Value Theory (EVT) is the mathematical framework to deal with the Extremes. It is based on two important theoretical results: the Fisher-Tippet theorem and the Balkema-De Haan-Pickands. The first one describes the distribution of maxima of an unknown dataset providing that maxima be independent and homogeneous. The second one gives a precise mathematical form to the distribution of the values above an high threshold. The two approaches (studying the maxima and studying the data above the threshold) are similar and asymptotically they give the same results while for finite samples it is common that the threshold method behaves slightly better than the method of maxima; indeed it considers more points (less variance). However the Peaks over threshold method, as it is called in Hydrology and Earth Sciences, is strongly dependent on the choice of the threshold level: low thresholds are associated to inconsistencies in the fitting procedure producing bias and other errors, high threshold levels conversely reduce the bias but increase the variance of the estimate. Then a correct choice for a threshold is the balance of Bias and Variance of the estimators. However in real datasets things are far more complicated. The volatility of the estimators and the sensitivity to data artefacts like outliers, spurious rounding-off greatly enhances the level of uncertainty. For this reason we think that the statistical study of the Extreme will never arrive to be a precise forecast tool but it will furnish only the probabilities of an event accompanied with large confidence intervals.

Note that the problem of finding a good threshold level is similar to distinguish between extreme and non extreme values: in a real dataset this level is continuously changed by any new extreme recorded. Each Extreme, like the Black Swans of Taleb, is able to change the previous estimate of the parameters and the quantile levels. That is in the Extremistan the statistical models are extremely sensitive to new severe events. In the region of Mediocristan where the Normal distribution is a valid tool the estimates are far more tolerant to the variability of the data. In this case the mean and the variance are valid estimators of the characteristics of the population and the model requires few data for a complete description. However even in this world of mild events the presence of spurious data could be a problem specially for the traditional (non robust) estimators. It is a classical and simple exercise to demonstrate how the value of the mean is changed by the presence of a single outlier while the median is more resistant to data contamination having a breakdown point level of 50% (we need to modify at least the half of data to change significantly the median value). The statistical estimators like the median, that are resistant to data contamination and outliers are called robust estimators and are a valid alternative to other traditional tools (mean, variance, linear least square methods etc.)

In this thesis we analyse two vast datasets of precipitation from the Sardinian Hydrological service and from the NOAA-NCDC catalogue of weather data collections. We are interested in checking the data using Exploratory statistics tools, looking for the best statistical model and possibly extending the knowledge associated to these models. We use a data oriented approach (borrowing an expression from the ICT domain) that is an analysis framework that starts from the data without any a priori hypothesis and then it selects the most adapt statistical model using tools of the Exploratory Statistics. Working with real data we can confirm that the statistical models are powerful but fragile tools, they work in the ideal world of pure data, well distributed, homogeneous and with no artefacts. But real databases have artefacts, many of them was created manually and transcription errors are commons, furthermore systematic errors like rounding-off the values (i.e. 5.2 mm truncated to 5 mm, or 4.8 mm to 5 mm) can affect many datasets with different impacts on the statistical model estimation procedure. For these reasons we must start from data, checking them for outliers, for contamination and finally we can apply the statistical models always remembering that these models are statistical approximations of the reality and not forecasting tools. This process known as data quality assessment, is fundamental for the application of the models. At the same time, we want to stress the models to see if they are robust and efficient. Which is the best model ? the best fitting technique ? the best level of the threshold ? and moreover which are the criteria to establish that a model is good (or bad) ? The threshold selection for instance must be adapted to the aims of the experimenter, as we see in this work there

is no single optimal threshold but a set of optimal thresholds depending on our needs. If we want to describe with the most possible reliability the maxima of a distribution we can tune the threshold to improve the description of the maxima, vice versa if we are more interested in catching the behaviour of the central part of the distribution we can choose other different threshold levels.

Finally we need to consider the perspective of a changing world where databases of climatic and geophysical data are easily available, but they are scarcely categorized, and they are *huge* for number and size of datasets. We have the opportunity to extract a lot of information about, for instance, the real world of precipitation but we need tools to deal with Big Data. Learning from data (we call it *data intelligence*), improving the models for automatic optimal adaptation and spurious data description, and reducing the huge-data complexity it will be in my opinion one of the most important research topic for the next years.

Outline of the work

Aims

In this thesis we want to investigate two databases of precipitation using Exploratory statistics tools, we want to improve the statistical models that best describe the datasets. We are interested in assessing the quality of data, looking for outliers and other artefacts. We discovered that the values from the Sardinian database were rounded-off at fixed levels. This artefact causes a big loss in the quality of the fit with the Generalized Pareto Distribution GPD (or other statistical models) and must be fixed modifying GOF (goodness of fit) tests and using robust statistical tools. In principle we do not make any hypothesis on the best statistical model for the databases, but using the L-Moments ratio diagram (a tool introduced by Hosking) we confirmed the optimal performance of the GPD model. When data are rounded we extensively check the performances of the GPD estimators demonstrating how all these ones (included the robust estimators) suffer from the rounding artefact. For the outliers we introduced a test statistics conceived to detect the GPD candidate outliers and we checked it against the real databases.

To correctly fit the GPD model an optimal threshold must be determined: we used for this task several methods from the literature and a new one based on a quality function that estimates the differences between the quantiles of the fitted model and of the true data. This new solution has the notable property to find the optimal threshold in function of a set of points (the pivots) that can be chosen to select the region of

interest. In other words we can find the optimal threshold for the tail, for the center of the distribution or for a mixed subset of points.

All the knowledge developed in the work with hydrological databases will be useful in future research with the output of climate models. In this case the huge amount of data must be checked for artefacts and analysed with automatic and expert tools.

Outline

We briefly sketch here the outline of the work, anticipating the most interesting results and the main topics of this research. In the next lines we deliberately omit the references to the literature. These references can be found along the work in the corresponding chapters.

Ordinary statistical tools

The simplest and most used statistical tools are probably the mean and the variance. If data are distributed with a Normal distribution or a Normal-like (that is a symmetric and compact distribution) the mean and the variance are able to catch the most of the complexity of the data. The situation is different when data come from a skewed distribution, in this case the Median and the Mean Absolute Deviation (MAD) and other quantile based measures are the correct way to describe the distribution.

The Extreme Value Theory (EVT) is the mathematical framework to deal with the extreme values of a dataset. The Generalized Extreme Value distribution (GEV) is conceived to work with all the independent maxima of the dataset arising from the true unknown distribution of values. The only condition to be respected is that maxima be independent, and identically distributed (homogeneous).

Working with values above an high threshold is done with a distribution closely related to the GEV: the Generalized Pareto Distribution (GPD). These two distributions have the same shape parameter (that is the same slope at the distribution tail). If the data above the high threshold are independent and homogeneous the convergence of the GPD to the data is satisfactory.

To fit the GPD model several methods were developed by the statisticians, the most notable one is the Maximum Likelihood Estimator that is also the asymptotically best fitting technique. For small samples and for contaminated data the MLE is outperformed by other methods like the Probability Weighted Moments (PWM) that are computationally more stable specially for small samples.

Tools for exploratory statistics

The Exploratory statistics aims to check the data before taking any decision on the best statistical model to apply. Data are plotted in several ways that allow to extract informations about the presence of outliers, rounded-off values and multivariate populations. The most common ES tools are the scatter plot, the quantile-quantile plot, the density plot and the empirical cumulative distribution, these graphical tools are useful to find outliers, to detect artefacts and they can be coupled, for instance, with numerical tests for outlier detection.

To detect the best statistical model for the extreme values we introduce the L-Moments ratio diagram. This tool explore the position of the real data in the space of two L-Moments (the L-skewness and the L-kurtosis) and it allows to select between several candidate statistical models. For the Sardinian and the NOAA-NCDC daily cumulated precipitation databases it is evident that the best statistical model is the Generalized Pareto Distribution (Figg. 1 and 2).

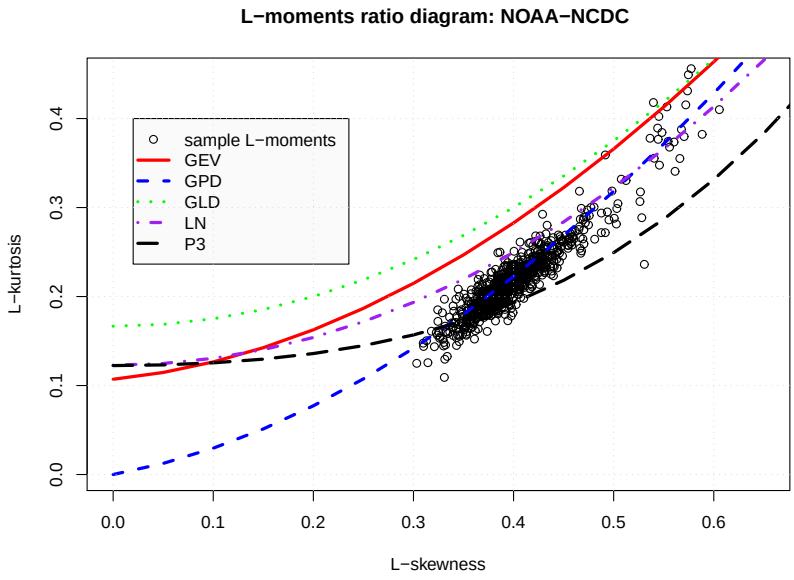

FIG. 1: L-Moments ratio diagram for the NOAA-NCDC stations. The theoretical curves are obtained with a polynomial interpolation (references in the 3rd chapter).

Exploratory statistics at work: the Sardinian and the NOAA-NCDC hydrological databases.

Applying the tools of the Exploratory statistics to the Sardinian and to the NOAA-NCDC databases we found several artefacts. In the American database several stations

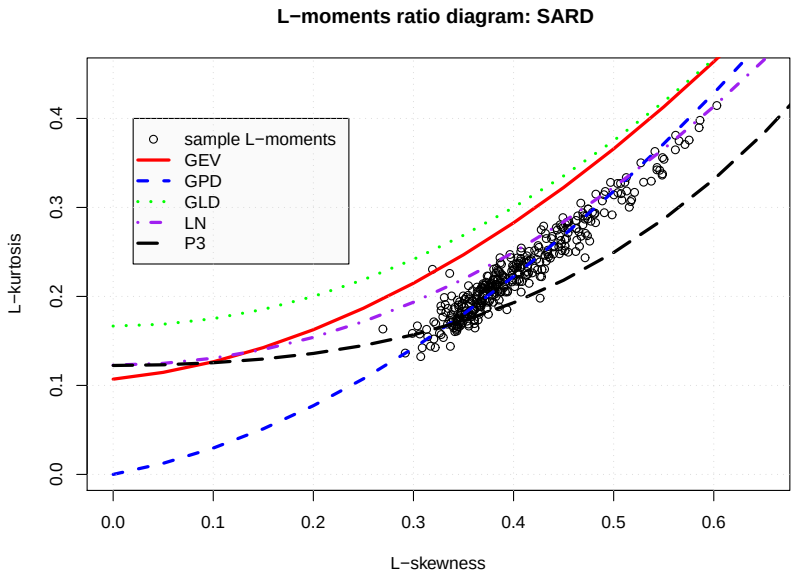

FIG. 2: L-Moments ratio diagram for the Sardinian stations. The theoretical curves are obtained with a polynomial interpolation (references in the 3rd chapter).

have spurious datasets derived from transcription errors or mixed data sources (Fig. 3). In the Sardinian database the values are rounded-off to a mix of rounding-off resolutions: the datasets show a characteristics "zig-zag" behaviour (Fig. 4).

Working with artefacts requires to adapt the statistical tests (like the fit methods that must be robust to the data contamination) and the goodness of fit (GOF) tests like the Anderson-Darling.

Statistical inference and improvements on the models

We try to stress the performances of the GPD model with the real databases and with synthetic datasets with added errors (rounding-off or outliers). We want to investigate the efficiency and the robustness of the GPD parameter estimation techniques in presence of rounding-off. We did this task with Monte Carlo simulations with synthetic datasets.

To use the GPD model a suitable threshold must be chosen in advance before to fit the dataset. If we choose a too low threshold it is common that the estimation is biased, vice-versa if we choose a too high threshold the bias of the estimation is absent but the variance is elevated because we are fitting a small residual portion of data above the threshold. In the work we explore several techniques, described in literature, for the optimal threshold selection.

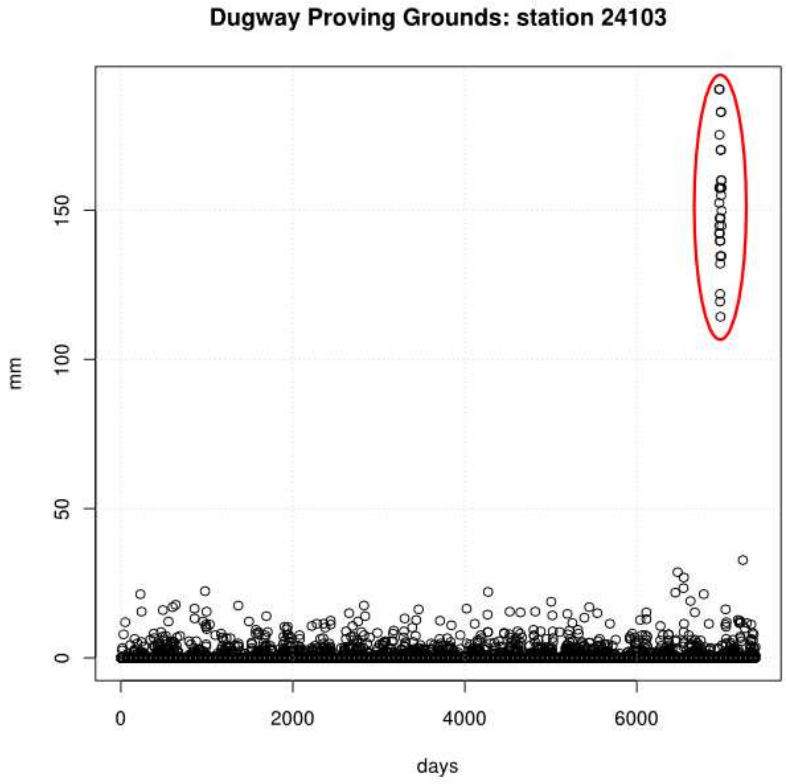

FIG. 3: The Dugway station from the NOAA-NCDC database shows a spurious population (red circle).

Graphical methods are often used with small databases, Goodness of fit tests and tests based on the exponentiality of the tails allow to find a range of candidate optimal thresholds for the datasets with numerical methods. However all these techniques estimate different ranges of the optimal threshold. In other words it seems evident that it not exist an optimal absolute threshold but it varies in function of the used technique. We can generalize this result introducing a quality function based on the sum of quantile differences between real data and the fitted model at several threshold levels. The optimal threshold will be found in function of a pivoting points set, that is a subset of the dataset where we focus our attention. For instance, if we choose the ten highest values as pivoting points the quality function $Q(u)$, where u is the threshold level, shows a minimum for a given value of u (Fig. 5 and Fig. 6). This technique allows to improve the estimation of the GPD model and consequently the determination of the extreme events.

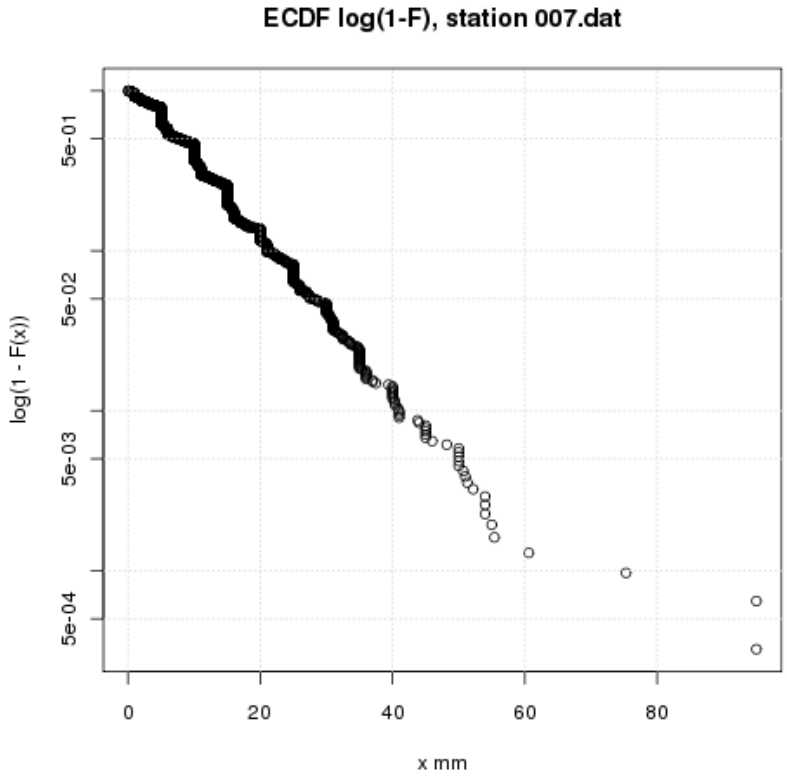

FIG. 4: Empirical Cumulative Distribution (in 1-F and log scale) of the Lunamatrona station from the Sardinian database. Note the presence of anomalous steps in the curve due to the rounding-off.

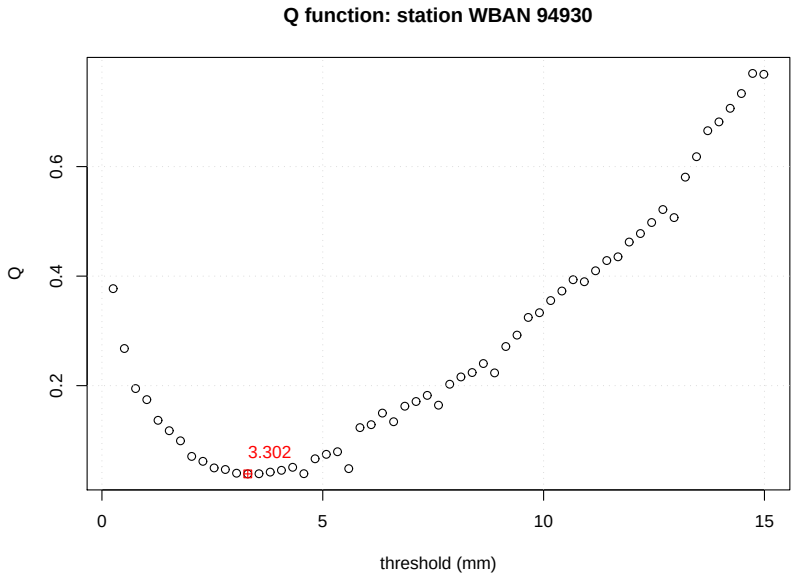

FIG. 5: Quality function $Q(u)$ in function of the threshold u. The pivoting points in this case are the last 10 greatest points.

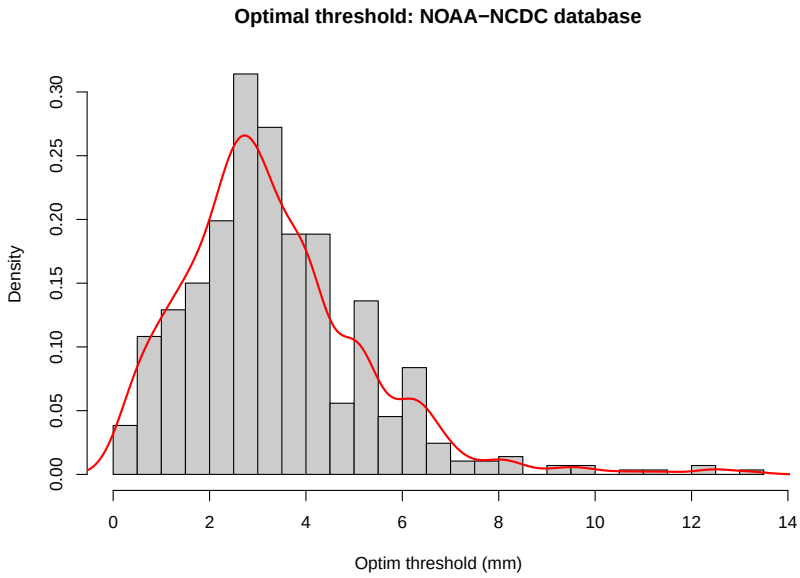

FIG. 6: Estimating the optimal threshold with the quality function. We select as pivoting points the 10 greatest points.

The big data challenge and future researches

The modern climatology is based on the climate models. These numerical tools produce as output fields of precipitation, vector winds, temperatures and so on. The models can be used as a prognostic tool if they are conceived for weather forecast, and as a long time averaged statistical tool for future forecasts of the Earth climate. The output of of these models is nowadays of huge dimensions, dozens of TB are produced for single sessions of simulation. To work with this amount of data we need robust software techniques and systems able to classify the information, to download and select data and finally to visualize it in an effective way. Tools from the Exploratory statistics help to assess the quality of data, to identify artefacts and probably to extract the extremes and the trends in the extreme events along countries or continents. The use of climate databases mixed with ES tools will be the topic of a future research.

Chapter 1

Mathematical theory of the Extreme events

1.1 Introduction

The Theory of the Extreme Events (EVT) is the mathematical framework that describes the occurrence of the events of exceptional size and exceptional rarity. It is based on strong theoretical results: the Fisher-Tippet theorem and the Balkema-De Haan-Pickands theorem. These important statements describe two different approaches to the study of the extreme values: a) studying the maxima of a dataset b) studying all the values of a dataset above an high threshold. These theorems are analogous to the law of large numbers that is a natural introduction to the Gaussian distribution (the distribution of the means tends asymptotically to the Normal distribution). Theorems of EVT apply to extreme values as a *probability* framework. In fact, even if the conditions of these theorems (data independence and homogeneity of the values) are mild it is possible to find datasets where the description of the extreme events with this tools seems contradictory. The random fluctuations or the presence of outliers or spurious data impossible to eliminate from the series leads sometimes to values of the distribution parameters that describe datasets of infinite variance: a non sense in Physics[1].

In this chapter we present the mathematical theory of the distribution of maxima or of the distribution of excesses over a threshold. These techniques in Hydrology are called respectively (Annual Maxima AM) and (Partial Duration Series PDS). Then we show

[1]Note however that if we are working with the non equilibrium physics, like the physics describing the phase transitions (like solid-liquid, magnetic-non magnetic), the variance of several quantities depending on the size of the sample are infinite as stated by the renormalization theory

the relationships between the AM and PDS theories, and in next chapter we summarize the modern methods of fit for the distributions related to these tools.

Intuitively an extreme event is an event that largely deviates from the median of the sample distribution. But an extreme event could be also an event of exceptional rarity occurring once in *long* period of time T. The average time T between two events of the same size x is called *return period* and it is a function of the size of the event: big events are rare. The correct determination of the return period function $T = T(x)$ is of fundamental importance for Engineering where the size of a bridge, for instance, must be chosen keeping in mind the possibilities of severe flooding occurring in probability once in a T years. The relation between the return period and the probability distribution of events of size x can be found in the paragraph 1.3.

In Hydrology several statistical distributions have been applied, and sometimes specifically conceived, to characterize the behaviour of severe rainfalls and the flood discharge in rivers, as well as other climatological variables. Among the proposed distributions we can remember the Lognormal,the Exponential, the Gamma, the Pearson Type III, the log-Pearson Type III, the Gumbel and the Weibull, Two Component Extreme Value TCEV, Generalized Extreme Value GEV and Generalized Pareto Distribution GPD are all distributions that received consideration in statistical hydrology. Referring the reader to Chow et al. (1988) and Stedinger et al. (1993) for more details on these statistics and their applications in hydrology, along this thesis we'll show the importance of the last two distributions, that appear particularly attractive for their asymptotic statistical properties: namely the GPD and the GEV.

If we define the extreme values as the maxima within time blocks (usually assumed 1-year long in the Earth Sciences), it can be proved that, if there exists a limit distribution of these maxima, this distribution belongs to the GEV family (the original seminal papers of this theory are Fisher and Tippett (1928) and Gnedenko (1943)). Conversely, when looking at the exceedances above a threshold, it can be proved that the GPD is the expected distribution (Pickands, 1975). Moreover, if a process follows a GPD with a given shape parameter, the block maxima follow a GEV distribution with the same shape parameter (relation derived by Pickands following by the Balkema - De Haan - Pickands theorem). The importance of the GPD / GEV framework is proven empirically by several authors that describe the performances of the GPD in the study of several hydrological quantities like the cumulated rainfall and the flood discharge. A detailed discussion can be found in the work of Madsen et al. (1997).

1.2 Recalls of probability theory

In this section we briefly recall some concept from probability and statistics, gradually introducing the notation used in all the thesis.

The axioms of the probability can be expressed using the theory of sets or with the more intuitive concept of frequency. Basically the probability is a measure of the number of the occurrences of an event in the space of all possible events Z. In terms of frequencies then the probability of an event x is the number of times that x appears in the ensemble Z divided by the number of events of the entire space Z. If we call P this frequency then using basic properties of set theory on the space Z we can introduce the probability axioms:

$$\begin{cases} P(A) \subset [0,1] \quad \forall A \subset Z \\ notA = P'(A) = 1 - P(A) \\ P(A\ or\ B) = P(A \bigcup B) = P(A) + P(B) - P(A \bigcap B) \\ P(A\ and\ B) = P(A \bigcap B) = P(A|B)P(B) = P(A)P(B) \quad if\ A, B\ \ are\ independent \\ P(A\ given\ B) = P(A|B) = P(A \bigcap B)/P(B) \end{cases} \tag{1.1}$$

in plain words:

1. the probability is always in the interval $[0,1]]$
2. the probability to not find an event A is $1 - P(A)$
3. the probability to find two events is the sum of the two events probability minus the intersection that represents the probability to find at the same time A and B
4. the conditional probability $P(A|B)$ is the probability to find at the same time A and B divided by the probability to find the event B

Note that if the events are independent (they are sets with no overlap) the total probability to find the two events is the product of the single probabilities.

If we have n events then the sum of the probabilities is by definition the unity:

$$\sum_{i=1}^{n} P(x_i) = 1$$

in the limit of $n \to \infty$ the sum can be transformed into an integral sum:

$$\int_Z p(x)dx = 1.$$

Using this property we can define the cumulative distribution function $F(x)$ as the integral:

$$P(x < X) = \int_a^X p(x)dx \tag{1.2}$$

where the integral is computed in the interval $[a, X]$ assuming that the value $a = \inf(Z)$ is the minimum of Z and Z has some kind of metrics. From this rule we can then define the probability density function $p(x)$ as the derivative of the cumulative density function:

$$p(x) = \frac{dP(x)}{dx} \tag{1.3}$$

To characterize a probability distribution and its behaviour it has been introduced the concept of the distribution moments. The well known first and the second order moments are the mean and the variance:

$$\begin{aligned} \mu &= \textstyle\int_Z xp(x)dx, \\ \sigma^2 &= \textstyle\int_Z (x-\mu)^2 p(x)dx \end{aligned} \tag{1.4}$$

for higher order moments m^n we have

$$m^n = \int_Z (x-\mu)^n p(x)dx, \tag{1.5}$$

The most famous and probably important distribution is the Normal distribution, usually indicated with $N(\sigma, \mu)$, its main properties are summarized in the table 1.1

Note that to *fit* (the process of adapting the theoretical curve to the data) the Normal distribution we need to compute the variance and the mean of the data. With these values we have μ and σ valid in the limits where we are confident that our data can be described by a Normal distribution. Many tests (called *Normality tests*) are present in literature to assess the correctness of the Normal model.

The normal distribution has the important property to have all moments finite and it is used as a test toolbox for other distributions. Every time it is possible the statisticians try to reduce the complexity of the model using the well known properties of that

$f(x)$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
$F(x)$	$\frac{1}{2}\left[1 + erf\left(\frac{(x-\mu)}{\sqrt{2\sigma^2}}\right)\right]$
$mean$	μ
$median$	μ
$variance$	σ^2

TAB. 1.1: Normal distribution properties

distribution. For instance the non-parametric methods use the rank of each data of the sample (distributed with an unknown statistical model) to build tests based on the normality assumption.

1.3 The *T*-year level (return period)

In Earth sciences and other disciplines the time between two events of the same magnitude is called *return period* or *T-year level*. It represents the probability of having an event of size u in a period of time T. Following Reiss and Thomas (2007) pag. 12 in a period of T years there is an average probability equal to 1 to have an event of size u:

$$E[F(x \geq u)] = 1 \tag{1.6}$$

where $F(x)$ is the probability distribution and E the mean value operator. Let $F^{-1}(P)$ the inverse function of F representing[2] the event in function of the probability P , if the we expect one event in T years then the probability is by definition $1/T$ and the probability to have an event of size u is:

$$P(X > u) = 1 - F\left(u(t)\right) = \frac{1}{T} \tag{1.7}$$

hence

$$u(T) = F^{-1}(1 - 1/T) \tag{1.8}$$

the 1.8 links the return level to the inverse of the probability distribution F.

[2]As we see in detail in the next chapter the $F^{-1}(P)$ is the *quantile function*

1.4 Introducing the extreme values theory from the ordinary distributions

1.4.1 Observing the exceedances over a threshold

Let $x_1, x_2, \ldots, x_n$ a set of random variables, if we reorder the set ascending we can describe the behaviour of the values above a threshold u. The data above the level u are the *exceedances* (or excesses). Using a notation from Hydrology and Earth Sciences these values are the Peaks over the threshold (POT).

If we have K exceedances then our set is divided in two parts: $X_i < u$ and $X_j \geq u$. Now it is an well known fact that the probability of events having *two* possible outcome values (above and below the threshold) can be described using the binomial distribution:

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad k = 0, 1, \ldots, n \tag{1.9}$$

where p is the probability to find a value above the threshold and $1-p$ the probability to find a value below the threshold. The mean of this distribution is

$$m = np = n\,(1 - F(u))$$

and it is a decreasing function: when the threshold increases the number of exceedances decreases.

The natural approximation for long series of the binomial distribution is the Poisson distribution. This limit is valid for $np = \lambda$ and $n \to \infty$, where λ is the parameter of the following equation:

$$P_\lambda(k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad k = 0, 1, 2, 3, \ldots \tag{1.10}$$

the parameter λ is equal to the mean and these two distributions apply to discrete events. The importance of this tools is linked to the probability of finding a value bigger than the threshold level $k \equiv u$. The Poisson equations states that the total set of values above the threshold decreases exponentially with the threshold level u . That is extreme events become exponentially rare for high thresholds.

1.4.2 Introducing the distribution of the maxima

Suppose to have data from an unknown distribution F, and to be interested in describing the behaviour of the highest (or the lowest) value from F. Let $X_1, X_2, \ldots, X_n$ some subset of data from F, if the X_i are independent consider the maxima of each set:

$$(m_1, m_2, \ldots, m_n) = (max(X_1), max(X_2), \ldots, max(X_n)) \tag{1.11}$$

Following the axioms of the probability, the total probability of having all n independent maxima is the product of the single probability of each maxima:

$$G(x) = F(X_1)F(X_1)\ldots F(X_n) = [F(X)]^n \tag{1.12}$$

when $n \to \infty$ the probability $G(x)$ converges toward the unique maximum of the distribution then

$$G(x) = lim_{n\to\infty}[F(X)]^n = \begin{cases} 1 \quad if \quad F(x) = 1 \\ \\ 0 \quad if \quad F(x) = 0 \end{cases} \tag{1.13}$$

this is the degenerate case, the limiting distribution could be obtained if they exists some constants a_n and b_n that allow to normalize the maxima:

$$\frac{m_n - b_n}{a_n} \qquad if \quad a_n > 0 \tag{1.14}$$

Fisher-Tippet theorem (1928)

Given a set of independent random variables $X_1, X_2, \cdots, X_n \in F(X)$ *and the distribution of independent maxima* $F(X)^n = \max\{X_1, X_2, \cdots, X_n\}$ *if they exist the constants* $a_n > 0$ *and* b_n *then it exists a non degenerate limiting distribution* $G(X)$ *:*

$$|[F(X)]^n - G(a_n X + b_n)| \to 0 \tag{1.15}$$

the equation $G(X)$ is called *Generalized Extreme Value* (GEV) distribution and has three mathematical types (see also Fig. 1.1) :

Type 1.(Gumbel-type distribution)

$$G(x) = e^{-e^{\frac{x-\mu}{\alpha}}} \tag{1.16}$$

Type 2. (Fréchet-type distribution)

$$G(x) = \begin{cases} 0 & x < \mu \\ e^{-\left(\frac{x-\mu}{\alpha}\right)^{-\gamma}} & x \geq \mu \end{cases} \tag{1.17}$$

Type 3. (Weibull-type distribution)

$$G(x) = \begin{cases} e^{-\left(\frac{x-\mu}{\alpha}\right)^{\gamma}} & x \leq \mu \\ 0 & x > \mu \end{cases} \tag{1.18}$$

where γ is the shape parameter, μ a location parameter (not to be confused with the mean) and $\alpha > 0$ is the scale parameter (it controls the size of the distribution like the variance in the Normal distribution).

These three distributions have completely different behaviours: the Gumbel distribution is unlimited, the Fréchet distribution has a lower limit while the Weibull distribution an upper one. Note that the Fréchet and the Weibull differ only for a sign, while the first one is used to study quantities that have a lower bound (like the precipitation that is always positive) the second one is used to study phenomena with an upper bound. The Weibull distribution is often used to study the behaviour of the minima.

The three types of the distributions could be unified in a single one with the Von Mises transformation ($1/\xi = -\gamma$) obtaining the *Generalized Extreme Values* (GEV) distribution:

$$F(x) = \exp\left\{\left[1 + \xi\left(\frac{x-\mu}{\alpha}\right)\right]^{-1/\xi}\right\} \quad , \quad 1 + \xi\left(\frac{x-\mu}{\alpha}\right) > 0 \quad , \quad -\infty < \xi < \infty \quad , \quad \alpha > 0. \tag{1.19}$$

This equation is of fundamental importance in the extreme events theory. The conditions of the Fisher Tippet theorem 1.14 are wide enough to satisfy a large number of marginal

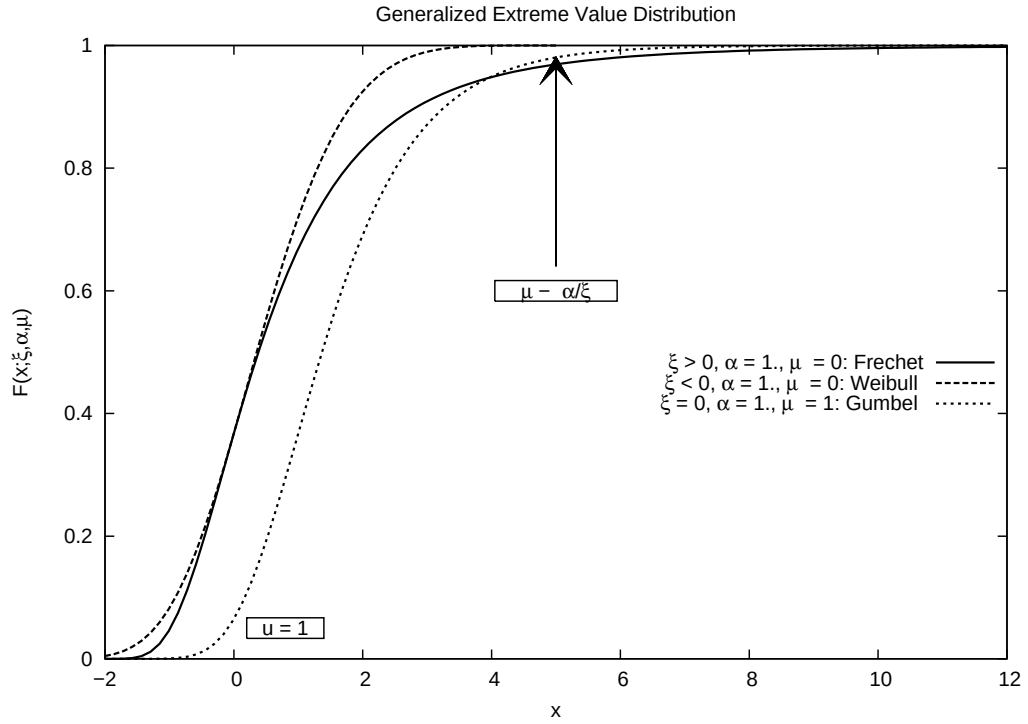

FIG. 1.1: The cumulative probability distribution of the GEV shown with different values of the parameters.

distributions F furnishing the values for the extremes.

Note that the 1.19 has 3 parameters to estimate and only in the degenerate case ($\xi = 0$ corresponding to the Gumbel distribution) the ξ parameter does not need to be estimated.

1.5 GEV moments and its properties

The estimation of the moments (mean, variance, skewness etc.) of the GEV is simplified using the Gamma function. The general expression of this function is:

$$\Gamma(\lambda) = \int_0^\infty x^{\lambda-1} e^{-x} dx, \qquad \lambda > 0 \tag{1.20}$$

and its properties useful for our calculations are $\Gamma(\lambda+1) = \lambda\Gamma(\lambda)$ and $\Gamma(1) = 1$.

By definition the j-th moment of a distribution (with the location parameter equals to 0 for simplicity) is

$$m_{j,g} = \int x^j g(x) dx$$

if the $g(x)$ is of Weibull 1.18 or Fréchet type 1.17 it is quite simple to demonstrate (rearranging the $g(x)$) that the mean is for the Fréchet distribution:

$$m_f = \Gamma(1-\xi) \tag{1.21}$$

and for the Weibull

$$m_w = -\Gamma(1-\xi) \tag{1.22}$$

as expected from the symmetry of the two distributions $m_{weib} = -m_{frec}$.

The expression for the variance of the GEV is a little bit more complicated. Let

$$g_j = \Gamma(1-j\xi)$$

then it is possible to demonstrate (see Reiss and Thomas (2007)) where $\gamma = -1/\xi$ and $\alpha = \sigma$ that the variance is

$$Var = \begin{cases} \alpha^2 \frac{g_2 - g_1^2}{\xi^2} & if \;\; \xi \neq 0, \quad \xi < 1/2 \\ \alpha^2 \frac{\pi^2}{6} & if \;\; \xi = 0 \\ \infty & if \;\; \xi \geq 1/2 \end{cases} \tag{1.23}$$

note that if the shape parameter is high $\xi \geq 1/2$ the variance is not a finite quantity. This possibility is rarely encountered in Physics and Engineering, although it is possible, as we see, to find datasets where the fitted model has $\xi > 1/2$, in this case an accurate check of the data usually allows to find errors in the dataset or more complex features like bivariate or multivariate populations [3].

1.6 Convergence and stability of the GEV

The Fisher-Tippet theorem introduces two constants a_n and b_n to normalize the maxima of the distribution $F(X)$, this condition 1.14 could be expressed as a linear transformation where the parameters depend only on the size n of the sample and not on value of

[3] in this case other populations mixed to the original data can lead to an erroneous estimate of the distribution parameters

the variable x:

$$[F(X)]^n = F(a_n X + b_n) \tag{1.24}$$

Remember that the F^n distribution expresses the distribution of the maxima (it is the product of the n probabilities of the independent maxima each one described by F). This simple formula expresses the *max-stability* feature: a distribution is called *max-stable* when a linear transformation of its maxima does not change the distribution itself. The stability postulate is a condition necessary and sufficient to ensure the convergence to the real maxima distribution. Gnedenko (1943) demonstrated that to obtain the asymptotic convergence of the maxima to the GEV distribution the following condition needs to be satisfied. Let $l(x)$ a positive and measurable distribution if

$$\lim_{t\to\infty} \frac{l(xt)}{l(x)} = 1 \quad \forall \; x \in \quad \Re \tag{1.25}$$

then the function $l(x)$ is a *slow varying* function. Given now a distribution function $F(x)$ if we define the survival function $1 - F$ the Gnedenko condition gives the result:

$$1 - F(x) = x^{-1/\xi} l(x), \quad x > 0 \tag{1.26}$$

in other words if the survival function follows a *power law* limited by a slow varying function l_F the limiting distribution for its maxima is a GEV.

The 1.26 is a *first order condition* ensuring the convergence but not giving informations about the rate of convergence of the limiting distribution. For practical purposes the convergence of the maxima distribution to the GEV (or to the Generalized Pareto distribution) must arise for finite samples and not only at infinity. To ensure an useful rate of convergence (see Goegebeur et al. (2008) and for a detailed explication the book of Beirlant et al. (1996) pag. 90) a *second order* condition for the slow varying function needs to be introduced:

$$\lim_{x\to\infty} \frac{\log l(xt) - \log l(x) + \log(t)/\xi}{b(t)} = \frac{t^\rho - 1}{\rho} \tag{1.27}$$

where $b(x) \to 0$; as $x \to \infty$, with $t \geq 1$ and $\rho \leq 0$.

If the 1.27 is respected then the convergence of the limiting distribution to the GEV is asymptotically Normal.

1.7 The distribution of the exceedances and the POT methods

The GEV distribution is the general framework to work with maxima when the mild conditions of the Fisher-Tippet theorem are satisfied (data homogeneous and independent). In Earth Sciences for instance it is common to consider one maximum per year to avoid problems with the seasonality that introduces correlation between data. Using only one value per year results in short datasets that are generally difficult to fit correctly, especially using distributions like the GEV that has 3 parameters. For this reason we can try to simplify our statistical model (using models with a single parameter like the exponential distribution or the Gumbel function) or we can try to extend the dataset considering more than one extreme value per year using a different mathematical model describing *all* the values above a given *high* threshold.

The theory of the *excesses* over an high threshold is funded on a theorem named on the discovers Balkema, de Haan, and Pickands. It states a relationship between linear transformations of the random variable X and the limiting distribution, the enunciate of the theorem is:

Balkema-de Haan - Pickands theorem.

Given a probability distribution $F(X)$ if it exists a couple of constants a_n and b_n depending only on the length n of the dataset $X_1, X_2, \cdots, X_n$ then the limiting distribution of $F(X)$ as u tends to the right endpoint is

$$F(a_n + b_n X > u) \to G(\xi, \mu, \sigma) \tag{1.28}$$

where $G(\xi, u, \alpha)$ is known as the Generalized Pareto distribution (GPD) and ξ,u and α are parameters: respectively the *shape*, the *location* (or *threshold*) and the *scale* parameter.

The mathematical form of the Generalized Pareto Distribution is the following

$$F(x; \mu, \alpha, \xi) = \begin{cases} 1 - \left(1 + \xi \dfrac{x-u}{\alpha}\right)^{-1/\xi} & \xi \neq 0 \\ \\ 1 - \exp\left(-\dfrac{x-u}{\alpha}\right) & \xi = 0 \end{cases} \tag{1.29}$$

where ξ,u, and α are the shape, the location and the scale parameter. Note that some references uses the $-k = \xi$ parameter for the shape and other ones use the symbol σ for the scale parameter.

The density function of the 1.29 is given by:

$$f(x;\mu,\alpha,\xi) = \begin{cases} \frac{1}{\alpha}\left(1+\xi\frac{x-u}{\alpha}\right)^{-1/\xi-1} & \xi \neq 0 \\ \\ \frac{1}{\alpha}\exp\left(-\frac{x-u}{\alpha}\right) & \xi = 0 \end{cases} \tag{1.30}$$

depending on the value of the shape ξ parameter there are different possibilities (refer to Fig.1.2):

- $\xi > 0$ the distribution has a long right tail much bigger than a Normal distribution tail (distributions with this property are often referred as "heavy tail distributions"). In this case, conventional moments of order greater than or equal to $1/\xi$ are degenerate (i.e. if $|\xi| > 1/2$ the ordinary variance is infinite)
- $\xi = 0$ the distribution has the ordinary exponential form.
- $\xi < 0$ the distribution is short tailed with an upper bound value $(u - \alpha/\xi)$

When the value of the shape parameter is zero then the 1.29 is a simple exponential distribution; when the shape parameter is not null then there are two different possibilities: a) the shape is negative and the distribution has a right positive endpoint b) the shape is positive and the distribution is unbounded Fig. 1.2 .

NOTE. The u location parameter or threshold could be expressed as an *upper order* statistics: in the reordered sample

$$X_1 < X_1 < \cdots < X_k = u < \cdots < X_n$$

the level u coincides with the k-th value of the reordered sample.

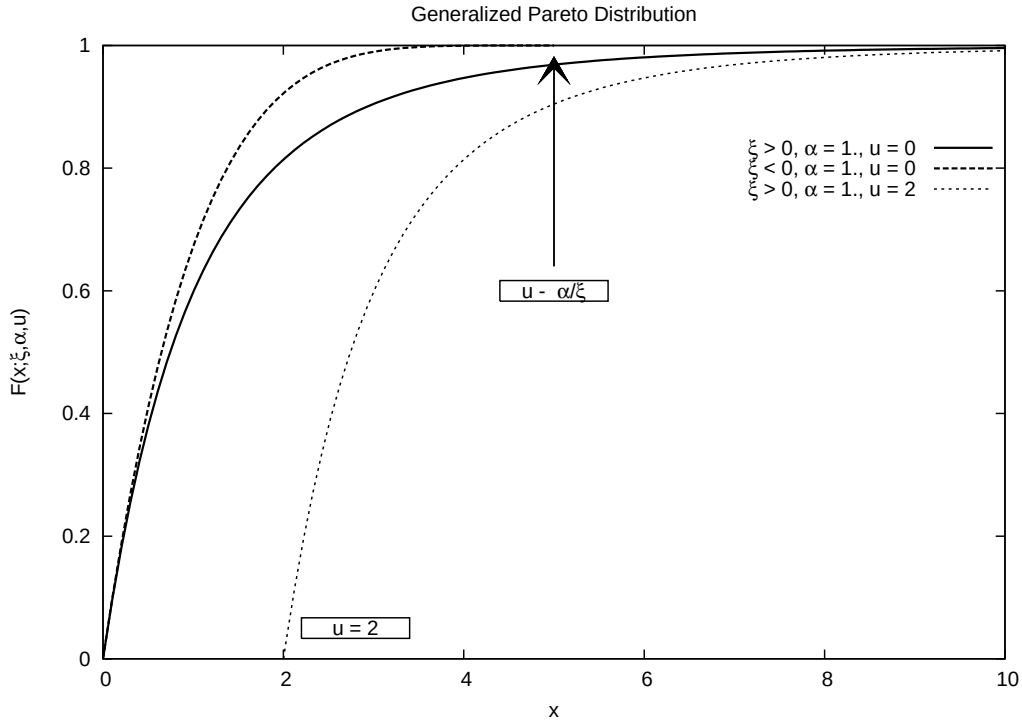

FIG. 1.2: The cumulative probability distribution of the GPD shown with different values of the parameters scale and shape.

1.8 The GPD threshold

The mathematical form of the GPD distribution 1.29 has a parameter u that plays the role of a location parameter. However for the GPD, unlike from other distributions as the GEV, the location parameter is chosen *before* fitting the curve and not obtained by the fit procedure of the distribution. Moreover in working frameworks like the Partial Duration Series (PDS) or the Peaks Over the threshold (POT)[4], the fit of the curve to data is conducted by a *left censoring* procedure that discards all data below the location parameter of the GPD. In this case the location parameter is called *threshold*. The procedure of fitting then consists in choosing a value for the threshold and *after* estimating the other parameters. How to choose the best threshold level is the argument of the Chapter 4: as we see the choice of this value it is not trivial. Here we highlight the extreme sensitivity of the fit to the choice of the right threshold level. In theory using the GPD, the threshold does not affect the value of the other parameters (except the scale parameter that shows a linear transformation with the threshold level see 1.10) but for real datasets the sensitivity to the threshold is evident and it affects the quantile

[4]PDS and POT methods are the same technique mainly used in Hydrology to study floods and severe rainfalls and daily cumulated precipitations.

levels estimated with the GPD model. Erroneous quantile levels mean an imprecise description of the extreme events.

1.9 Relationship between GPD and GEV

The GPD is linked to the GEV by a simple mathematical transformation as in Reiss and Thomas (2007) pag. 23

$$GPD = 1 + \log(GEV) \quad if \quad \log(GEV) > 1 \tag{1.31}$$

the similarity between GPD and GEV means that the shape and the scale (see 1.29 and 1.19) parameters are the same for both distributions. A good fit for the Pareto distributions leads to a good fit for the GEV. But unlike the GEV, with the GPD only the fraction of data below the threshold is discarded, while the GEV considers only the few independent maxima.

1.10 Stability of the GPD and POT domain of attraction

The Balkema-De Haan-Pickands theorem introduces stability conditions for the values. When data are transformed with a linear combination

$$X' = a_u + b_u X$$

where a_u and b_u are evaluated at the u threshold the shape of the distribution does not change, this fact is important to rescale and transform data and coefficients. Let define the distribution of the exceedances $F^{[u]}$ as

$$F^{[u]}(X) = \frac{F(X) - F(u)}{1 - F(u)} \tag{1.32}$$

where u is the threshold and the factor $1 - F(u)$ is needed to normalize the probability distribution. The POT stability condition states that

$$F^{[u]}(a_u + b_u X) = F(X) \tag{1.33}$$

the main consequence of this formula is the property that a GPD distribution shape does not change selecting different threshold. Moreover it is possible to demonstrate that

when the threshold u is modified the value of the scale parameter α changes following a linear transformation (Coles, 2001 pg. 83):

$$\alpha_u = \alpha_{u_0} + \xi(u - u_0) \tag{1.34}$$

Where u_0 is the threshold of the level "0" and u a different threshold level.

Chapter 2

Exploratory statistics

2.1 Introduction

In this chapter we focus our attention on the most important statistical tools for the Exploratory Data Analysis. This statistical set of techniques is of fundamental importance in the *pre*-analysis of data. Looking for outliers, for errors in the datasets, for inconsistency in the results of the numerical estimators (even the widely used mean could have some problems working with some kind of data) and considering the robustness and the efficiency of an estimator we can check the data before passing them to a theoretical model. In Informatics and Engineering this process is known as *data quality assessment*, and for Meteorological databases where data has been collected manually for a long period of time this check is indispensable.

2.2 Tool and methods for exploratory statistics

The statisticians have invented several tools and techniques to explore data using plots or statistical indicators. With plots, for instance, it is possible to rapidly check the presence of multivariate distributions [1]: task not trivial with numerical techniques.

We introduce here the empirical probability density function plot and the empirical cumulative density plot. In the next lines we'll call the empirical cdf with the acronym ECDF; the empirical pdf will be the *histogram* if the sample is discrete or the *density* if the sample is continuous or approximated using continuous functions.

[1] usually if the sample arises from a multivariate distribution when plotting the empirical pdf the graph shows more than one peak

2.2.1 The empirical cumulative distribution

Given a set of random variables $x_i \;\; i = 1, 2, \cdots, n < \infty$, let $I_j(x < x_j)$ the number of values of x less than x_j then the empirical cumulative distribution or simply the sample distribution is defined as:

$$\hat{F}_n(x) = \frac{1}{n} \sum_{j=1}^{n} I_j(x < xj) \tag{2.1}$$

the plot of $\hat{F}_n(x)$ is easy to realize with a *plotting position* rule:

- reorder the sample in ascending order $\mathbf{x}_{sorted} \equiv x_1 < x_2 < \cdots x_n$
- let j the rank of the x_j value, then define a plotting position rule as

$$p_j = \frac{j}{n+1}$$

- plot the vectors $(\mathbf{x}_{sorted}, \mathbf{p})$

note that the values of the y axis are limited in the interval $[0, 1]$ while the x axis is the interval $[\min(x), \max(x)]$.

The plotting position rules are still an active theme of research (see for instance Makkonen (2008a)) with a dispute lasting one century from the first pp-rule introduced by the hydrologist Allen Hazen in 1914. The simple rule $j/n + 1$ produces biased plots if the sample length is odd.

For instance let the sample be of the length of 7 then the positions are

$$1, 2, 3, 4, 5, 6, 7$$

; the value $4/7$ is not the center of the distribution nor the value $3/7$. For this reason Hazen introduced the rule $(j - 0.5)/(n + 1)$ that is one of the most used plotting position rules. More generally the pp-rules are members of the family:

$$p_j = \frac{j - a}{n - 2a + 1}$$

When $a = 0.5$ this formula gives the Hazen rule, other values of a are possible. For our purposes working with EV distributions, and following the suggestions of Gumbel and Weibull ($a = 0$), we'll choose the simple form

$$p_j = \frac{j}{n+1}$$

2.2.2 Histograms and density plots

The probability distribution of the sample could be investigated using the popular tool of the histogram , also called frequency plot. To obtain this graph the data must be organized in this way:

- reorder the sample in ascending order $x_1 < x_2 < \ldots < x_n$
- classify the values of x in *bins*, that is divide the interval $[\min(x), \max(x)]$ in $m < n$ equal parts $a_1 < a_2 < \cdots < a_m$ and count the number of x_i belonging to each interval class (bin)

  $$[a_k, a_{k+1}] \quad 1 < k < m$$

  let p_j/n this quantity (*frequency* or empirical probability).
- plot the vectors $(\mathbf{a}, \mathbf{p})$

the classical representation of this tool is the *bar plot* where each interval is represented by a vertical bar (Fig. 2.1).

The common criticism to this tool is the arbitrariness of the bin length l (length of each interval $[a_k, a_{k+1}]$) while depending on the choice of this classification several artefacts appear. If we choose a too small value for the bin length the plot is confused and the histogram presents a number of unrealistic peaks, conversely if we choose a greater interval l the finer details are lost. Moreover if the distribution is symmetric the same problems seen with the plotting position rule are present (false bias), see for instance Fig. 2.1 where the distribution (a Gaussian centred on zero) must be symmetric but the choice of the number of bins changes the appearance and the symmetry of the plot. To overcome this problems the statistician invented the *density* plot obtained with the continuous kernel functions superimposed to the discrete set of data.

The approximation of the kernel function is based on this simple idea. Observing an histogram made of bar plots we note the rough approximation of the y value with constant values (the bars), if we substitute the bars with continuous functions like the bell-shape curve with an adjustable width b, we can obtain a smoother representation of the distribution. This approximation is called *density* plot and it is a common tool in many statistical packages (like `R`, `MATLAB`, `Origin` etc).

Let $x_1 < x_2 < \cdots x_n$ the reordered sample, on every point define a function $f_{n,b}(x)$ of the form:

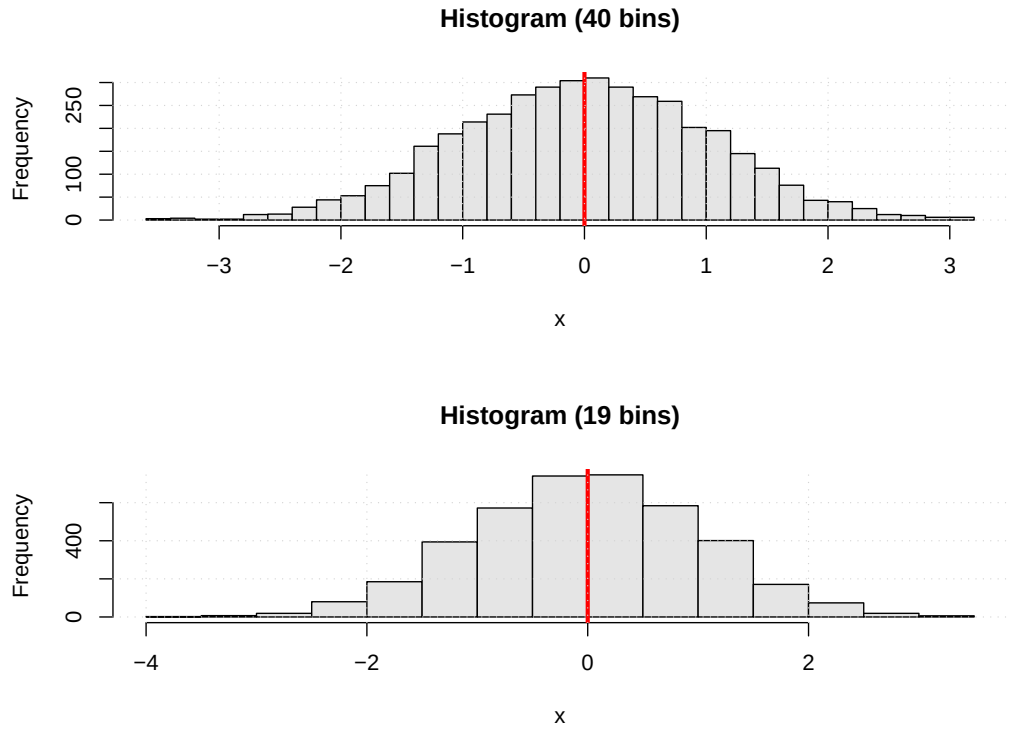

FIG. 2.1: An histogram of a random Normal sample is plotted with different number of bins.

$$f_{n,b}(x) = \frac{1}{nb}\sum_{i=1}^{n} k\left(\frac{x - x_i}{b}\right) \tag{2.2}$$

where the parameter $b > 0$ is the *bandwidth* (equivalent to the width of the bin for the histogram), the $k(x)$ function, centred around the value x_i, is the *kernel* function and has the following property:

$$\int k(x)dx = 1$$

Note that the contributions of the functions $k(x_i)$ evaluated at the point x_i are summed on the entire interval $[\min(x), \max(x)]$, for this reason where the values x are more concentrated, the value of $f_{n,b}(x)$ is higher (see Fig. 2.2) and gradually falls down far from the points x_i. This behaviour is different from the histogram where the step function falls down immediately at the end of each bin. However as in the case of the histogram binning width the choice of the bandwidth parameter b is crucial: a too big value of b leads to a plot loosing many details, a too small value for b produces peaks and artefacts as in Fig. 2.3 .

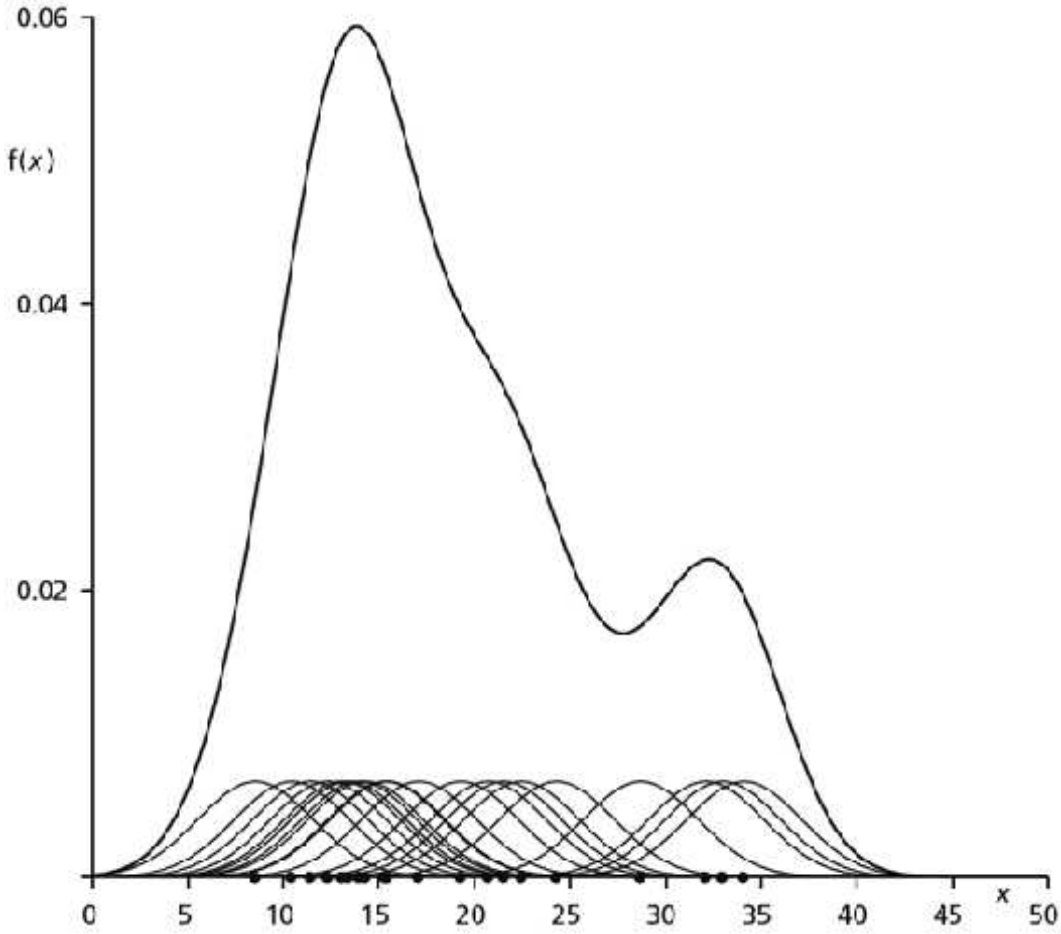

FIG. 2.2: Sum of kernel Gaussian functions. The zones where the points are more concentrated have a greater sum of kernel functions.

The $k(x)$ function could be of many different types, the most common choice is the Gaussian equation, but it possible even to reproduce the histogram using the step function for k:

$$k(x_j) = \begin{cases} p_j/n & if \;\; x \in (-b/2 + x_j, b/2 + x_j) \\ \\ 0 & if \;\; x \notin (-b/2 + x_j, b/2 + x_j) \end{cases}$$

other choices are triangular functions, polynomials and trigonometric functions in bounded intervals (to avoid problems with the periodicity). The theory of the kernel function introduces tools for the automatic optimal bandwidth selection and the most common statistical packages implement these algorithms, however a visual inspection of the density function superimposed to the histogram (see Fig. 2.4) is of capital importance to validate the plot.

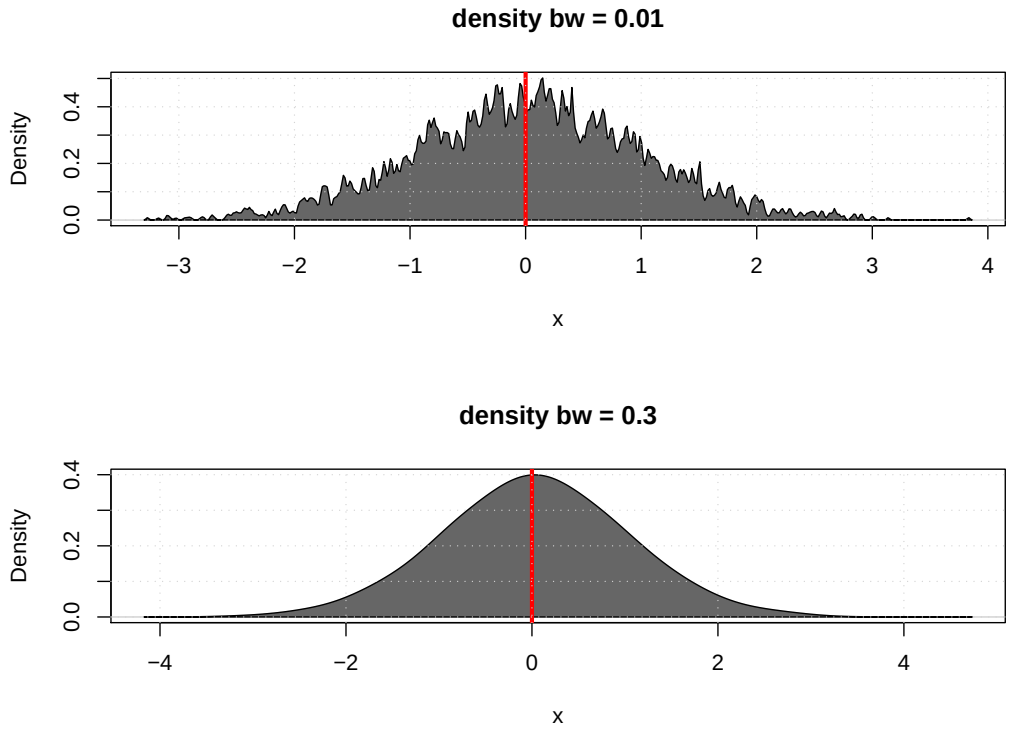

FIG. 2.3: Density plots (pdf) of the same random sample of Fig. 2.1 with different values of bandwidth. Note the same phenomena of oversampling when the bandwidth is too small.

2.3 Quantile methods

2.3.1 Definitions

Let $x_1 < x_2 < \ldots < x_n$ a reordered sample if we divide the interval $[x_1, x_n]$ in m classes the value of x belonging to the class k is called the k-th *quantile*. For instance consider the sample

$$1, 2, 4, 5, 7, 9, 13, 21, 22$$

divide the interval in 5 parts corresponding to $0\%, 25\%, 50\%, 75\%, 100\%$ of the interval of values: the corresponding quantiles will be $1, 4, 7, 13, 22$. Note that the value $7 \equiv 50\%$ is the median of the distribution, the computed mean is $m = 9.33$ then the distribution is not symmetric (in the case of symmetric distributions the mean is the center of the distribution and by definition it coincides with the median). Furthermore the value $22 \equiv 100\%$ is the maximum while $1 \equiv 0\%$ is the minimum.

It is common in statistics to characterize a distribution of random variables reporting not only the mean and the variance but also giving the extreme quantiles (the 95th of

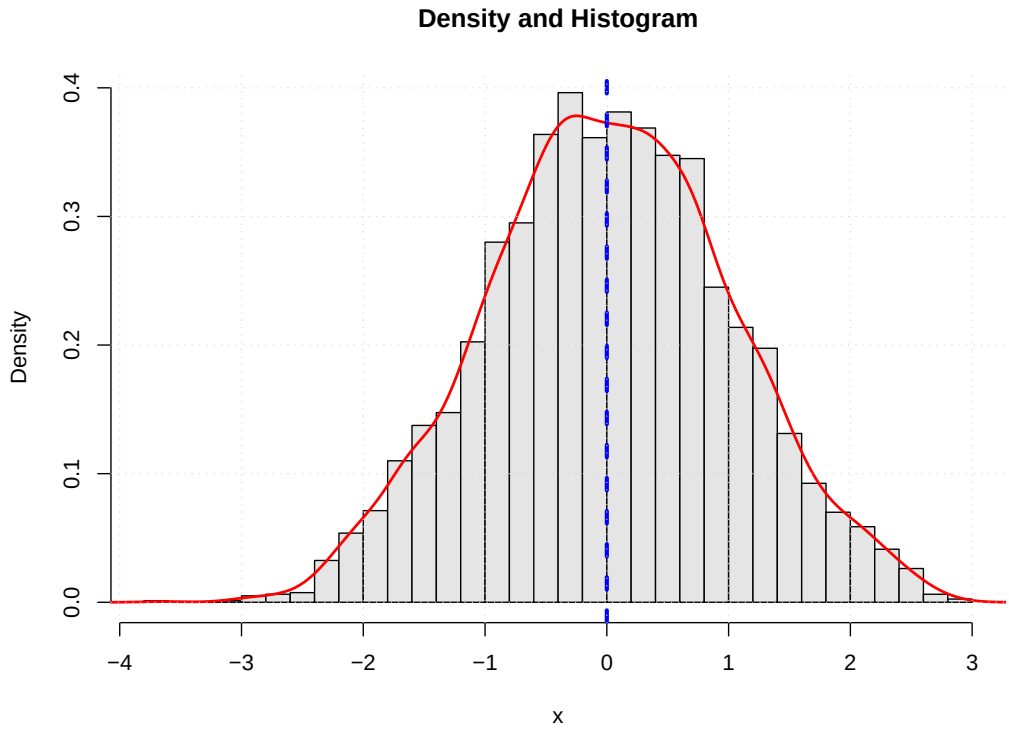

FIG. 2.4: Density plot (pdf) and Histogram superimposed of the same random sample of Fig. 2.1

100, the 99th of 100 and sometimes the 999th of 1000) and the maximum (minimum). Remember that the $50th$ (or 50%) quantile is the median.

2.3.2 Quantile functions

The quantile level binds the probability (ex. $50\% = 0.5$) to the value of the random variable x. The mathematical problem is the same of inverting the cumulative probability distribution $p = F(x)$, computing the quantile function $F^{-1}(p)$ we obtain:

$$x = F^{-1}(p) \quad p \in [0, 1] \tag{2.3}$$

where the variable p is the probability or the quantile *level* (also known as the p-value) and x is the quantile *value*. The 2.3 expresses the relationship between the probability and the values of a distribution.

For the Normal distribution the quantile function is defined by:

$$x(p) = \mu + \sigma\sqrt{2}\, Z(2p - 1) = N^{-1}(p), ; \quad p \in [0, 1] \tag{2.4}$$

where $Z(p) = erf^{-1}$ is the inverse of the error function

$$erf\,(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt. \tag{2.5}$$

the numerical (or the analytical) approximation of erf^{-1} could be found in many textbooks of statistics and probably in every statistical package.

2.3.3 Quantile plots

The quantile plot is the tool to visually check if an empirical distribution belongs to a known family of theoretical distributions. Suppose for instance to make the hypothesis that data are distributed like a Normal distribution. The fit of a Normal distribution is straightforward: just compute the mean m and the standard deviation s of the sample and substitute them in the quantile function of the Normal N^{-1} 2.4 on the place of the theoretical mean μ and standard deviation σ. The quantile function allows to create another set of theoretical values based on the probability levels:

$$x'_j = N^{-1}\left(p(x_j)\right), \quad j = 1, \ldots, n$$

where $p(x_j)$ is the probability level of x_j. The plot of the set of the theoretical (fitted) $\mathbf{x}'$ versus the original sorted sample $\mathbf{x}$ is the *quantile plot* (see Fig. 2.5).

The quantile levels by definition increase with the values of x and usually the graph is displaced around the straight diagonal line dividing the I and III quadrant of the Cartesian coordinates plane. Observing the quantile plot 2.6 (also called *qq-plot*) we can deduce the following properties:

- If the points are nicely placed on a straight diagonal line then the chosen distribution for the fit is working good.
- The dispersion of data around the diagonal line give us an indication of the goodness of the fit and of the sample variance: an elevated variance spreads the points in the plane, a low variance concentrates the values around the line.
- In the extreme regions of the graph the points fluctuate more and they leave the straight line [2].

[2] this is a common phenomena in exploratory statistics where the incertitude in the tail estimation is due to the low level of probability of extreme values and it depends also by the choice of ordinary tools that privilege the *core* of the distribution. The *robust* statistical tools could be conceived to work better with the tails giving a bigger (or a smaller) importance to extremal data if desired.

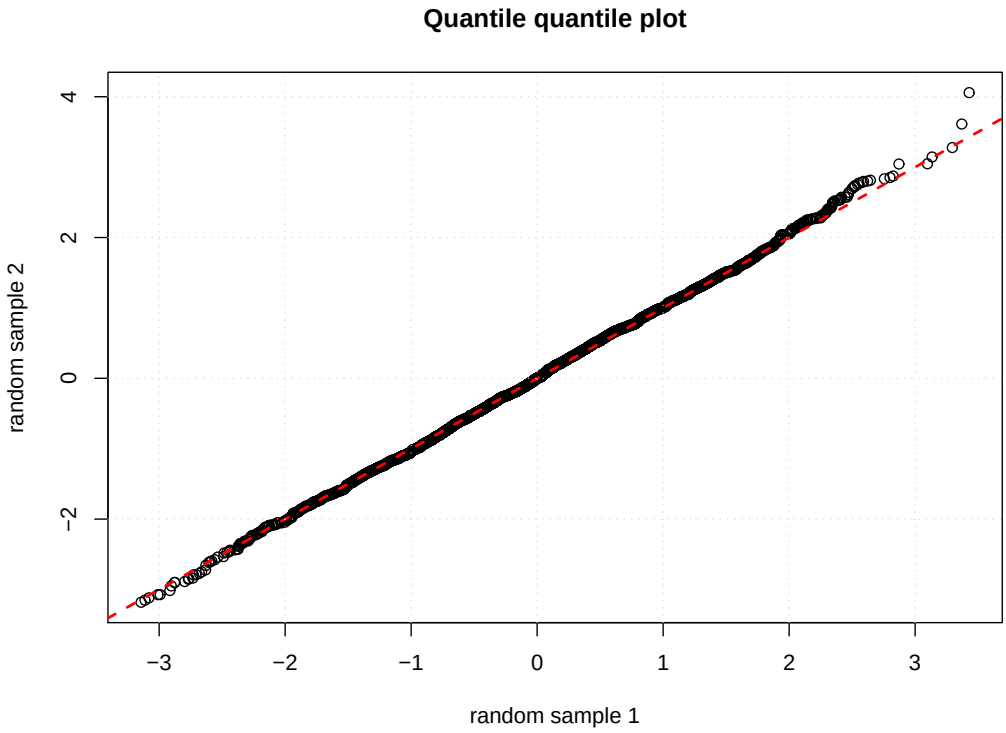

FIG. 2.5: Quantile-quantile plot of two Normal random samples with the same parameters σ, μ.

The quantile plot is not limited to the Normal distribution but it is possible to apply it on every CDF that we can invert analytically or numerically. In the next chapter we'all apply the qq-plot to a database of precipitation making the hypothesis that the GPD (Generalized Pareto Distribution) is the best (or one of the best) statistical model to describe the average behaviour of the precipitation.

2.4 Confidence intervals and bootstrap methods

In Physics, Engineering and Science it is common to give the outcome value of an experiment with the relative errors: a measure of the incertitude in expressing the physical quantities. For instance

$$G = 6.693 \cdot 10^{-11} \pm 0.027 \cdot 10^{-11} m^3 kg^{-1} s^{-2}$$

is the latest (Science pag. 74, issue of the 5th January 2007) best known value of the gravitational constant G. The value of the constant is expressed with a symmetric confidence interval. In most cases the error of a measure is a symmetric interval, that is there is the same probability to find the *true* unknown measure below of the proposed

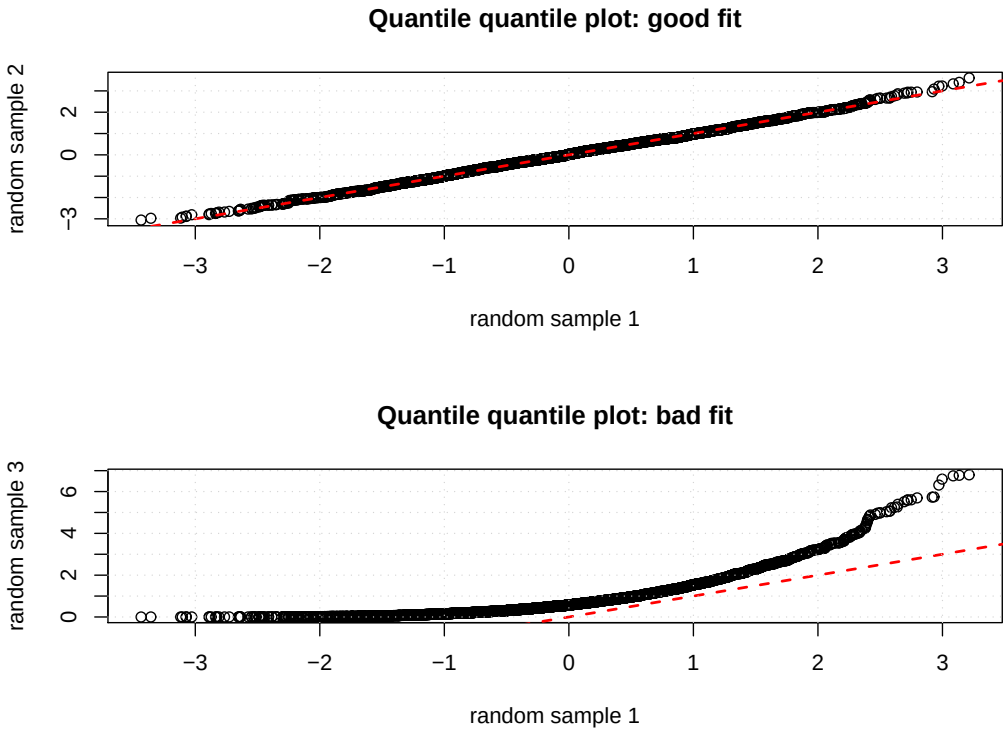

FIG. 2.6: Quantile-quantile plot in case of good fit (top) or bad fit(bottom) of the statistical model. In the bad case we confront an Exponential dataset with a Normal one.

value or above of it. In the EV theory quantities like wind absolute speeds, precipitation, flood levels are strictly positive with monotonic decreasing pdfs then the probability of greater values around the measure is different than the probability of smaller ones. The correct representation of a measure for non symmetrical distributions is the *interval of confidence* $a < X < b$ with an associated *confidence level*. This quantity expresses the interval (a, b) in terms of probability. Recalling the previous sections on quantiles we can express the confidence interval as a quantile interval, for instance choosing $a \equiv 5\%$ and $b \equiv 95\%$ the interval $[a, b]$ summarizes the 90% of the variability of the sample and we affirm that the measure X, $a < X < b$ is expressed with the 90% of *confidence*.

Calculating the confidence interval is not difficult if we can answer in a correct way to the question: what is the theoretical distribution of data ? In some case we are confident of the answer. For instance if our experiment produces values that are the means $\mathbf{m}$ of *independent* subsets of data $\mathbf{x}_i$ then these means tend to converge to the theoretical Normal mean μ and they are distributed as a Normal random variables whatever be the *true* unknown distribution of data [3]. In this case (data arise from a

[3]This result is known as the **law of large numbers** and has two versions a) the *strong* version that affirms the absolute convergence of the empirical mean m to the theoretical value μ b) the *weak* version

Normal distribution) we can express the confidence interval using the theoretical result described in every statistics textbook:

$$-2s + X \simeq -2\sigma + X < x < X + 2\sigma \simeq X + 2s. \tag{2.6}$$

where s is the empirical standard deviation that approximates σ the theoretical value. The relative confidence level is of the 95% (also known as the 2σ *rule*).

2.4.1 Bootstrap methods

In many cases it is possible to calculate the confidence interval for the theoretical distributions. Whenever the adaptation of the theoretical distribution to the data distribution (the fit procedure) is good the theoretical equations for the confidence intervals on estimated parameters (like the mean and the variance) are enough to describe the statistics. Otherwise if we are not confident on the statistical model to use with we can try to derive the confidence interval directly from the data with no other hypothesis than the outcomes of an experiment (a random variable) can appear more than once in the sample. In other words we assume the possibility that the empirical random values can be duplicated and replaced (in order to conserve the original length of sample) *without* changing the statistical properties of the sample. In this way we create *perturbed* copies of the original sample each one with one or more data randomly replaced. This hypothesis is the basis of the *Bootstrap* procedure that belongs to the family of Monte Carlo methods [4].

The bootstrap technique (re-sampling with substitution of the initial set of data) can be used to build confidence intervals. For instance, if we need to compute the confidence level of the mean for the dataset $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ we can use the following procedure:

- create k datasets

  $$\mathbf{x_j} \quad j \in [1, k]$$

  by randomly replacing one (or more than one) value x_i with another one extracted from the same dataset $\mathbf{x_j}$.

- compute the mean of each perturbed dataset $\mathbf{x_i}$, let m_i this value.

that affirms that the mean converges *in probability* to μ. The strong version of this law is followed by another strong second order hypothesis: the finiteness of the variance of every subset of data. This last condition is known as *Kolmogorov's strong law* see Sen and Singer (1993)

[4] the Monte Carlo methods use random number generators to simulate statistical phenomena like random walks, point processes, energy minimization, numerical multidimensional integral calculation and random variables generation.

- compute the empirical 5% and 95% quantiles of the distribution of the means $F(m)$ and use these values as confidence interval.

Note that if the sample originates from a non symmetric distribution then the confidence interval computed with the bootstrap procedure will be non symmetric too.

The advantage of the bootstrap procedure is that we do not assume any hypothesis on the distribution of real data, but we derive the variability directly from the original dataset. Note that this method in itself is not capable to give us *more* information than which contained in the original dataset, that is increasing the number of bootstrapped datasets we reduce the random fluctuations but we not have any *new* information on the true distribution that originates the random variables.

2.5 Outliers and scatter plots

One of the simplest analysis of an one dimensional dataset that we can perform is the scatter plot diagram. This technique consists in plotting the data using as abscissa the index (not to be confused with the ordered rank position) of each value and as ordinate the value itself. Data are not rearranged, sorted or modified. The plot is a point plot where the position of a point in the dataset is the x coordinate and the value the y coordinate. On the vertical axis (the axis of values) we can draw a line indicating the mean and one or more quantile levels (see Fig. 2.7). If we note that several points are far from the others and stay on the extreme quantiles region we can suspect that these ones are *outliers* or even that these values are erroneous (for instance they derive from a wrong transcription).

The concept of outlier is an open question of the modern statistics, when a value must be considered an outlier or conversely a big but ordinary value it is not clear. Sometimes we are confident about the nature of a point strongly exceeding the other values. For instance if we have a temperature field on the Earth land surface and in the dataset it is present the value 12000 C then we can no doubt consider this point as an error and correct it. Other examples from Earth Sciences are negative precipitations or values representing non physical quantities like a daily cumulated rainfall of 10 km.

2.5.1 Tests for outliers

Working with outliers we often have no clear evidence of the nature of these values: we cannot simply decide to eliminate them without the risk to loose important informations

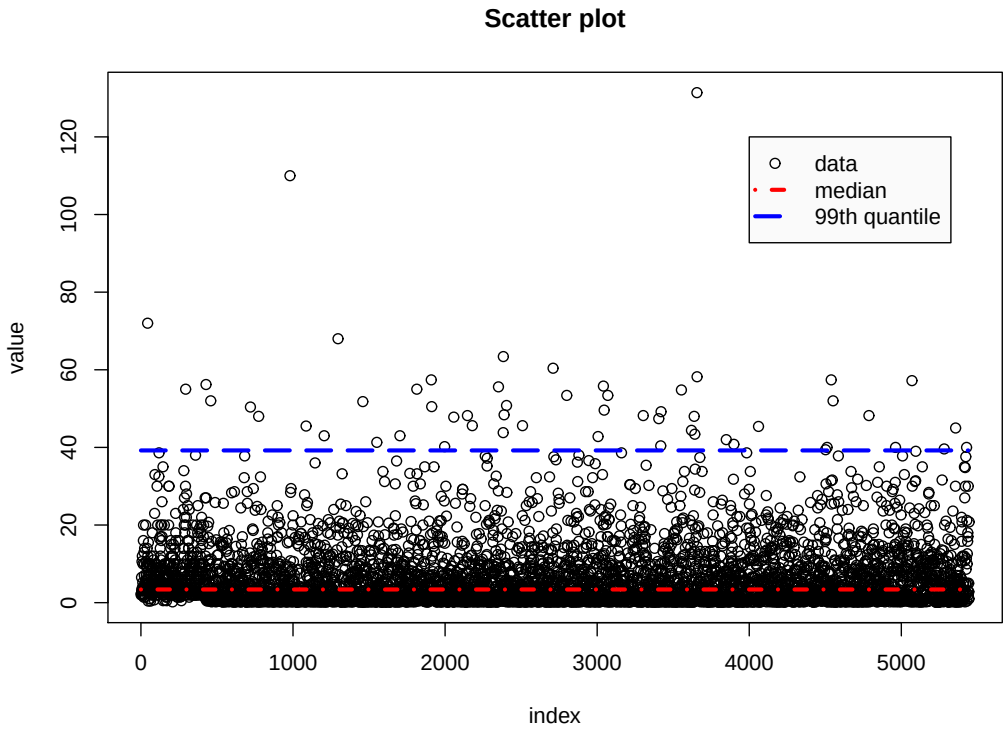

FIG. 2.7: Scatter plot with the 98th quantile and the median level lines superimposed. The plot shows two possible high outliers.

of the dataset. In this case we can use two different approaches: trying to detect the outliers and perform further analysis, or ignore them and use statistical techniques resistant to outliers. The latter solution known as Robust Statistics is briefly presented in the next section. Here we describe several classical automatic tests for outlier detection. Note that visual inspection of data is possible only when we work with small databases.

We follow the suggestion of the NIST (part of the US. commerce department) software for statistics and its Engineering Statistics Handbook (freely available online at `http://www.itl.nist.gov/div898/handbook/`. They cite the reference Iglewicz and Hoaglin (1993) . The handbook presents this techniques to work with outliers

- **outlier labelling** - flag the detected outliers for further investigation (i.e. are the potential outliers erroneous data ?, indicative of an inappropriate distributional model ? and so on).
- **outlier accommodation** - use robust statistical techniques that are resistant to the effects of outliers.
- **outlier identification** - test whether observations are outliers.

the first approach is known as *masking* the outliers. If we have clear indications about the distribution of data (i.e Normal distribution) the outlier detection is simple, just remove or mask the values that have very low probability ($> 3\sigma$). If we do not have indications about the distribution of data we can still try to accommodate outliers (with robust statistics methods), in this case outliers are still present but they do not affect our estimations. Finally we can decide to perform some test for the deviation of these values from the centrality measures (mean, median). The idea is that if a value is too far from the central measure it will be candidate as outlier.

Working with data we can try to mask a single point in order to detect one or more outliers. This technique is nicely described in the handbook:

> Masking can occur when we specify too few outliers in the test. For example, if we are testing for a single outlier when there are in fact two (or more) outliers, these additional outliers may influence the value of the test statistic enough so that no points are declared as outliers. On the other hand, swamping can occur when we specify too many outliers in the test. For example, if we are testing for two or more outliers when there is in fact only a single outlier, both points may be declared outliers (many tests will declare either all or none of the tested points as outliers). Due to the possibility of masking and swamping, it is useful to complement formal outlier tests with graphical methods. Graphics can often help identify cases where masking or swamping may be an issue. Swamping and masking are also the reason that many tests require that the exact number of outliers being tested must be specified.
>
> Also, masking is one reason that trying to apply a single outlier test sequentially can fail. For example, if there are multiple outliers, masking may cause the outlier test for the first outlier to return a conclusion of no outliers (and so the testing for any additional outliers is not performed).

One of the most practical test for outliers is the Grubbs test (Grubbs, 1950). It is based on the test statistics:

$$Z_i = \frac{\max |x_i - m|}{s} \tag{2.7}$$

where x_i is the set of data, m the mean and s the standard deviation of the sample. The critical levels of Z are tabulated and presented in many statistical software (`R`, `S-plus` for instance). The quantity 2.7 is called Z-score and several authors (Iglewicz and Hoaglin, 1993) propose to substitute it with a robust version:

$$Z_i = \frac{0.6745(x_i - \nu)}{MAD} \tag{2.8}$$

where ν is the median, and MAD the mean absolute deviation 2.10 . A rule of thumb for the 2.8 is that a value will be considered an outlier if its Z-score value is 3 times greater than the previous value in the Z-score ordered vector. Note that this test exists in different versions (one sided for skewed distributions, two sided for symmetric distributions) to detect one or more outliers with symmetric and asymmetric distributions.

2.6 Short notes on Robust Statistics

The Robust Statistics is a discipline of the modern statistics based on the idea that the statistical tools must be resistant to the presence of outliers or more generally to data contamination. A classical introductory example of the robust statistics is the different behaviour of the mean and of the median in response to an outlier (see for instance the review paper of Daszykowski et al. (2007)).

2.6.1 The median and other robust estimators

Consider a set of random variables:

$$1, 2, 4, 5, 7, 9, 13, 21, 22.$$

with the median $\nu = 7$ and mean $m = 9.33$. Now let change the value 13 and double it, the mean is now 10.33 while the median has always the same value 7. A single value modification has changed the value of the mean while the median remained unchanged.

The mean then is a non robust estimator in fact it has a *breakdown point* (percentage of points to be modified to significantly change the value of the estimator) around 0% of the sample. Conversely to influence the median one needs to contaminate around the 50% of the values. The breakdown point cannot pass the 50% value because if more than the half of data is contaminated the original distribution of data is completely lost.

A common measure of dispersion is the standard deviation

$$s = \sqrt{\frac{1}{n-1}\sum_{i \leq n}(x_i - m)^2}$$

where m is the mean and $n - 1$ is a correction for the bias due to the estimation of the mean m. If the mean is a non robust estimator it is obvious that neither the standard

deviation could be a robust estimator. For this reason other measures of dispersion were introduced, the most common ones are:

- **The interquartile difference**. This estimator is described as the difference from the 3th and the 1th quartile, a *quartile* is the 25% quantile, its formula is:

$$IQ = q_{3/4} - q_{1/4} \tag{2.9}$$

  where the $q_{3/4}$ is the quantile level at $3/4 = 75\%$ of the sample and $q_{1/4}$ at 25% .

- **The median absolute deviation MAD**. This estimator computes the absolute difference between the values and the median:

$$MAD(x) = \frac{1}{n}\sum_{j=1}^{n} |x_j - \nu| \tag{2.10}$$

  if ν is the median

2.6.2 The trimmed mean

The last robust estimator that we present is the trimmed mean $m_{X\%}$ indicating that the mean is trimmed at the level $X\%$. The idea of this estimator is to trim the lower $X/2\%$ and the upper $X/2\%$ of the sample and then compute the mean. We can express this estimator with the formula:

$$m_{X\%} = \frac{1}{k}\sum_{i=1}^{k} x_i \qquad if \;\; x_i \in [\min(x) + X/2, \max(x) - X/2] \tag{2.11}$$

Note that the sum ends at the index $k < n$, that is the number of x points falling in the interval $[\min(x) + X/2, \max(x) - X/2]$. By construction the breakdown point of the trimmed mean is the trim level $X\%$.

2.6.3 The efficiency of an estimator

The robust estimators are attractive, they are able to resist to the outliers and they are quite simple to calculate. However there is a problem with these estimators: they are scarcely efficient. To clarify this concept we need to introduce an inequality known as the Cramér-Rao inequality (in honour of H. Cramér and C.R. Rao that first derived it in 1945), this inequality fixes a lower bound to the variance of the estimators. More

precisely it states that for an unbiased estimator θ the variance is bounded to the Fisher information function by the inequality

$$Var(\theta) \geq \frac{1}{I(\theta)} \tag{2.12}$$

where $I(\theta)$ is the Fisher Information function:

$$I(\theta) = E\left\{\left(\frac{\partial \ln(L)}{\partial \theta}\right)^2\right\} \tag{2.13}$$

where L is the likelihood function (see in the section 2.8 for more details). The information function gives the total amount of information that a sample can furnish on an unknown estimator θ, and it fixes a limit on the precision of the estimate. Finally we can define the *efficiency* of an estimator as the rate of its variance with the inverse of the Fisher information function:

$$\eta = \frac{Var(\theta}{1/I(\theta)} \tag{2.14}$$

The rate between the efficiency of two estimators t_1 and t_2 of an unknown parameter θ is the *relative efficiency* and can be expressed as:

$$\eta_r = \frac{E(t_1 - \theta)^2}{E(t_2 - \theta)^2} \tag{2.15}$$

If the variance of the estimator t_1 is less than the variance of the estimator t_2 then the first estimator is more performing than the other one. It is possible to demonstrate that the variance of the MAD and the variance of the median are greater than the respective measures of dispersion (s standard deviation) and location (m mean): the price to pay for robustness is a loss of efficiency.

2.7 Critical remarks on estimators

The tools introduced in this section allow to describe the distribution of data in terms of robust, efficient, and biased/unbiased estimators. The reader however must be careful to avoid common mistakes in exploratory analysis. Let consider for instance the sample:

$$1, 2, 3, 4, 20, 21, 22, 22$$

the mean of this sample is 11.875, the median (a robust estimator) 12, the standard deviation 10.07. But it is clear that no one of these estimators have true sense, in fact

there is no data in the interval (4,20). A simple plot (Fig. 2.8) of data shows that the points are clustered in two subsets centred on 2.5 and on 21, in this case it is possible (although not certain) that data arise from *two* different populations one for the lower values and the other one for upper values. A smarter statistics will divide the sample in two samples and will give us information about each subset.

This simple example illustrates the necessity of the exploratory statistics *before* any application of even the most robust estimators to the datasets. If we suspect to have more than one population in our sample several tests to separate the populations in clusters are available (like the PCA, the principal component analysis).

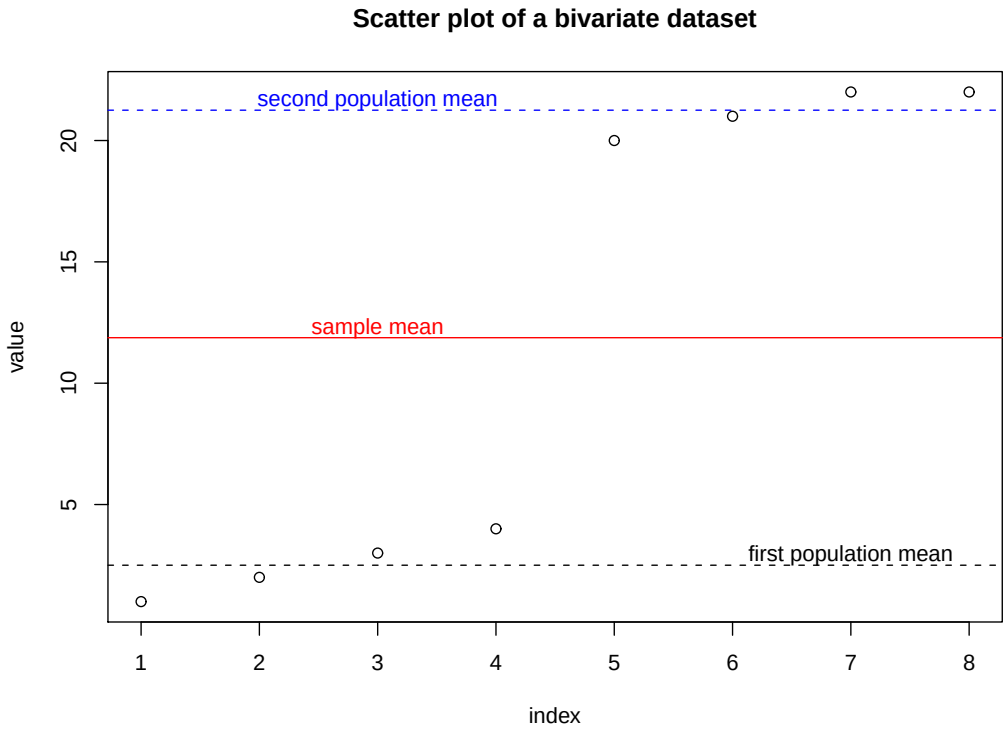

FIG. 2.8: Scatter plot of a bivariate sample dataset.

2.8 The Maximum Likelihood function and estimator

The Maximum Likelihood function (ML) - indicated as $L(x, \theta)$ - and the subsequent Maximum Likelihood Estimator (MLE) technique, are based on the simple idea to find the best set of parameters θ of a distribution $F(\mathbf{x}; \theta)$ that maximize the probability to find again the set of data $\mathbf{x}$. In other words the ML expresses the best parameters for the given set of data.

Let now $f(x_1;\theta), f(x_2;\theta), \ldots, f(x_n;\theta)$ the probability levels of $x_1, x_2, \ldots, x_n$ where $f(x,\theta)$ is the density function of F. If the values are independent the ML function is the product of joint probabilities:

$$L(\mathbf{x};\theta) = f(x_1;\theta) \cdot f(x_2;\theta) \cdot \ldots, \cdot f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta) \tag{2.16}$$

taking the logarithm of $L(\mathbf{x};\theta)$ we can convert the product into a sum:

$$\log\left(L(\mathbf{x};\theta)\right) = \log\left(\prod_{i=1}^{n} f(x_i;\theta)\right) = \sum_{i=1}^{n} f(x_i;\theta) \tag{2.17}$$

The logarithm is a monotone increasing function then the maxima of $\log(L)$ are the same of L. To find the best set of θ that maximize the joint probability L we use the classical result of analysis defining the *score* functions as derivatives:

$$V(\theta) = \frac{\partial\big(\log L(\mathbf{x};\theta)\big)}{\partial\theta} = \frac{1}{L(\mathbf{x};\theta)}\frac{\partial L(\mathbf{x};\theta)}{\partial\theta} \tag{2.18}$$

and imposing the condition $V(\theta) = 0$. Note that to calculate the maxima of 2.17 we need differentiable Likelihood functions and by consequence differentiable density functions $f(x;\theta)$. This condition is not always respected, we can try to use numerical optimisation but sometimes the maxima of the likelihood function are non existent [5]. Obtaining the set of values θ that optimises L is equivalent to perform a good fit of the theoretical distribution of the data.

The most interesting properties of the estimators based on the Maximum Likelihood are:

- **Consistency**. The parameters estimated with the ML converge in probability to the true unknown parameters of the distribution.
- **Asymptotic Normality**. The parameters estimated with the ML tend asymptotically to be distributed like a Normal distribution.
- **Efficiency**. The ML estimators are asymptotically equivalent to the Cramér-Rao inequality bound [6] then the MLE is asymptotically the best possible estimator or

[5] For instance the score function respect to the threshold parameter u of the GPD distribution is unbounded with no candidate values for the maxima. This is the reason that forces us to use different techniques than the fit to estimate the optimal threshold level.

[6] See also the section 2.6.3

more precisely no asymptotically unbiased estimator has lower asymptotic mean squared error than the MLE.

these properties are valid in the asymptotic limit: for real datasets it is quite common that other estimators perform better than the MLE.

2.9 Tools for EV model choice

In this section we introduce a powerful tool (the L-moments ratio diagram) to guess the best distribution to describe the data. Indeed we have theoretical results showing that the limiting distribution of the excesses is the GPD and the limiting distribution of maxima is the GEV, but in exploratory statistics it is an useful exercise to assess the nature of a dataset *without* any prior hypothesis on the data distribution. In the next chapter we'll apply the L-moments ratio diagram to ensure that the GPD is the best candidate to describe the distribution of the excesses for two different databases of precipitation. For reference purposes we introduce now other distributions having an heavy tail,they may be considered competitors of the GPD/GEV in the description of the extremes.

2.9.1 Heavy tail distributions

The Generalized logistic distribution

This distribution is used in Hydrology to describe the extremes (see for instance Ashkar and Mahdi (2006)), it is similar to the GEV. The logistic distribution (a reduced version of the GLD) could be obtained studying the differences of two random variables distributed as a Gumbel random variable: see Kotz and Nadarajah (2004) and Kotz and Nadarajah (2001) for a detailed review. The form chosen for the GLD equation is reported in the web page
`http://docs.scipy.org/doc/scipy/reference/tutorial/stats/continuous.html` [7].

$$f(x;c) = \frac{ce^{-x}}{[1+e^{-x}]^{c+1}} \quad c>0, \;\; x>0 \tag{2.19}$$

where c is a positive parameter and $f(x;c)$ is the pdf.

[7] we choose this page dedicated to an important statistical software since the mathematical form of the distribution is ready to use for computational purposes.

The Log Normal distribution

The Log Normal distribution is the ordinary Normal distribution where the random variable x is substituted by the logarithm $\ln(x)$, the mathematical form of the pdf is:

$$f(x;\sigma,\mu) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}, \qquad x > 0 \tag{2.20}$$

note that in contrast with the Normal distribution the x variable appears outside the exponential and it is strictly positive.

The Pearson Type III distribution

This distribution has a skewed shape and its pdf is:

$$f(x;c,s) = s(x+c)^{c^2-1}e^{-xc} \quad c \geq 0. \tag{2.21}$$

references for this distribution can be found in the page from `mathworld` (the website of the creator of the Mathematica software):

`http://mathworld.wolfram.com/PearsonTypeIIIDistribution.html`

2.9.2 L-moments theory

Hosking (1990) proposed the use of L-moment ratio diagram for the choice of the distribution that best represent sample data. Sillitto (1969) derived first L-moments (calling them with a different name), as coefficients in the approximation of a quantile function by polynomials. The probability weighted moments (PWM) defined by Greenwood et al. (1979) are precursors of L-moments. Specifically, the PWM's of a continuous random variable X with cumulative distribution function F are the quantities:

$$M_{p,r,s} = E[X^p \{F(X)\}^r \{1-F(X)\}^s] = \int_0^1 \{x(F)\}^p F^r \{1-F\}^s \, dF \tag{2.22}$$

where $x(F)$ is the inverse of F and p, r, s are real numbers. When $r = s = 0$ the 2.22 reduces to the expectation of simple moments of order p. Usually the PWM's are used letting $p = 1$ and one of the other exponents (s or r) equal to zero:

$$\begin{aligned} \alpha_s &= M_{1,0,s} = E[X\,\{1-F(X)\}^s] \\ \beta_r &= M_{1,r,0} = E[X\,\{F(X)\}^r] \end{aligned} \tag{2.23}$$

When fitting a parametric distribution to a data set, it is a common practice to estimate parameters by equating the sample moments to those expected for the fitted distribution. In a similar manner also PWM's can be used with this aim with many empirical and theoretical advantages (Hosking et al. (1985); Hosking and Wallis (1987); Hosking (1990)).

The following linear combinations have been proposed for unbiased estimation of sample PWM's Landwehr et al. (1979); Hosking and Wallis (1995)):

$$\begin{aligned} a_0 &= \frac{1}{n}\sum_{j=1}^{n} x_j \quad ; \quad a_s = \frac{1}{n}\sum_{j=1}^{n} \frac{(n-j)(n-j-1)\cdots(n-j-s+1)}{(n-1)(n-2)\cdots(n-s)} x_j \\ b_0 &= \frac{1}{n}\sum_{j=1}^{n} x_j \quad ; \quad b_r = \frac{1}{n}\sum_{j=1}^{n} \frac{(j-1)(j-2)\cdots(j-r)}{(n-1)(n-2)\cdots(n-r)} x_j \end{aligned} \tag{2.24}$$

where a_s and b_r are sample estimators of α_s and β_r PWM's respectively, x_j are sample values arranged in increasing order and n the sample size.

Several linear combinations of PWM's can be directly interpreted as measures of location, scale and shape of probability distribution. These are the L-moments defined by Hosking (1990). The first sample L-moments can be computed with the following linear combinations of (sample) PWM's:

$$\begin{aligned} \ell_1 &= a_0 && = b_0 \\ \ell_2 &= a_0 - 2a_1 && = 2b_1 - b_0 \\ \ell_3 &= a_0 - 6a_1 + 6a_2 && = 6b_2 - 6b_1 + b_0 \\ \ell_4 &= a_0 - 12a_1 + 30a_2 - 20a_3 && = 20b_3 - 30b_2 + 12b_1 - b_0 \end{aligned} \tag{2.25}$$

whose coefficients are those of the "shifted Legendre polynomials".

The first L-moment ℓ_1 is the sample mean, a measure of location, the second L-moment ℓ_2 is (a multiple of) Gini's mean difference statistics: a measure of the dispersion of the data values around their mean. By dividing the higher-order L-moments by the dispersion measure, we finally obtain the L-moments ratios:

$$\tau_r = \frac{\ell_r}{\ell_2} \tag{2.26}$$

As the simple moments ratios, L-moments ratios in equation (2.26) are dimensionless quantities that characterize the behaviour of the distribution. Specifically, τ_3 is a measure of skewness and τ_4 is a measure of kurtosis: they are respectively referred to as the *L-skewness* and *L-kurtosis*. They take values between -1 and +1 (values smaller than -1 may exceptionally arise from some even-order L-moment ratios computed from very small samples).

The L-moment analogue of the coefficient of variation is the L-CV (it can only take values between 0 and 1) that is defined by:

$$\tau = \frac{\ell_2}{\ell_1} \tag{2.27}$$

Hosking (1990) derived the first expected L-moments and L-moments ratios as function of parameters for several distributions. In the next chapter we'll use this ratio diagram to identify the best distribution candidate for hydrological datasets.

2.10 Fit methods for the EV distributions

The following section is dedicated to the methods used to fit the theoretical distribution on the real data. The fit procedure consists in finding the best values of the distribution parameters that describe the sample. It is also possible to describe the fit procedure in terms of probability of finding the best set of parameters θ that model the data; this approach is expressed by the theory of the Maximum Likelihood estimator (see 2.8 for an introduction and below for the GPD/GEV case). We focus our attention on the fit of the GPD because, as we'll see in the next chapter, this distribution seems to be the best candidate to describe our data. However we stress here the concept of *best fit* because the fit performances change in function of our needs in exploratory statistics. For instance if we are more interested in fitting the tail of the distribution we can use an weighting function $w(x)$ to add importance to the tails, conversely if we are more interested in fitting the core of the distribution we can discard (trim) the extreme tails.

Several tools described below are conceived to be robust (but still quite efficient) in order to deal with outliers and erroneous data, other like the MLE are the most efficient but only in the asymptotic limit. We recall again the principle of the exploratory statistics: never trust naively to data and numerical tools.

2.10.1 Summary of fit methods for the GPD

In last years many estimators have been developed to improve the efficiency and the robustness of the fitting techniques. The estimators can be of several classes: tail index (as Hill or De Haan asymptotic estimators), maximum likelihood functions, simple moments and probability weighted moments, medians and goodness of fit based estimators. Every estimator class has some drawbacks and some advantages. The maximum likelihood estimator MLE is thought on the maximization of the likelihood function L: this method is very important because it is asymptotically (i.e. for large samples) the best. Nevertheless when the sample is small or is contaminated by spurious data it could be very unrealistic. In a classical work Hosking and Wallis (1987) compare the MLE performances of the methods of moments and probability weighted moments. They show how the MLE could give inaccurate results for small samples. Juárez and Schucany (2004) describe the performances of the MLE estimator on contaminated GPD samples showing that the MLE method lacks of robustness; then they introduce a robustified MLE method known as minimum divergence power density(MDPD) estimator.

The tail index estimator class finds the asymptotic slope of the distribution tail using a plot position rule of log transformed data. The method is useful for a large variety of distributions like Gamma, Gumbel, Student and GPD. By combining the tail index with other order statistics it is possible to estimate the GPD parameters as shown in Pickands (1975). The estimator has generally good performances for large samples of pure data, but must be tuned using a suitable threshold. Moreover the numerical algorithms for tail index estimation of GPD often suffer of convergence problems.

The accuracy of the GPD fit depends on the kind of data and must be demonstrated every time. In fact, as shown by Rosbjerg et al. (1992), it is possible that, for moderate tails (values of the shape parameter near zero) and small datasets, the description with the ordinary exponential distribution may be more accurate of the GPD one. This is an expected result since the exponential distribution requires the estimation of a single parameter, while the GPD needs the estimation of two parameters. Obviously similar results can be obtained, more generally, by comparing the GPD performances in the case of ξ known and α unknown with the case of both parameters unknown.

For the numerical implementation of the first three widely applied methods the reader is referred to Hosking and Wallis (1987) and Stedinger et al. (1993). For the other (and more recent) methods, the references are provided in the paragraph of each method.

- **Maximum Likelihood Estimator (MLE)**. The MLE is a standard parameter estimation technique applicable to any statistical distribution. It is based on the idea to find the set of θ parameters that maximize the likelihood function $L(X, \theta)$ evaluated on the sample X (for more details see 2.8) . We remark that the location parameter u could not be obtained by the MLE, in fact the score function $\partial L/\partial u$ is unbounded (no extremes). For the GPD distribution the MLE could be estimated in different ways leading to slightly different results. The MLE can be evaluated in an univariate way (with a smart substitution of scale and shape parameters: see Grimshaw (1993)) or in the classical bi-variate way. Many authors proved that the MLE is the best estimator in presence of large samples ((Hosking and Wallis, 1987)), the asymptotic behaviour of this method is known to be the best possible. But for small samples (≤ 100 values) the fit is not always good and the method is outperformed by other techniques ((Hosking and Wallis, 1987)). Moreover another problem of the MLE is that sometimes the numerical algorithm used to estimate the maxima fails to converge to the local maxima. A robust and powerful computational method must be used to find the maxima avoiding convergence problems. The MLE estimator is described by the log of the maximum likelihood function: $L(\mathbf{x}; \alpha, \xi) = \sum_{i=1}^{n} \log f(x_i; \alpha, \xi)$:

$$L(\mathbf{x}; \alpha, \xi) = \begin{cases} -n \log \alpha - \dfrac{(1+\xi)}{\xi} \displaystyle\sum_{i=1}^{n} \log\left(1 + \xi \frac{x_i}{\alpha}\right), & \xi \neq 0 \\ \\ -n \log \alpha - n \dfrac{\bar{x}}{\alpha} & \xi = 0 \end{cases} \tag{2.28}$$

  the derivatives by respect to ξ and α for fixed u are the *score* functions. Solving the system of equations of the score functions permits to find the values of ξ and α. Numerical methods are required to maximize the L function. Some numerical improvements, useful to estimate parameters with MLE, was proposed by Grimshaw (1993).

- **Moments estimator (MOMENTS)**. It represents the simplest GPD estimator, it is based on the mean and the variance of the distribution. The method is theoretically applicable only for values of $\xi < 1/2$ since for $\xi \to 1/2$ the variance tends to be infinite. Hosking and Wallis (1987) suggest to use the MOMENTS

estimator for $\xi < 1/4$ where the estimator is unbiased. When $\xi \approx 0$ the accuracy of the method is close to the MLE estimator. Nevertheless we must note that the moments method is very sensible to outliers, in fact the mean and the variance are statistical quantities lacking of robustness: a single outlier could dramatically change all these quantities. Given the mean

$$m = \frac{1}{n}\sum_i x_i$$

and the variance of data

$$s^2 = \frac{1}{n-1}\sum_i (x_i - m)^2$$

the relation with the shape and scale parameters is

$$\xi = \frac{1}{2}\left[\left(\frac{m^2}{s^2}\right) - 1\right], \tag{2.29}$$

$$\alpha = \frac{1}{2}m\left[\left(\frac{m^2}{s^2}\right) + 1\right] \tag{2.30}$$

The method is useful for values of $\xi < 1/4$ an valid when $\xi < 0.5$, outside this interval the variance s^2 is infinite and the method is not applicable. For values of $\xi \approx 0$ the accuracy of the method is comparable to the MLE estimator. Confidence intervals for ξ and α are given by Hosking and Wallis (1987).

- **Probability Weighted Moments estimators (PWMU and PWMB)**. The probability-weighted moments PWM were introduced by Greenwood et al. (1979) and represent an alternative to the ordinary moments. As for the MOMENTS estimator, the parameters can be expressed as a function of PWMs. The PWM estimator is particularly advantageous for small datasets because the probability weighted moments have a smaller uncertainty than the ordinary moments. The best performance is reached for $\xi \approx 0.2$ with light breakdown for negative shape values. Hosking and Wallis (1987) give two definitions of PWM, unbiased (PWMU) and biased (PWMB), but the difference can be detected only for small samples.

  The probability-weighted moments PWM are defined by

$$M_{p,r,s} = E\{X^p[F(X)]^r[1-F(X)]^s\} = \int_0^1 [x(u)]^p u^r (1-u)^s du \tag{2.31}$$

where E is the expectation value, and $F(X)$ a distribution of a random variable X, $x(u)$ is the quantile function. Linear combinations of the quantities

$$\alpha_r = M_{1,0,r} = \int_0^1 (1-u)^r x(u) du \tag{2.32}$$

$$\beta_r = M_{1,r,0} = \int_0^1 (u)^r x(u) du \tag{2.33}$$

are used to define the L moments; introduced by Hosking (1990) they are directly related to the measures of scale, location and shape of the probability distribution. For the GPD the L-moments of shape and scale parameters are obtained by the simple relations:

$$\xi = \frac{m}{m-2w} - 2, \tag{2.34}$$

$$\alpha = \frac{2wm}{m-2w} \tag{2.35}$$

where

$$w = \frac{1}{n}\sum_{j=1}^{n}(1-\pi_j X_{(j)}), \quad \pi_j = \frac{j+\gamma}{n+\delta}, \quad j = 1, 2, ..., n$$

and $X_{(1)}, X_{(2)}, ..., X_{(n)}$ are the order statistics of random variable $\mathbf{X}$. Note that π_j is a plotting position rule. The empirical values $\gamma = -0.35$ and $\delta = 0$ are given by Hosking and Wallis (1987) and they are chosen with the purpose to minimize the bias of the estimator specially for small samples. The method is usable for values of ξ less than 0.5. Confidence intervals can be found in Hosking and Wallis (1987). In the same paper two versions of the PWM estimators are proposed: the PWMU (unbiased estimator) and PWMB (biased estimator). The main difference is in the definition of w:

$$w_r = \frac{1}{n}\sum_{j=1}^{n}\frac{(n-j)(n-j-1)...(n-j-r+1)}{(n-1)(n-2)...(n-r)}X_j$$

with $r = 1$. The bias of the estimator is present only for very small samples ($n < 50$), asymptotically the PWMU and the PWMB are perfectly equivalent.

- **Maximum penalized likelihood (MPLE)**. Coles and Dixon (1999) introduced a weight function for the maximum likelihood function $L(X, \theta)$ valid for $\xi > 0$. This estimator corrects the tendency of MLE to diverge for small samples.

The authors introduced a penalizing function for ML defined by:

$$P(\xi) = \begin{cases} 1 \quad if\ \xi \leq 0 \\ exp\left[-\lambda\left(\frac{1}{1-\xi} - 1\right)^{\alpha}\right] \quad if\ 0 < \xi < 1 \\ 0 \quad if\ \xi \geq 1 \end{cases} \tag{2.36}$$

where the authors suggest $\alpha = \lambda = 1$. Note that for $\xi \leq 0$ the MPLE corresponds to the ordinary MLE.

- **Minimum density power divergence (MDPD)**. This robust estimator has been introduced by Juárez and Schucany (2004) and it is derived from the MLE using a special function of divergence between the fitted function and the data. A constant is introduced to control the trade-off between robustness and efficiency. This property could be very attractive when the datasets are contaminated. The robust estimator defines the *density power divergence* between the densities $f = f(X;\theta)$ and $g = g(X)$ defined as

$$d_{\alpha}(g, f) = \int_{\chi} \left\{ f^{1+\beta}(x) - \left(1 + \frac{1}{\beta}\right) g(x) f^{\beta}(x) + \frac{1}{\beta} g^{1+\beta}(x) \right\}\ dx \tag{2.37}$$

where for fixed $\beta > 0$ the minimum density power divergence is defined as the point in the space of parameters $\theta \in \Theta$ that minimizes the distance between the f function and the empirical density g. The parameter θ could be obtained minimizing with respect to θ the previous formula expressed for the f family:

$$H_{\beta}(\theta) = \int_{\chi} f^{1+\beta}(x) dx - \left(1 + \frac{1}{\beta}\right) \frac{1}{n} \sum_{i=1}^{n} f^{\beta}(X_i;\theta) \tag{2.38}$$

finally the estimator for the GPD is obtained minimizing with respect to the shape and scale parameters ξ, α the equation:

$$H_{\beta}(\xi, \alpha) = \frac{1}{\alpha^{\beta}(1 + \beta - \beta\xi)} - \left(1 + \frac{1}{\beta}\right) \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\alpha^{\beta}} \left(1 - \xi \frac{X_i}{\alpha}\right)^{(\frac{1}{\xi} - 1)\beta} \tag{2.39}$$

Juarez and Scuchany proved that the MLE has the highest efficiency when GPD data are uncontaminated, but when data arise from a combination of different distributions then the MDPD is more efficient than the MLE. They suggest to use $\beta = 0.1$ for a good estimate of moderate contaminated samples.

- **Likelihood moment estimator(LME)**.

This method is recent (Zhang, 2007) and was proposed to be a replacement of the PWM and MOMENTS techniques. It is based on the well known relation for the

mean of the excesses:

$$E[(1+bY)^r] = \frac{1}{1-r\xi}, \quad with \quad 1-r\xi > 0 \tag{2.40}$$

where r is a constant (with $r = 1$ and $b = -\frac{\xi}{\alpha}$) and $Y = X - u$ are the excesses. Fixing a parameter r this last equation could be solved by respect to b using numerical methods. Then with the relation derived from the maximum likelihood score function:

$$\xi_{est} = \frac{1}{n}\sum_{i=1}^{n}\log(1 + b_{est}x_i)$$

and the definition of b we can solve it to obtain α and ξ. The method promises to be efficient, robust and simple although it is slow (at least in the R POT implementation). The parameter r must be fixed a priori and Zhang suggests to use a value of $r = 0.5$.

- **Median estimator (MED)**. This estimator is the most CPU time intensive, it is designed to be resistant to outliers (see Peng and Welsh (2001)). Nevertheless for pure GPD data its performances are very poor as shown by Juárez and Schucany (2004). References for the median estimator could be found on Peng and Welsh (2001). They compute the GPD parameter iteratively solving the group of equations

$$\alpha = -\frac{\xi}{2^{-\xi}-1} median\{x_i\} \tag{2.41}$$

$$z(\xi) = median\left\{\frac{1}{\xi^2}\log\left(1-\xi\frac{X_i}{\alpha}\right) - \frac{(1-\xi)X_i}{\alpha\xi - \xi^2 x_i}\right\} \tag{2.42}$$

where $z(\xi)$ is subject to several special conditions and can converge if some starting point ξ, α is provided close to the real values.

Numerical packages containing the implementation of these estimation techniques can be found in the R package POT created and maintained by M. Ribatet (more references in Ribatet (2007)).

2.10.2 Summary of fit methods for the GEV

In this section we'll expose some remarks on the fit methods available for the GEV distribution. The methods for fitting the GEV are similar to those used to fit the GPD. The same theory of PWM and MLE could be found in the cited works of Hosking, or with greater detail for the PWM in Diebolt et al. (2008). Note that the fit of the GEV has not the simple method of Moments, then it is necessary to use numerical tools. For

instance for the PWM method (page 36 of Diebolt et al. (2008)) one needs to solve the system:

$$\begin{cases} \beta_0 = \mu - \frac{\alpha}{\xi}\left[1 - \Gamma(1-\xi)\right] \\ 2\beta_1 - \beta_0 = \frac{\alpha}{\xi}\Gamma(1-\xi)(2^{\xi} - 1) \\ \frac{3\beta_2 - \beta_0}{2\beta_1 - \beta_0} = \frac{3^{\xi}-1}{2^{\xi}-1} \end{cases} \tag{2.43}$$

where the β_r must be computed with the estimator

$$\hat{\beta}_r = \frac{1}{n}\sum_{i=1}^{n}\left(\prod_{l=1}^{r}\frac{j-l}{n-l}\right)X_{j,n}$$

and $X_{j,n}$ are the reordered data of the sample. Note that in the 2.43 we have changed the GEV parameters to be coherent with the notation of this work (in the reference $\xi = \gamma$ and $\alpha = \sigma$).

Chapter 3

Exploring hydrological databases

3.1 Basic Exploratory Statistics

In this chapter we analyse two different hydrological databases of cumulated daily precipitation. The first one is derived from the former Servizio Idrografico Italiano, section of Cagliari Sardegna (Italian Hydrographic service) and consists of 394 stations with 60 years of daily data (from 1920 to 1980) [1] , the second one is an American database freely available from the web site of the NOAA (National Oceanic and Atmospheric Administration) and it consists in 1582 stations situated mainly in US and recorded in a period of 130 years (see below for details).

Our analysis will be an application of Exploratory Statistics to data, avoiding, if possible, any a priori hypothesis with a critical approach to estimators and statistical modelling. The starting point is the description of the databases with basic statistical tools (mean, median, variance) and ensure the quality of data looking for outliers, artefacts or errors in the datasets. Then we'll perform a test to choose the most suitable statistical model. We have theoretical arguments to prefer distributions like the GEV and the GPD for the data description, nevertheless we'll use a test (L-moment ratio diagram) to find *experimentally* the best distribution for the datasets. Finally, in Chapter 4, using the *best* stations among these databases we'll stress the performances of the chosen statistical model .

In detail the analysis will follow this schema:

- Basic exploratory statistics (mean, variance, ...)
- test for GPD/GEV choice

[1] The national service was dismissed in 2002 and its competences passed to the regional authorities

- Graphing techniques to find outliers and geographical distributions
- Selecting a subset of stations to test some estimator
- test for rounded data, test of GOF
- tests for threshold selection

We'll try to make our analysis in parallel side by side for the two databases. The reader is invited to refer to the previous chapters for the theory of the tools used in the analysis.

3.2 The NOAA-NCDC daily precipitation database

The NOAA-NCDC daily rainfall database ("NOAA-NCDC" or simply "NCDC" in the rest of the thesis) is available at

`http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCDC/.DAILY/.FSOD/.PRCP/`.

The archive consists of 1582 weather stations of US army/navy mainly from the United States and parts of South Pacific and Europe. Each station is identified by its WBAN number (Weather Bureau of Army and Navy), all data samples have the same length: $N = 47846\ days = 131\ years$ starting from 1 Jan 1869 to 31 Dec 1999, but only a few stations have more than a century of data. The station list, their geographical coordinates and some useful metadata could be found in the file:

`http://mi3.ncdc.noaa.gov/mi3report/MISC/wbanmasterlist.xls`

The data format is NetCDF (`http://www.unidata.ucar.edu/software/netcdf/`), the precipitation is measured in inches, using units of 0.01 inches. In our analysis we converted the values to mm. Several stations have small series of precipitation data because they worked only during the world wars and then they were dismissed. However 403 stations have more than 3000 rainy days. The table (3.1) summarizes some general basic statistics on the NCDC database.

The NCDC daily precipitation database have the 90% of stations with a median precipitation lower than 6 mm, the means are more elevated (the difference between median and mean is typical of skewed distributions like the GPD). The maximum daily recorded rainfall is around 1 m of water. The 90th quantile of the precipitation is around 200 mm of daily cumulated precipitation indicating that strong rainfalls are not unusual in

Qlevel	elevation	length	nozeros	mean	median	90th quantile	max
0%	-13.12	675	405	1.31	0.51	11.89	13.21
10%	9.15	1960	613	3.53	1.78	46.18	51.31
20%	18.91	3255	834	4.58	2.03	61.72	68.58
30%	39.04	4843	1199	5.68	2.54	77.04	85.60
40%	84.48	6267	1752	6.63	3.05	91.67	101.85
50%	149.15	8028	2268	7.78	3.56	108.81	120.90
60%	228.14	11381	2910	8.78	4.06	124.13	137.92
70%	311.40	14369	3746	9.69	4.57	141.27	156.97
80%	471.23	17192	4824	10.96	5.08	170.08	188.98
90%	1124.84	18241	6058	12.33	5.84	209.17	232.41
100%	3049.70	37852	15701	27.85	10.41	845.82	939.80

TAB. 3.1: The NOAA-NCDC database quantiles levels (increasing top-down) for the quantities : the elevation (m) of the stations, the length of the sample and the number of rainy days, the mean (mm), the median (mm), the 90th quantile and the maximum of each station (mm).

the NCDC stations. Note that a 10% of stations are in mountain over 1000 m, while the highest station is at 3050 m.

3.2.1 Outliers check

Following the techniques to find the outliers (cfr. 2.5.1) we perform the check of two outliers per single dataset using the Z-score value test. The criteria will be finding the values that have a Z-score value more than 3 times the previous Z-score value in the sorted sample [2]. For the Z-score computation we use the robust version with median and MAD estimators on the place of the mean and the standard deviation.

[2]We can give to this ratio a more precise meaning. If we make some hypothesis on the probability of the excesses (theoretically a GPD) we can compute, using Monte Carlo techniques the probability that for finite samples the Z-score ratio between the two greatest values will be more than 3. In the appendix A we perform this computation and give a table of confidence.

WBAN	rainy days	max(mm)	value ratio	Z-score ratio
13729	1462	857.25	7.09	7.30
13807	2607	534.92	6.42	6.66
14609	2249	812.80	6.30	6.44
14895	889	584.20	6.71	6.89
14911	3162	762.00	3.00	3.01
15620	7406	655.32	6.62	6.81
22604	6838	647.70	3.48	3.51
23072	1044	773.43	14.85	15.42
23160	969	762.00	13.16	13.66
24132	3863	939.80	21.14	21.98
26407	568	121.92	2.96	3.06
34051	1610	226.06	5.46	5.78
3969	1925	762.00	9.26	9.55
94224	6059	375.16	3.09	3.19
94728	5794	685.80	5.35	5.49

TAB. 3.2: The potential outliers of the NOAA-NCDC database. The first column is the WBAN code of the station, the second the number of rainy days in the time series, the third the maximum of the station, the fourth the rate of two extreme maxima, and the fifth the Z-score ratio of the two extreme values

We found 15 potential outliers, however only the values from the stations (23072, 23160, 24132, 3969) are possible outliers; note that the Z-score ratio between the two extreme values of each time series is greater than 9, that is the outlier is more than 9 times the rest of the sample.

In a special case (station 24132) we have other informations that allow to confirm that the big value (almost 1 m of daily cumulated precipitation) is an error. The station is in Bozeman Montana, another database (with a graphical output) at the web page

`http://www.wrcc.dri.edu/cgi-bin/cliMAIN.pl?mtboze`

from the Western Regional Climate Centre reports for this station no more than 2.6 inches = 66 mm in a period of time covering 1892 - 2005. Then we can be confident that the value exceeding 21 times the mean precipitation (or 14 times the maximum data from the other database) is an error (probably a transcription error) and can be removed (see also the Fig. 3.1).

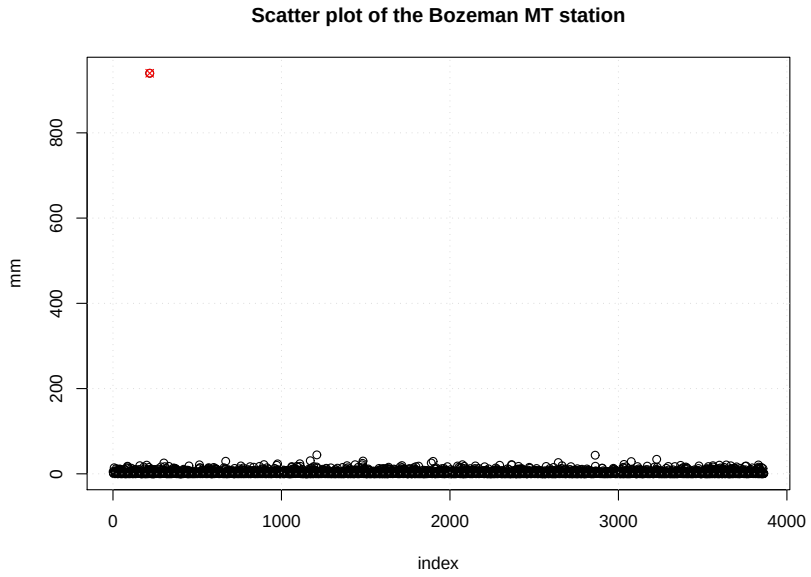

FIG. 3.1: Daily cumulated rainfall for the Bozeman (MT) station, with the 939 mm outlier marked in red. Note that the value is actually 14 times greater than the next maximum.

A part from these specific cases (4 severe outliers) we are confident that using the entire database and over 400 station with at least 3000 rainy days, the influence of similar outliers will be weak, specially using robust tools like median and MAD that not suffers of the presence of outliers [3].

3.2.2 Geographical distribution of heavy rainfalls

The geographical distribution of heavy rainfalls (Fig. 3.2) is quite complex but it is clear that the tropical and oceanic effects (hurricanes in the Gulf of Mexico) are responsible of the strongest events. In fact the daily cumulated averages are very high in the tropical or subtropical region of US decreasing from south to north along the

[3] in the rest of the work we will use the median to estimate the shape and the scale parameter for the GPD model specially in the cases where the correct determination of the threshold it is not a trivial task.

Mississippi: from Alabama to the Great Lakes region. The Rocky Mountains stop the rain and in this region there's no severe events recorded on this database (cfr. also Fig. 3.3 where the 90th quantile identify the rainiest stations). Finally the pacific coast is more rainy, specially the Washington state and the northern California.

A confrontation with the historical mean annual cumulated precipitation (source Oregon Climate Service) in Fig. 3.4 shows that the behaviour of the median daily cumulated rainfall is similar to the annual averages: rainy regions have high daily precipitation and this rule is valid also for severe events expressed with the quantiles.

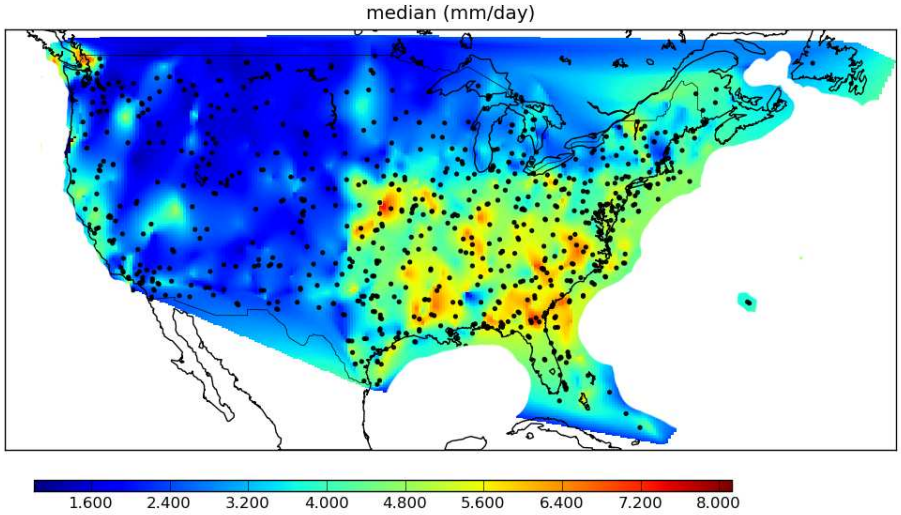

FIG. 3.2: The median (mm/day) of the daily cumulated precipitation for the NOAA database in the US region. The black points are the stations. Note the geographical patterns,the rainy regions are near the Gulf of Mexico,while the Rocky mountains stop the rain.

TECHNICAL NOTE. The geographical plots are obtained using a triangular tessellation of the plane such as Delaunay triangles. The missing values are interpolated linearly from the nearest neighbours available on a regular grid.

3.3 The Sardinian daily precipitation database

The Sardinian database ("SARD" in the the following chapters) is formed by 394 stations covering all the island with regularity (see the distribution of the black points in the Fig. 3.5). The classes of precipitation for median, quantiles and maxima are in the table 3.3.

Only the 10% of stations is above 700 m (mountain) with the highest station at 1071 m. The mean is always greater than the median confirming that the distribution is right

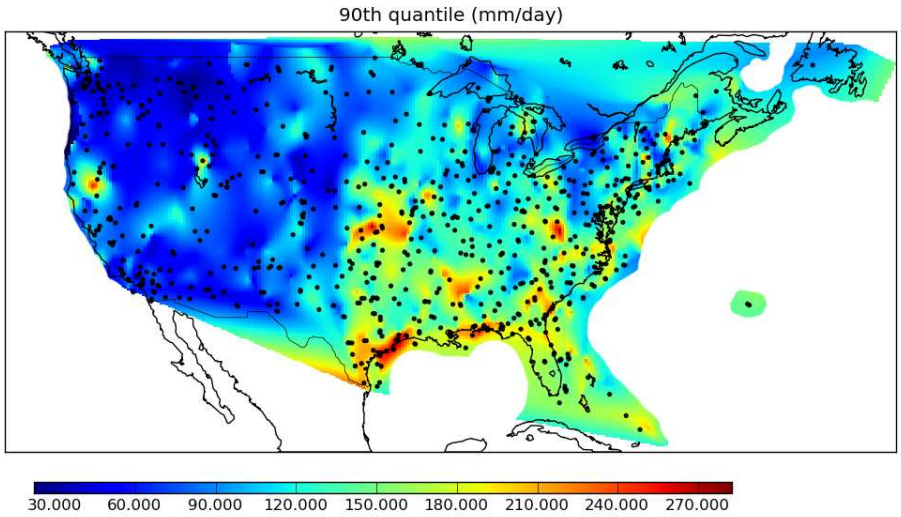

FIG. 3.3: The 90th quantile (mm/day) of the daily cumulated precipitation for the NOAA database in the US region.The black points are the stations. The rainiest stations have also the strongest events but the pattern it is not always respected.

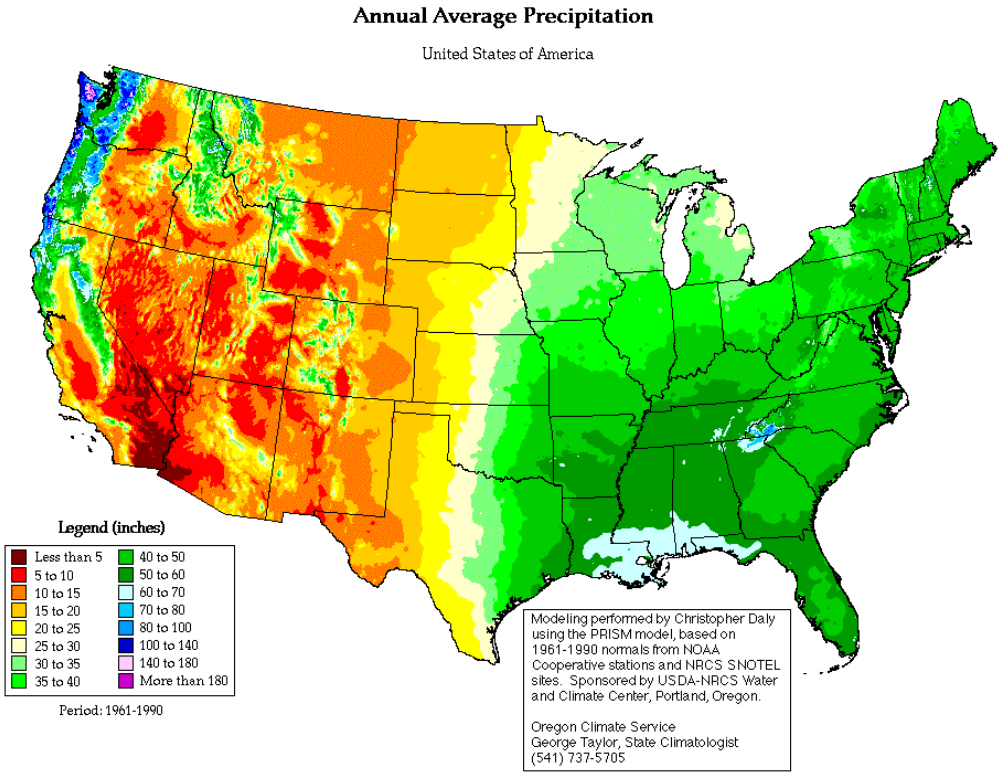

FIG. 3.4: The US averaged precipitation from the Oregon Climate Service. Note that this plot is the annual average (computed in a century) and not the daily cumulated average as the other plots.

Qlevel	elevation	length	nozeros	mean	median	90th quantile	max
0%	1	366	50	4.52	1.20	28.80	32
10%	17	3575	803	6.51	3.20	69.80	77.56
20%	50	6244	1256	7.35	4	80.86	89.84
30%	107.40	7233	1724.60	8.03	4.50	89.06	98.96
40%	186.40	10940	2572.80	8.58	5	97.54	108.38
50%	272	15645	3092	9.27	5.20	108	120
60%	350	17547	3369.60	9.84	6	126	140
70%	438.20	17988	3674.20	10.43	6.30	144.04	160.04
80%	559.80	18197	4212	11.26	7	177.59	197.32
90%	714.20	18263	4684.40	12.33	8	254.30	282.56
100%	1071	21550	7869	15.98	11	489.60	544

TAB. 3.3: The Sardinian database quantiles levels (increasing from top to bottom) for the quantities : the columns are the elevation of the stations (m), the length of the sample and the number of rainy days, the mean (mm), the median (mm), the 90th quantile (mm) and the maximum (mm) of each station.

skewed, like in the NOAA database there is a 10% of stations with daily cumulated precipitation with more than 200 mm of water. Finally the 90% of the stations have at least 800 rainy days.

3.3.1 Outliers check and geographical distribution

The check of the outliers follows the same rules of the NOAA analysis. But in this case we have a single station with a Z-score ratio greater than 3: the station 47 (Uta (CA)) has the maximum value of 400 mm of daily cumulated and the previous value 134 mm with a Z-score ratio of 3.01. We do not consider this value a true outlier [4].

The picture of the 90th quantile (see Fig. 3.6)gives us a clear indication of the climate in Sardinia. The eastern regions exposed to the influence of the Thyrrenian sea are the

[4]The station of Uta, Cagliari was interested in a severe flood in the 12th of November, 1999 when the daily cumulated rain was of 375 mm (see the report compiled by the regional agency `http://www.sar.sardegna.it/pubblicazioni/periodiche/analisi_10.2008-04.2009.pdf`)

rainiest and with the most extreme events. Another place of extreme events is the Gulf of Cagliari and particularly the region of Capoterra interested in several floods in the last years with daily cumulated rainfalls greater than 300 mm . Note that the median daily cumulated rainfall (Fig. 3.5) is not reflecting the climate of the island: in this case the internal regions have high values of daily cumulated that not coincide with the mountains where, as expected, the averaged annual precipitation is greater (see the report of the regional agency available at

`http://www.sar.sardegna.it/pubblicazioni/periodiche/analisi_10.2008-04.2009.pdf`

for a detailed discussion of the climate in the Sardinia island).

3.4 Empirical cumulative distributions

The empirical cumulative distributions (ECDF) and the probability density plots (cfr. respectively the section 2.2.1 and 2.2.2) are the standard tools to analyse the shape of a probability distribution. We apply these techniques to both databases but plotting only the most interesting cases of the (1976 = 1582 + 394) stations.

3.4.1 NOAA-NCDC database

In the NOAA-NCDC database we check the ECDF using the semilog plot of the survival function $y = 1 - F(x)$ (where $F(x)$ is the empirical cumulated frequency) in order to magnify the behaviour of the data on the distribution tail. We highlight the following cases of candidate errors and candidate outliers.

3.4.1.1 Station 24132 (Bozeman MT)

This station is the one described in the outliers check, the ECDF plot (Fig. 3.7) in y semilog scale confirms the presence of the outlier on the far right lower region of the plot.

3.4.1.2 Station 13873: Athens Clarke GA

The station 13973 of Athens Clarke, Georgia has an ECDF semilog plot (Fig. 3.8) with a straight diagonal line. This indicates a low shape parameter, that is the station could be described with an exponential distribution (in this case the fitted GPD parameter gives $\xi = 0.013$ close to the exponential form of the GPD $\xi = 0$, see the GPD definition on the first chapter).

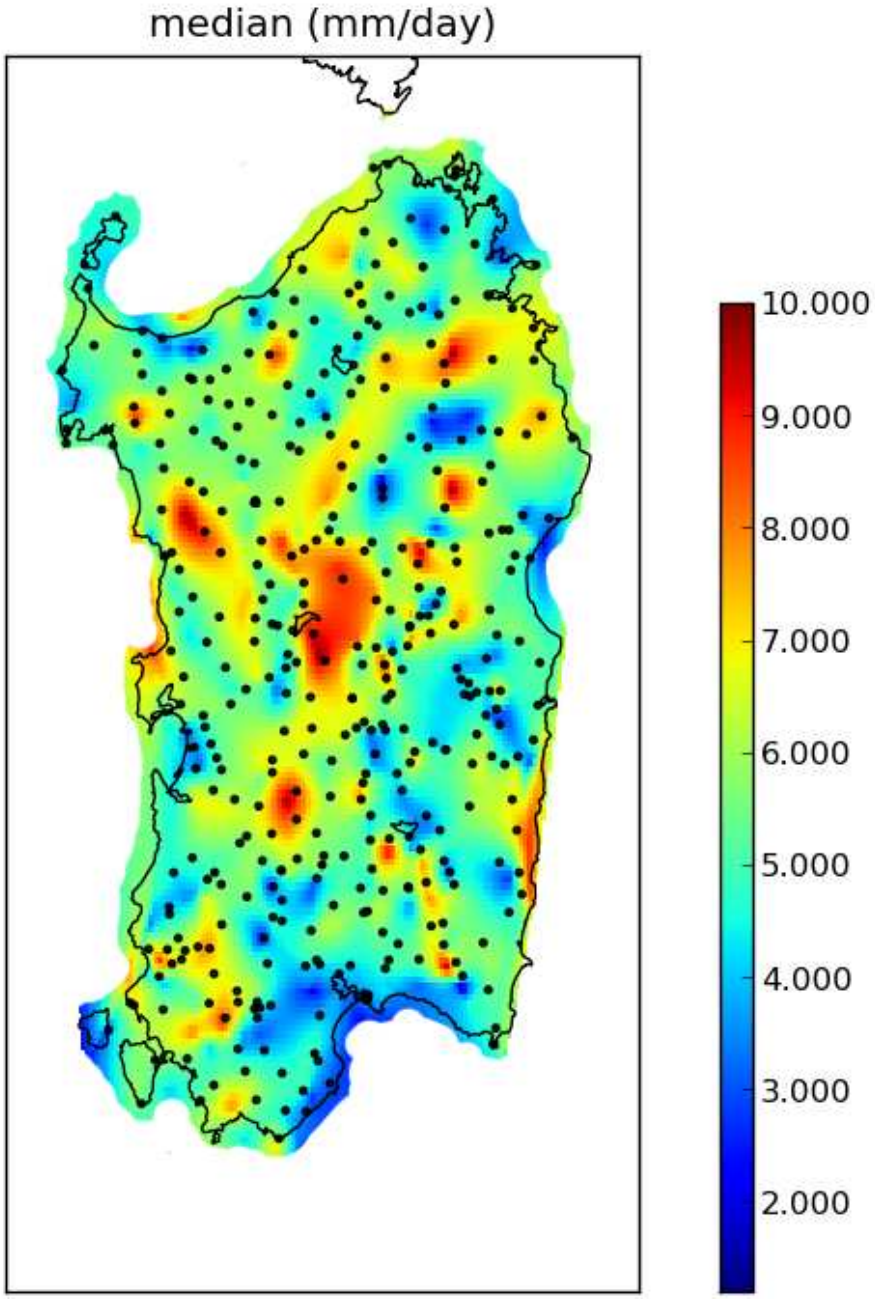

FIG. 3.5: The median (mm / day) of daily cumulated rainfalls in Sardinia. The black points are the stations.

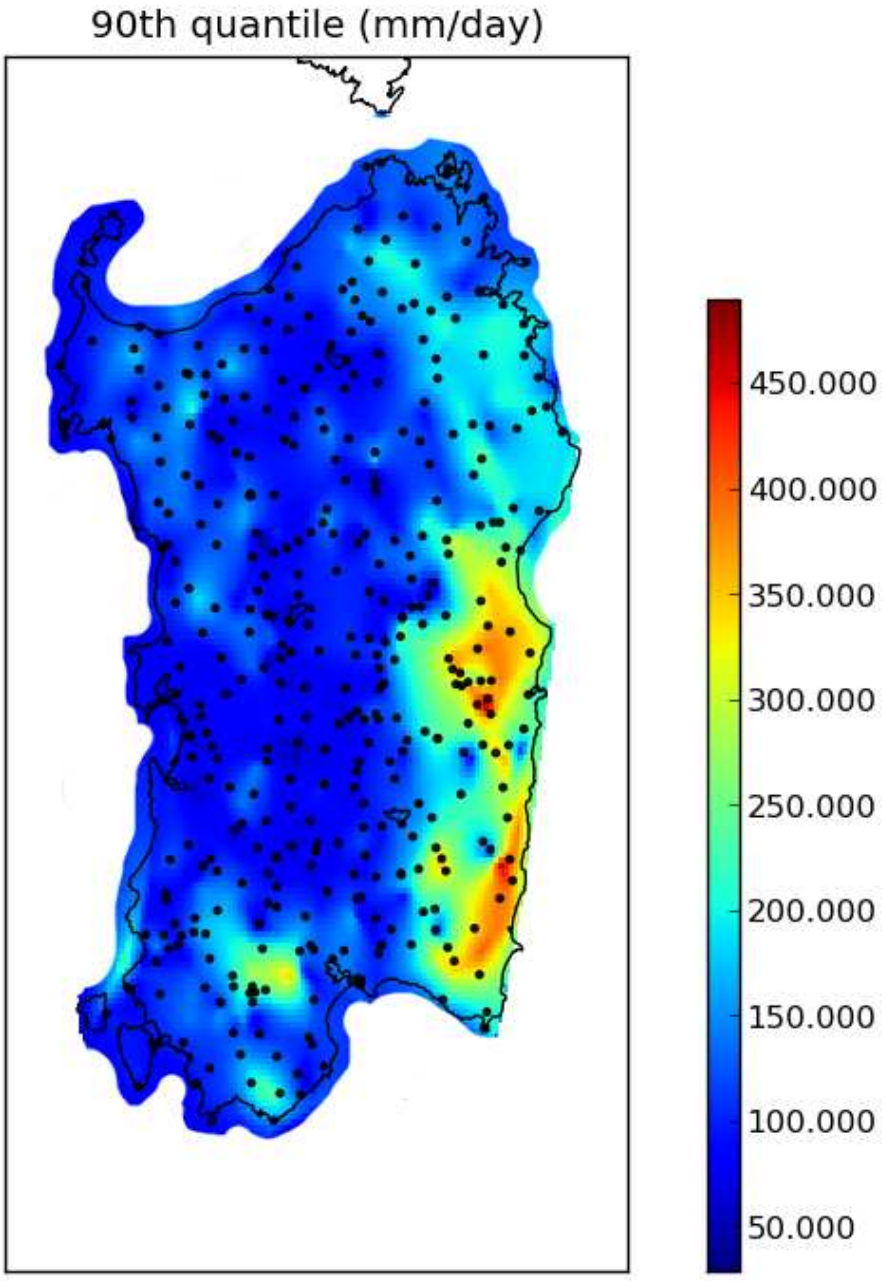

FIG. 3.6: The 90th quantile (mm/day)of daily cumulated rainfalls in Sardinia. The black points are the stations.

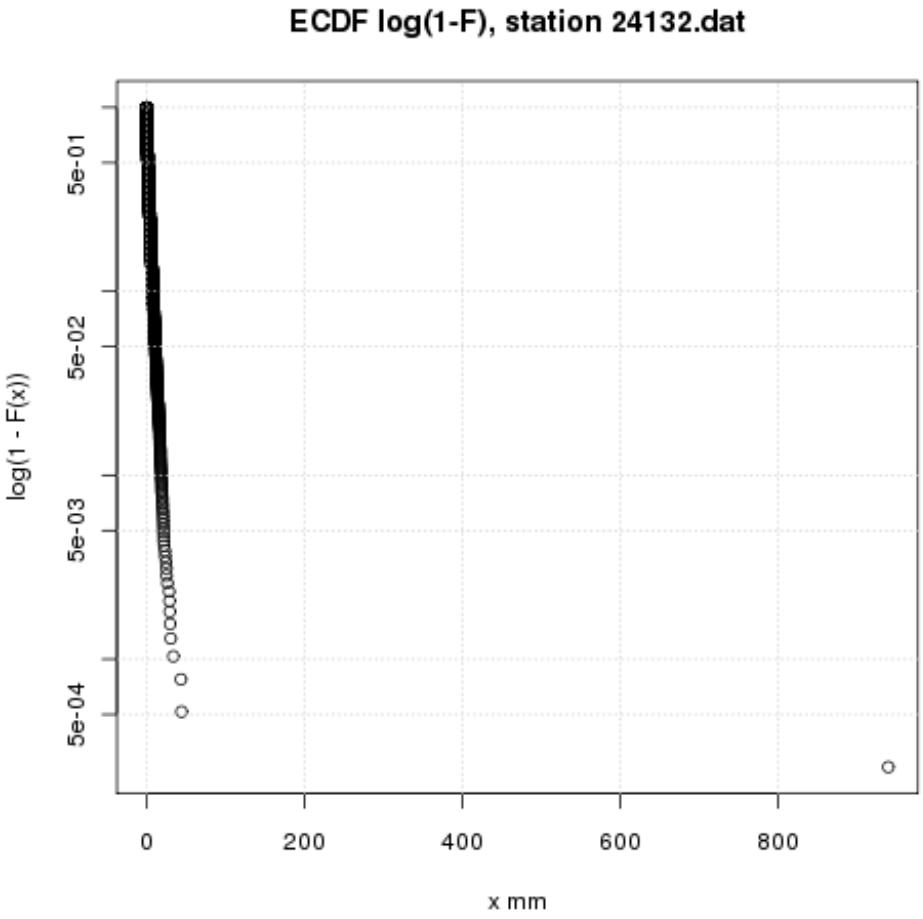

FIG. 3.7: ECDF in semilog scale of the Bozeman station with the candidate outlier on the extreme right region of the plot.

3.4.1.3 Station 24103: Dugway proving grounds UT

The station of Dugway proving grounds (Utah) has a quite strange ECDF semilog plot (Fig. 3.9). The plot shows a S curve with a set of detached points on the right of the plot that have a rapidly decreasing frequency. The form of the ECDF suggests the presence of more than one population, one for the lower data, the other one for the higher.

A check of the scatter plot of the station (Fig. 3.10) and the successive detail (Fig. 3.11) helps to understand the situation. The bigger population forms a cluster of 30 values (Fig. 3.11) with daily precipitation above 100 mm for 30 consecutive days. The Utah region is semiarid and far from the tropical zones exposed to Monsoon (not present in the continental US territory) where strong continuous precipitations are possible. In this case 30 days of consecutive heavy rain are certainly an error in the database.

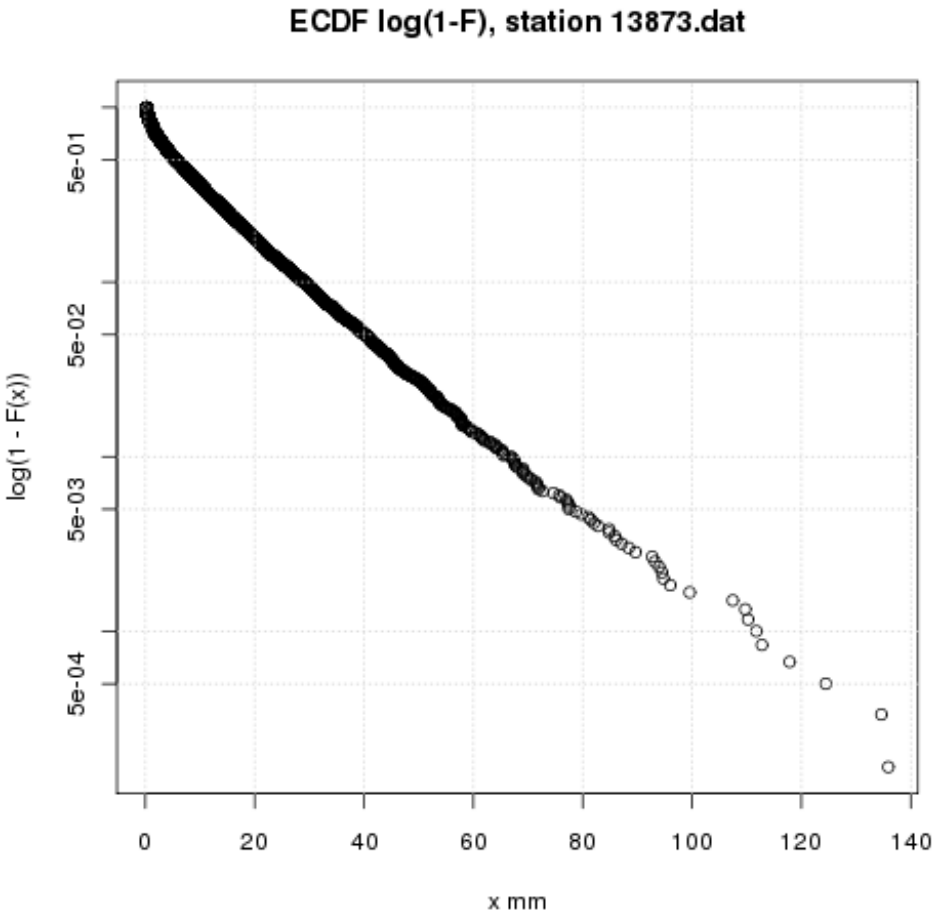

FIG. 3.8: ECDF in semilog scale of the Athens Clarke station. The GPD model could be substituted by a simple exponential law (shape parameter close to zero).

3.4.2 Robust statistics on contaminated data: an example

Using the data from the Dugway station we can try to investigate the power of robust estimators: median and MAD versus mean and standard deviation. We compute the influence of the data contamination on the estimator introducing the formula:

$$\eta_r = \frac{|\theta_c - \theta|}{\theta}$$

where θ_c is the estimator applied to contaminated data and θ the estimator applied to non-contaminated data. Note that η_r varies between $0/\infty$. For values near zero the estimator is not sensible to the contamination, for high values the estimator is particularly sensible. For the
5 Dugway station the results are summarized in Tab. 3.4 .

note that the median is unaffected by the outliers ($\eta_r = 0$) while the mean is changed by the 53% and the standard deviation more than 80%.

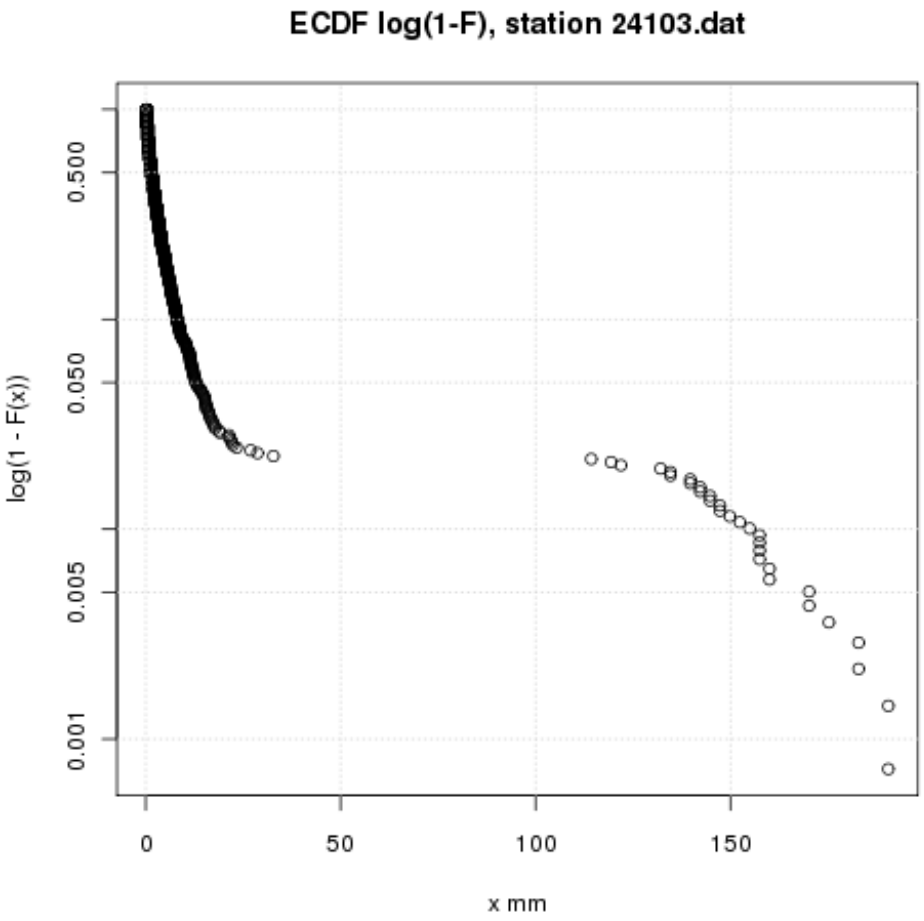

FIG. 3.9: ECDF in semilog scale of the Dugway Proving grounds station. Note the S for of the ECDF that can be interpreted only as a result of two populations one for the lower part of data and the other one for the higher part.

estimator	uncontaminated data	contaminated data	η_r
median (mm)	1.524	1.524	0%
mean (mm)	2.852	6.08	53%
s (mm)	3.71	22.24	83%
MAD (mm)	1.506	1.883	20%

TAB. 3.4: The result of the data contamination (from the Dugway station) for different estimators. Note the performances of the median: totally unaffected by the outlier.

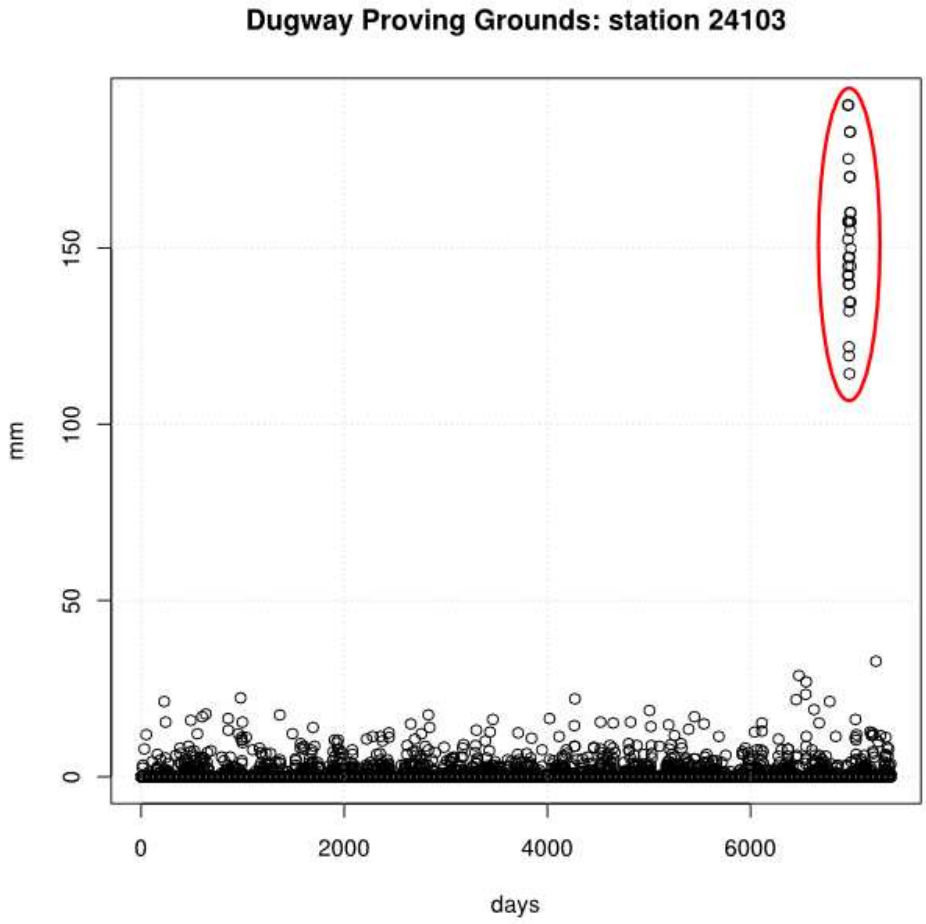

FIG. 3.10: Scatter plot of the Dugway station. The population on the far right is above 100 mm and it forms at this resolution a vertical line.

3.4.3 Sardinian database

The ECDFs of the Sardinian database are interesting for different reasons. While we do not find any evident artefacts like the Dugway station (Fig. 3.10) the plots (Fig. 3.12 and Fig. 3.13) of the Lunamatrona weather station show some interesting features: the points proceed with a zig-zag behaviour.

To understand this non trivial behaviour we need to introduce a fundamental characteristics of the rain gauges measuring the daily cumulated precipitation. All rain gauges, like pluviometers and pluviographs have a *sensitivity level* representing the minimum amount of measurable water. All measures are multiples of this value (0.1 mm or in the Italian and European devices and 0.01 inches in the US tools). The time series of a rain gauge (and of all measuring instruments) are discretized to the sensitivity level.

In the case of the Lunamatrona plots (Fig. 3.12 and Fig. 3.13) we are observing the effect of the discretization, however the sample is not discretized to the only level of 0.1 mm . As stated by

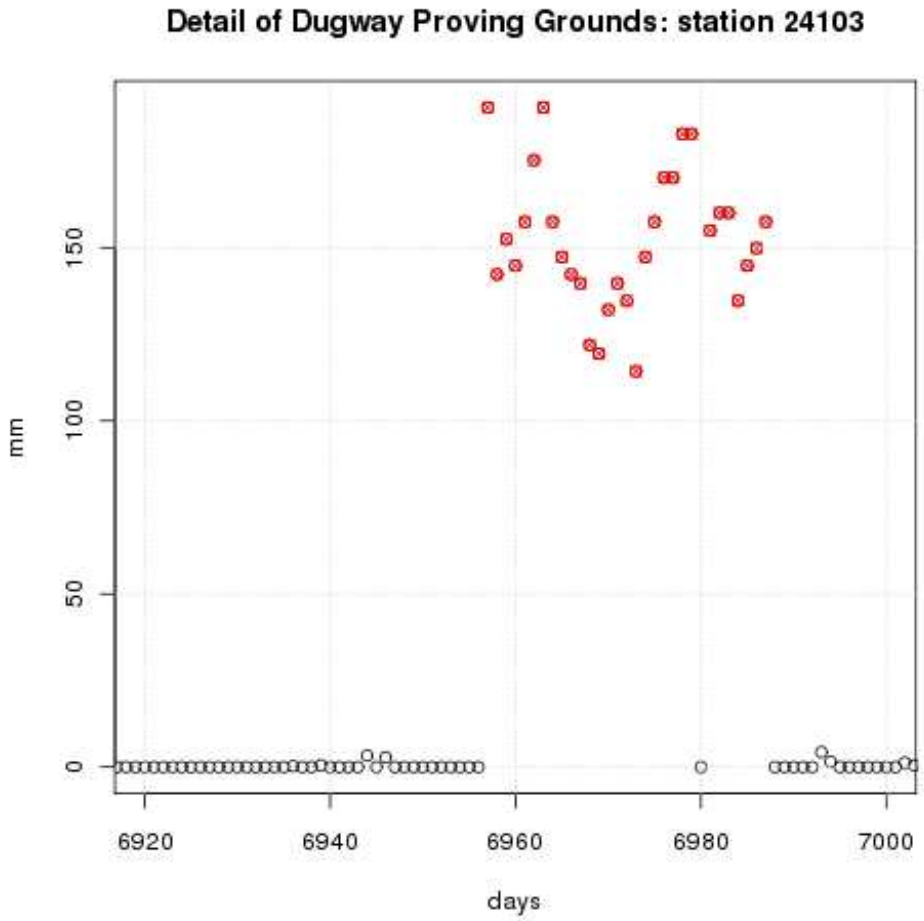

FIG. 3.11: Scatter plot of the Dugway station: detail. Note the presence of a suspect population of high values for 30 consecutive days.

Deidda and Puliga (2006) and Deidda (2007) the values of the dataset were discretized with a rounding-off rule, rounding the values to multiples of the following vector of *rounding-off rules*.

$$\Delta = (0.1mm, 0.2mm, 0.5mm, 1.mm, 5mm)$$

To visually check the discretization of the dataset we can plot a graph (Fig. 3.14) of the frequencies for each possible measure (a multiple of 0.1 mm) showing, with different symbols, the multiple values of 0.5 mm and 1 mm by the factor 10 (ex. 1.5 mm, 2.5 mm and so on).

Every value multiple of 0.5 mm has an anomalous frequency much higher than the theoretical expected one (Fig. 3.15). Moreover in the theoretical sample the values multiple of the rounding-off sensitivity (0.1 mm) are sometimes above the neighbours frequencies, other times below the neighbours ones (see Fig. 3.15). In the sample from the Sardinian database the frequencies of the Δ vector multiples are always above the values of their neighbours (Fig. 3.14): this simple feature allows to build an easy test to identify the presence of spurious rounding-off in the data (see next section).

The presence of the rounding-off is of relevance also for the fit of the statistical models over the data. We show in the section 3.5 that the most suitable model for the precipitation data is the GPD. However if we fit the GPD model on the Lunamatrona dataset using different thresholds we obtain the plot in Fig. 3.16 . In this case the variability of the fitting model is extreme, the shape parameter varies, with notable spikes, in the interval $\xi \in (-0.2, 0.6)$ while for theoretical reasons (stability of the GPD distribution) we expect that after an initial transient the shape parameter must stay almost constant. Indeed a confrontation with the station 235 (Ozieri SS) (Fig. 3.17) shows that the value of the shape is more constant at least for low values of the threshold (when the threshold is too high the variance of the estimator increases and the parameter shows a volatile behaviour). Note that neither in this case we can exclude the presence of the rounding-off for the dataset.

As a final plot we report the same plot for the shape versus threshold for a station of the NOAA-NCDC database (Fig. 3.18) in this case the rounding is less evident or perhaps absent.

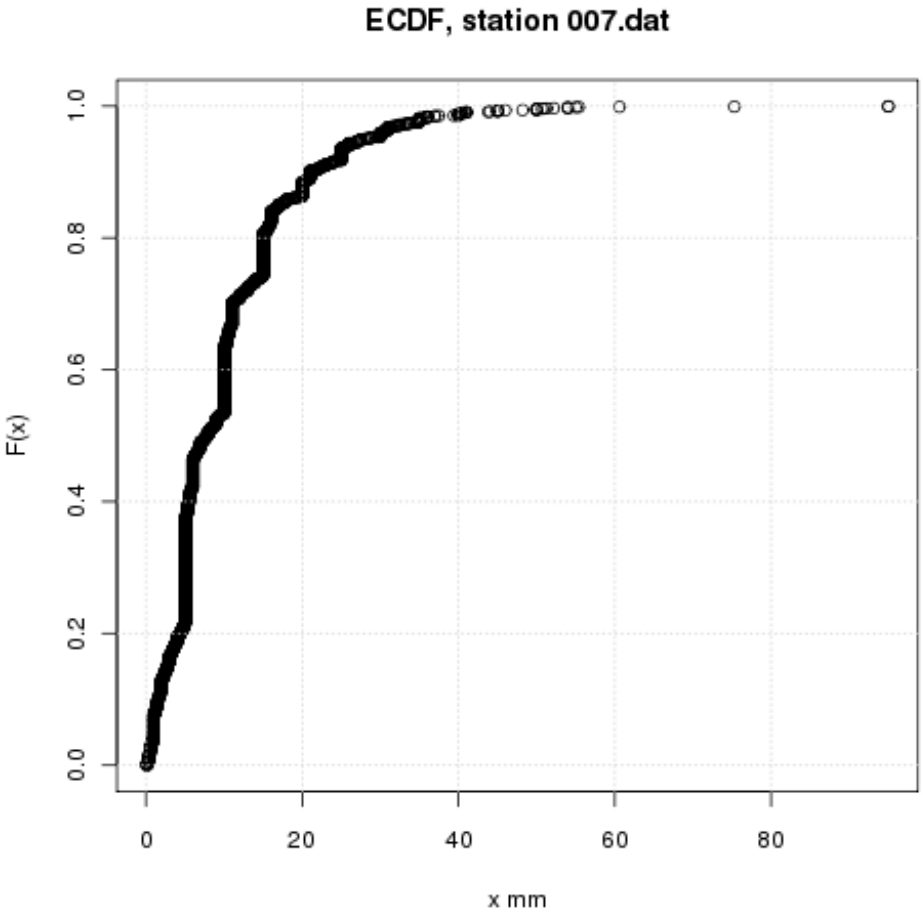

FIG. 3.12: ECDF plot of the Lunamatrona station. Note the "zig-zag" behaviour of the low values.

NOTE. To deal with the rounding-off rules we are forced to use robust estimators and a statistics obtained fitting the data with a *threshold range*. This technique described in Beguería (2005)

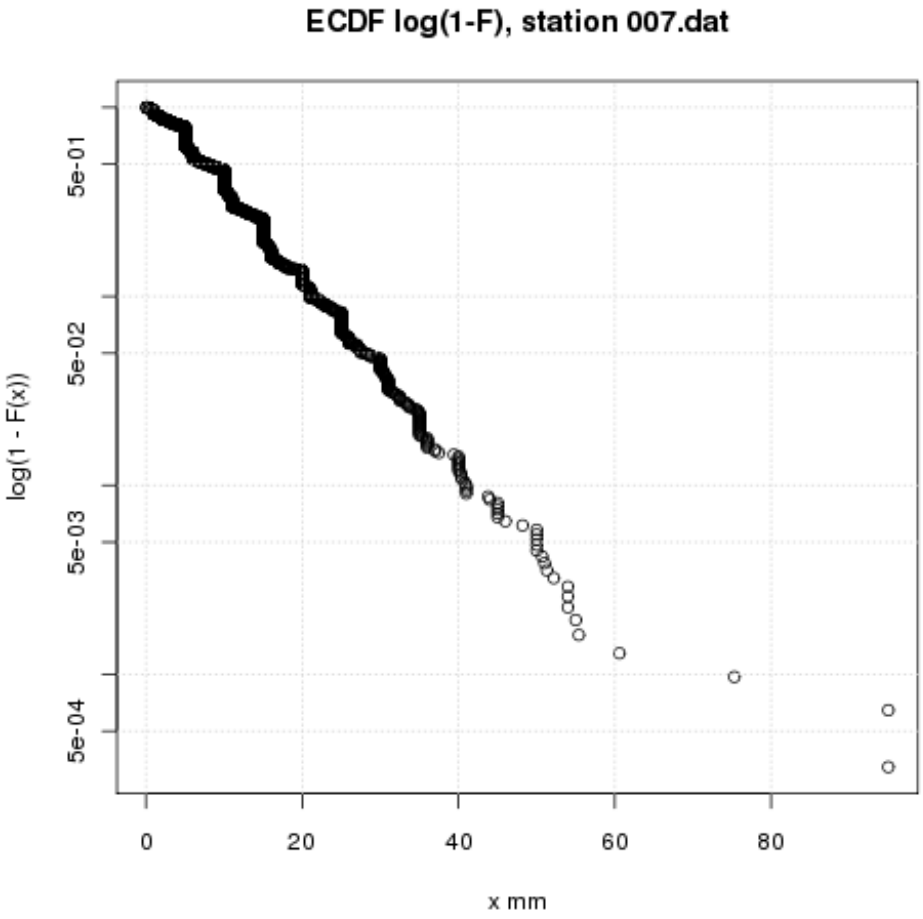

FIG. 3.13: ECDF plot of the Lunamatrona station (semilog scale). The "zig-zag" behaviour is more evident.

consists in choosing a range of values as a possible threshold and use them to improve the statistics of fitted parameters. That is instead to look for the best candidate value of the threshold we can use a range of values and make a statistics with medians [5] of the estimated parameters at the different levels of the threshold. Even in the Lunamatrona dataset if we use this approach we can still obtain valid informations about the fitted parameters .

3.4.4 Tests for rounding-off

In this section we'll describe a simple non parametric test to check the presence of anomalous frequencies in the dataset due to the rounding-off. Our goal is to identify the values of the precipitation that are artificially rounded-off with the result of an increasing in frequency.

Let

$$\Delta = (k \cdot a, 2k \cdot a, \ldots, nk \cdot a) \;\; a = 0.1\ mm$$

[5] The median estimator is specially suitable to deal with rounded-off datasets because it has a low sensitivity to the contaminated values.

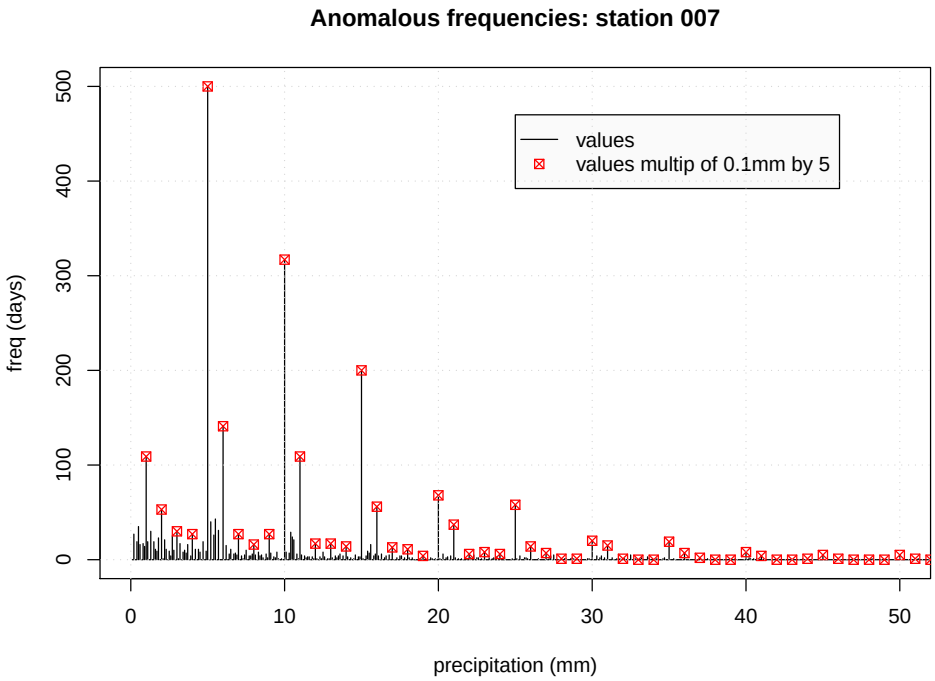

FIG. 3.14: Anomalous frequencies for the Lunamatrona station (007). The multiples of 0.5mm have anomalous frequencies.

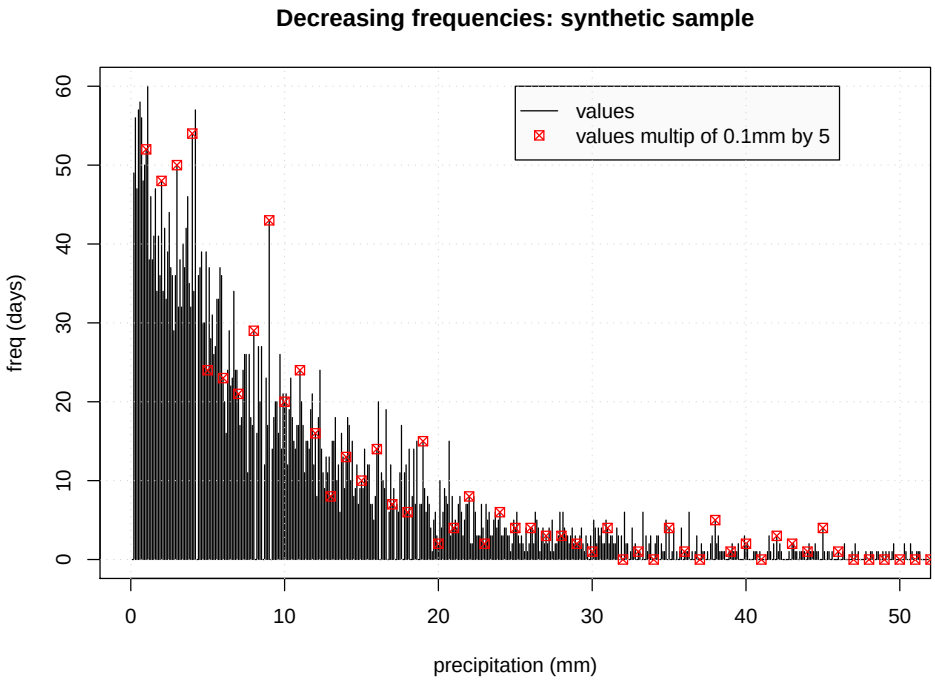

FIG. 3.15: Frequency of a synthetic samples rounded at 0.1 mm. The values are obtained with a simulated GPD distribution with similar parameter of the Lunamatrona station.

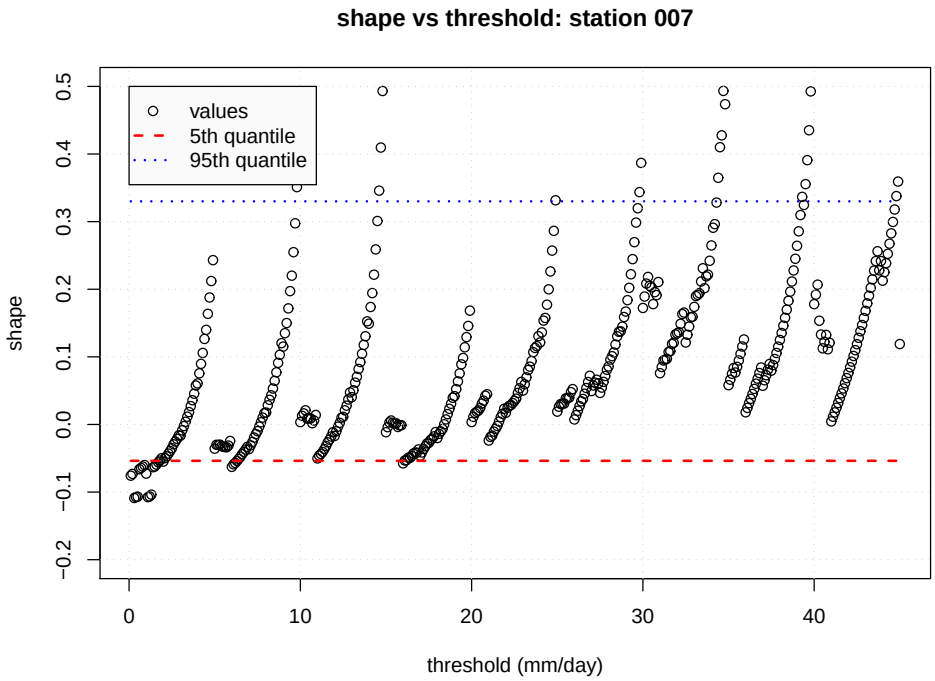

FIG. 3.16: Shape parameter fitted for the station 007 Lunamatrona at several thresholds. Note the high volatility of the graph specially for values corresponding to values multiple of 0.5mm.

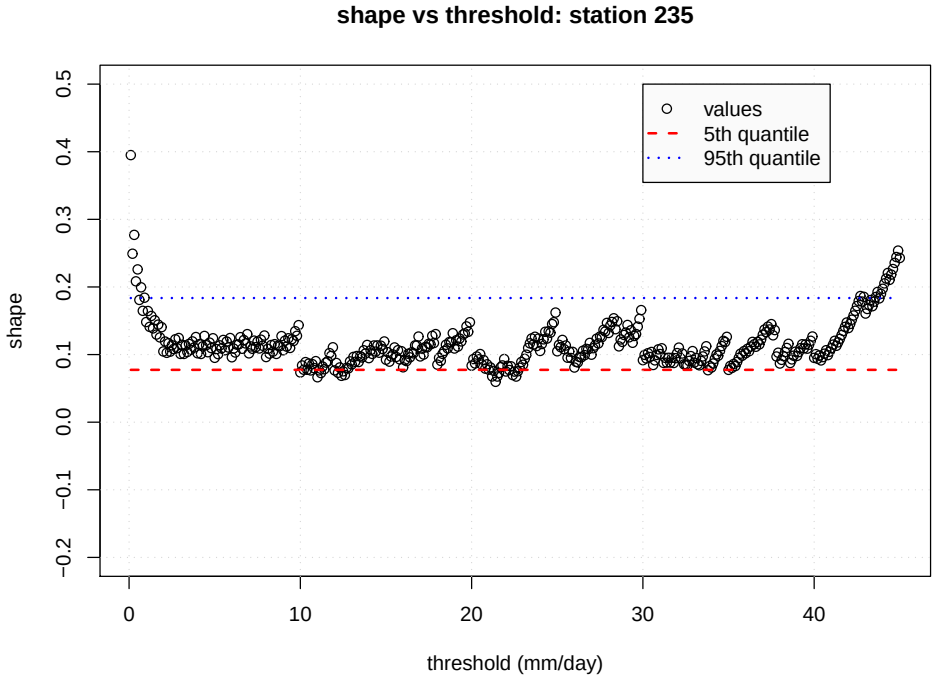

FIG. 3.17: Shape parameter fitted for the station 235 Ozieri at several thresholds. Note that neither in this case we can exclude the presence of rounding-off in the dataset.

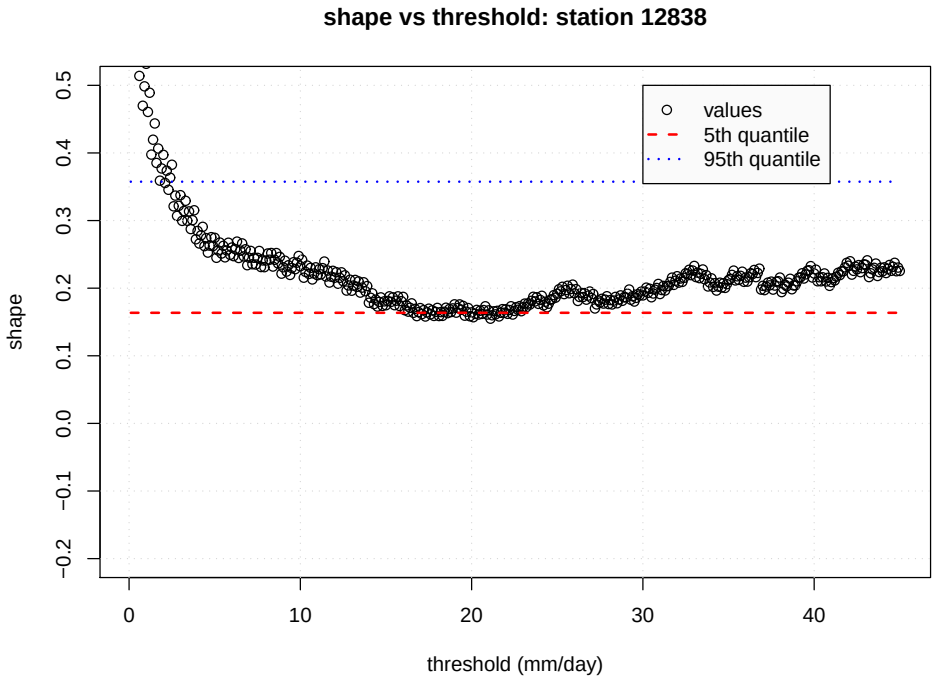

FIG. 3.18: Shape parameter fitted for the NOAA-NCDC station WBAN 12838 at several threshold levels.

where $k \in (1, 2, \ldots, n)$, where Δ is our vector of rounding-off rules. For instance our sample is rounded-off at multiples of $k = 5$ or multiples of $k = 10$ of the sensitivity $a = 0.1\ mm$. We want to discover the values of k where the sample is artificially rounded.

Our starting point we'll be a simple observation: the frequency of the values in phenomena like rainfalls is a monotonic *decreasing* function, that is the strongest events are less frequent than the mild or weak events. Looking to the Fig. 3.14, where we drawn, with different symbols, the points from the values of the rounding-off rules vector (and its multiples). Note that these values are peaks: the anomalous frequencies are almost always above the other neighbouring points (like in the Fig. 3.19 case a). Conversely for points with an ordinary decreasing frequency function we expect that *in probability* the point and its two neighbours are disposed like in the Fig. 3.19 case b).

We expect that counting the number of the occurrences of the case a) or case b) (refer to Fig. 3.19) to find significant differences for values artificially rounded and for values that are not rounded.

We can build a simple statistical test in this way:

1. choose a value of k the multiple factor of the sensitivity level 0.1 mm
2. build the frequency plot with value steps of 0.1 mm , let $f_1, f_2, \cdots, f_n$ the correspondent frequencies

3. evaluate the two cases of Fig. 3.19: $A + C < 2B$ or $A + C > 2B$ (we can call these two expressions **triangle rule**) where

$$\begin{cases} A = (j-1) \cdot k \\ B = j \cdot k \\ C = (j+1) \cdot k \end{cases} \tag{3.1}$$

4. count the number of cases a) and b) giving the value

$$c_a(j) = -1$$

and

$$c_b(j) = 1$$

to each case respectively, then divide the sum of the cases by the number m of multiples of k counted in the sample, let

$$S_k = \frac{1}{m} \sum_{j=1}^{m} c_{a,b}(j)$$

the value of this sum.

5. repeat from 1) with a different value of k.

If we compute the sums S_k we expect that in correspondence of the values of k artificially rounded the sum is negative, while for values of k that are not multiples of the unknown Δ rounding-off rules the S_k will be positive. In other words for rounded-off values the most frequent case is A (see Fig. 3.19) while for non rounded-off values the most frequent case is B.

The Fig. 3.20 shows the result of the test applied to different values of the multiple k to the Sardinian database. It is evident that the values multiple of 5 units (5,10,15 ...) have a different behaviour than the multiple of 6, and 7. In shaded colours we represent also the curve obtained using synthetic GPD samples rounded only at the base resolution of 0.1 mm [6]. The test is able to confirm the presence of rounding and to find the values of rounding (in this case every $5mm$). If we apply the same test to the NOAA database we obtain results similar but less evident (Fig. 3.21): it is less probable that data were rounded-off.

3.5 Choosing the statistical model for data analysis

Until this moment we did not make any explicit hypothesis on the data distribution. Following the rules of the Exploratory Statistics we checked the outliers, the ECDFs and the presence of artefacts like the rounding-off of the samples. To choose the most appropriate statistical model

[6] In the case of a theoretical sample, the frequencies are always decreasing leading to a sum value S_k statistically positive but with a marginal possibility to be negative.

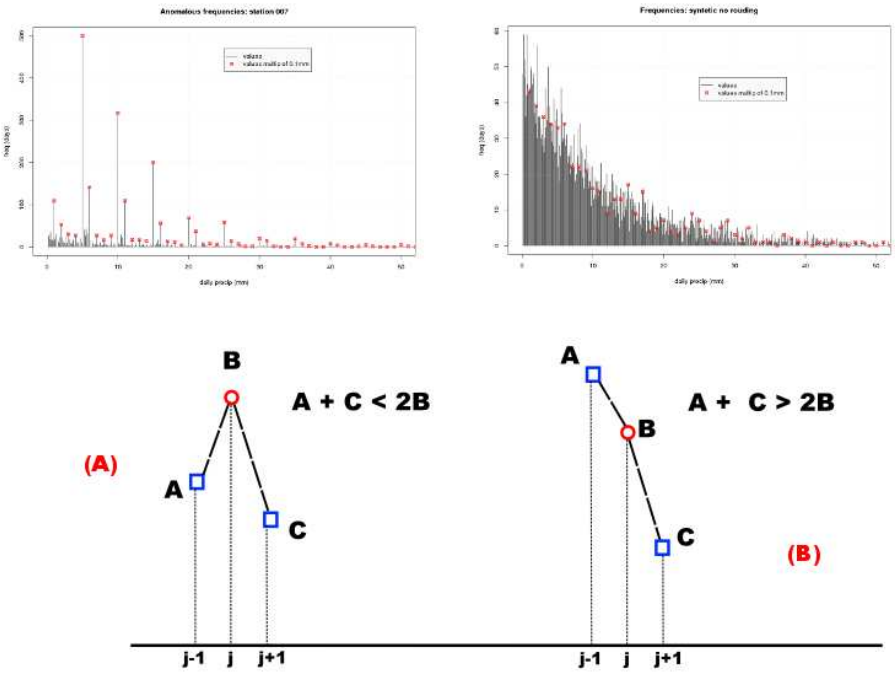

FIG. 3.19: Setup of the rounding test: A) the frequency of the rounded value is greater than frequency of the two neighbours, B) the frequency of the rounded value is lower than the frequency of the two neighbours

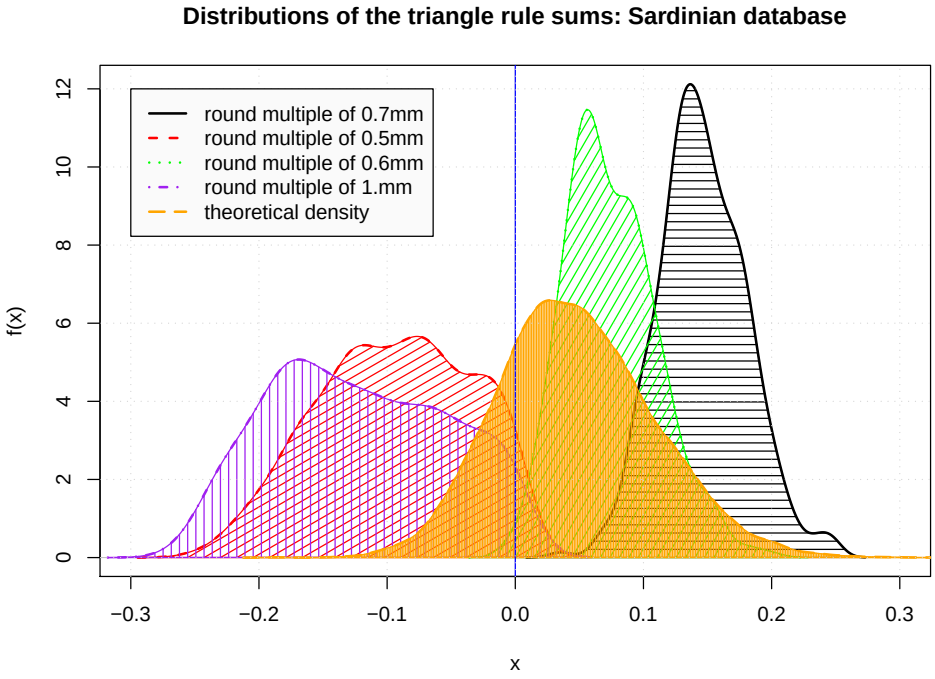

FIG. 3.20: Distribution of the sums of triangle rules for different values of the multiple factor k,SARD database.

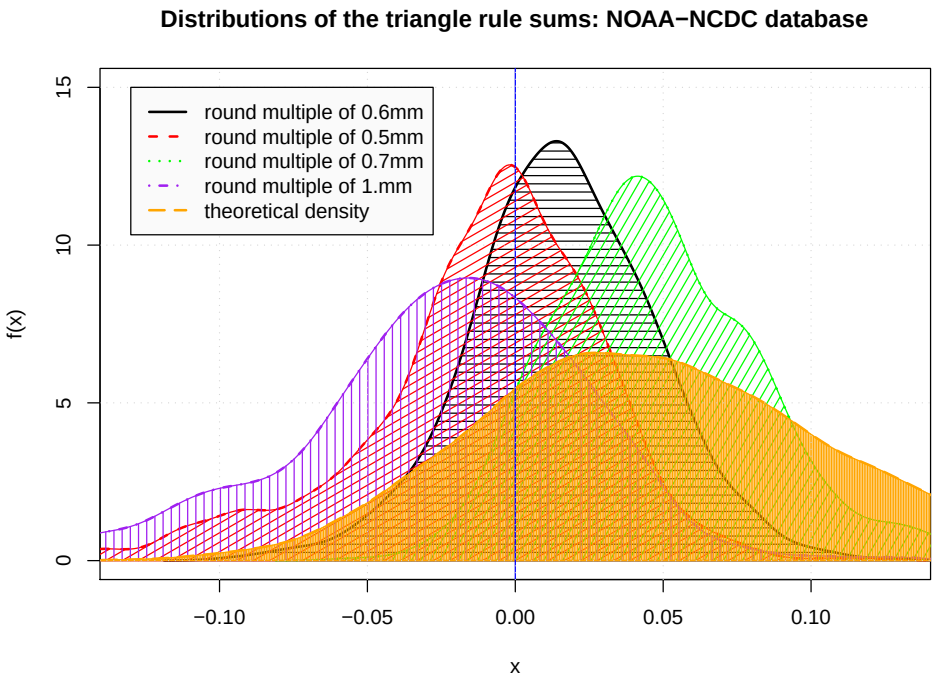

FIG. 3.21: Distribution of the sums of triangle rules for different values of the multiple factor k, NOAA-NCDC database

we can make use of a tool introduced in the previous chapter in the section 2.9: the L-Moments ratio diagram.

We apply this ratio diagram to the SARD and the NOAA-NCDC database confronting different distributions like the GLD (Generalized Logistic), the P3 (Pearson type III), the LN (LogNormal) and the GEV. The theoretical curves could be obtained with Monte Carlo tests (computing the L-Moments ratios at different values of the shape parameter) or using polynomial approximations (as in Deidda and Puliga (2006)).

The results of the test are shown in Fig. 3.22 and Fig. 3.23 for the two databases. The GPD behaviour is strictly confirmed by the plots. We'll use this fact as a general indication for the next analysis, assuming that our databases of daily cumulated rainfall follow a GPD distribution.

3.6 Applying the GPD model

The first tests to perform with the GPD model are the fits of the distribution parameters (shape and scale). For the threshold we cannot assume a fixed value, it is most convenient to use a range of values (as in Beguería (2005)) and make a simple statistics on the estimated parameters. For practical purposes we evaluate the GPD model in an range of thresholds (10th, 40th quantile), this interval is arbitrary (see the section 1.8 for a detailed discussion) but allows to capture most of the variability of the estimators avoiding the problems at the extreme regions. Applying this

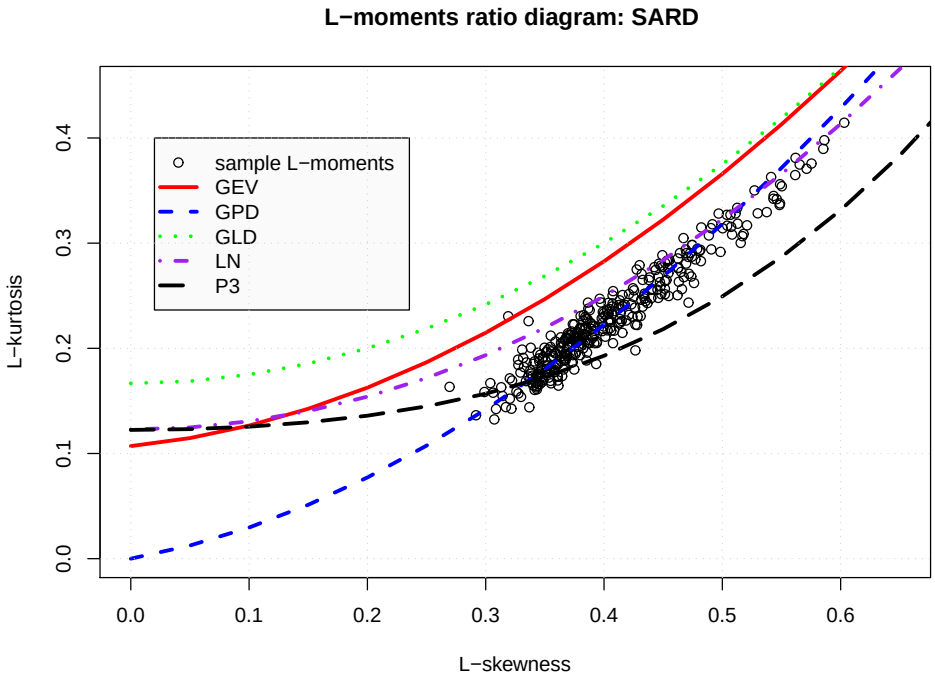

FIG. 3.22: L-Moments ratio diagram on the Sardinian database for several theoretical curves.

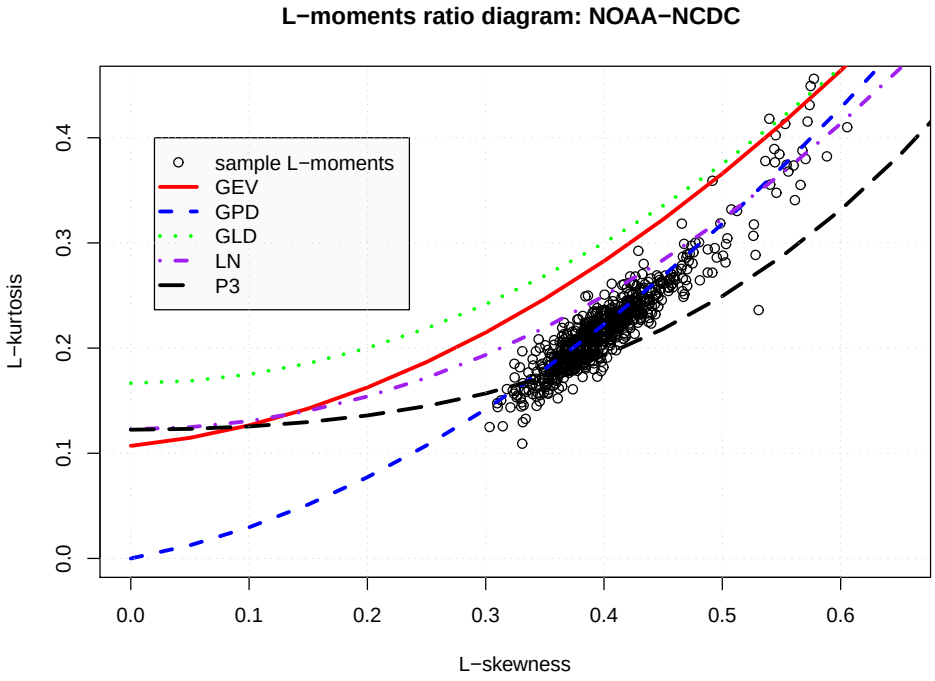

FIG. 3.23: L-Moments ratio diagram on the NOAA-NCDC database for several theoretical curves.

Qlevel	elevation	xi	dxi	alpha	dalpha	90th quantile	median	MAD
0%	-13.12	-0.05	0.01	1.05	0.09	11.89	0.51	0.38
10%	9.15	0.13	0.02	3.54	0.17	46.18	1.78	1.88
20%	18.91	0.16	0.03	4.49	0.21	61.72	2.03	2.26
30%	39.04	0.19	0.03	5.60	0.25	77.04	2.54	3.01
40%	84.48	0.21	0.03	6.56	0.29	91.67	3.05	3.77
50%	149.15	0.23	0.04	7.73	0.34	108.81	3.56	4.52
60%	228.14	0.25	0.05	8.66	0.39	124.13	4.06	4.90
70%	311.40	0.28	0.05	9.36	0.46	141.27	4.57	5.65
80%	471.23	0.33	0.06	10.28	0.56	170.08	5.08	6.40
90%	1124.84	0.42	0.08	11.46	0.71	209.17	5.84	7.16
100%	3049.70	0.98	0.37	22.30	1.79	845.82	10.41	14.31

TAB. 3.5: Quantiles levels (increasing from top to bottom) of the estimated GPD parameters and errors for the NOAA-NCDC database: elevation (m), shape, error on shape, scale (mm), error on scale (mm), 90th quantile (mm), median and MAD estimators (mm).

rule to the dataset we obtain the following tables (3.5 and 3.6) summarizing the quantiles for the GPD parameters of the two databases.

3.6.1 NOAA-NCDC database GPD estimated parameters

The Tab. 3.5 represents the quantiles of the GPD estimated parameters with the threshold u varying in the interval (10th,40th) quantiles. We note that the 90% of data are distributed with a shape parameter less than 0.42 and only a small percentage over 0.5. These values needs to be investigated because the values of shape above 1/2 indicate an infinite variance (or probably the presence of errors in the dataset like the case of the Dugway station) or perhaps these samples are too shorts for a good fit (see the section 3.6.3 for a detailed discussion).

Qlevel	elevation	xi	dxi	alpha	dalpha	90th quantile	median	MAD
0%	1.00	-0.49	0.02	4.99	0.24	28.80	1.20	1.48
10%	17.00	0.01	0.02	7.40	0.29	69.80	3.20	3.85
20%	50.00	0.04	0.02	8.08	0.31	80.86	4.00	4.45
30%	107.40	0.06	0.03	8.49	0.33	89.06	4.50	5.19
40%	186.40	0.08	0.03	8.93	0.36	97.54	5.00	5.78
50%	272.00	0.11	0.03	9.34	0.40	108.00	5.20	5.93
60%	350.00	0.14	0.03	9.73	0.44	126.00	6.00	6.23
70%	438.20	0.19	0.04	10.10	0.50	144.04	6.30	6.67
80%	559.80	0.24	0.05	10.68	0.59	177.59	7.00	7.12
90%	714.20	0.33	0.07	11.63	0.80	254.30	8.00	7.41
100%	1071.00	0.59	0.43	16.84	3.50	489.60	11.00	10.38

TAB. 3.6: Quantiles levels (increasing from top to bottom) of the estimated GPD parameters and relative errors for the Sardinian database: elevation (m), shape, error on shape, scale (mm), error on scale (mm), 90th quantile (mm), median and MAD estimators (mm).

3.6.2 Sardinian database GPD estimated parameters

The table 3.6 represents the quantiles of the GPD parameters estimated on the Sardinian database. Even in this dataset are present values of shape above 1/2 indicating the possible presence of errors in the datasets or samples too short for a good fit (see 3.6.3).

3.6.3 Exploring stations with an high shape parameter

To explore the stations with a shape parameter greater than 1/2 we extract them from the databases and we proceed in two directions: a) increasing the range of the explored thresholds b) studying the behaviour of the shape with the size of the dataset. In fact we suspect that a) the high shape value arises from an incorrect choice of the threshold b) we are fitting stations with too few points. However we remember that the GPD statistics is only a *model* of univariate data it cannot describe the complexity of the datasets far from the asymptotic limits. In fact in our analysis we assume that our data are homogeneous (that is they derive from a single type of

mathematical distribution) but we cannot exclude that in some station several types of climate are mixing different populations of precipitation.

In the NOAA-NCDC database there are 54 stations with shape parameter greater than 0.5, but only 34 have more than 1000 positive points and 16 have more than 2000 (in small samples the natural volatility of the EV values - that fluctuates for several orders of magnitude - can lead to inaccurate fits). A further analysis shows that several of high shape stations are the ones cited for incorrect values (like in Fig. 3.10) reducing the number to 8 candidates in 1582 stations. This number is small in confront of the total number of stations, then having no other informations about these datasets regarding other possible problems (like bivariate or multivariate populations) we are forced to conclude that even non physical GPD models (remember that for $\xi > 1/2$ the variance is infinite) are present in real databases. Note however that it is possible that *longer* time series from the same weather stations correct the problem as it can be an effect of random fluctuations.

In the Sardinian database, where the presence of the rounding make evident the fluctuations in frequency for low values, we found only 7 stations with shape greater than 1/2. This time the series have no evident errors (like the added values of the cited Dugway station), and they are associated to the rainiest stations in Sardinia (where the maximum daily recorded cumulated rainfall was greater than 300 mm). In this case the points are not spurious (they are associated to real floods reported in literature) then in absence of other indications we conclude that these values are due to random fluctuations of the extreme quantiles of these dangerous stations. Again it is possible that longer time series from these stations decrease the value of the shape parameter below the physical limit of $\xi \leq 1/2$.

3.7 Geographical distribution of GPD parameters

The geographical distribution of the GPD parameters is reported on the Figg. 3.24 - 3.27. Note that the shape parameter - at least for the Sardinian case - is similar to the distribution of the 90th quantile indicating that an high value of shape is associated to severe daily precipitations.

3.8 An alternative hypothesis

In this paragraph we explore a simple alternative hypothesis to the GPD distribution for data analysis. For its simplicity the exponential distribution

$$f(x, \lambda) = \lambda e^{-\lambda x} \quad \lambda, x > 0 \tag{3.2}$$

is widely used in EV theory. The fit of this distribution is straightforward in fact $\lambda = 1/m$ if m is the mean of the sample. Fitting a single parameter distribution allows to describe even stations with few data and we do not need to discard any point or to look for a threshold. Here we want to try a simple confrontation between the GPD and the EXP model in predicting the maxima of each dataset. For this reason we organized a simple experiment:

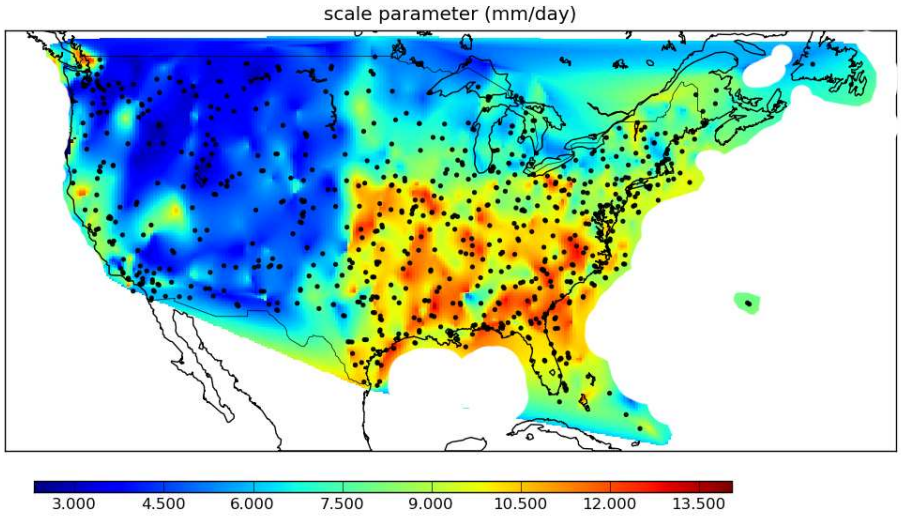

FIG. 3.24: Distribution of the GPD scale parameter in the NOAA-NCDC database.

- Fit the NOAA-NCDC datasets with the GPD and the EXP distributions, at a given threshold level $u = 5\ mm$, left censoring the data for both distributions in order to fit an homogeneous sample.
- Compute the quantile of the probability level $p = 1 - 1/n$ if n is the length of each sample. Let $q_{gpd}(p)$, $q_{exp}(p)$, $q_{data}(p)$ the computed quantile levels for the fitted distribution GPD and EXP and for the real data where $q_{data}(p) = max(data)$. We assume that the probability varies in the interval $[0, 1 - 1/n]$ in steps of $1/n$ (plotting position rule).
- Plot the distributions of $q_{gpd}(p)$, $q_{exp}(p)$, $q_{data}(p)$

the result is summarized in Fig. 3.28 where it appears clear that the GPD is more able than the EXP distribution to guess the true level of the extremes. Note that the median (or the mean) of both methods are similar (with a little positive bias for the GPD distribution). We realized a table 3.7 for median and mean. Note that median and mean of the EXP distribution are similar (indicating a symmetric distribution centred around the value of 133 mm), while the GPD and the data are not symmetric. Finally the GPD distribution overestimates the average value of the maxima.

We can conclude that even if the exponential model is not bad in describing the average maxima of the database the GPD model is more suitable to describe the extreme values of the rainfall databases. In the next chapter we investigate deeply the problem of the optimal threshold estimation that is critical to improve the reliability of the model.

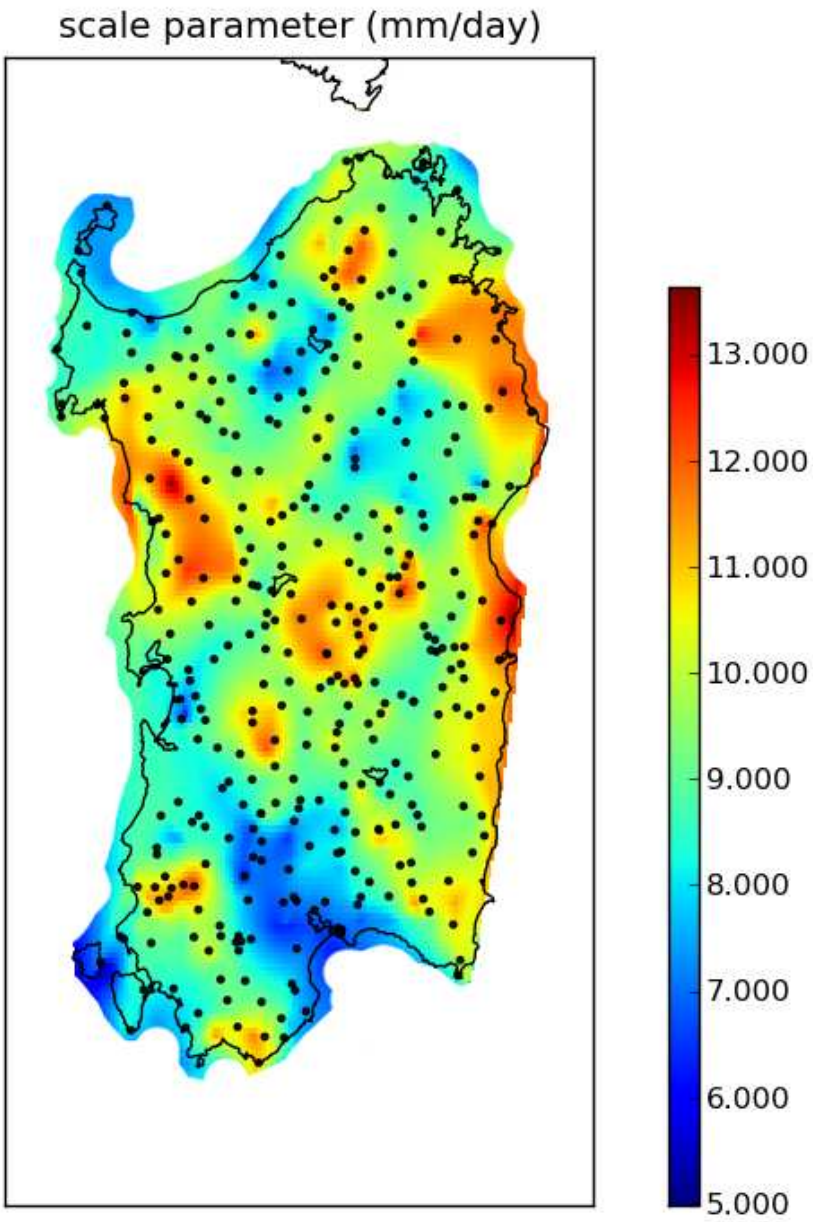

FIG. 3.25: Distribution of the GPD scale parameter in the SARD database.

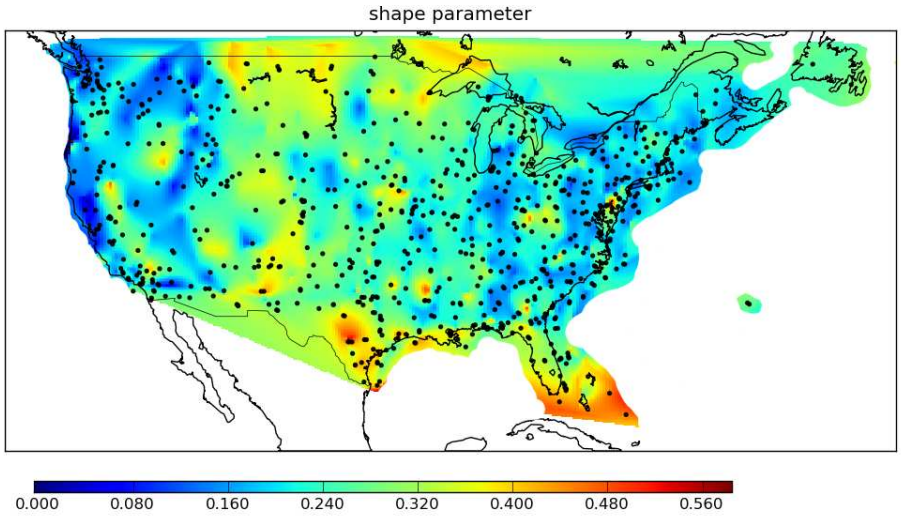

FIG. 3.26: Distribution of the GPD shape parameter in the NOAA-NCDC database.

Method	mean (mm)	median (mm)
data	155.2	133.3
GPD fit	213.5	164.7
EXP fit	134.3	133.3

TAB. 3.7: Mean and median of the maxima for the GPD and EXP fit methods and for real data from the NOAA-NCDC database. The GPD overestimate the mean value (remember that the GPD distribution is specially created for the extremes).

3.9 Final remarks on exploratory statistics

In this chapter we adopted a *data oriented approach* (borrowing an expression from the professional informatics) starting from the empirical distribution of data, checking for the outliers and choosing the model in function of an important tool (the L-Moments ratio diagram) to guess the most adapt statistical distribution. We noticed how the artefacts can affect the records, how they can change the statistics and influence the goodness of fit. The exploratory statistics is a powerful technique for the analysis of data, but must be used with care. In fact some artefact could be not evident and the confidence in ordinary tools like the mean and the variance of a population can lead to errors in the estimate. We suggest every time it is possible to use robust estimators that are less affected by outliers and other errors or features of the data. However even with these tools we still need to check the nature of the high value points. To do that we need a statistical model otherwise we have no tools to extract the candidate outliers. Finally we remember that a no doubt proof for detecting an outlier is a parallel check with another *different*

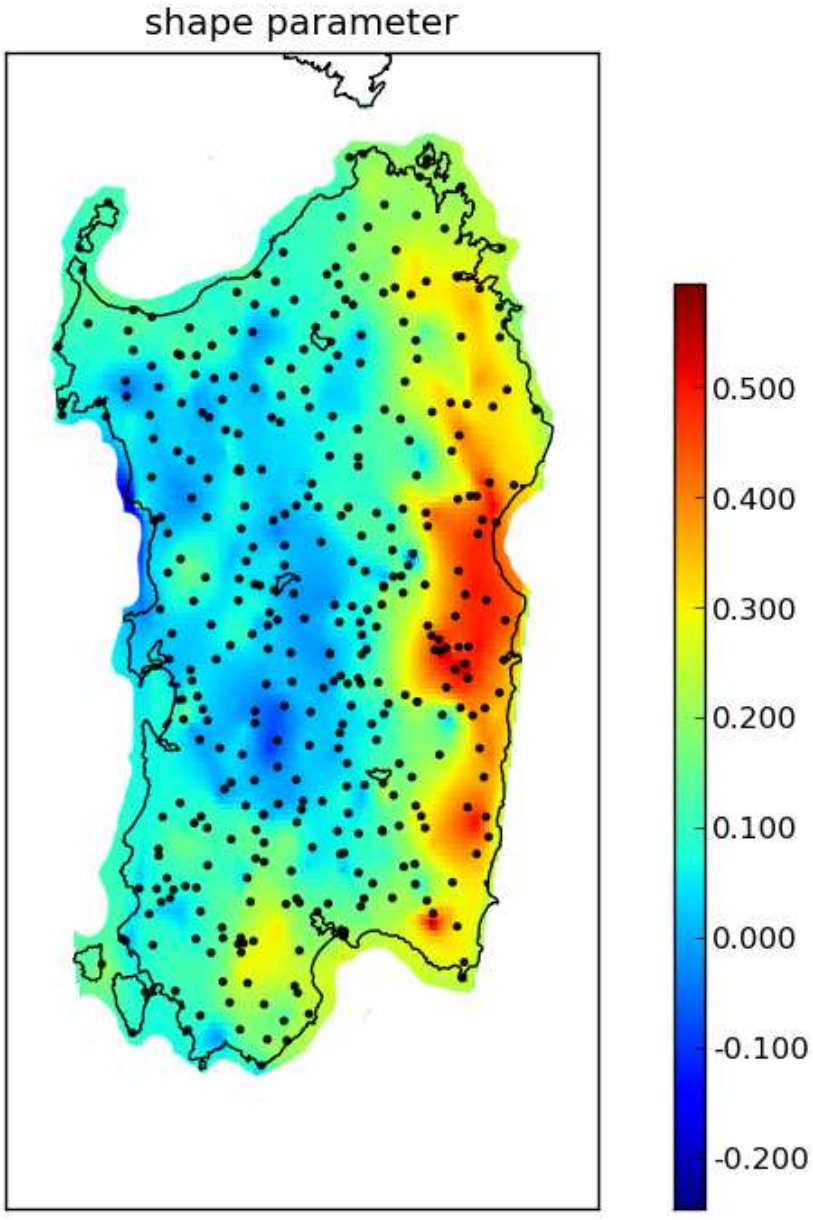

FIG. 3.27: Distribution of the GPD shape parameter in the SARD database. Note the evident orographic effect on the Thyrrenian sea side, where the strongest event appear.

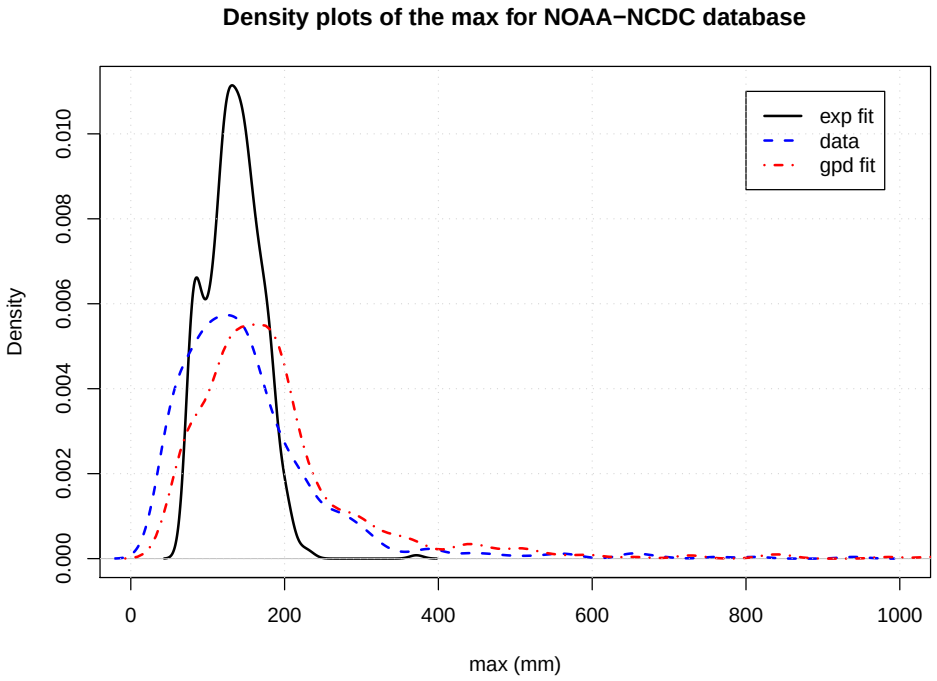

FIG. 3.28: Distribution of maxima for data, EXP fit and GPD fit. Database NOAA-NCDC

time series of the same station, lacking this information we can have only a proof in probability building confidence levels for the suspected data.

Chapter 4

Improving the statistical models

4.1 Introduction

In the previous chapter we explored the hydrological databases to find a good model to describe the data. We found that the GPD is the best candidate distribution for the datasets at least above an high threshold. As an alternative hypothesis we studied the performances of the simple exponential distribution in predicting the maxima confronting it with the GPD model. We encountered problems due to artefacts like transcription errors (true outliers) and corrupted datasets with rounding-off rules. However we left behind some fundamental questions: a) what are the performances of the estimators in the case of rounded-off data ? b) what is the best threshold to choose to apply the GPD model ? To answer to these questions we abandon for a moment the exploratory approach and assume that data *are* distributed following a GPD distribution with an unknown degree of contamination from spurious data or rounding. To study the performances of the estimators on rounded-off data we generate synthetic samples with rounding-off rules similar to those seen for real data in the previous chapter. Alternatively to find the optimal threshold we'll use the NOAA dataset (that has a lower level of rounding-off contamination) applying several tests described in literature. Finally we propose a new methodology for threshold estimation based on quantiles functions.

4.2 Methods for optimal threshold determination

The methods to determine the optimal threshold of the GPD distribution can be divided in categories. We describe first the non parametric methods for optimal threshold selection, these methods are interesting because they do not assume any hypothesis on the distribution of data, they simply try to identify the region where the behaviour of the data starts to change from non extreme to extreme. The graphical methods for threshold estimation are classical in the theory of the GPD, where the plot of the mean of the excesses and the stability of the GPD shape parameter are widely used (for small databases) to find the starting point where the GPD

model is valid. Then we describe the methods based on the goodness of fit statistics built on the idea to find the optimal threshold in function of a test statistics indicating the validity of the GPD model over the data . Finally a different category of tests is the kernel test statistics of the asymptotic tail slope estimators (like the Hill tail slope estimator), this category of tests checks the exponentiality of the log-transformed GPD data in the asymptotic region where the log-tail could be approximated with a straight line.

4.2.1 Non parametric methods

The Gerstengarbe plot (Gerstengarbe and Werner, 1989) is a simple graphical (but also numerical) method to find the region where the GPD could be used. This method does not estimate the best GPD fit nor uses a GOF statistics, but it divides the data distribution in two parts. The lower part is not suitable to be described as an extreme value distribution, while the upper part could be fitted with extreme values distribution. The Gerstengarbe plot is based on the sequential Mann-Kendall changing point test (Kendall, 1975).
Given n reordered random variables $X_1 < X_2 < ... < X_n$ and the rank sum:

$$t_i = \sum_{j=1}^{i} R_j = \sum_{j=1}^{i} rank(X_j) \quad i < n$$

The Mann-Kendall statistics could be expressed by:

$$\begin{aligned} u(t_i) &= \frac{t_i - E(t_i)}{\sqrt{\sigma^2(t_i)}}, \quad i = 1, ...n \\ E(t_i) &= \frac{i(i-1)}{4} \\ \sigma^2(t_i) &= \frac{i(i-1)(2i-5)}{72} \end{aligned} \tag{4.1}$$

where $E(t_i)$ and $\sigma^2(t_i)$ are respectively the mean and the variance of the sample. The series of the statistics u_i could be computed using an ascending order of the series $\mathbf{X}$ or a reversed one. Calling $u_i^{'}$ the reversed series the intersection between $u_i^{'}$ and u_i (as in Fig. 4.1) defines a *changing point* as proposed in Gerstengarbe and Werner (1989). The main interest in this method is due to the fact that this one does not make any hypothesis on the empirical distribution and on the goodness of the fit. However the estimated GPD thresholds are usually lower than other methods .

4.2.2 Graphical methods

Graphical methods to find the best threshold are widely used in literature, specially when the analysis is limited to a few samples and data have no outliers or spurious data. The most used one is the Mean Excess Plot (also known as *mean residual life* plot see Fig. 4.2 and the book of Coles

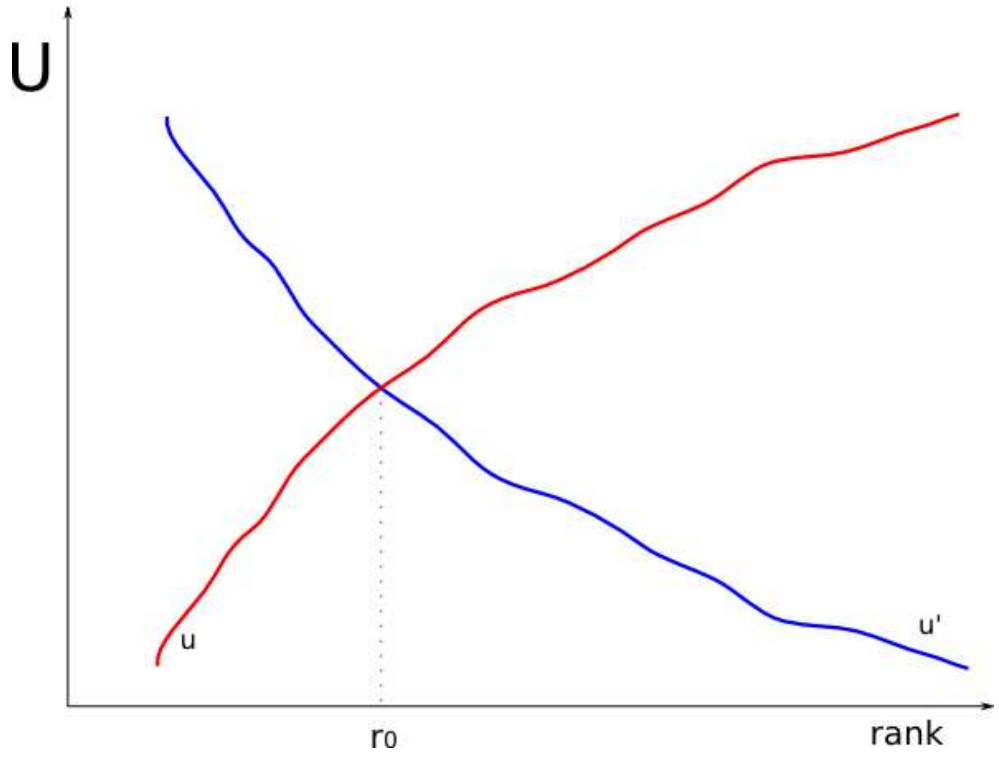

FIG. 4.1: Gertensgarbe plot for two statistics: u and u' of rank ascending end descending samples.

) where the mean of the excess above the threshold is plotted versus the threshold. When the points start to behave like a straight line the lower abscissa of this line is the correct threshold. In the same figure the stability plot of the GPD shape parameter shows the characteristic *plateau* of the shape at different threshold levels. The plateau starting point coincides with the one estimated by the mean excess plot. However it is clear that the choice of the correct threshold is a bit subjective. To avoid this problem specially when the number of samples to analyse is big numerical methods must be preferred.

4.2.3 Mean Square Error: the AD and CV statistics

The methods of the Goodness of fit (with the relative test statistics) are based on the idea to find the *lower* acceptable value of threshold satisfying the test statistics at the preferred confidence level. We introduce the mean square error MSE for an estimated parameter. $\hat{\theta}$ is the sum of the Bias and the Variance:

$$MSE(\hat{\theta}) = E[(\theta - \hat{\theta})^2] = Var(\hat{\theta}) + \Big(Bias(\theta, \hat{\theta})\Big)^2 \tag{4.2}$$

The MSE and the Bias are measures of the goodness of the fit for a given distribution. In the GPD model the variance increases with the threshold, while vice-versa the bias decreases (fitting

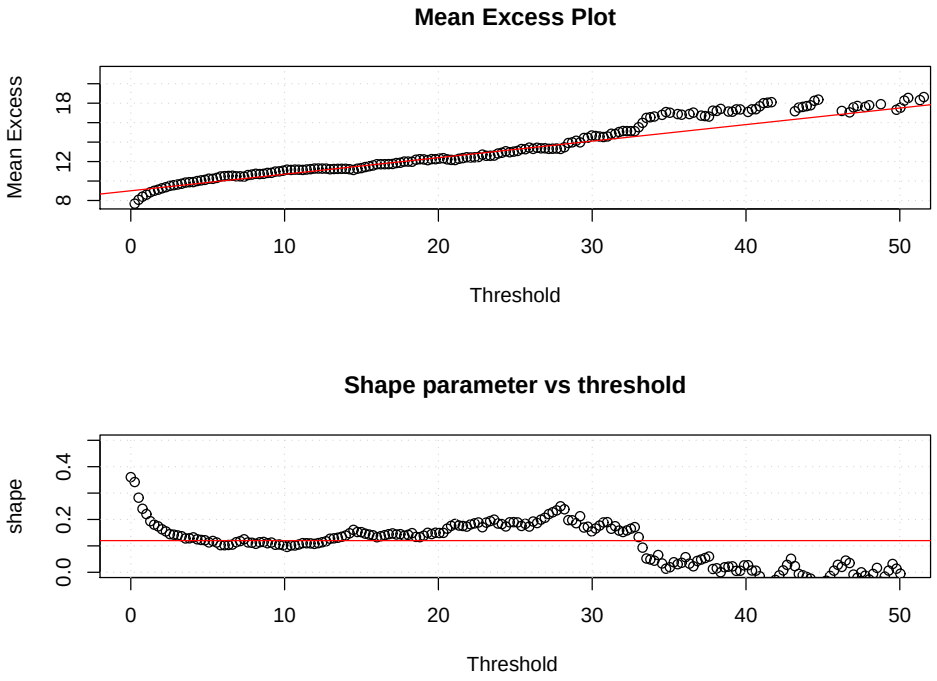

FIG. 4.2: Graphical methods for threshold estimation: left, mean excess plot; right, shape stability versus threshold. Both plots with GPD example data

the tail of the distribution leads to a more accurate estimate of the shape parameter but the estimation has higher variance).

The Cramér von Mises statistics is based on the MSE and for discretized data is given by:

$$W^2 = \sum_{i=1}^{n} \left[F(X_i) - (2i-1)/(2n)\right]^2 \tag{4.3}$$

where F is the empirical cumulative distribution function in the analytical or empirical form directly obtained from the dataset using plotting position rules.The Anderson Darling statistics is similar but introduces an weighting function to improve the tail estimation and uses the logarithmic differences instead of the quadratic ones:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\log(F(X_i)) - \log(1 - F(X_{n+1-i}))\right] \tag{4.4}$$

Unfortunately there's no explicit formula for the limiting distribution of $A^2_\infty = \lim_{n\to\infty} A^2_n$ and $W^2_\infty = \lim_{n\to\infty} W^2_n$ although they exist their characteristics complex function (for details Anderson and Darling (1952)).

In literature there are several numerical approximations, notable the work of Marsaglia and Marsaglia (2004) that uses a powerful random number generator and a numerical infinite sum

approximation of the statistics to evaluate A^2_∞ with an high precision level. Other simulation or numerical tables were built using Monte Carlo methods. In this work we used the Monte Carlo methods described in Choulakian and Stephens (2001) in order to produce the critical level tables for A^2 and W^2 with different confidence levels.

The main details of this simulation are the following:

1. Generate m GPD samples of length n with fixed scale parameter (the scale parameter is not acting on the A^2 and W^2 statistics) and shape variable in the interval of interest (in our case $\xi \in (-0.5, 0.5)$).
2. For each m samples compute the estimated (with the preferred fit method) GPD parameters ξ_{est} and α_{est}, keep these values to build the probability function $F(X_i)$ (obtained inverting the GPD function from the values X_i).
3. Compute the A^2 and W^2 statistics of each m sample then build the relative quantiles levels that will become our confidence table.

NOTE. The fit methods must be *unbiased* unless we want to obtain the test statistics for biased estimators like the simple moments.

4.2.4 Monte Carlo methods and Contingency tables of A^2 and W^2

We present here the confidence tables obtained with the same procedure of Choulakian and Stephens (2001). The table (Tab. 4.1) is computed for 4 different confidence levels.

The accuracy of this table is greater for lower quantiles and high levels of ξ rather than for $\xi \to -0.5$ where the estimated value is less accurate and the relative errors are about 5%. Note that in Choulakian and Stephens (2001) the values are corrected with asymptotic methods. For samples ($n > 50$) the statistics is satisfactory and the missing ξ values in the table could be obtained using linear interpolation. A correction for small samples is described in Anderson and Darling (1952). Finally remember that the procedure described here assumes that we do not know the two GPD parameters, if we know a value (shape or scale) of the GPD the values of the statistics changes and the test must be adapted.

4.2.5 Kernel statistics

The method described here considers the optimal threshold as the lower level of the data where the log transformed tail could be confused with an exponential distribution. The test, based on Kernel statistics, computes the region of acceptance of the exponential hypothesis. The kernel density statistics approximate the empirical distribution with an weighted sum of *kernel* functions (same technique described in the Chapter 2 for the density plots). The most used kernel function is the Gaussian curve, but the choice of the most appropriate function is suggested by the experience and by the data distribution.

	A^2				W^2			
ξ	0.8	0.9	0.95	0.99	0.8	0.9	0.95	0.99
-0.5	0.738	0.978	1.251	2.058	0.111	0.146	0.184	0.285
-0.4	0.696	0.898	1.120	1.690	0.106	0.138	0.172	0.261
-0.3	0.677	0.863	1.068	1.574	0.103	0.135	0.167	0.250
-0.2	0.656	0.844	1.028	1.492	0.100	0.131	0.162	0.237
-0.1	0.639	0.818	1.010	1.470	0.096	0.126	0.157	0.231
0.0	0.617	0.788	0.966	1.392	0.093	0.120	0.149	0.220
0.1	0.601	0.764	0.927	1.352	0.089	0.116	0.142	0.209
0.2	0.581	0.738	0.898	1.298	0.086	0.111	0.136	0.200
0.3	0.571	0.720	0.877	1.273	0.084	0.107	0.132	0.195
0.4	0.558	0.707	0.855	1.208	0.081	0.104	0.128	0.186
0.5	0.545	0.687	0.833	1.190	0.079	0.101	0.124	0.178

TAB. 4.1: MLE computed table for A^2 (left part of the table) and W^2 (right part of the table) critical values at different confidence level. Note that the sign of the GPD shape parameter used in Choulakian and Stephens (2001) is positive while we use the minus sign.

The log-transform of a GPD random variable behaves like an exponential distribution and in this case the kernel Lewis statistics and the kernel Jackson statistics are useful to test for the exponentiality of data, see for instance Henze and Meintanis (2002). From the log transformed data it is possible to define an important estimator of the shape parameter: the Hill estimator
5 (Hill, 1975):

$$1/\xi = H_{h,k} = \frac{1}{k}\sum_{j=1}^{k} log(x_{n-j+1,n}) - log(x_{n-k,n}), \quad k = 1, ..., n-1 \tag{4.5}$$

where $x_1, x_2, \ldots, x_n$ are the reordered data (quantiles). The Hill estimator measures the increase of mean logarithmic quantile minus the so-called *fixed point* (lower point of the distribution or left censoring level). In other words this estimator evaluates the reciprocal of the slope measured on the tail of the distribution. The approximation is exact in an asymptotic sense.

When the Pareto model is valid then the survival function $1 - F(x)$ could be expressed as a power law with the Hill exponent $(1/\xi)$:

$$1 - F(x) = x^{-1/\xi}\ell_F(x), \quad x > 0 \tag{4.6}$$

where $\ell_F(x)$ denotes a *slow varying* function at infinity, that is a function with the property

$$\frac{\ell_F(\lambda x)}{\ell_F(x)} \to 1 \quad x \to \infty \ \lambda > 0. \tag{4.7}$$

The kernel GOF tests stress the hypothesis that the Hill approximation of the slow varying function holds with a given confidence level. In a logarithmic scale the tail of a GPD distribution behaves like a straight line, the threshold is selected at the level where the linearity is acceptable. From this point of view this method is completely different from the other ones, in fact it starts from the extreme tail of the distribution, lowering the threshold until the chosen confidence level. Conversely graphical, parametric or quadratic GOF tests start with low or even zero threshold and rise the level until the GOF statistics is satisfactory. We expect to find higher threshold levels with the kernel statistics.

The Jackson statistics is derived from the general kernel estimation formula:

$$T_J = \frac{\sum_{j=1}^{n} t_{j,n} Y_{j,n}}{\sum_{j=1}^{n} Y_j} \tag{4.8}$$

where $Y_j = \log X_j$ are exponentially distributed random variables and

$$t_{j,n} = \sum_{i=1}^{j} \frac{1}{n - i + 1} \tag{4.9}$$

is the weighting function. Similarly the Lewis statistics is derived from the 4.8 using the following transformation:

$$V_j = (n - j + 1)(Y_{j,n} - Y_{j-1,n})$$

and defined by the equation:

$$T_L = \frac{\sum_{j=1}^{n} \frac{1}{n+1} V_{n-j+1}}{\sum_{j=1}^{n} Y_j} \tag{4.10}$$

In the work of Goegebeur et al. (2008) the Jackson (JK) and the Lewis (LW) statistics are used to implement a new best fit technique to estimate the optimal threshold. To achieve this goal the exponential random variables (Y_i) were transformed back to Pareto variables X_j then the

T_L and T_J statistics are changed in

$$\begin{aligned} T^J_{k,n} &= \sqrt{k}\frac{\frac{1}{k}\sum_{j=1}^{k} K_J\left(\frac{j}{k+1}\right) Z_j}{H_{k,n}}, \\ K_J &= -1 - log(u) = -1 - log\left(\frac{j}{k+1}\right), \\ Z_j &= j\left(log(X_{n-j+1,n} - log(X_{n-j,n})\right), \\ H_{k,n} &= \frac{1}{k}\sum_{j=1}^{k} Z_j. \end{aligned} \tag{4.11}$$

where $H_{k,n}$ is the Hill function and K_J the kernel.

The Lewis 4.10 is transformed in the following manner:

$$\begin{aligned} T^L_{k,n} &= \sqrt{k}\frac{\frac{1}{k}\sum_{j=1}^{k} K_L\left(\frac{j}{k+1}\right) Z_j}{H_{k,n}}, \\ K_L &= u - 0.5 = \frac{j}{k+1} - 0.5, \end{aligned} \tag{4.12}$$

The transformations in 4.11 and 4.12 have the notable property that the limiting distribution is the normal $N(\mu, \sigma)$. For this reason the confidence levels are straightforward. For instance to build a two tail test with a 5% of total confidence the critical level is 1.96σ as from the well known properties of the Normal statistics. A combination of Lewis and the Jackson statistics could be used to find the best fit (JKLWBF) for the Hill estimator. The method is summarized in the paper of Goegebeur et al. (2008).

4.3 Evaluating the optimal threshold from the data

To evaluate the optimal GPD threshold the methods described in the previous section were applied to the NCDC database. First we computed the table of critical values for the Anderson Darling and the Cramér von Mises GOF tests Tab. 4.1. Using these critical values we can avoid to estimate for each threshold level the values of A^2 and W^2 given the fitted shape and scale parameters. When the estimated shape parameter is not present in the table (4.1) a linear interpolation is applied. All other methods do not require any critical value table because the Jackson and Lewis limiting distributions are normal and the Gerstengarbe plot is based on a non parametric technique while its distribution (the Mann-Kendal statistics) is also normal.

To evaluate the performances of these tests we use several heuristic considerations. All statistics (except the Gerstengarbe plot) refer to parametric GOF tests evaluating the threshold at the lower acceptable value of the MSE (mean square error) or the Exponentiality of the sample. We do not know the *true* optimal thresholds then we cannot evaluate the performances of the estimators using again a MSE (Mean Square Error) criteria, we need other considerations.

The GPD statistics has the property that if $\xi < 0$ the distribution is bounded to an upper value and, if $|\xi| > 1/2$, the variance is not a finite quantity (Hosking and Wallis, 1987). A part from statistical fluctuations there is no physical reason to conceive upper bounded distributions for rainfalls (at least in a region far from the physical limits of the atmospheric evaporation-precipitation process), while the infinite variance have no physical meaning at all.

Using these heuristic criteria the Fig. 4.3 (where the Boxplot for the shape parameters are plotted at the estimated optimal thresholds) shows that the only methods that do not have non-physical values are the AD and the CV tests. They do not have negative values, they have narrow distributions and no shape values over 0.5. In the same plot the failure of the LWJKBF method is evident (unrealistic high shape values, high distribution spread). This failure is due to the Lewis statistics that shows the same failure pattern.

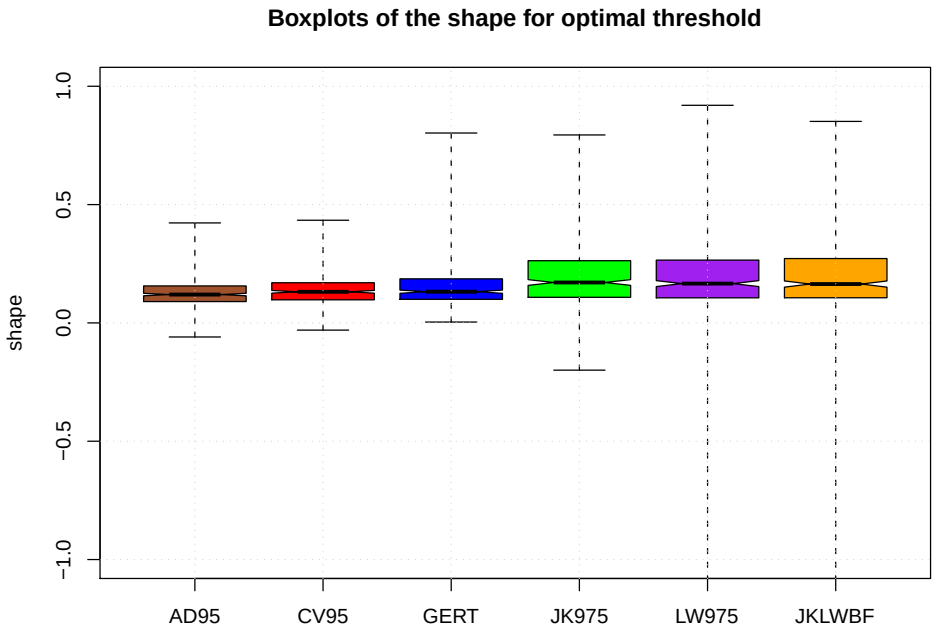

FIG. 4.3: Boxplot of the GPD shape parameter computed for NOAA-NCDC database, using different methods at the best estimated threshold. The acronyms stands for AD Anderson-Darling, CV Cramér Voin Mises, GERT Gerntensgarbe plot, JK95 Jackson kernel statistics with 95% confidence, LW95 Lewis kernel statistics with 95% confidence, JKLWBF Jackson Lewis mixed method for best fit.

The percentage threshold (`thrprc` in the plots) represents the fraction of data above the optimal threshold divided by the length of the sample (considering only positive values: i.e. only rainy days), or equivalently the *fraction of discarded data*. As shown by the Fig. 4.4 the Gerstengarbe plot has the lower percentage od discarded data ($0.4 \equiv 40\%$), while the JK and LW have the highest values: to pass the test with these two last methods we need to discard the 95% of data. For the AD and CV statistics the discarded part is around 70% of the total sample.

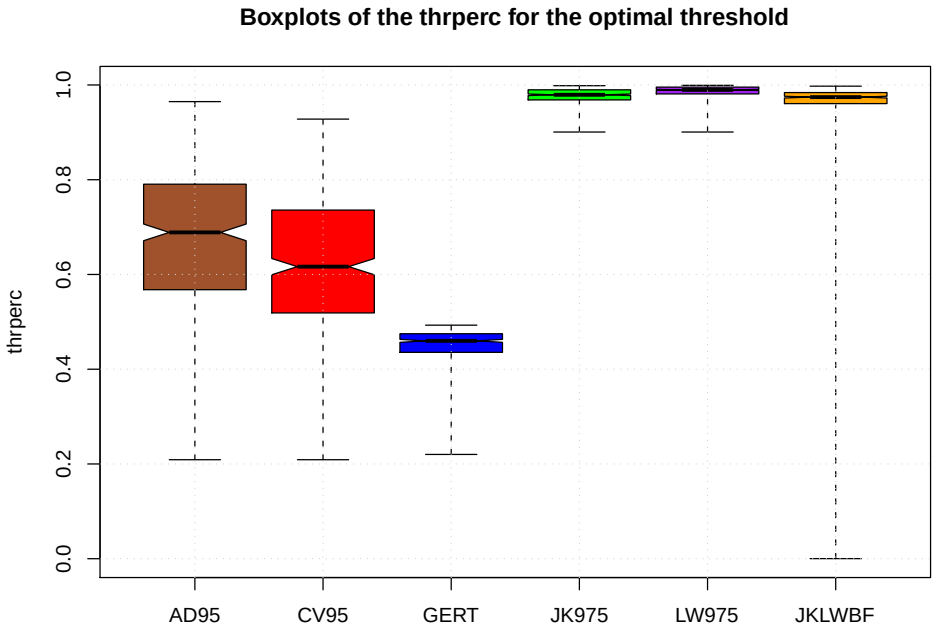

FIG. 4.4: Boxplot of the percentage threshold parameter computed for NOAA-NCDC database, using different methods. The acronyms are AD Anderson-Darling, CV Cramér Voin Mises, GERT Gerntensgarbe plot, JK95 Jackson kernel statistics with 95% confidence, LW95 Lewis kernel statistics with 95% confidence, JKLWBF Jackson Lewis mixed method for best fit.

Another criteria to evaluate the performance of the optimal threshold tests is looking at the distribution spread of the estimated thresholds (Fig. 4.5). Supposing that all stations are described in the same correct way by the GPD model then the broader is the distribution of values the less powerful will be the test and vice-versa. Indeed we expect that the best method will be the one able to find a narrower interval of optimal thresholds. With this rule the AD and CV statistics are the most powerful while the JK and LW are the less accurate.The Gerstengarbe plot have the lowest mean estimated threshold. The most representative values for the estimated thresholds are in the interval 4-8 mm of cumulated rainfall (see Fig.4.5). As a rule of thumb we can assume that a good initial threshold for further analysis will be 5 mm (value given by the best method the Anderson Darling).

We are interested in finding non trivial relationships between the threshold and the estimated GPD parameters, the median, the mean and the variance of data. We look for the linear correlation between each possible couple of variables (for instance, the mean and the threshold, the mean and the median and so on). Some linear relationship are trivial like the one between the mean and the median, other like the relationship between the mean (or the median) with the thrprc (percentage of discarded data at the optimal threshold) are more interesting.

We found that there is a negative linear correlation between the percentage of discarded data (thrpc) and the average daily cumulated rainfall. If we assume that the lower is the percentage

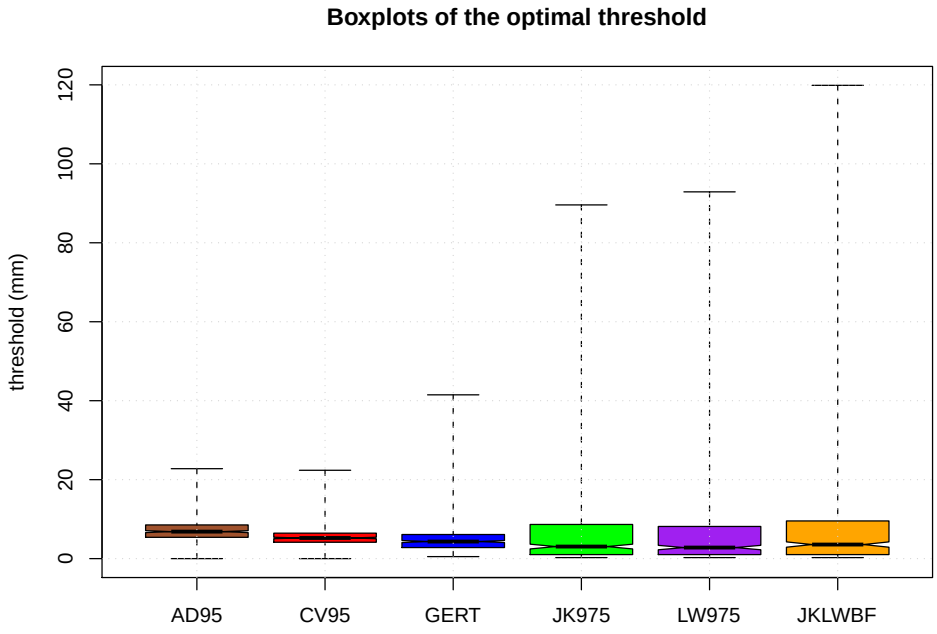

FIG. 4.5: Boxplot of the best estimated threshold computed for NOAA-NCDC database, using different methods . The acronyms are AD Anderson-Darling, CV Cramér Voin Mises, GERT Gerntensgarbe plot, JK95 Jackson kernel statistics with 95% confidence, LW95 Lewis kernel statistics with 95% confidence, JKLWBF Jackson Lewis mixed method for best fit.

of discarded data (thrprc) the better will be the GPD model in describing the dataset then the linear dependence with the average daily cumulated precipitation means that *the most rainy stations are better described by the GPD model than the driest one.* The values of the correlation coefficient are in the interval $(-0.81, -0.6)$ where the lower values is the one of the Gerstengarbe plot, the higher value is of the CV method.

A plot of the linear relationship between the percentage of discarded data (thrprc) and the mean daily cumulated rainfall is shown (for the Gertensgarbe statistics) in Fig. 4.6. This linear relationship proves that the GPD model becomes most effective when a station has an higher level of precipitation.

4.3.1 Geographical distribution of the best estimated threshold

The geographical analysis of the shape parameter estimated at the optimal threshold does not show any special pattern in the US region (Fig. 4.7 and Fig. 4.8) . Stations with high shape are mixed with stations with low shape in a quite uniform way, for the Gerstengarbe plot (Fig. 4.8) a little pattern is shown with higher shaped stations in the inland regions of the United states.

Conversely an evident geographical pattern is shown in the Fig. 4.9 and in the Fig. 4.10 where the percentage of discarded data at the optimal threshold is plotted in the US region. The most

FIG. 4.6: Linear relationship between the percentage of discarded data (thrprc) and the averaged daily cumulated rainfall (mm) for the NOAA-NCDC database. The negative correlation shows that the most rainy stations are the optimally described by the GPD model.

rainy states of southern US (Louisiana, Georgia, Alabama, Florida ...) are well fitted by the GPD model and the percentage of discarded data at the optimal level (thrprc) is around the 40%, while in the driest regions it can arrive to the 90%. The pattern for the Gertensgarbe plot of the same variables (Fig. 4.10) is the same.

4.3.2 Notes on the Hill estimator

The Hill estimator based methods like the JK and LW kernel statistics discard more than the 90% of data, leading to a great variance of the estimated GPD parameters (Fig. 4.3). To study the reasons of this behaviour we briefly analyse the theory of the Hill estimator.

Given a set of i.i.d random variables $x_1, x_2, \cdots, x_n$ described by a distribution F. Suppose that holds the relation:

$$1 - F(x) = x^{-1/\xi} l(x), \quad c > 0, \xi > 0 \tag{4.13}$$

where $l(x)$ is a slow varying function that has the property:

$$lim_{t\to\infty} \frac{l(tx)}{l(x)} = 1, \quad \forall\ t \neq 0$$

When the condition 4.13 is satisfied and it exists an index $k < N$ it is possible to estimate the shape parameter ξ using the Hill estimator :

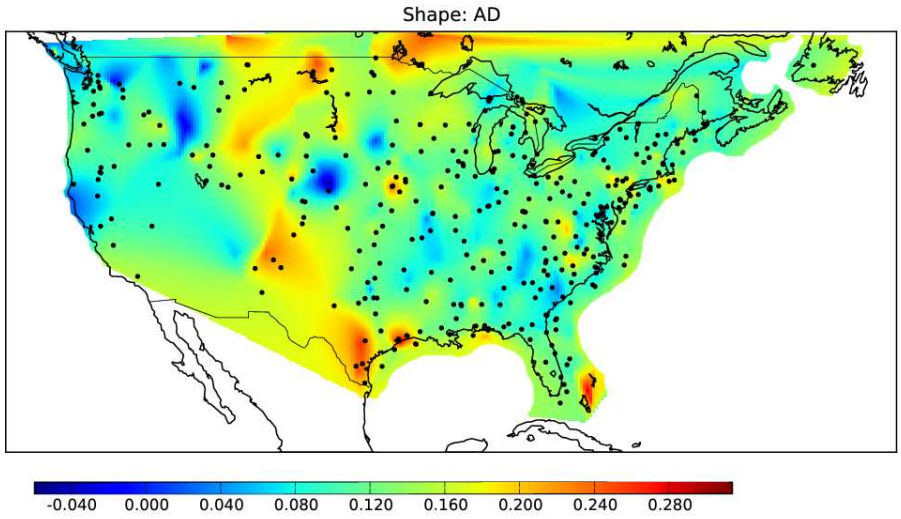

FIG. 4.7: Geographical distribution of the shape parameter for the optimal threshold estimated with the AD method. The shape parameter is distributed in a quite uniform way over the country.

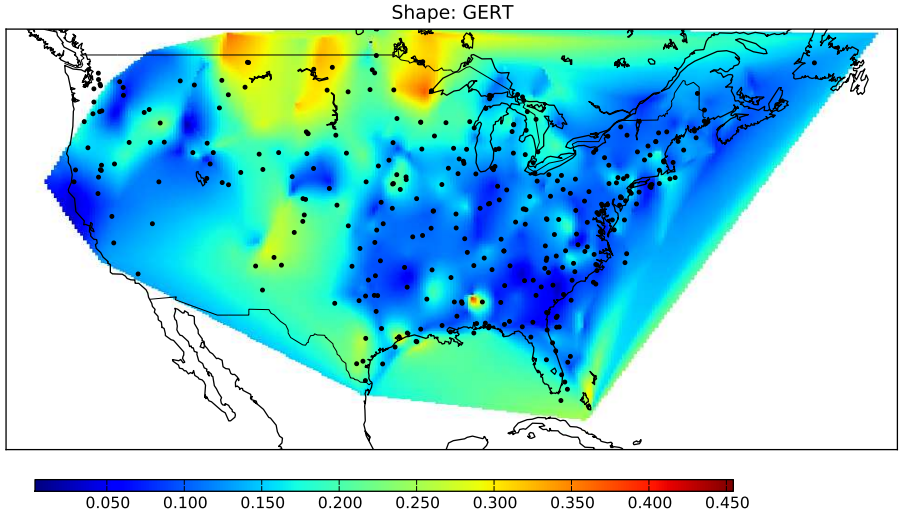

FIG. 4.8: Geographical distribution of the shape parameter for the optimal threshold estimated with the GERT method.The shape parameter is distributed in a quite uniform way over the country but a little pattern is shown for the inland and central regions.

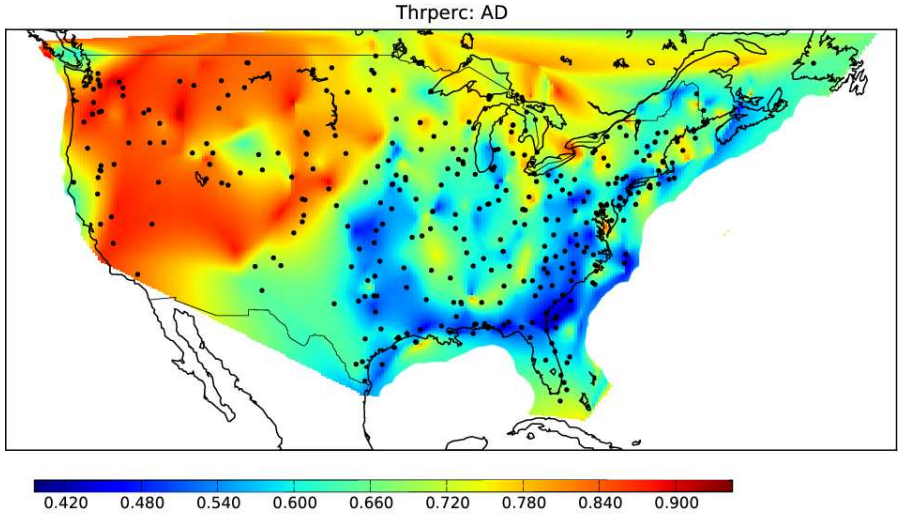

FIG. 4.9: Geographical distribution of the fraction of the discarded data (thrpc) at the estimated optimal threshold level. Medhod AD.

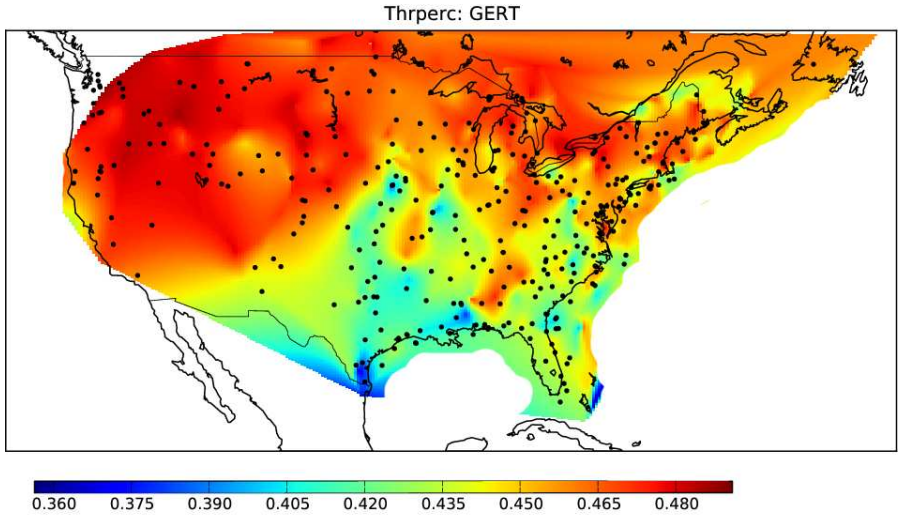

FIG. 4.10: Geographical distribution of the fraction of the discarded data (thrpc) at the estimated optimal threshold level. Medhod GERT.

$$\xi_{k,N} = \frac{1}{k}\sum_{i=1}^{k}(\log(x_{i,N}) - \log(x_{k,N})) \tag{4.14}$$

this equation holds in the asymptotic limits

$$k < N, \quad k \to \infty \quad and \;\; k/N \to 0$$

where the k/N ratio is called *truncation parameter* and for finite samples it is crucial to find an optimal value k_{opt} that ensures the correct estimation of the ξ parameter.

The Hill estimator has the serious drawback that it is not location invariant (see for instance Neves and Alves (2008)): the estimated shape parameter changes in function of the selected threshold. Furthermore the Hill plot in function of the k parameter is extremely volatile, and the convergence to the exact shape parameter is ensured in probability only if second order conditions are respected:

$$lim_{t\to\infty}\frac{\log(Q(tx)) - \log(Q(x)) - \xi\log(x)}{A(t)} = \frac{x^{\rho}-1}{\rho} \tag{4.15}$$

where the Q function is the quantile function of F and $A(t)$ is a slow varying function. The ρ parameter controls the convergence speed of the Hill parameter: if $|\rho|$ is high the convergence is faster. These conditions are exact in the asymptotic limit,that is for finite samples there's no clear indication how to find an optimal k_{opt} parameter.

To better investigate the efficiency of the Hill estimator and the relative GOF tests based on the kernel statistics of log-transformed data, we can perform several simple Monte Carlo simulations. The main idea is to explore the Normality of the Hill estimator and the Jackson and Lewis GOF statistics in the limiting region where the estimators hold $k < N, \quad k \to \infty \quad and \;\; k/N \to 0$. The procedure will be the following:

- Extract data from stations
- Choose an interval $k_a < k_b$ of upper extremes.
- Compute the Hill estimator or the GOF Lewis/Jackson statistics in the interval $k_a < k_b$.
- Perform a normality test with given confidence level for the Hill estimated parameter.

We use the Anderson Darling normality test (with both unknown normal parameters σ and μ) with the confidence level of 95%, see Anderson and Darling (1952). The results for the NOAA-NCDC stations are summarized in the table Tab. 4.2.

We can observe that

- the percentage number of stations passing the 95% AD normality test for the Hill estimator is low ($< 60\%$).

Estimator	k(25;75) %	k(25;125) %
Hill estimator	17.8	2.5
Jackson statistics	43.7	28.9
Lewis statistics	52.1	29.8

TAB. 4.2: Percentage of stations passing the normality test for Hill and Kernel GOF statistics with different sample size intervals

- the smoothing effect of the Kernel functions increases the fraction of the Hill estimate of the stations following a normal distribution.
- the Lewis statistics is slightly better than the Jackson one in smoothing data toward a normal distribution
- for finite samples shrinking the interval of upper extremes $k_a < k_b$ the fraction of stations passing the normality test increases, that is lesser variability

To study the convergence speed of the estimator toward the normality condition we can build simple Monte Carlo tests using samples of different size. We expect that using pure GPD data and bigger size samples the number of statistic samples of Hill, Lewis, Jackson type, passing the normality test increases with the size of the sample. In the 4.11 the fraction of pure GPD samples passing the normality test increases with the size of the sample but stay below of 50% (for a given interval $k_a < k_b$) even for samples with more than 40'000 points. From these simple tests we can affirm that for samples of finite size (less than 40'000 points) even pure GPD data show some problems in converging to a normal statistics for the Hill (or Hill-like) estimators at the tail of the distribution. Very likely the failure of the kernel statistics in finding the optimal threshold is due to this slow convergence to the real estimator value and also to the great volatility of the Hill estimator.

4.4 Final remarks on optimal threshold estimation

The methods described in the previous sections for optimal threshold estimation show a great variability. The Gerstengarbe plot give us the lower values of the best threshold while the asymptotic methods furnish much higher values. We highlighted the good performances of the Anderson Darling and Cramér Von Mises GOF techniques in finding a more compact set of candidate optimal thresholds. In all cases the prediction is depending on the characteristics focused by the test. If a test, like the JK or LW, gives more importance on the remote tail of the distribution it is logical to obtain higher thresholds.

We stress here the fact that the optimal threshold level is strongly dependent by the method used for the estimation. This concept could be generalized by the proposal of a new method for

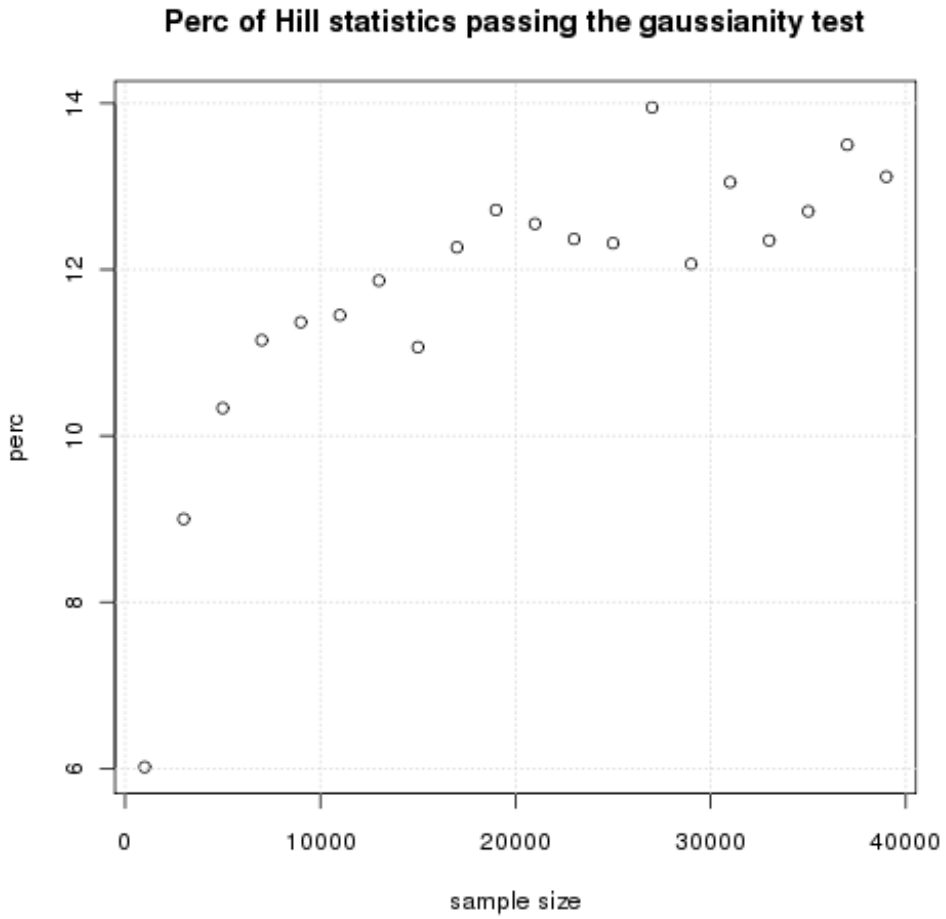

FIG. 4.11: Percentage of Hill statistics passing the normality test. Note the scarce efficiency of the test. The second order conditions - normality of the estimator sample - are not respected even for big samples.

optimal threshold estimation. We renounce to find the best *absolute* threshold, and we look for an optimal threshold depending on our needs. In the next section we propose a new method based on the quantile estimation that gives us the possibility to choose the area or areas of major interest in the distribution: that is we can choose to privilege the extreme values or to give more
5 importance to the central part of the distribution or even we can choose a mix of extreme or moderate values to control.

4.4.1 Proposal for a new method

To find the optimal threshold of a GPD distribution we have to choose first our interest areas in the data distribution. Let $x_1, x_2, \ldots, x_n$ the sample to study with length n. Define a subset of the x_i points:

$$\hat{\mathbf{x}} = \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m \quad m < n$$

we call these points of interest *pivots*. Let

$$Q(\hat{x};u) = \frac{1}{m}\sum_{i=1}^{m} w(\hat{x}_i)|q_{est,u}(\hat{x}_i) - q(\hat{x}_i)| \tag{4.16}$$

the averaged sum of the absolute values of the difference between the estimated quantiles $q_{est,u}$ (with the GPD fitted curve at the threshold u) and the real quantiles q of the data, both quantiles evaluated at the point $\hat{x}_i$. $w(\hat{x}_i)$ is an weighting function. In other words the function Q controls the distance of the fitted curve from the real data at the preferred pivot points (see Fig. 4.12). We define the optimal threshold u_{opt} for the pivot set $\hat{\mathbf{x}}$ as the minimum of the quantile sum:

$$u_{opt} \equiv \min\left(Q(\hat{\mathbf{x}})\right) = \frac{dQ(\hat{\mathbf{x}},u)}{du} = 0 \tag{4.17}$$

To estimate the u_{opt} we can take the derivative of the Q quantile sum function or simply, if the set of pivoting values is a finite quantity, we can check for every $u \in [a > \min(x), b < \max(x)]$ all the possible values of Q and pick up the minimum value. Remember that the dataset from rain gauges are discretized to the sensitivity of the instrument (0.1 mm or 0.254 mm for our databases) then the number of possible values of u to search is limited.

The introduced quantile estimator Q is a measure of the goodness of fit of our EV data in function of the threshold for a given set of pivoting points. Note that the method is not restricted to the GPD but could be useful for all distribution having a left censoring parameter.

The choice of the pivots points is critical and depend on the needs of the experimenter. To fit the tail of the distribution we can choose the extreme values, to fit the central part we choose pivots in the central part of the distribution, and we can give to each part a different weight. The weighting function plays an important role, because it allows to filter the influence of undesired values. Finally note that the fit technique (MLE, simple moments) must be robust and unbiased, moreover this tool to find the optimal threshold is not conceived to deal with contaminated data. In this case it is possible we need to choose best weighting functions or more robust versions of the Q and w functions.

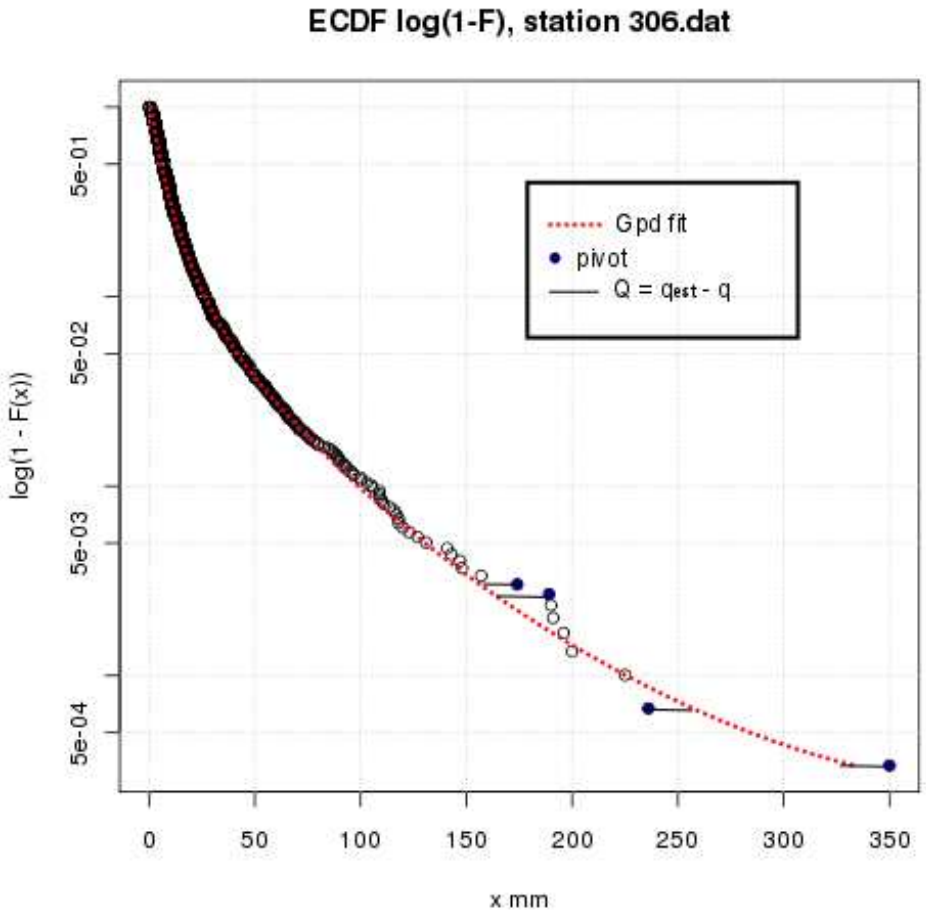

FIG. 4.12: An example of pivoting points (in blue) for the station 306 (Sardinian Database).

4.4.2 A practical example

As an example of this method we choose the following pivot points:

$$\hat{\mathbf{x}} = (x_{n-k}, x_{n-k+1}, \ldots, x_n) \quad with \quad x_1 < x_2 < \ldots < x_n$$

that is the greatest k points of the dataset. We estimate the optimal threshold using the Q quantile sum 4.16 with the weighting function equals to the unity $w(x) = 1$ for each threshold u level.

The Fig. 4.13 shows the behaviour of Q in function of the threshold for the NOAA-NCDC station WBAN 94930. The Fig. 4.14 is the distribution of the estimated optimal threshold on the NOAA-NCDC database with $k = 20$ pivoting points (the 20 greatest values of each recorded data). It is interesting to note that with this rule the optimal threshold of the stations has a net peak for 3 mm that is the smallest value of threshold among all the proposed methods.

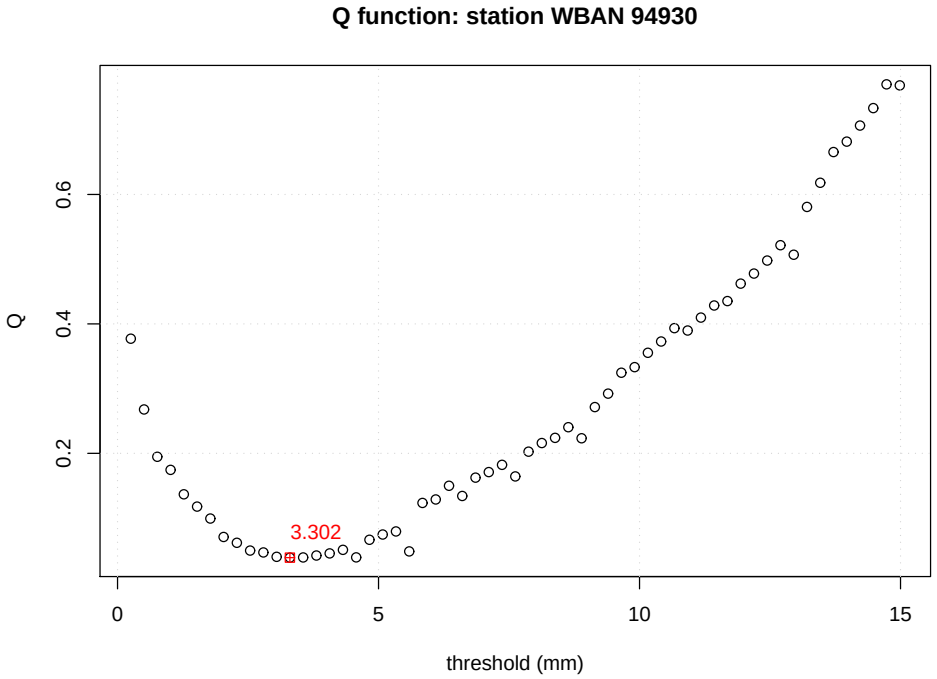

FIG. 4.13: Behaviour of the $Q(u)$ function in function of the threshold u

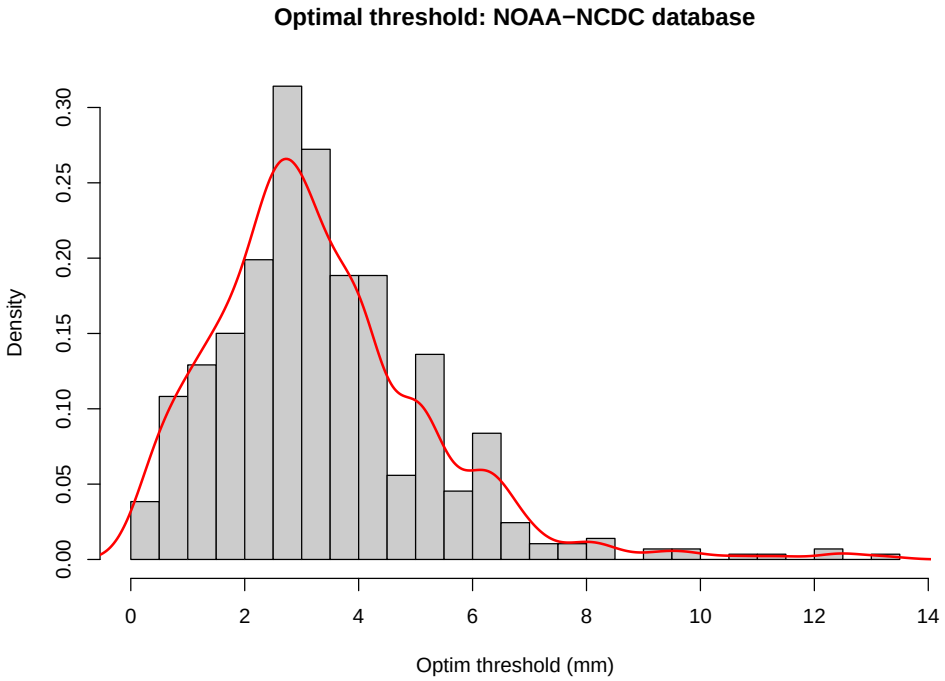

FIG. 4.14: Distribution of the optimal threshold for the NOAA database with $k = 20$ extreme pivoting points.

4.5 Testing the robustness of the GPD estimator in presence of rounding

In this section we show how the estimators of the GPD model are influenced by the presence of the rounding-off. We generate synthetic samples artificially rounded at various resolutions and then we fit the samples with several techniques in order to evaluate the estimator bias and variance in function of the rounding-off magnitude. Our goal is to find the best (and more robust) estimators.

The fitting techniques are those described in the Chapter 2.

4.5.1 Performance of the estimators

The performances of a parameter estimator depend on many factors: the sample size, the presence of spurious, multivariate or trended (not stable) data, the internal dependence (autocorrelation), the algorithmic stability and the precision of the estimation technique. For instance the MLE is efficient only for big samples while the PWM estimator is able to estimate the GPD parameters for samples of few points, but this one has a consistent bias for values of $\xi < 0$.

In order to compare the performances of the estimating techniques described in the Chapter 2, two groups of tests were carried out using Monte Carlo techniques. The first group has the aim to compare the performances of the estimators on continuous samples, while the second one aims to evaluate the performances on rounded-off records.

The performances are evaluated by the Bias and the root mean square error RMSE:

$$Bias = E(\theta_{est} - \theta_{true}) \tag{4.18}$$

$$RMSE = \sqrt{E\left[(\theta_{est} - \theta_{true})^2\right]} \tag{4.19}$$

where $\theta_{est}, \theta_{true}$ are the estimated and the true (i.e. used to generate the synthetic samples) values of the parameter respectively. In our case θ can be the ξ and/or the α parameter of the GPD.

Bias and RMSE are thus computed on synthetic samples generated by the GPD model with threshold $u = 0$ and (ξ, α) couples of parameters that can be considered representative of daily rainfall records, while the size of the synthetic sample is 500.

In the second group of tests the GPD synthetic samples are rounded off accordingly to different discretization magnitudes. In order to reduce the sampling errors in the Bias and RMSE computation a large number of samples will be necessary for each test (that is for each couple of ξ and α parameters and for each discretization value). Nevertheless, as a compromise with the computational time required by the tests, every Monte Carlo simulation is limited to 1000

synthetic samples. The consequent variability of Bias and RMSE estimates were filtered out by a Gaussian kernel smoother (Cleveland, 1979), run over the x-axis of each figure.

4.5.2 Tests over continuous GPD samples

In this group of tests, Bias and RMSE (see 4.19) were computed with Monte Carlo techniques on continuous samples generated by GPD with threshold $u = 0$, parameters $\alpha = 7$ and $\xi \in (-0.5, 0.5)$. In the plots of the Bias 4.15 and of the RMSE 4.16 we can evaluate the performances of the considered methods. For pure GPD data the reference estimator is the MLE, thus lines close to the MLE one give evidence of good performances. For instance, the MOMENTS estimator has a severe breakdown for $\xi > 0.3$ and a good accuracy for $\xi \approx 0$. The LME estimator fails for $\xi < -0.3$, but this region is of scarce importance in rainfall and flood time series analysis since distributions are usually unbounded. The performance of MDPD is very good for negative shape values, on the contrary the MED estimator, as highlighted by Juárez and Schucany (2004), has a poor efficiency (less than 20% of the MLE estimator) and an higher Bias: it becomes competitive only on contaminated samples. The PWM estimators (biased and unbiased) perform well for $\xi \approx 0.1$ where the efficiency is the best possible, nevertheless below this value the RMSE is worse than for the other methods. Finally for the penalized maximum likelihood MPLE we have good performances close to MLE ones.

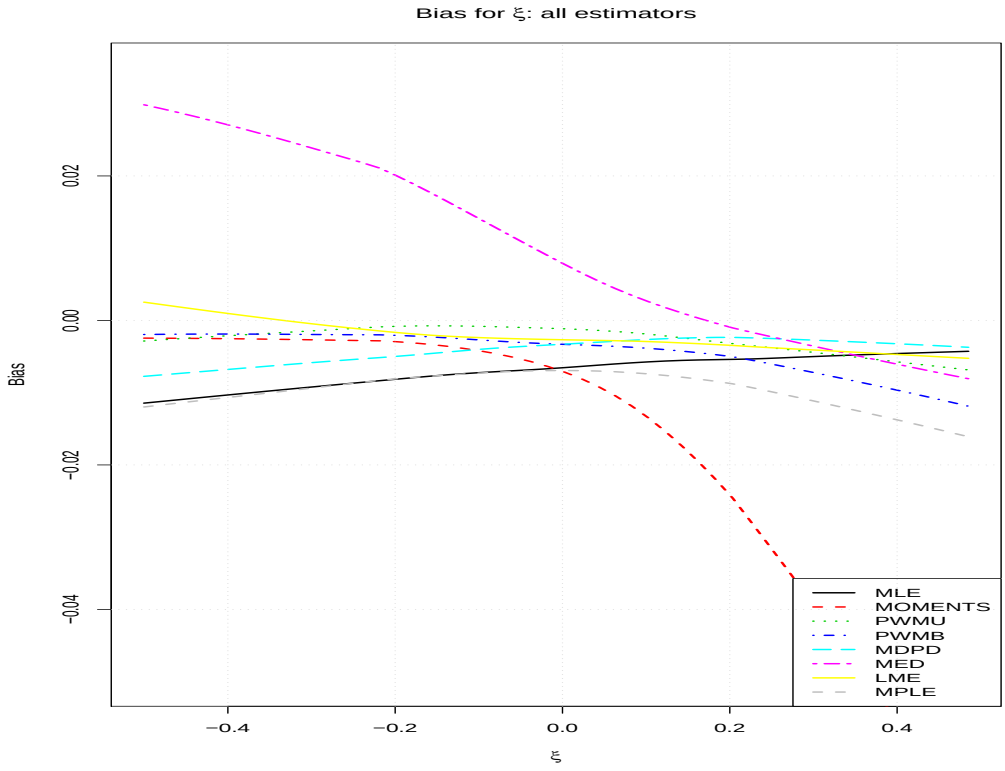

FIG. 4.15: Bias of shape parameter ξ for different GPD estimators. The Bias is computed with Monte Carlo techniques over continuous samples of size 500, generated by a GPD with threshold $u = 0$, $\alpha = 7$ and ξ in the range $(-0.5, 0.5)$. The final result is filtered by a robust Gaussian kernel smooth function (Nadaraya, 1964).

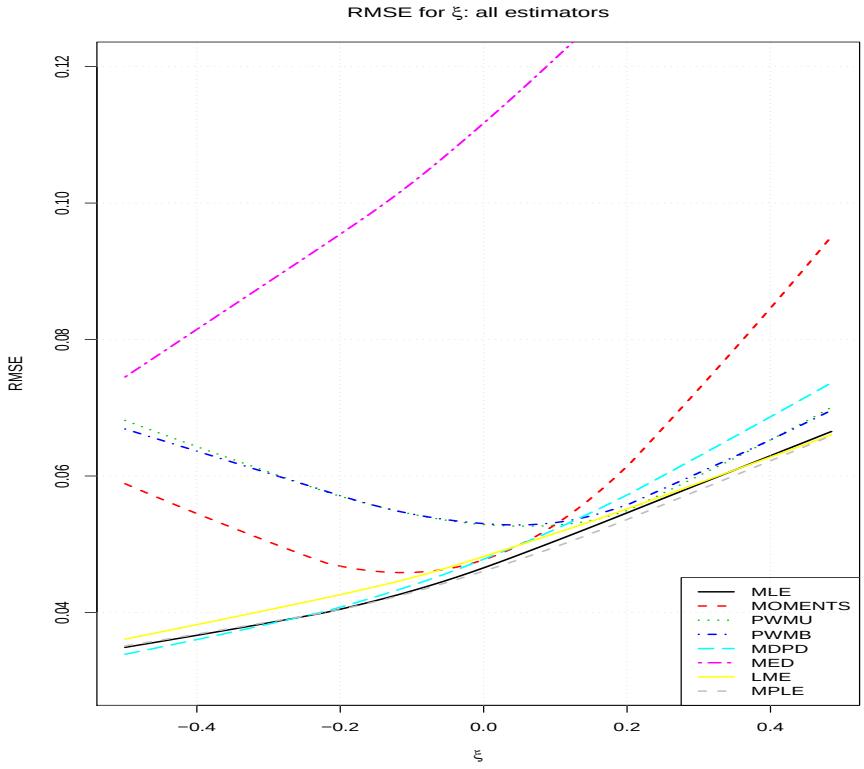

FIG. 4.16: As in Figure 4.15, but for RMSE of the shape parameter ξ.

4.5.3 Tests and performances over rounded-off GPD samples

This group of tests investigates the performances of the estimators over rounded-off samples. We have already noted the presence of rounded-off records with a mixture of different resolutions (0.1, 0.2, 0.5, 1.0 and 5.0 mm) in the Sardinian database. To simplify the interpretation of the results of our analysis we consider here a single rounding-off δ for each test. The magnitude of δ explores the entire range from 0 to 5 mm with increments of 0.1 mm. We expect that the Bias and the RMSE (4.19) increase with the magnitude of the rounding-off. This result could be explained with geometrical arguments: the empirical cumulative distribution function of a rounded-off sample displays a step-like behaviour that produces a big uncertainty in the estimate: then bigger steps will produce greater incertitude in the estimate. To explore parameters in the domain of interest of daily rainfall records we tested the Bias and the RMSE for $\xi \in \{0., 0.15, 0.3\}$ and $\alpha \in \{7., 12.\}$ that are representative of parameter values estimated over the 200 longest datasets of the Sardinia Region (Deidda, 2007).

Figures 4.17 and 4.18 show the Bias of ξ and α respectively as a function of the rounding-off resolution: each subplot is obtained by Monte Carlo generations of GPD samples with fixed (ξ, α) parameter couples, obtained by the combination of the values reported above. Both for ξ and α the Bias increases with the magnitude of the discretization. A similar behaviour is displayed by Figures 4.19 and 4.20 where the RMSE of ξ and α is plotted again versus the rounding-off resolution. The comparison of RMSE and Bias for rounded-off samples shows noticeable differences with respect to the continuous case (Figures 4.15 and 4.16). The LME

estimator has an evident failure in the ξ estimates for rounding-off larger than 1 mm while in the continuous case it has very good performances close to (or even better than) the MLE ones. Similar behaviour is displayed by the PWM estimators: we highlight that the tests on rounded samples are performed with the values $\xi \in \{0., 0.15, 0.3\}$ where the PWM performances in the continuous case are very close to MLE. In Figures 4.19 and 4.20, displaying the RMSE, we can observe that the MOMENTS estimator performs better for high ξ values showing an opposite behaviour with respect to the continuous case where the MOMENTS efficiency becomes worst for large ξ values. The MOMENTS efficiency in some cases is even better than the MLE one.

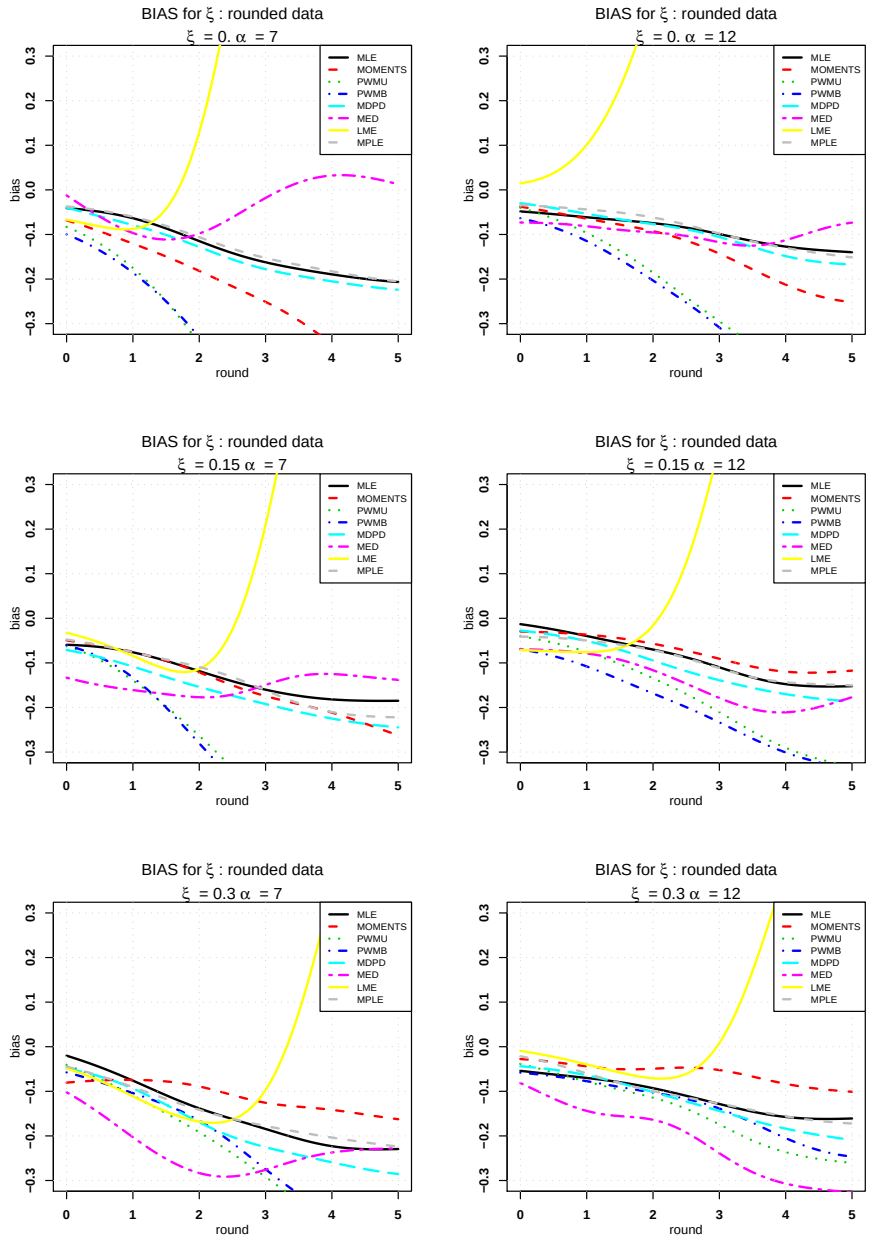

FIG. 4.17: Bias for the shape parameter ξ estimated with different techniques on rounded-off samples. Results are presented as a function of rounding-off magnitude that ranges from 0 to 5 mm. Subplots refer to different couples of shape and scale parameters (see subtitles) selected in the range of representative values of daily time series.

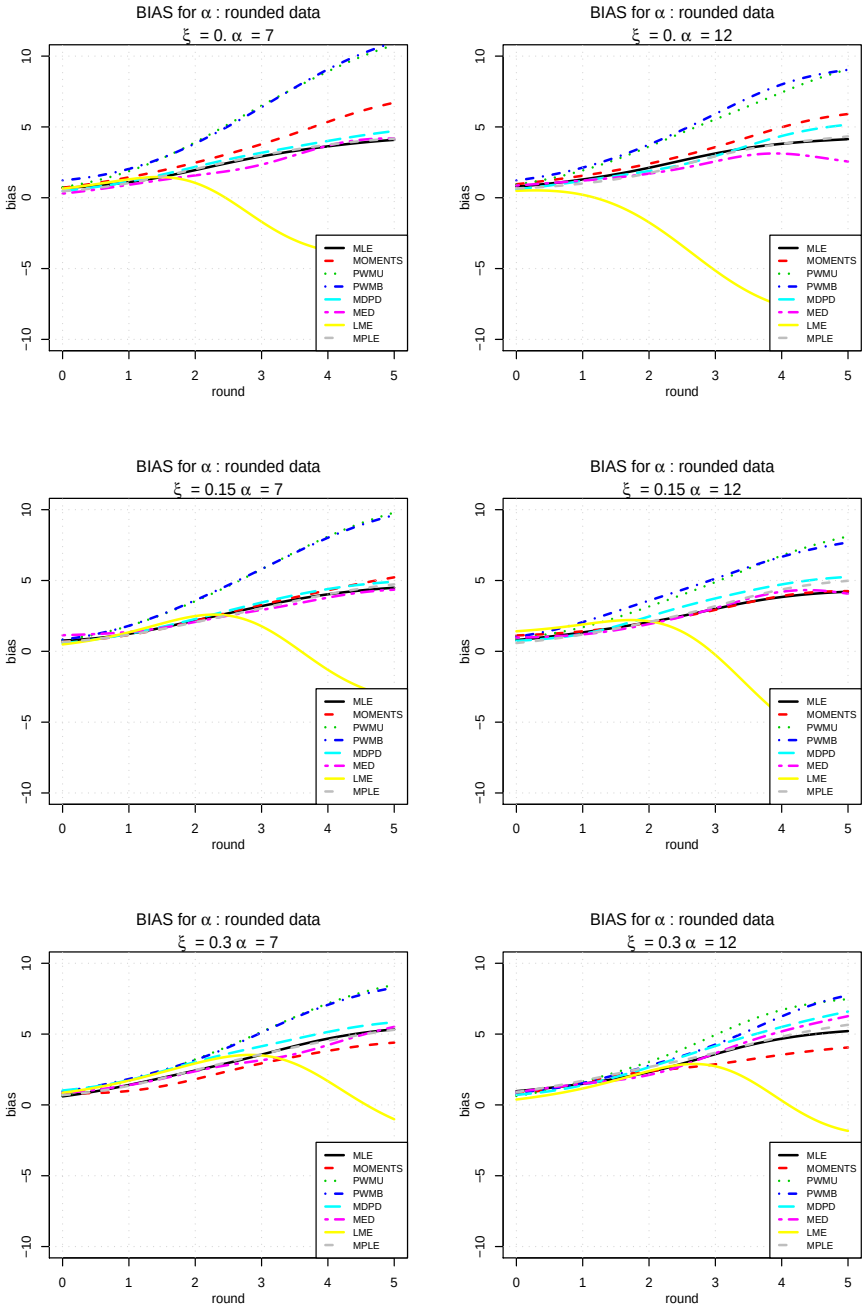

FIG. 4.18: As Figure 4.17, but for Bias of the scale parameter α.

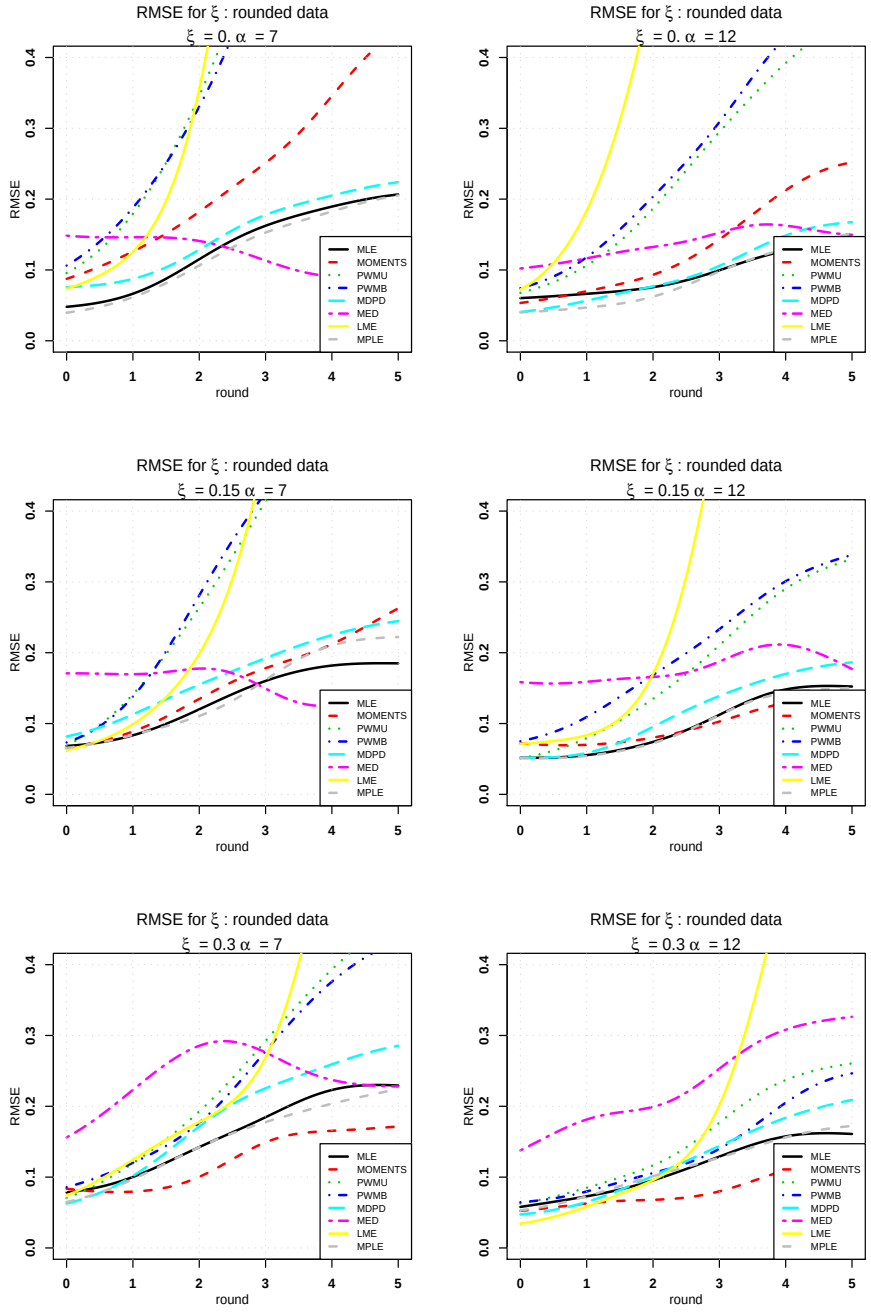

FIG. 4.19: As Figure 4.17, but for RMSE of the shape parameter ξ.

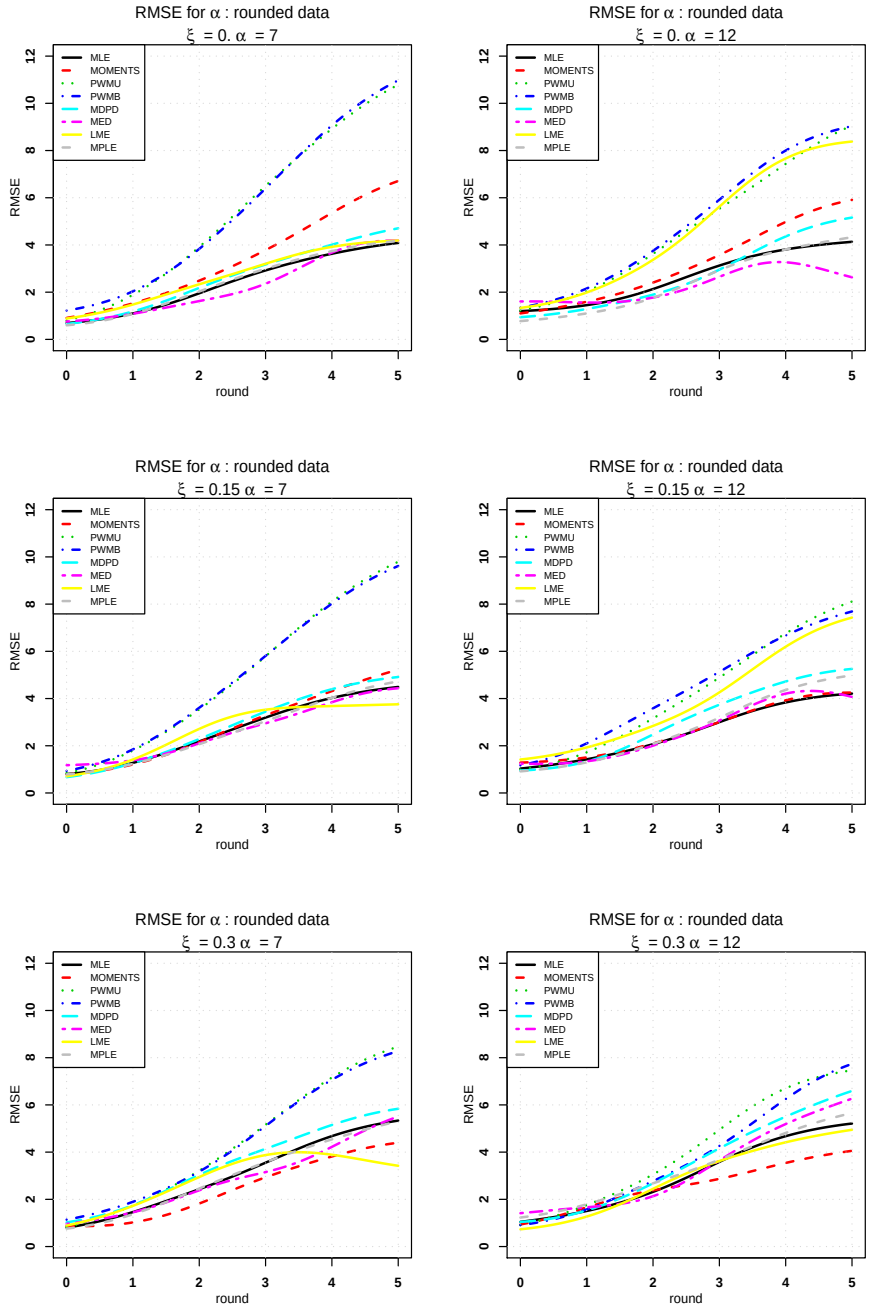

FIG. 4.20: As Figure 4.17, but for RMSE of the scale parameter α.

Finally, besides the above considerations on the relative performances of a given estimator with respect to the others, we highlight that *all methods provide parameter estimates affected by unacceptable errors*. In some cases the Bias and the RMSE are larger than the 100% of the parameter value of the parent distribution, even when we select the best estimator for the specific case. For instance looking at the RMSE that is the most significant index (it includes also the Bias) we can observe from the plots in Fig. 4.19 that even for the best estimator the error is often around $0.1 \div 0.2$ for the ξ parameter that usually assumes values between 0.0 and 0.5 in hydrological applications. Similar considerations hold also for Fig. 4.20 where RMSE of

α is often about 4 mm and more, while α estimates for daily rainfall depths are usually of order of 10 mm.

4.6 Final remarks on GPD estimators for rounded-off values

The performances of the estimators in presence of rounding-off are not particularly good. Neither the robust estimators are safe to the rounding-off: when the amplitude of the rounding increases the bias and the variance become soon unacceptable. This result encourages the creation of more specific and resistant (robust) estimators to deal with rounded-off distributions. Remember that every experimental distribution is rounded at least at the sensitivity level, then we need to consider this source of moderate/strong bias and perform a check for the presence of rounding (as done in the previous Chapter).

Finally we recommend to estimate the optimal threshold and *after* to fit the distribution deciding a priori the areas of interest in the dataset. In this way it is possible to filter some undesired features like the strong influence of the rounding for the lower values. Moreover remember to prior check the outliers otherwise the extreme quantiles will be overestimated and the pivoting method will fail.

Chapter 5

Conclusions and future researches

5.1 Toward Big Data analysis

Climate change is a popular and most debated topic of contemporary research. The datasets of precipitation, ice coverage, snow falling , temperatures and other physical variables are stored in datacenters of several intergovernmental institutions. The most famous one is the IPCC (Intergovernmental Panel for Climate Change) that collects data from measured weather stations in the world and data from numerical climatic models. In Europe the EU financed a project called ENSEMBLES to collect and store the simulations of the climate models arising from several Universities and Research Institutions of Europe, Japan and USA. After few years of data collection these databases became of huge dimensions. Just for the precipitation datasets, the Regional Climate models (that perform climate simulation on the scale of the Europe with a spatial resolution around 25 km at the medium latitudes) the ENSEMBLE datacenter `http://ensemblesrt3.dmi.dk/` stores almost 180 GB of data. If we consider that the variables involved in the models are several dozens then the total size of the database passes the hundreds of TB and approaches to the fantastic quantity of a PB of total information. Now just imagine to download, store and analyse even a small fraction of this database, you need days of continuous download and several days of effective analysis. You have to realize a good system to extract, load and evaluate this extraordinary amount of data. This is not a cut and paste work on Excel tables, we need powerful tools, massive parallel computing and smart data extraction tools. We need to deal with the Big Data challenge.

Furthermore there is another problem with these datacenter, they have a poor level of data classification. Each file has its metadata, that is informations about the contents, the climatic model, the unities of measure, the data structure, and so on but you can access to these informations only downloading the entire file. Downloading the file to explore the metadata is time consuming and for this reason many data center implements a protocol (OpenDap `http://opendap.org`) that extracts a single portion of the data or simply display the metadata of each file. However this effort is not enough, they do not provide a systematic access of the information of the same

level like the one that you can obtain with a database query tool. From these considerations a possible new research project will be the realization of a software able to:

- explore the huge datasets of climatic models to collect information
- explore the metadata of the files of the datasets generating reports and searchable catalogues
- generating a map and a comprehensive file list of each file of the models present in the data center
- downloading only the parts of interest of the data
- performing simple analysis on the data without downloading them to the local hard disk
- mapping the analysis in parallel to several computers and reconstructing the result in a final summary
- cutting the climatic data to specific areas of interest like for instance watersheds

Hopefully some prototype of this system will be ready in the next months and as a possible link with the extremes value research this tool can be used to extract the weather configuration that generates the extreme events.

5.1.1 How the Exploratory statistics could be useful to work with simulated (and real) climatic data

The materials and the knowledge on Exploratory statistics and Hypothesis test accumulated during the thesis work could be used to provide an analysis framework for the climate data. Data from models must be checked for possible artefacts, for instance the presence of several mixed populations in the precipitation database. A separate tool of analysis could be realized to perform the following task of data assessment :

- outliers check: candidate points must be further analysed with statistical tools
- confrontations with real data (tools for test of hypothesis)
- trend detection using the variance theory (ANOVA methods)
- extreme events detection: is it possible to detect extremes in the output of the climate models ?
- studying the variations in the extreme events during the future

the structure of the software could implement each task in a separate plugin. The starting point will be the data availability.

5.2 Conclusions

In this thesis we analysed two hydrological databases, our approach was data oriented: we renounced to do too many a priori hypothesis on the quality of data, on their distribution on the homogeneity and lack of outliers and artefacts. We discovered the problems that a rounded-off dataset can cause on the estimation of the very sensitive models for EVT. Finally we addressed the problem of finding an optimal threshold level for the GPD distribution with a change of perspective no more an absolute optimal threshold but a threshold chosen to satisfy some desired features of the studied dataset (i.e. a good fit to the most extreme data). The importance of this exploratory analysis is now ready to be used in the Big Data challenge where the size of the databases is so huge that we must be smart enough to consider only the needed data but just only after the assessment of the data quality.

Appendix A

Test for GPD outliers

In this section we describe a test to check the candidate GPD outliers with a given confidence level. The method is based on the idea to find the averaged ratio between the two biggest values of a GPD distribution, let x_{max} and x_{max-1} these values. We are interested in finding confidence levels for the ratio:

$$r = \frac{x_{max}}{x_{max-1}} \tag{A.1}$$

To obtain the confidence levels for r we create a Monte Carlo simulation with the following steps:

- Create m samples of length n and shape parameter varying in the range $\xi \in [a, b]$ where for practical purposes $(a, b) = [-0.8, 0.8]$.
- Compute the m ratios r for each value of ξ, let $r_\xi(i)\ \ i = 1, 2, \ldots, m$ these values.
- Reorder the distribution of r_ξ and save the quantile levels, 90th, 95th and 99th. We'll obtain a curve for each couple (q, ξ) where q is the quantile level.

The Fig. A.1 shows the curves for the ratio r at several values of the shape ξ. The values are reported in the table A.1 for use and reference. The missing values for ξ could be interpolated linearly or cubically.

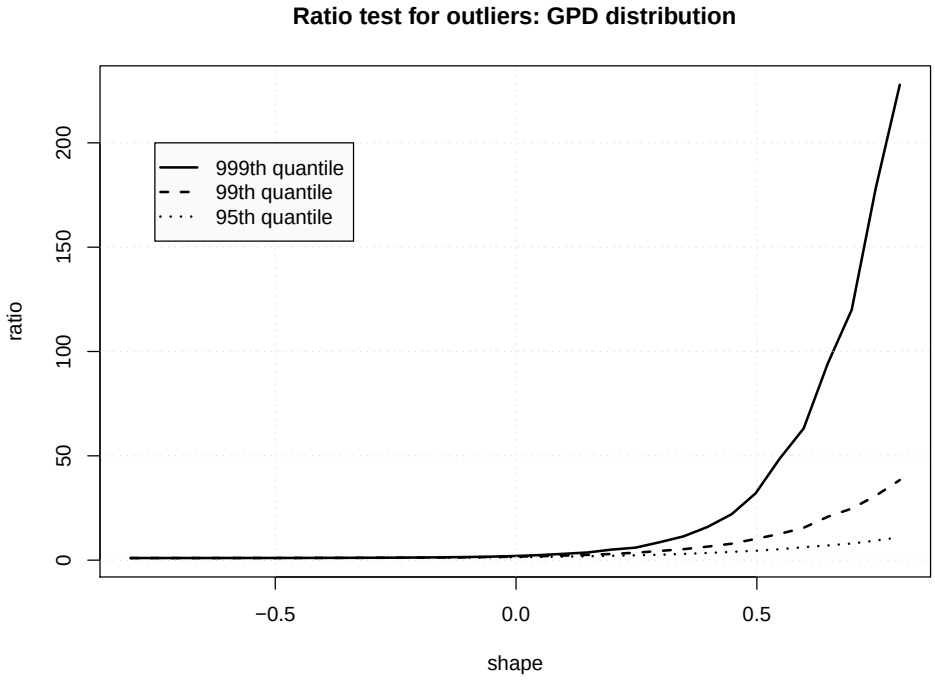

FIG. A.1: Quantile curves of the ratio r of the two biggest values for a GPD distribution in function of the shape parameter ξ

ξ	95%	99%	99.9%	ξ	95%	99%	99.9%
-0.80	1.00	1.01	1.01	0.05	1.54	1.87	2.42
-0.75	1.01	1.01	1.01	0.10	1.69	2.17	3.03
-0.70	1.01	1.01	1.02	0.15	1.86	2.53	3.67
-0.65	1.01	1.02	1.02	0.20	2.09	2.96	5.05
-0.60	1.02	1.02	1.03	0.25	2.31	3.56	5.99
-0.55	1.02	1.03	1.04	0.30	2.60	4.40	8.57
-0.50	1.03	1.04	1.05	0.35	2.97	5.27	11.40
-0.45	1.04	1.05	1.06	0.40	3.46	6.48	15.92
-0.40	1.05	1.07	1.09	0.45	3.96	7.87	22.00
-0.35	1.06	1.09	1.11	0.50	4.51	10.19	32.03
-0.30	1.08	1.11	1.15	0.55	5.22	12.58	48.61
-0.25	1.11	1.15	1.20	0.60	6.24	15.53	63.14
-0.20	1.15	1.21	1.27	0.65	7.05	20.78	94.15
-0.15	1.19	1.28	1.36	0.70	7.98	24.72	120.02
-0.10	1.25	1.37	1.50	0.75	9.39	30.89	178.32
-0.05	1.33	1.49	1.68	0.80	10.98	38.40	227.83
-0.00	1.42	1.66	2.00				

Tab. A.1: Ratio test for the outliers of the GPD distribution in function of the shape parameter ξ

Bibliography

[1] Anderson T.W. and Darling D.A. Asymptotic theory of certain goodness-of-fit criteria based on stochastic processes. *Annals of Mathematical Statistics*, 23:193–212 (1952).

[2] Ashkar F. and Mahdi S. Fitting the log-logistic distribution by generalized moments. *Journal of Hydrology*, 328(3-4):694 – 703 (2006). doi:DOI:10.1016/j.jhydrol.2006.01.014.

[3] Beguería S. Uncertainties in partial duration series modelling of extremes related to the choice of the threshold value. *J. Hydrol.*, 303:215–230 (2005).

[4] Beirlant J., Teugels J.L., and Vynckiee P. Practical analysis of Extreme Values. Leuven University Press (1996).

[5] Choulakian V. and Stephens M.A. Goodness-of-fit tests for the generalized pareto distribution. *Technometrics*, 43:478–484 (2001).

[6] Chow V.T., Maidment D.R., and Mays L.W. Applied hydrology. McGraw-Hill, Singapore (1988).

[7] Cleveland W.S. Robust locally weighted regression and smoothing scatterplots. *Journal of American Statistics Association*, 74:829–836 (1979).

[8] Coles S. and Dixon M. Likelihood-based inference for extreme value models. *Extremes*, 2(1):5–23 (1999).

[9] Daszykowski M., Kaczmarek K., Heyden Y.V., and Walczak B. Robust statistics in data analysis - a review basic concepts. *Chemometrics and Intelligent Laboratory Systems*, 85:203–219 (2007).

[10] Deidda R. An efficient rounding-off rule estimator: Application to daily rainfall time series. *Water Resour. Res.*, 43:W12405 (2007). doi:10.1029/2006WR005409.

[11] Deidda R. and Puliga M. Sensitivity of goodness of fit statistics to rainfall data rounding off. *Phys. Chem. Earth*, 31:1240–1251 (2006). doi:10.1016/j.pce.2006.04.041.

[12] Diebolt J., Guillou A., Naveau P., and Riberau P. Improving probability-weighted moment methods for the generalized extreme value distribution. *Statistical Journal*, 6:33–50 (2008).

[13] Fisher R.A. and Tippett L.H. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In Mathematical Proceedings of the Cambridge Philosophical Society, volume 24, pages 180–190 (1928). doi:10.1017/S0305004100015681.

[14] Gerstengarbe F.W. and Werner P.C. A method for the statistical definition of extreme-value regions and their application to meteorogical time series. *Zeitschrift für Meteorologie*, 30:224–226 (1989).

[15] Gnedenko B.V. Sur la distribution limite du terme maximum d'une série aléatoire. *Ann. Math.*, 44:423–453 (1943).

[16] Goegebeur Y., Beirlant J., and de Wet T. Linking pareto-tail kernel goodness-of- fit statistics with tail index at optimal threshold and second order estimation. *Statistical Journal*, 6:51–69 (2008).

[17] Greenwood J.A., Landwehr J.M., Matalas N.C., and Wallis J.R. Probability weighted moments: definition and relation to parameters of several distributions expressable in inverse form. *Water Resour. Res.*, 15(5):1049–1054 (1979).

[18] Grimshaw S.D. Computing maximum likelihood estimates for the generalized pareto distribution. *Technometrics*, 35:185–191 (1993).

[19] Grubbs F.E. Sample criteria for testing outlying observations. *Annals of Mathematics and Statistics*, 21:27–58 (1950).

[20] Henze N. and Meintanis S.G. Tests of fit for exponentiality based on the empirical laplace transform. *Statistics*, 36:147–161 (2002).

[21] Hill B.M. A simple general approach to inference about the tail of a distribution. *Ann. Stat.*, 3(5):1163–1174 (1975).

[22] Hosking J.R.M. L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *J. Roy. Stat. Soc. B Met.*, 52(1):105–124 (1990).

[23] Hosking J.R.M. and Wallis J.R. Parameter and quantile estimation for the generalized pareto distribution. *Technometrics*, 29:339–349 (1987).

[24] Hosking J.R.M. and Wallis J.R. A comparison of unbiased and plotting-position estimators of l-moments. *Water Resour. Res.*, 31:2019–2025 (1995).

[25] Hosking J.R.M., Wallis J.R., and Wood E.F. Estimation of the generalized extreme value distribution by the method of probability-weighted moments. *Technometrics*, 27:251–261 (1985).

[26] Iglewicz B. and Hoaglin D. How to Detect and Handle Outliers, volume 16. The ASQC Basic References in Quality Control: Statistical Techniques (1993).

[27] Juárez S. and Schucany W.R. Robust and efficient estimation for the generalized pareto distribution. *Extremes*, 7(3):237–251 (2004).

[28] Kendall M. Rank Correlation Methods. Griffin, London (1975).

[29] Kotz S. and Nadarajah S. Extreme Value Distributions: Theory and Applications. Imperial college press (2001).

[30] Kotz S. and Nadarajah S. A generalized logistic distribution. *Int. Journal of Mathematics and Mathematical sciences*, 19:3169–3174 (2004). doi:10.1155/IJMMS.2005.3169.

[31] Landwehr J.M., Matalas N.C., and Wallis J.R. Probability weighted moments compared with some traditional techniques in estimating gumbel parameters and quantiles. *Water Resour. Res.*, 15:1055–1064 (1979).

[32] Madsen H., Pearson C.P., and Rosbjerg D. Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events 2. regional modeling. *Water Resour. Res.*, 33(4):759–770 (1997).

[33] Makkonen L. Bringing closure to the plotting position controversy. *Communication in Statistics, Theory and Methods*, 37:460–467 (2008a).

[34] Marsaglia G. and Marsaglia J. Evaluating the anderson-darling distribution. *Journal of statistical software*, 9:2–6 (2004).

[35] Nadaraya E.A. On estimating regression. *heory of Probability and its Applications*, 9(1):141–142 (1964). doi:doi:10.1137/1109020.

[36] Neves C. and Alves M.I.F. Testing extreme value conditions - an overview and recent approaches. *Statistical Journal*, 6(1):83–100 (2008).

[37] Peng L. and Welsh A.H. Robust estimation of the generalized pareto distribution. *Extremes*, 4(1):53–65 (2001).

[38] Pickands J. Statistical inference using extreme order statistics. *Ann. Stat.*, 3:119–131 (1975).

[39] Reiss R.D. and Thomas M. Statistical Analysis of Extreme Values. Birkhäuser Basel (2007).

[40] Ribatet M. Pot: Modelling peaks over a threshold. *R News*, (7):34–36 (2007).

[41] Rosbjerg D., Madsen H., and Rasmussen P.F. Prediction in partial duration series with generalized pareto-distributed exceedances. *Water Resour. Res.*, 28(11):3001–3010 (1992).

[42] Sen P.K. and Singer J.M. Large sample methods in statistics. Chapman & Hall, Inc. (1993).

[43] Sillitto G.P. Derivation of approximants to the inverse distribution function of a continuous univariate population from the order statistics of a sample. *Biometrika*, 56:641–650 (1969).

[44] Stedinger J.R., Vogel R.M., and Foufoula-Georgiou E. Frequency analysis of extreme events. In D.R. Maidment, editor, Handbook of Hydrology, chapter 18. McGraw-Hill (1993).

[45] Zhang J. Likelihood moment estimation for the generalized pareto distribution. *Aust. Nz. J. Stat.*, 49(1):69–77 (2007).