

Estimation of upper bounds for return levels: Methodology and illustrations on univariate random sequences

Pascal Alain Dkengne Sielenou, Stéphane Girard

▶ **To cite this version:**

Pascal Alain Dkengne Sielenou, Stéphane Girard. Estimation of upper bounds for return levels: Methodology and illustrations on univariate random sequences. [Research Report] Inria - Research Centre Grenoble – Rhône-Alpes. 2022. hal-03562999

HAL Id: hal-03562999
https://inria.hal.science/hal-03562999

Submitted on 9 Feb 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Estimation of upper bounds for return levels

Methodology and illustrations on univariate random sequences

Pascal Alain Dkengne Sielenou & Stéphane Girard

Inria Grenoble Rhône-Alpes

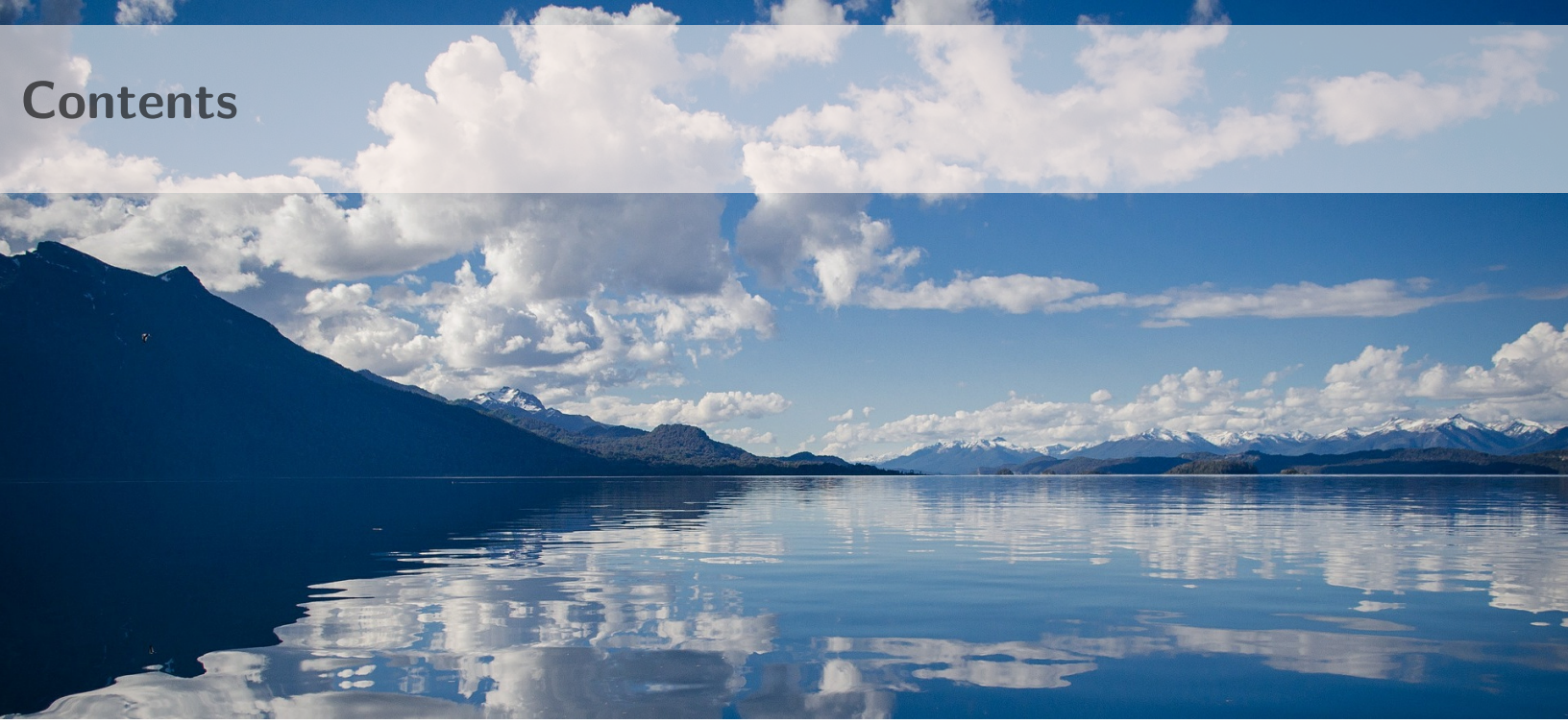
Copyright © 2021 Pascal Alain Dkengne Sielenou & Stéphane Girard

Open License

Art. No xxxxx
ISBN xxx–xx–xxxx–xx–x
Edition 1.1

Published by Inria Grenoble Rhône-Alpes
Printed in France

Contents



1 Introduction 5

2 Basics on GEV model for extreme values analysis 6

2.1 Method of block maxima 6

2.2 Estimation of return levels 10

2.3 Estimation of extremal index 11

2.4 Selection of block size 13

3 Determination of upper bounds for return levels 18

3.1 Routine procedure 18

3.2 Main algorithm 21

3.3 Diagnostic test 22

3.4 Heuristic justification 23

4 Illustrations on univariate stochastic processes 28

4.1 IID cases 28

4.2 Non-IID cases 34

4.3 Cautions 38

5 Illustrations on univariate time series in practice 48

5.1 Operating framework 48

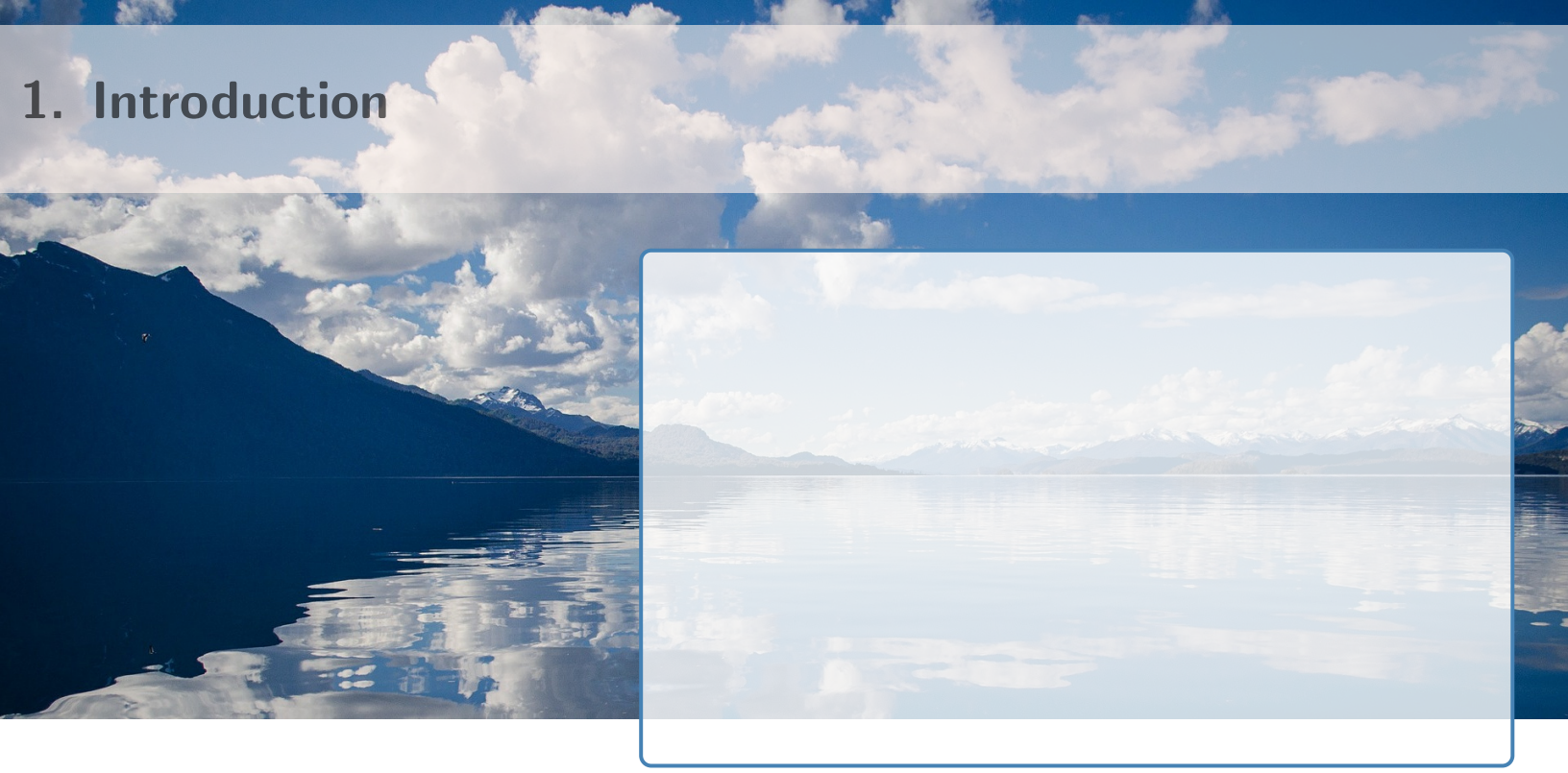
5.2 Data and goals 49

5.3 Main results 51

6 Conclusions 57

Literature 57

1. Introduction



This work addresses the issue of estimating return levels associated with return periods longer than the observation duration of a random phenomenon.

Equivalently, the problem is to estimate the probabilities associated with the occurrence of unobserved extreme values of a random phenomenon.

Extreme value theory offers statistical models from which estimators of these quantities of interest can be constructed.

However, the resulting estimators depend on a parameter which is essential to extract a sample of extreme values to model.

According to the used extreme value probability distribution family, this parameter can be either a threshold corresponding to the value beyond which an observation is considered as extreme, or an integer value corresponding to the number of consecutive observations whose maximum is considered as extreme.

Moreover, the choice of such a parameter is quite tricky in practice since an inappropriate value, even large enough, can lead to an underestimation of the target critical values.

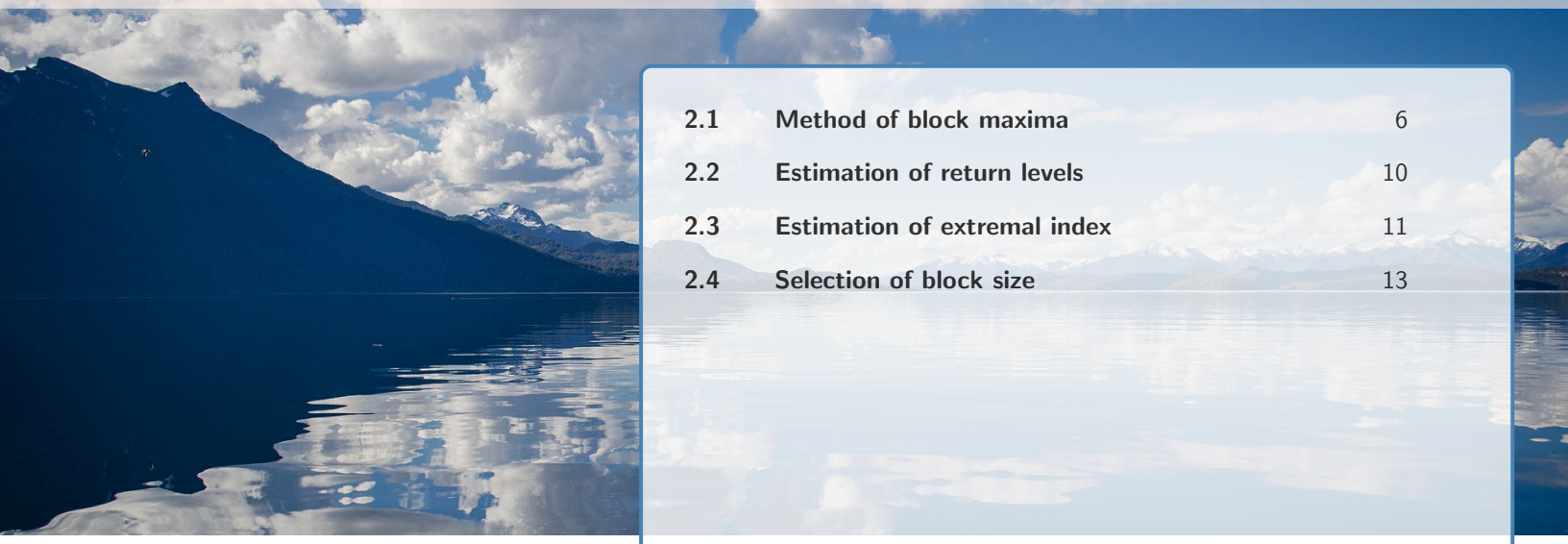
This study addresses this problem by setting the objective of constructing estimators for upper bounds of return levels.

This deviation from the initial problem can be justified by the fact that an appropriate strategy for managing a high risk remains effective for managing lower risks.

The rest of this document is organized as follows.

- *Chapter 2 briefly presents some results from extreme value theory while introducing the theoretical foundation of the proposed strategy.*
- *Chapter 3 contains not only a detailed description of our algorithmic strategy but also its theoretical justification.*
- *Chapter 4 shows the validity of this strategy on simulated data from some continuous random variables and stationary processes.*
- *Chapter 5 illustrates this strategy on real data in which the goal is to quantify the upper bounds of extrapolations associated with the localization errors of a vehicle on the road.*

2. Basics on GEV model for extreme values analysis



2.1	Method of block maxima	6
2.2	Estimation of return levels	10
2.3	Estimation of extremal index	11
2.4	Selection of block size	13

Introduction

The block maxima approach is one of the main methodologies in extreme value theory to obtain a suitable model for extrapolations.

In this approach, the block size is usually selected in order to reflect the possible intrinsic periodicity of the studied phenomenon.

The generalization of this approach to data from non-seasonal phenomena is not straightforward.

To address this problem, we propose in this chapter a data-driven method to identify the candidate block sizes to focus on.

2.1 Method of block maxima

Let X be a random variable (associated with the phenomenon of interest) for which we want to assess the probability of extreme events. Let $X_1,\dots,X_n$ be n independent copies of X. Define the sample maximum by $M_n = \max\{X_1,\dots,X_n\}$. The main goal of extreme value analysis is to appropriately estimate for a large value $x \geq M_n$ the following probability

$$\mathbb{P}\{X > x\}. \tag{2.1}$$

The inverse of the probability (2.1) is defined as the return period T of x. In other words, T is the time period during which X is expected to exceed on average once the value x. It is clear that classical statistical methods are not applicable to solve the above problem. Indeed, for $x \geq M_n$ the empirical estimation of the probability (2.1) is equal to zero as there is no observation beyond the sample maximum. Moreover, a parametric estimation may not be reliable either since a good fit in the distribution bulk does not necessarily yield a good fit in the tail. For instance, both Gaussian and Student distributions can fit very well a given set of observations whereas the behavior of large values from the fitted Student distribution is significantly different from the behavior of large values from the fitted Gaussian distribution. Extreme value theory provides the solid fundamentals needed for the statistical modeling of extreme events and the computation of probabilities such as (2.1). The strength of extreme

value theory is that, ideally, the original parent distribution function of X needs not to be known, because the maximum term M_n, up to linear normalization, asymptotically follows a distribution nowadays called generalized extreme value (GEV) family [e.g. 10, 11, 13, 7, 6, 2]. Consequently, a sample of M_n (also called block maxima) where the nonnegative integer n (referred to as block size) approaches infinity can be approximated by the GEV distribution as stated in Theorem 2.1 from [6].

Theorem 2.1 If there exist sequences of constants $a_n > 0$ and $b_n \in \mathbb{R}$ such that

$$\lim_{n\to+\infty} \mathbb{P}^n\left\{\frac{X-b_n}{a_n} \le x\right\} = \lim_{n\to+\infty} \mathbb{P}\left\{\frac{M_n-b_n}{a_n} \le x\right\} = G(x) \tag{2.2}$$

or equivalently

$$\lim_{n\to+\infty} n\,\mathbb{P}\left\{\frac{X-b_n}{a_n} > x\right\} = -\log G(x) \tag{2.3}$$

for a non-degenerate distribution function G, then G belongs to the Generalized Extreme Value (GEV) family

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

defined on $\left\{x \in \mathbb{R} : 1+\gamma\left(\frac{x-\mu}{\sigma}\right) > 0\right\}$, where $\gamma \neq 0$, $\mu \in \mathbb{R}$, $\sigma > 0$.

The distribution G includes three parameters: the location parameter μ, the scale parameter σ and the shape parameter γ also referred to as the extreme value index. The GEV family can be divided into three families, namely the Fréchet family, the Weibull family and the Gumbel family. The Fréchet and the Weibull families correspond respectively to the cases where $\gamma > 0$ and $\gamma < 0$. The Gumbel family with $\gamma = 0$ is interpreted as the limit of (2.4) as $\gamma \to 0$, leading to the distribution

$$G(x) = \exp\left\{-\exp\left\{-\left(\frac{x-\mu}{\sigma}\right)\right\}\right\}, \quad x \in \mathbb{R}. \tag{2.5}$$

By Taylor expansion, one can observe that the Fréchet family has a power law decaying tail whereas the Gumbel family has an exponentially decaying tail [7]. Consequently, the Fréchet family suits well heavy tailed distributions (e.g. the Pareto and the Loggamma distributions) while the Gumbel family characterizes light tailed distributions (e.g. the Gaussian and the Gamma distributions). Finally, the Weibull family is the asymptotic distribution of finite right endpoint distributions such as the Uniform and the Beta distributions.

In Theorem 2.1, one can see that the sequences of constants a_n and b_n are strongly related to the limiting GEV distribution parameters μ, σ and γ. The next theorem provides explicit expressions of these relationships in the particular case where the random variable under study follows a GEV distribution.

Theorem 2.2 Norming constants for GEV distributions If X is a random variable having the GEV distribution function $G(x;\mu_0,\sigma_0,\gamma_0)$, then the limit in (2.2) is satisfied

with the sequence a_n and b_n defined by

$$b_n = \mu_0 + \sigma_0 \left(\frac{n^{\gamma_0} - 1}{\gamma_0} \right), \quad a_n = \sigma_0 \, n^{\gamma_0} \tag{2.6}$$

unless $\gamma_0 = 0$, in which case

$$b_n = \mu_0 + \sigma_0 \log(n), \quad a_n = \sigma_0. \tag{2.7}$$

Furthermore, all other sequences of constants $\widetilde{a}_n > 0$ and $\widetilde{b}_n \in \mathbb{R}$ for which the limit in (2.2) also holds with X are related to the sequences a_n and b_n by

$$\lim_{n \to +\infty} \frac{a_n}{\widetilde{a}_n} = a, \quad \lim_{n \to +\infty} \frac{\widetilde{b}_n - b_n}{a_n} = b \tag{2.8}$$

for some constants $a > 0$ and $b \in \mathbb{R}$.

Proof. of Theorem 2.2. Easy computations show that

$$\lim_{n \to +\infty} G^n(a_n x + b_n; \mu_0, \sigma_0, \gamma_0) = \exp\left\{ -(1 + \gamma_0 x)^{-1/\gamma_0} \right\}$$

unless $\gamma_0 = 0$, in which case

$$\lim_{n \to +\infty} G^n(a_n x + b_n; \mu_0, \sigma_0, \gamma_0) = \exp\left\{ -e^{-x} \right\}.$$

The proof ends noticing that limits in (2.8) are obtained from the result of Khintchine (see Thoerem 1.2.3 in [13]). □

Each of the extreme value models derived so far has been obtained through mathematical arguments that assume an underlying process consisting of a sequence of independent random variables. However, for some data to which extreme value models are commonly applied, temporal independence is usually an unrealistic assumption. Extreme conditions often persist over several consecutive observations, bringing into question the appropriateness of models such as GEV distributions. A detailed investigation of this question is given in [13]. The dependence in stationary series can take many different forms, and it is impossible to develop a general characterization of the behaviors of extremes unless some constraints are imposed. These conditions aim to ensure that the gap to independence between sets of variables that are far enough apart is sufficiently close to zero to have no effect on the limit laws for extremes. A summary of the obtained results is given in Theorem 2.3 from [6].

Theorem 2.3 Let $X_1, X_2, \ldots$ be a stationary process and $X_1^\star, X_2^\star, \ldots$ be a sequence of independent variables with the same marginal distribution. Define $M_n = \max\{X_1, \ldots, X_n\}$ and $M_n^\star = \max\{X_1^\star, \ldots, X_n^\star\}$. Under suitable regularity conditions,

$$\lim_{n \to +\infty} \mathbb{P}\left\{ \frac{M_n^\star - b_n}{a_n} \leq x \right\} = G_1(x)$$

for normalizing sequences $a_n > 0$ and $b_n \in \mathbb{R}$, where G_1 is a non-degenerate distribution function, if and only if

$$\lim_{n \to +\infty} \mathbb{P}\left\{ \frac{M_n - b_n}{a_n} \leq x \right\} = G_2(x),$$

where

$$G_2(x) = G_1^{\theta}(x) \tag{2.9}$$

for some $\theta \in (0, 1]$.

Since the marginal distributions of the X_i and $X_i^{\star}$ are the same, any difference in the limiting distribution of maxima must be attributable to the dependence of the X_i series. The parameter θ defined by (2.9) is called the extremal index. This quantity summarizes the strength of dependence between extremes in a stationary sequence. Theorem 2.3 implies that, if maxima of a stationary series converge, provided that an appropriate condition is satisfied, the limit distribution is related to the limit distribution of an independent series according to equation (2.9). The effect of dependence in stationary series is simply a replacement of G_1 as the limit distribution, which would have arisen for the associated independent series with same marginal distribution, with G_1^{θ}. This is consistent with Theorem 2.1, because if G_1 is a GEV distribution, so is G_1^{θ}. According to the foregoing, if the limiting distribution of a random sequence $M_n = \max\{X_1, \cdots, X_n\}$ from a stationary sequence $X_1, X_2, \cdots$ is non degenerate, then the probability distribution of the sample maxima M_n can be approximated by the continuous GEV distribution family for large values of n. One of the practical methodologies for statistical modeling of extreme values consists to apply the block maxima approach. In this method, data are split into sequences of observations of length n, for some large value of n, generating a series of m block maxima, $M_{n,1}, M_{n,2}, \ldots, M_{n,m}$, say, to which the generalized extreme value distribution can be fitted. The choice of a block size n is equivalent to the choice of the number m of block maxima. The delicate point of this method is the appropriate choice of the time periods defining blocks. Indeed, a too high value of n results in too few block maxima and consequently high variance estimators. For too small n, estimators become biased. A similar issue is the selection of threshold in the peak over threshold (POT) method for fitting the generalized Pareto distribution to excesses [19, 17, 22, 23].

The block maxima method has been widely used in extreme value modeling of seasonal data such as wind speeds, floods and rainfalls by setting for example, with a year as block size when data are daily observed. For non seasonal data from other fields such as vehicle engineering, the selection of an optimal block size is still a problem. Some recent studies in the literature have attempted to solve this issue [20, 4, 5]. The method proposed by [4] and [5] can be summarized as follows. The last 10% part of the actual data is reserved as test data. GEV models are fitted to different samples of block maxima from the first 90% part of the actual data. The estimated GEV models are used to generate samples (also referred to as predicted data) of size equal to that of test data. The selected block size is associated with the GEV model for which the highest similarity is observed between large values from the predicted and test data. Our main comment about this method is that the use of only one test data may not be enough to guarantee that the resulting GEV model is suitable to characterize large values from future data. To continue reviewing the literature, one can sum up the method developed by [20] as follows. GEV models are fitted on different samples of block maxima from the actual data. The goodness-of-fit (g.o.f.) of the estimated GEV models is evaluated by means of an entropy based indicator which includes three g.o.f. measures, namely the Kolmogorov Smirnov, the Chi-square and the average deviation in probability density function. The selected block size is associated with the GEV model for which the smallest value of the above mentioned g.o.f. indicator is observed. Our main comment about this method is that the resulting GEV model exhibits the best fitting result with respect to

the considered criterion. However, the selected GEV distribution does not necessarily have desired property of being sufficiently accurate when extrapolating.

It is worth noticing that the estimators of the GEV model parameters depend on the order of observations. Consequently, permuting the observations will in general not lead to the same estimates. To overcome this limitation, new estimates have been constructed based either on sliding blocks [3, 16] or on (all) permutation blocks [14, 15] for reducing the estimation uncertainty in estimators from disjoint blocks under quite general conditions. However, all of these approaches still require specifying an appropriate block size in practice. The rest of this study is designed to explain the theoretical and practical aspects of the methodology we propose to achieve this block size selection goal.

2.2 Estimation of return levels

Let $p \in (0, 1)$. The quantile z_p of the GEV family is obtained by solving the equation

$$G(z_p) = 1 - p, \tag{2.10}$$

where G is the GEV distribution function. For $\gamma \neq 0$, the solution of equation (2.10) is

$$z_p = \mu - \frac{\sigma}{\gamma}\{1 - [-\log(1-p)]^{-\gamma}\} = \mu + \sigma\left(\frac{w_p^{\gamma} - 1}{\gamma}\right), \tag{2.11}$$

where $w_p = [-\log(1-p)]^{-1} \geq 0$ and for $\gamma = 0$, the solution of equation (2.10) is

$$z_p = \mu - \sigma \log\{-\log(1-p)\} = \mu + \sigma \log(w_p). \tag{2.12}$$

In common terminology, z_p is the return level associated with the return period $T = p^{-1}$. This means that the level z_p is expected to be exceeded on average once every T blocks of observations. It is easy to see that the return level z_T is strictly increasing with the return period T. Consequently, one can estimate the frequency of events associated with values larger than the highest observation of the studied random variable. Let us denote by $(\widehat{\mu}, \widehat{\sigma}, \widehat{\gamma})$ the maximum likelihood estimate of the GEV distribution parameters (μ, σ, γ) obtained when fitting a sample of m block maxima z_i, $i = 1, \ldots, m$ with a GEV distribution, where the block size is equal to n. By substituting $(\widehat{\mu}, \widehat{\sigma}, \widehat{\gamma})$ into (2.11) and (2.12), the maximum likelihood estimate of z_p is obtained for $\gamma \neq 0$ as

$$\widehat{z}_p = \widehat{\mu} + \widehat{\sigma}\left(\frac{w_p^{\widehat{\gamma}} - 1}{\widehat{\gamma}}\right) \tag{2.13}$$

and for $\gamma = 0$, the maximum likelihood estimate of z_p is obtained as

$$\widehat{z}_p = \widehat{\mu} + \widehat{\sigma} \log w_p. \tag{2.14}$$

Furthermore, by the delta method,

$$Var\left(\widehat{z}_p\right) \approx \nabla z_p^T V \nabla z_p, \tag{2.15}$$

where V is the asymptotic variance-covariance matrix [6] of the joint estimate $(\widehat{\mu}, \widehat{\sigma}, \widehat{\gamma})$ of the parameter (μ, σ, γ) and

$$\begin{aligned} \nabla z_p^T &= \left[\frac{\partial z_p}{\partial \mu}, \frac{\partial z_p}{\partial \sigma}, \frac{\partial z_p}{\partial \gamma}\right] \\ &= \left[1, \gamma^{-1}\left(w_p^\gamma - 1\right), \sigma\gamma^{-2}\left(1 - w_p^\gamma + w_p^\gamma \log w_p^\gamma\right)\right] \end{aligned}$$

is evaluated at $(\widehat{\mu}, \widehat{\sigma}, \widehat{\gamma})$. In the particular case where $\gamma = 0$, V stands for the asymptotic variance-covariance matrix [6] of the joint estimate $(\widehat{\mu}, \widehat{\sigma})$ of the parameter (μ, σ) and

$$\nabla z_p^T = \left[\frac{\partial z_p}{\partial \mu}, \frac{\partial z_p}{\partial \sigma}\right] = \left[1, \log w_p\right].$$

2.3 Estimation of extremal index

In many practical cases that include common environmental or financial applications, the underlying observations are not independently generated from the same probability distribution, but possibly from a strict stationary process. In the latter case, the block maxima method still "works" because the block maxima are still approximated by a GEV distribution (see Theorem 2.3). However, the location and scale parameters attached to block maxima from a strict stationary process are different from those attached to block maxima from independent random variables following the marginal distribution of the underlying stationary process. Nevertheless, the location and scale parameters attached to both samples of block maxima are non-linearly related by the extremal index, a parameter $\theta \in (0, 1]$ capturing the tendency of extreme observations from the stationary process to occur in clusters. Consequently, the additional step to perform when estimating return levels that depend on the marginal distribution of a stationary process is the estimation of extremal index. Regarding the estimation of extremal index, a large variety of estimators has been proposed (see, e.g., [8, 9, 12, 16, 18, 21] and references therein). This section provides a brief discussion on the most common approaches to estimate θ, namely the runs, blocks and intervals methods.

Let $X_1, \cdots, X_n$ be a sample of n consecutive observations from a stationary process with marginal distribution function F. We consider as extreme observations those exceeding a chosen high threshold u. Let $N = N_n(u) = \sum_{i=1}^n \mathbb{1}(X_i > u)$ be the number of observations exceeding u and let $1 \leq S_1 < \cdots < S_N \leq n$ be the exceedance times. The inter-exceedance times are $T_i = S_{i+1} - S_i$ for $i = 1, \cdots, N-1$. The blocks and runs estimators are based on their own clusters identification procedure and both correspond to the ratio between the number of clusters and the number of exceedances above the threshold u. More precisely, if $C_n(u)$ is the number of found clusters of extreme observations then the blocks and runs estimators are expressed as

$$\widehat{\theta} = \frac{C_n(u)}{N_n(u)}. \tag{2.16}$$

The relationship (2.16) shows that the extremal index can also be interpreted as the reciprocal of the mean number of exceedances in a cluster. The blocks method to identify clusters consists in choosing a block length, say b, then partitioning the sequence $X_1, \cdots, X_n$ into disjoint blocks of b consecutive observations, and assigning to the same cluster any extreme observations (exceedances) lying within the same block. The runs method to identify clusters consists in choosing a run length, say r, and assigning to the same cluster any extreme observations

separated by fewer than r non-extreme observations. In addition to the threshold choice, the blocks and runs methods require the block and run lengths as respective parameters which can have significant impact on the estimate of extremal index. The intervals estimators [9] is introduced to overcome this limitation. It is completely prescribed by the data and it provides a consistent estimate of the extremal index. The intervals estimators is defined by the following formula, namely

$$\widehat{\theta} = \min\left\{1, \frac{2\left(\sum_{i=1}^{N-1} T_i\right)^2}{(N-1)\sum_{i=1}^{N-1} T_i^2}\right\} \tag{2.17}$$

unless $\max\{T_i : 1 \le i \le N-1\} > 2$, in which case

$$\widehat{\theta} = \min\left\{1, \frac{2\left(\sum_{i=1}^{N-1} (T_i - 1)\right)^2}{(N-1)\sum_{i=1}^{N-1} (T_i-1)(T_i-2)}\right\}. \tag{2.18}$$

Here, T_i denotes the inter-exceedance times defined above.

We end this section by showing how to estimate the return levels associated with the unknown marginal distribution function F of a stationary process having an extremal index $\theta > 0$. Let $z_i = \left(z_{i,1}, \cdots, z_{i,m(i)}\right)$ be the sample maxima from disjoint blocks of i consecutive observations, where $z_{i,j}$ is the maximum of the observations $X_1, \cdots, X_n$ within the j-th block. Assume that the block size i is such that the sample maxima z_i is well approximated by a GEV model and denote by $(\widehat{\gamma_i}, \widehat{\sigma_i}, \widehat{\mu_i})$ the estimated vector of parameters. Now, consider the threshold u defined by $u = F_n^{-1}(1 - 1/i)$, where $F_n(x) = (1/n)\sum_{j=1}^{n} \mathbb{1}(X_j \le x)$ for all $x \in \mathbb{R}$. Then, use the formula (2.17)-(2.18) to find an estimate $\widehat{\theta}$ of the extremal index θ. It follows from Theorem 2.3 that the approximation

$$F^i(x) \approx GEV^{1/\widehat{\theta}}(x; \widehat{\gamma_i}, \widehat{\sigma_i}, \widehat{\mu_i}) \tag{2.19}$$

holds for all $x \ge u$. That lead us to consider $F(x)$ as approximately GEV-distributed when $x \ge u$ since in this case, (2.19) can also be written as

$$F(x) \approx GEV(x; \widehat{\gamma}, \widehat{\sigma}, \widehat{\mu}),$$

where

$$\widehat{\gamma} = \widehat{\gamma_i}, \qquad \widehat{\sigma} = \widehat{\sigma_i}\left(i \times \widehat{\theta}\right)^{-\widehat{\gamma_i}}, \qquad \widehat{\mu} = \widehat{\mu_i} + \widehat{\sigma_i}\left(\frac{\left(i \times \widehat{\theta}\right)^{-\widehat{\gamma_i}} - 1}{\widehat{\gamma_i}}\right). \tag{2.20}$$

It is important to note that unless $\widehat{\theta} = 1$ or $\widehat{\gamma} = 0$, the parameters $\widehat{\sigma}$ and $\widehat{\mu}$ are different from $\widehat{\sigma_i}$ and $\widehat{\mu_i}$. It results that for the marginal distribution F, the return level z_p associated with the probability $p \in [1 - 1/i, 1)$ can be estimated by the formula

$$\widehat{z_p} = \widehat{u} + \widehat{\sigma}\left(\frac{w_p^{\widehat{\gamma}} - 1}{\widehat{\gamma}}\right), \tag{2.21}$$

where $w_p = -1/\log(1-p)$.

2.4 Selection of block size

In this section, we try to provide an answer to the following natural question which arises in practice:

Problem 2.1 *Given a continuous stationary sequence X_1, X_2, $\cdots$, how can we choose the value of n which guarantees that the GEV model fitted to a sample from the random variable $M_n = \max\{X_1, \cdots X_n\}$ is appropriate for extrapolation?*

In the sequel, we exploit Theorem 2.4 to provide an objective answer to the above question which is valid for both continuous and discrete random variables.

Theorem 2.4 Let $X_1, X_2, \ldots$, be a continuous stationary sequence. Let $M_n = \max\{X_1, \ldots, X_n\}$. Under suitable regularity conditions, suppose that for large n, there are constants $a_n > 0$ and $b_n \in \mathbb{R}$ such that for all $x \in \mathbb{R}$

$$\lim_{n \to +\infty} \mathbb{P}\{M_n \leq a_n x + b_n\} = G(x; \mu, \sigma, \gamma),$$

for some constants $\mu \in \mathbb{R}$, $\sigma > 0$ and $\gamma \in \mathbb{R}$, where G is the GEV distribution function. Then for all non-negative integer $j > 1$, we have

$$\lim_{n \to +\infty} \mathbb{P}\{M_{j \times n} \leq a_n x + b_n\} = G(x; \mu_j, \sigma_j, \gamma_j), \tag{2.22}$$

where for $\gamma \neq 0$,

$$\mu_j = \mu + \sigma \left(\frac{j^\gamma - 1}{\gamma} \right), \quad \sigma_j = \sigma\, j^\gamma, \quad \gamma_j = \gamma \tag{2.23}$$

and for $\gamma = 0$,

$$\mu_j = \mu + \sigma \log(j), \quad \sigma_j = \sigma, \quad \gamma_j = 0.$$

Proof. of Theorem 2.4. Let $X_1^\star, X_2^\star, \ldots$ be a continuous sequence of independent and identically distributed random variables whose common distribution is the marginal distribution of the stationary sequence $X_1, X_2, \ldots$. Define $M_n^\star = \max\{X_1^\star, \ldots, X_n^\star\}$. The idea is to consider $M_{j \times n}^\star$, the maximum random variable in a sequence of $j \times n$ variables for some large value of n, as the maximum of j maxima, each of which is the maximum of n observations. From Theorem 2.3, there exists $\theta \in (0, 1]$ such that the following equality holds true for all $j > 1$.

$$\lim_{n \to +\infty} \mathbb{P}\{M_{j \times n} \leq a_n x + b_n\} = \left[\left(\lim_{n \to +\infty} \mathbb{P}\{M_n^\star \leq a_n x + b_n\} \right)^\theta \right]^j.$$

Hence, one can write

$$\lim_{n \to +\infty} \mathbb{P}\{M_{j \times n} \leq a_n x + b_n\} = \left(\lim_{n \to +\infty} \mathbb{P}\{M_n \leq a_n x + b_n\} \right)^j = \left(G(x; \mu, \sigma, \gamma) \right)^j.$$

The conclusion results from the following straightforward algebraic computation

$$\left(G(x; \mu, \sigma, \gamma) \right)^j = G(x; \mu_j, \sigma_j, \gamma).$$

□

A natural technique to identify potential candidates for the optimal block size consists in fitting the GEV distribution at a range of block sizes, and to look for stability of parameter estimates. The argument is as follows. By Theorem 2.4, if a GEV distribution is a reasonable model for block maxima of a block size n_0, then block maxima of block size $n_j = j \times n_0$ for any integer $j > 1$, should also follow a GEV distribution with the same shape parameters. However, the location parameter μ_j and the scale parameter σ_j are expected to change with j as in formula (2.22) and (2.23). By reparametrizing the GEV distribution parameters when $\gamma \neq 0$ as

$$\mu^\star = \mu_j + \sigma_j \left(\frac{(1/j)^\gamma - 1}{\gamma} \right), \quad \sigma^\star = \sigma_j (1/j)^\gamma \tag{2.24}$$

and when $\gamma = 0$ as

$$\mu^\star = \mu_j + \sigma_j \log(1/j), \quad \sigma^\star = \sigma_j \tag{2.25}$$

the estimates $\widehat{\gamma}$, $\widehat{\sigma^\star}$ and $\widehat{\mu^\star}$, of γ, $\sigma^\star$ and $\mu^\star$ should be constant (up to estimation uncertainty) if n_0 is a valid block size for sample maxima to follow the GEV distribution. This argument suggests plotting $\widehat{\gamma}$, $\widehat{\sigma^\star}$ and $\widehat{\mu^\star}$, together with their respective confidence intervals, and selecting for each normalized parameter an integer n_0 as the lowest value for which these estimates remain approximately constant for almost all $n_j = j \times n_0$ with $j \geq 1$. Uncertainty in the estimation of the normalized GEV distribution parameters $\mu^\star$ and $\sigma^\star$ can be assessed by using the delta method as follows. For $\gamma = 0$, the asymptotic variance of the rescaled location parameter is

$$Var\left(\widehat{\mu^\star}\right) = \left(\nabla \widehat{\mu^\star}\right)^T V(\widehat{\mu_j}, \widehat{\sigma_j}) \nabla \widehat{\mu^\star}, \tag{2.26}$$

where $V(\widehat{\mu_j}, \widehat{\sigma_j})$ is the asymptotic variance-covariance matrix of the joint estimate $(\widehat{\mu_j}, \widehat{\sigma_j})$ of the parameter (μ_j, σ_j). Here, the gradient is calculated by the following formula

$$\left(\nabla \widehat{\mu^\star}\right)^T = \left[\frac{\partial \widehat{\mu^\star}}{\partial \widehat{\mu_j}}, \frac{\partial \widehat{\mu^\star}}{\partial \widehat{\sigma_j}} \right] = [1, -\log(j)].$$

Similarly, for $\gamma \neq 0$, the asymptotic variances of the rescaled location parameter and the rescaled scale parameter are

$$\begin{cases} Var\left(\widehat{\mu^\star}\right) = \left(\nabla \widehat{\mu^\star}\right)^T V(\widehat{\mu_j}, \widehat{\sigma_j}, \widehat{\gamma_j}) \nabla \widehat{\mu^\star} \\ Var\left(\widehat{\sigma^\star}\right) = \left(\nabla \widehat{\sigma^\star}\right)^T V(\widehat{\mu_j}, \widehat{\sigma_j}, \widehat{\gamma_j}) \nabla \widehat{\sigma^\star} \end{cases} \tag{2.27}$$

where $V(\widehat{\mu_j}, \widehat{\sigma_j}, \widehat{\gamma_j})$ is the asymptotic variance-covariance matrix of the joint estimate $(\widehat{\mu_j}, \widehat{\sigma_j}, \widehat{\gamma_j})$ of the parameter $(\mu_j, \sigma_j, \gamma_j)$. Here, the gradients are calculated by the following formula in which $\widehat{\gamma_j}$ is denoted by $\widehat{\gamma}$ for the sake of clarity

$$\left(\nabla \widehat{\mu^\star}\right)^T = \left[\frac{\partial \widehat{\mu^\star}}{\partial \widehat{\mu_j}}, \frac{\partial \widehat{\mu^\star}}{\partial \widehat{\sigma_j}}, \frac{\partial \widehat{\mu^\star}}{\partial \widehat{\gamma}} \right] = \left[1, -j^{-\widehat{\gamma}} \left(\frac{j^{\widehat{\gamma}} - 1}{\widehat{\gamma}} \right), \widehat{\sigma_j} \left(\frac{1 - j^{-\widehat{\gamma}} (\widehat{\gamma} \log(j) + 1)}{\widehat{\gamma}^2} \right) \right],$$

and

$$\left(\nabla \widehat{\sigma^\star}\right)^T = \left[\frac{\partial \widehat{\sigma^\star}}{\partial \widehat{\mu_j}}, \frac{\partial \widehat{\sigma^\star}}{\partial \widehat{\sigma_j}}, \frac{\partial \widehat{\sigma^\star}}{\partial \widehat{\gamma}} \right] = \left[0, j^{-\widehat{\gamma}}, -\widehat{\sigma_j} j^{-\widehat{\gamma}} \log(j) \right].$$

It follows from the foregoing that a large set of potential candidates for the optimal block size can be obtained.

Given a sample $\mathcal{X} = (x_1, \ldots, x_n)$ of size n, denote the sample of maxima from disjoint blocks of i consecutive observations by $z_i = \left(z_{i,1}, \ldots, z_{i,m(i)}\right)$, where $z_{i,j}$ is the maximum of the $\mathcal{X}$'s observations within the j-th block of size i. It is easy to show that the normalized parameters $\mu^\star$, $\sigma^\star$ and $\gamma^\star$ defined in (2.24) satisfy for all block size i the relation

$$G\left(x; \mu^\star, \sigma^\star, \gamma^\star\right) = G^{1/i}\left(x; \mu_i, \sigma_i, \gamma_i\right) \approx \mathbb{P}\{X \leq x\}. \tag{2.28}$$

Moreover, one can see in the formula (2.11) that return levels increase with each of the three parameters μ, σ and γ. These results suggest the following strategy to select six candidate block sizes $i^\star$ including the related estimates of the GEV model parameters which can be used for extrapolation.

Procedure 2.1
Let S be the set of block sizes for which the fitted GEV model, namely $GEV\left(x; \widehat{\gamma}_{X,z_i}, \widehat{\sigma}_{X,z_i}, \widehat{\mu}_{X,z_i}\right)$ is in adequation with sample of univariate block maxima z_i associated with the block size i. We take $\widehat{\gamma}^\star_X \equiv \widehat{\gamma}^\star_{X,z_{i^\star}}$, $\widehat{\sigma}^\star_X \equiv \widehat{\sigma}^\star_{X,z_{i^\star}}$ and $\widehat{\mu}^\star_X \equiv \widehat{\mu}^\star_{X,z_{i^\star}}$ where

$$i^\star \in \left\{i^\star\left(\gamma^\star\right), i^\star\left(\sigma^\star\right), i^\star\left(\mu^\star\right), \widetilde{i}^\star\left(\gamma^\star\right), \widetilde{i}^\star\left(\sigma^\star\right), \widetilde{i}^\star\left(\mu^\star\right)\right\} \tag{2.29}$$

with

$$i^\star\left(\gamma^\star\right) = \underset{i \in S}{\arg\max}\left\{\widehat{\gamma}_{X,z_i}\right\} \tag{2.30}$$

$$i^\star\left(\sigma^\star\right) = \underset{i \in S}{\arg\max}\left\{\widehat{\sigma}_{X,z_i}\right\} \tag{2.31}$$

$$i^\star\left(\mu^\star\right) = \underset{i \in S}{\arg\max}\left\{\widehat{\mu}_{X,z_i}\right\} \tag{2.32}$$

and

$$\widetilde{i}^\star\left(\gamma^\star\right) = \underset{i \in S}{\arg\min}\left\{\left|\widehat{\gamma}^\star_{X,z_i} - \sum_{i \in S} w_\gamma\left(\widehat{\gamma}^\star_{X,z_i}\right) \times \widehat{\gamma}^\star_{X,z_i}\right|\right\} \tag{2.33}$$

$$\widetilde{i}^\star\left(\sigma^\star\right) = \underset{i \in S}{\arg\min}\left\{\left|\widehat{\sigma}^\star_{X,z_i} - \sum_{i \in S} w_\sigma\left(\widehat{\sigma}^\star_{X,z_i}\right) \times \widehat{\sigma}^\star_{X,z_i}\right|\right\} \tag{2.34}$$

$$\widetilde{i}^\star\left(\mu^\star\right) = \underset{i \in S}{\arg\min}\left\{\left|\widehat{\mu}^\star_{X,z_i} - \sum_{i \in S} w_\mu\left(\widehat{\mu}^\star_{X,z_i}\right) \times \widehat{\mu}^\star_{X,z_i}\right|\right\} \tag{2.35}$$

in which for all $i \in S$,

- the normalized parameters are quantities given by

$$\widehat{\gamma}^\star_{X,z_i} = \widehat{\gamma}_{X,z_i} \tag{2.36}$$

$$\widehat{\sigma}^\star_{X,z_i} = \widehat{\sigma}_{X,z_i}(1/i)^{\widehat{\gamma}_{X,z_i}} \tag{2.37}$$

$$\widehat{\mu}^\star_{X,z_i} = \widehat{\mu}_{X,z_i} + \widehat{\sigma}_{X,z_i}\left(\frac{(1/i)^{\widehat{\gamma}_{X,z_i}} - 1}{\widehat{\gamma}_{X,z_i}}\right) \tag{2.38}$$

- the weights are quantities defined by

$$w_\gamma\left(\widehat{\gamma}^\star_{X,z_i}\right) = \frac{\exp\left\{\widehat{\gamma}^\star_{X,z_i}\right\}}{\sum_{j\in S}\exp\left\{\widehat{\gamma}^\star_{X,z_j}\right\}} \tag{2.39}$$

$$w_\sigma\left(\widehat{\sigma}^\star_{X,z_i}\right) = \frac{\exp\left\{\widehat{\sigma}^\star_{X,z_i}\right\}}{\sum_{j\in S}\exp\left\{\widehat{\sigma}^\star_{X,z_j}\right\}} \tag{2.40}$$

$$w_\mu\left(\widehat{\mu}^\star_{X,z_i}\right) = \frac{\exp\left\{\widehat{\mu}^\star_{X,z_i}\right\}}{\sum_{j\in S}\exp\left\{\widehat{\mu}^\star_{X,z_j}\right\}} \tag{2.41}$$



- *Note that in the proposed procedure, we give more weight to GEV models with large values of parameters in order to reduce the risk of selecting the worst model for extrapolations.*
- *Note that the survival levels, also called the return levels, can be estimated by means of the formula (2.21) along with (2.20).*
- *Note that the above procedure is described under the main assumption that observations are independently generated from the same probability distribution. The case where observations are generated from a strict stationary process will be studied in Chapter 3.*

In practice, the proposed procedure can be very expensive in terms of execution time for long sequence of observations. To overcome this limitation, we introduce Theorem 2.5 which justifies that our procedure is expected to yield the same result when a sub-sequence of the largest values is used.

Theorem 2.5 Let X be a continuous random variable having the distribution function F_X. For any real number u smaller than the right endpoint $x_X \leq +\infty$ of X, denote by F_{X_u} the distribution function of the random variable X given that $X > u$, that is

$$1 - F_{X_u}(x) = \mathbb{P}\{X > x \,|\, X > u\} = \frac{\mathbb{P}\{X > x, X > u\}}{\mathbb{P}\{X > u\}}, \quad x \in \mathbb{R}. \tag{2.42}$$

If there are sequences of constants $a_n > 0$ and $b_n \in \mathbb{R}$ as well as some parameters $\gamma \in \mathbb{R}$, $\mu \in \mathbb{R}$ and $\sigma > 0$, such that for all $x \in \mathbb{R}$, we have

$$\lim_{n\to+\infty} n\left[1 - F_X(a_n x + b_n)\right] = -\log G(x;\mu,\sigma,\gamma) \tag{2.43}$$

in which G belongs to the GEV distribution function family, then for all $x \in \mathbb{R}$, we have

$$\lim_{n\to+\infty} n\left[1 - F_{X_u}\left(\widetilde{a}_n x + \widetilde{b}_n\right)\right] = -\log G(x;\mu,\sigma,\gamma), \tag{2.44}$$

where

$$\widetilde{a}_n = a_{\frac{n}{\mathbb{P}\{X>u\}}} > 0, \quad \widetilde{b}_n = b_{\frac{n}{\mathbb{P}\{X>u\}}} \in \mathbb{R}. \tag{2.45}$$

Proof. of Theorem 2.5.

$$\begin{aligned}
\lim_{n\to+\infty} n\left[1-F_{X_u}\left(\widetilde{a}_n x+\widetilde{b}_n\right)\right] &= \lim_{n\to+\infty} n\left(\frac{\mathbb{P}\left\{X>\widetilde{a}_n x+\widetilde{b}_n,\, X>u\right\}}{\mathbb{P}\{X>u\}}\right)\\
&= \lim_{n\to+\infty}\left(\frac{n}{\mathbb{P}\{X>u\}}\right)\mathbb{P}\left\{X>\left(a_{\frac{n}{\mathbb{P}\{X>u\}}}\right)x+\left(b_{\frac{n}{\mathbb{P}\{X>u\}}}\right),\, X>u\right\}\\
&= \lim_{m\to+\infty} m\,\mathbb{P}\{X>a_m x+b_m,\, X>u\},\quad (m=n/\mathbb{P}\{X>u\})\\
&= \lim_{m\to+\infty} m\,\mathbb{P}\{X>a_m x+b_m\},\quad (a_m x+b_m>u,\ m\to+\infty)\\
&= -\log G(x;\mu,\sigma,\gamma) \qquad (2.46)
\end{aligned}$$

The proof ends making use of the following properties.

1. If $\gamma>0$, then Theorem 3.3.7 in [7] allows to see that $b_n=0$ and

$$\begin{aligned}
\widetilde{a}_n &= a_{\frac{n}{\mathbb{P}\{X>u\}}}\\
&= \inf\left\{x\in\mathbb{R}:\ F_X(x)\geq 1-\frac{\mathbb{P}\{X>u\}}{n}\right\}\\
&\geq \inf\left\{x\in\mathbb{R}:\ F_X(x)\geq 1-\frac{1}{n}\right\}\\
&= a_n. \qquad (2.47)
\end{aligned}$$

2. If $\gamma<0$, then Theorem 3.3.12 in [7] allows us to see that $b_n=x_X$ and

$$\begin{aligned}
\widetilde{a}_n &= a_{\frac{n}{\mathbb{P}\{X>u\}}}\\
&= x_X-\inf\left\{x\in\mathbb{R}:\ F_X(x)\geq 1-\frac{\mathbb{P}\{X>u\}}{n}\right\}\\
&\leq x_X-\inf\left\{x\in\mathbb{R}:\ F_X(x)\geq 1-\frac{1}{n}\right\}\\
&= a_n. \qquad (2.48)
\end{aligned}$$

3. If $\gamma=0$, then Theorem 3.3.26 in [7] allows us to see that

$$\begin{aligned}
\widetilde{b}_n &= b_{\frac{n}{\mathbb{P}\{X>u\}}}\\
&= \inf\left\{x\in\mathbb{R}:\ F_X(x)\geq 1-\frac{\mathbb{P}\{X>u\}}{n}\right\}\\
&\geq \inf\left\{x\in\mathbb{R}:\ F_X(x)\geq 1-\frac{1}{n}\right\}\\
&= b_n \qquad (2.49)
\end{aligned}$$

and

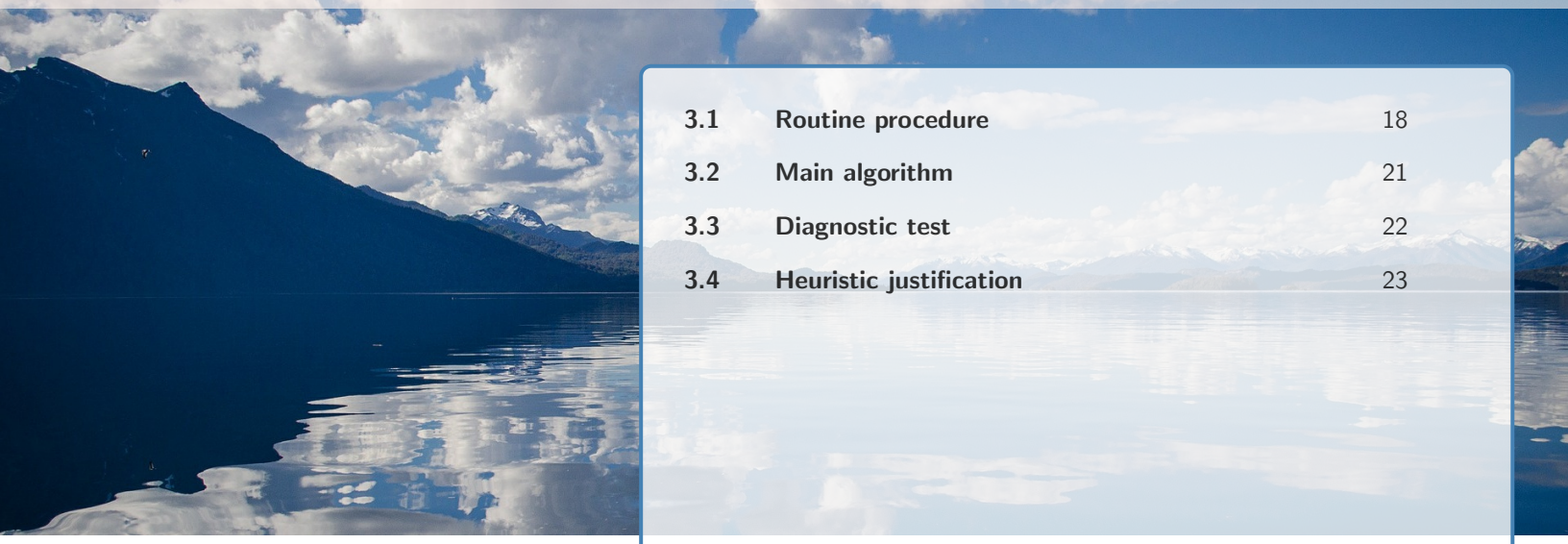
$$\widetilde{a}_n=a_{\frac{n}{\mathbb{P}\{X>u\}}}=a\left(b_{\frac{n}{\mathbb{P}\{X>u\}}}\right)=a\left(\widetilde{b}_n\right), \qquad (2.50)$$

where

$$a(x)=\int_x^{x_X}\frac{1-F_X(t)}{1-F_X(x)}\,dt,\quad \forall\, x<x_X. \qquad (2.51)$$

□

3. Determination of upper bounds for return levels



3.1	Routine procedure	18
3.2	Main algorithm	21
3.3	Diagnostic test	22
3.4	Heuristic justification	23

Context

Consider a data set $\mathcal{X} = (x_1, \ldots, x_n)$ of n consecutive observations from a stationary process having an unknown marginal cumulative distribution function F.

Denote by $x_F(T)$ the true return level from F associated with a return period T, namely

$$x_F(T) = F^{-1}\left(1 - \frac{1}{T}\right).$$

Objective

Denote by $\widehat{x}_{GEV}(T)$ the return level of the GEV distribution estimated on block maxima from the sample $\mathcal{X}$ associated with a return period T, namely

$$\widehat{x}_{GEV}(T) = GEV^{-1}\left(1 - \frac{\widehat{i}}{T}; \widehat{\gamma}, \widehat{\sigma}, \widehat{\mu}\right).$$

Here, the quantities $\widehat{i}$, $\widehat{\gamma}$, $\widehat{\sigma}$, $\widehat{\mu}$ are respectively called the block size, the shape, the scale and the location parameters.

The goal of this Chapter is to provide a guideline to find GEV model parameters $(\widehat{i}, \widehat{\gamma}, \widehat{\sigma}, \widehat{\mu})$ which guarantee that the true return level $x_F(T)$ is always smaller than its estimate $\widehat{x}_{GEV}(T)$, namely

$$\widehat{x}_{GEV}(T) \geq x_F(T)$$

for any return period $T > n$.

3.1 Routine procedure

Procedure 3.1

Stage 1: Given a sample $\mathcal{X} = (x_1, \ldots, x_n)$ of size n, extract the samples of maxima from disjoint blocks of consecutive observations, denoted by $z_i = \left(z_{i,1}, \ldots, z_{i,m(i)}\right)$ for $i = 1, 2, \ldots$, where $z_{i,j}$ is the maximum of the $\mathcal{X}$'s observations within the j-th block of size i.

Stage 2: For $i = 1, 2, \ldots$ perform the following tasks.

- Carry out the Augmented Dickey-Fuller (ADF) stationary test on the sample maxima z_i and record the p-value, denoted by $p_{i,ADF}$, of the test statistic.
- Carry out the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) stationary test on sample maxima z_i and record the p-value, denoted by $p_{i,KPSS}$, of the test statistic.
- Use the maximum likelihood estimation method to fit the GEV distribution with non zero shape parameter to each sample maxima z_i. Denote the fitted cumulative distribution function by GEV_{z_i}.
- Carry out the Anderson-Darling (AD) test to check the goodness-of-fit of the sample maxima z_i with the GEV distribution. Then record the p-value, denoted by $p_{i,AD}$, of the test statistic.
- Construct a $100 \times (1-\alpha)$%-confidence interval for the normalized shape parameter $\gamma_i^\star = \gamma_i \neq 0$ and denote it by $C(\gamma_i^\star)$.
- Construct a $100 \times (1-\alpha)$%-confidence interval for the normalized scale parameter $\sigma_i^\star > 0$ and denote it by $C(\sigma_i^\star)$.
- Construct a $100 \times (1-\alpha)$%-confidence interval for the normalized location parameter $\mu_i^\star \in \mathbb{R}$ and denote it by $C(\mu_i^\star)$.

Stage 3: Compute the subset S of block sizes defined by

$$S = \{i = 1, 2, \ldots : \; p_{i,AD} \geq \alpha \; \cap \; (p_{i,ADF} < \alpha \; \cup \; p_{i,KPSS} \geq \alpha)\},$$

where $\alpha \in (0, 1)$ is the significance level for the tests. The set S contains all block sizes i for which the sample maxima z_i is stationary and is in adequacy with the GEV distribution.

Stage 4: Perform the following tasks.

- Construct the largest subset $S(\gamma^\star)$ of S containing block sizes i for which the associated confidence intervals $C(\gamma_i^\star)$ overlap in the sense that they have some elements in common, namely

$$\bigcap_{i \in S(\gamma^\star)} C(\gamma_i^\star) \neq \emptyset.$$

  This means that the estimates of normalized GEV distribution parameters $\gamma_i^\star \neq 0$ are not significantly different for all block sizes in the set $S(\gamma^\star)$.
- Construct the largest subset $S(\sigma^\star)$ of S containing block sizes i for which the associated confidence intervals $C(\sigma_i^\star)$ overlap in the sense that they have some elements in common, namely

$$\bigcap_{i \in S(\sigma^\star)} C(\sigma_i^\star) \neq \emptyset.$$

  This means that the estimates of normalized GEV distribution parameters $\sigma_i^\star > 0$ are not significantly different for all block sizes in the set $S(\sigma^\star)$.

- Construct the largest subset $S(\mu^\star)$ of S containing block sizes i for which the associated confidence intervals $C(\mu_i^\star)$ overlap in the sense that they have some elements in common, namely

$$\bigcap_{i \in S(\mu^\star)} C(\mu_i^\star) \neq \emptyset.$$

This means that the estimates of normalized GEV distribution parameters $\mu_i^\star \in \mathbb{R}$ are not significantly different for all block sizes in the set $S(\mu^\star)$.

Stage 5: Perform one and only one of the two tasks a) and b).

a) Perform the following tasks.
   - Select the quantities $\left(i(\gamma^\star), \gamma_{i(\gamma^\star)}, \sigma_{i(\gamma^\star)}, \mu_{i(\gamma^\star)}\right)$ in which

$$i(\gamma^\star) = \underset{i \in S(\gamma^\star)}{\arg\max} \left\{\gamma_i^\star\right\}$$

   is a block size and $\left(\gamma_{i(\gamma^\star)}, \sigma_{i(\gamma^\star)}, \mu_{i(\gamma^\star)}\right)$ are the GEV model parameters fitted to the sample maxima $z_{i(\gamma^\star)}$.
   - Select the quantities $\left(i(\sigma^\star), \gamma_{i(\sigma^\star)}, \sigma_{i(\sigma^\star)}, \mu_{i(\sigma^\star)}\right)$ in which

$$i(\sigma^\star) = \underset{i \in S(\sigma^\star)}{\arg\max} \left\{\sigma_i^\star\right\}$$

   is a block size and $\left(\gamma_{i(\sigma^\star)}, \sigma_{i(\sigma^\star)}, \mu_{i(\sigma^\star)}\right)$ are the GEV model parameters fitted to the sample maxima $z_{i(\sigma^\star)}$.
   - Select the quantities $\left(i(\mu^\star), \gamma_{i(\mu^\star)}, \sigma_{i(\mu^\star)}, \mu_{i(\mu^\star)}\right)$ in which

$$i(\mu^\star) = \underset{i \in S(\mu^\star)}{\arg\max} \left\{\mu_i^\star\right\}$$

   is a block size and $\left(\gamma_{i(\mu^\star)}, \sigma_{i(\mu^\star)}, \mu_{i(\mu^\star)}\right)$ are the GEV model parameters fitted to the sample maxima $z_{i(\mu^\star)}$.

b) Perform the following tasks.
   - Select the quantities $\left(i(\gamma^\star), \gamma_{i(\gamma^\star)}, \sigma_{i(\gamma^\star)}, \mu_{i(\gamma^\star)}\right)$ in which

$$i(\gamma^\star) = \underset{i \in S(\gamma^\star)}{\arg\min} \left\{\left|\gamma_i^\star - \sum_{i \in S(\gamma^\star)} w_\gamma\left(\gamma_i^\star\right) \times \gamma_i^\star\right|\right\}, \quad w_\gamma\left(\widehat{\gamma}_i^\star\right) = \frac{\exp\left\{\widehat{\gamma}_i^\star\right\}}{\sum_{j \in S(\gamma^\star)} \exp\left\{\widehat{\gamma}_j^\star\right\}}$$

   is a block size and $\left(\gamma_{i(\gamma^\star)}, \sigma_{i(\gamma^\star)}, \mu_{i(\gamma^\star)}\right)$ are the GEV model parameters fitted to the sample maxima $z_{i(\gamma^\star)}$.
   - Select the quantities $\left(i(\sigma^\star), \gamma_{i(\sigma^\star)}, \sigma_{i(\sigma^\star)}, \mu_{i(\sigma^\star)}\right)$ in which

$$i(\sigma^\star) = \underset{i \in S(\sigma^\star)}{\arg\min} \left\{\left|\sigma_i^\star - \sum_{i \in S(\sigma^\star)} w_\sigma\left(\sigma_i^\star\right) \times \sigma_i^\star\right|\right\}, \quad w_\sigma\left(\widehat{\sigma}_i^\star\right) = \frac{\exp\left\{\widehat{\sigma}_i^\star\right\}}{\sum_{j \in S(\sigma^\star)} \exp\left\{\widehat{\sigma}_j^\star\right\}}$$

   is a block size and $\left(\gamma_{i(\sigma^\star)}, \sigma_{i(\sigma^\star)}, \mu_{i(\sigma^\star)}\right)$ are the GEV model parameters fitted to the sample maxima $z_{i(\sigma^\star)}$.

– Select the quantities $\left(i(\mu^\star), \gamma_{i(\mu^\star)}, \sigma_{i(\mu^\star)}, \mu_{i(\mu^\star)}\right)$ in which

$$i(\mu^\star) = \underset{i \in S(\mu^\star)}{\arg\min} \left\{ \left| \mu_i^\star - \sum_{i \in S(\mu^\star)} w_\mu\left(\mu_i^\star\right) \times \mu_i^\star \right| \right\}, \quad w_\mu\left(\widehat{\mu}_i^\star\right) = \frac{\exp\left\{\widehat{\mu}_i^\star\right\}}{\sum_{j \in S(\mu^\star)} \exp\left\{\widehat{\mu}_j^\star\right\}}$$

is a block size and $\left(\gamma_{i(\mu^\star)}, \sigma_{i(\mu^\star)}, \mu_{i(\mu^\star)}\right)$ are the GEV model parameters fitted to the sample maxima $z_{i(\mu^\star)}$.

Stage 6: Select the estimated GEV model parameters $(i^*, \gamma^*, \sigma^*, \mu^*)$ where

$$\begin{aligned} i^* &= \max\left\{i(\gamma^\star), i(\sigma^\star), i(\mu^\star)\right\}, \\ \gamma^* &= \max\left\{\gamma_{i(\gamma^\star)}, \gamma_{i(\sigma^\star)}, \gamma_{i(\mu^\star)}\right\}, \\ \sigma^* &= \max\left\{\sigma_{i(\gamma^\star)}, \sigma_{i(\sigma^\star)}, \sigma_{i(\mu^\star)}\right\}, \\ \mu^* &= \max\left\{\mu_{i(\gamma^\star)}, \mu_{i(\sigma^\star)}, \mu_{i(\mu^\star)}\right\}. \end{aligned}$$

3.2 Main algorithm

Procedure 3.2

Stage 1: Given a sample $\mathcal{X} = (x_1, \ldots, x_n)$ of n observations. Denote the identity permutation of the set $\{1, \cdots, n\}$ by π_1, that is $\mathcal{X}_{\pi_1} = \mathcal{X}$. Define the permutations π_k of the set $\{1, \cdots, n\}$ for $k = 2, \ldots, K$ by $\mathcal{X}_{\pi_k} = \left(x_{m_k+1}, \ldots, x_n, x_1, \ldots, x_{m_k}\right)$, where $m_k = (k-1) \times m$ with $m = \lceil n/K \rceil$ and $K \leq \sqrt{n}$.

Stage 2: For $k = 1, \ldots, K$, perform the following tasks.

- Apply the routine procedure described in Section 3.1 to the sample $\mathcal{X}_{\pi_k}$. Denote the selected GEV model parameters by $\left(i_k^*, \gamma_k^*, \sigma_k^*, \mu_k^*\right)$.
- Use the method of [9] to estimate the extremal index θ_k^* of the input (stationary) sequence $\mathcal{X}$ above the threshold u_k^* defined as the empirical quantile of $\mathcal{X}$ whose order is $1 - 1/i_k^*$.

Stage 3: Estimate the required GEV model parameters $(\widehat{i}, \widehat{\gamma}, \widehat{\sigma}, \widehat{\mu})$ as follows.

- Calculate the block size $\widehat{i}$ by means of the following formula

$$\widehat{i} = \frac{1}{K} \sum_{k=1}^{K} i_k^*. \tag{3.1}$$

- Calculate the extremal index $\widehat{\theta}$ by means of the following formula

$$\widehat{\theta} = \frac{1}{K} \sum_{k=1}^{K} \theta_k^*. \tag{3.2}$$

- Calculate the pseudo parameters $\widetilde{\gamma}$, $\widetilde{\sigma}$ and $\widetilde{\mu}$ of the GEV model by means of the formula

$$\widetilde{\gamma} = \frac{1}{K}\sum_{k=1}^{K}\gamma_k^*, \qquad \widetilde{\sigma} = \frac{1}{K}\sum_{k=1}^{K}\sigma_k^*, \qquad \widetilde{\mu} = \frac{1}{K}\sum_{k=1}^{K}\mu_k^*. \tag{3.3}$$

- Calculate the scale and the location parameters $\widehat{\sigma}$ and $\widehat{\mu}$ of the GEV model by means of the following formula

$$\widehat{\sigma} = \widetilde{\sigma}\times\left(\widehat{\theta}\right)^{-\widetilde{\gamma}}, \qquad \widehat{\mu} = \widetilde{\mu} + \widetilde{\sigma}\times\frac{\left(\widehat{\theta}\right)^{-\widetilde{\gamma}} - 1}{\widetilde{\gamma}} \tag{3.4}$$

- Calculate the shape parameter $\widehat{\gamma}$ of the GEV model using either formula (3.5) or (3.6) depending on whether **Stage 5-a)** or **Stage 5-b)** is respectively considered in the **Routine procedure**.

$$\widehat{\gamma} = 0.8379\times\widetilde{\gamma} \tag{3.5}$$
$$\widehat{\gamma} = 0.8628\times\widetilde{\gamma} \tag{3.6}$$



- *The Procedure 3.2 integrates a scheme to consistently estimate the two important quantities, namely the block size i_0 and the extremal index θ as they are generally unknown in practice.*
- *In practice, the Procedure 3.2 can be very expensive in terms of execution time for long sequence of observations. Making use of Theorem 2.5, one can reduce this latency by dealing with a sub-sequence of the largest values.*

3.3 Diagnostic test

Diagnostic test to the obtained GEV distribution with parameters $\widehat{i}$, $\widehat{\mu}$, $\widehat{\sigma}$ and $\widehat{\gamma}$ for return levels upper bounds estimation can be carried out by checking whether quantiles of generalized Pareto (GP) distribution having scale parameter $\sigma = \widehat{\sigma} + \widehat{\gamma}(\widehat{u} - \widehat{\mu})$ and shape parameter $\gamma = \widehat{\gamma}$ are significantly greater than those of the empirical excesses over the threshold $\widehat{u}$. Here, $\widehat{u}$ is the quantile of observations whose order is $1 - 1/\widehat{i}$. Recall that the generalized Pareto distribution function is defined on the set $\{x \in \mathbb{R} : 1 + \gamma x/\sigma > 0\}$, by

$$H(x) = 1 - \left(1 + \gamma\frac{x}{\sigma}\right)$$

where $\sigma > 0$ is the scale parameter and $\gamma \in \mathbb{R}$ is the shape parameter. In practice, the GEV model obtained from the Procedure 3.2 could be either too pessimistic or more realistic depending on whether **Stage 5-a)** or **Stage 5-b)** is respectively considered in the **Routine procedure**. Thus, we suggest users to consider the GEV distribution resulting from the flow including **Stage 5-b)** unless the above diagnostic test fails.

3.4 Heuristic justification

Let $x = x_1, \cdots, x_n$ be a sequence of n observations from a stationary process having an extremal index $\theta \in (0, 1]$. Assume that there is a block size i_0 such that the function $GEV\left(x; \mu_{i_0}, \sigma_{i_0}, \gamma_{i_0}\right)$ approximates well the limit distribution for maxima of the underlying stationary process. Here, the parameters μ_{i_0}, σ_{i_0} and γ_{i_0} are obtained by applying the block maxima modeling approach to the sequence x. According to Theorem 2.3, the function $GEV^{1/\theta}\left(x; \mu_{i_0}, \sigma_{i_0}, \gamma_{i_0}\right)$ will also approximate well the limit distribution for maxima of the marginal distribution associated with the above stationary process. Making use of (2.28), it results that this marginal distribution is well approximated for all $x \geq F_n^{-1}(1 - 1/i_0)$ by the function $GEV\left(x; \mu^\star, \sigma^\star, \gamma^\star\right)$, where F_n is the empirical cumulative distribution function associated with the sequence x and

$$\mu^\star = \mu_{i_0} + \sigma_{i_0} \times \left(\frac{(i_0 \times \theta)^{-\gamma_{i_0}} - 1}{\gamma_{i_0}}\right), \quad \sigma^\star = \sigma_{i_0} \times (i_0 \times \theta)^{-\gamma_{i_0}}, \quad \gamma^\star = \gamma_{i_0}. \tag{3.7}$$

On the other hand, the application of Procedure 3.2 to the sequence x gives an estimate of the above marginal distribution defined by the function $GEV\left(x; \widehat{\mu}^\star, \widehat{\sigma}^\star, \widehat{\gamma}^\star\right)$, where

$$\widehat{\mu}^\star = \widetilde{\mu}^\star + \widetilde{\sigma}^\star \times \left(\frac{\theta^{-\widetilde{\gamma}^\star} - 1}{\widetilde{\gamma}^\star}\right), \quad \widehat{\sigma}^\star = \widetilde{\sigma}^\star \times \theta^{-\widetilde{\gamma}^\star}, \quad \widehat{\gamma}^\star = \widetilde{\gamma}^\star. \tag{3.8}$$

In (3.8), the parameters $\widetilde{\mu}^\star$, $\widetilde{\sigma}^\star$ and $\widetilde{\gamma}^\star$ are though as the normalized versions of $\widetilde{\mu}$, $\widetilde{\sigma}$ and $\widetilde{\gamma}$ defined in (3.3). It is clear that the inequalities in (3.9) hold true by the construction of the Procedure 3.1 since the GEV model having the largest values of the normalized parameters is selected at each round, that is

$$\mu^\star \leq \widehat{\mu}^\star, \quad \sigma^\star \leq \widehat{\sigma}^\star, \quad \gamma^\star \leq \widehat{\gamma}^\star. \tag{3.9}$$

This yields the following expected inequality between the related return levels defined in (2.11), namely

$$z_p = \mu^\star + \sigma^\star \left(\frac{w_p^{\gamma^\star} - 1}{\gamma^\star}\right) \leq \widehat{\mu}^\star + \widehat{\sigma}^\star \left(\frac{w_p^{\widehat{\gamma}^\star} - 1}{\widehat{\gamma}^\star}\right) = \widehat{z}_p. \tag{3.10}$$

It remains to show that the shape parameter $\widehat{\gamma}$ defined in (3.4) and (3.5) is always greater than the true value of the shape parameter γ associated with the considered stationary process. This statement can be verified by analyzing the results of the following procedure.

Procedure 3.3

Stage 1: Consider a trivariate sequence $(\gamma_1, \sigma_1, \mu_1), \cdots, (\gamma_J, \sigma_J, \mu_J)$ of $J = 500$ GEV model parameters defined as follows.

- The univariate sequence $\gamma_1, \cdots, \gamma_J$ is a sample from the continuous uniform probability distribution on the interval $(-0.5, 1)$.
- The univariate sequence $\sigma_1, \cdots, \sigma_J$ is a sample from the continuous log-normal probability distribution having 0 as mean and 1/4 as standard deviation.
- The univariate sequence $\mu_1, \cdots, \mu_J$ is a sample from the continuous standard normal probability distribution.

Stage 2: For $j = 1, \cdots, J$, consider the sequence $\mathcal{X}_{j,1}, \cdots, \mathcal{X}_{j,K}$ of $K = 128$ samples, where each sample $\mathcal{X}_{j,k} = \left(x_{j,k,1}, \cdots, x_{j,k,n}\right)$ consists of $n = 10,000$ observations from the continuous GEV probability distribution having the parameter vector $(\gamma_j, \sigma_j, \mu_j)$.

Stage 3: For $j = 1, \cdots, J$ and for $k = 1, \ldots, K$, apply the **Routine procedure** to the sample $\mathcal{X}_{j,k}$. Denote the selected GEV model parameters by $\left(i^*_{j,k}, \gamma^*_{j,k}, \sigma^*_{j,k}, \mu^*_{j,k}\right)$.

Stage 4: For $j = 1, \cdots, J$, estimate the shape parameter $\widetilde{\gamma}_j$ by means of the formula

$$\widetilde{\gamma}_j = \frac{1}{K} \sum_{k=1}^{K} \gamma^*_{j,k}. \tag{3.11}$$

The plots in Figures 3.1-3.2 show that there is a linear relationship between the true GEV model shape parameters γ_j and the estimated ones $\widetilde{\gamma}_j$ resulting from the Procedure 3.3. Moreover, this relationship is very strong as indicated by the characteristics of the associated simple linear regression model displayed in Figures 3.3-3.4. It is clear from these results that the estimated shape parameters $\widehat{\gamma} = 0.8379 \times \widetilde{\gamma}$ or $\widehat{\gamma} = 0.863 \times \widetilde{\gamma}$, represented by the yellow lines in Figures 3.1-3.2, are always greater than the true γ as expected.

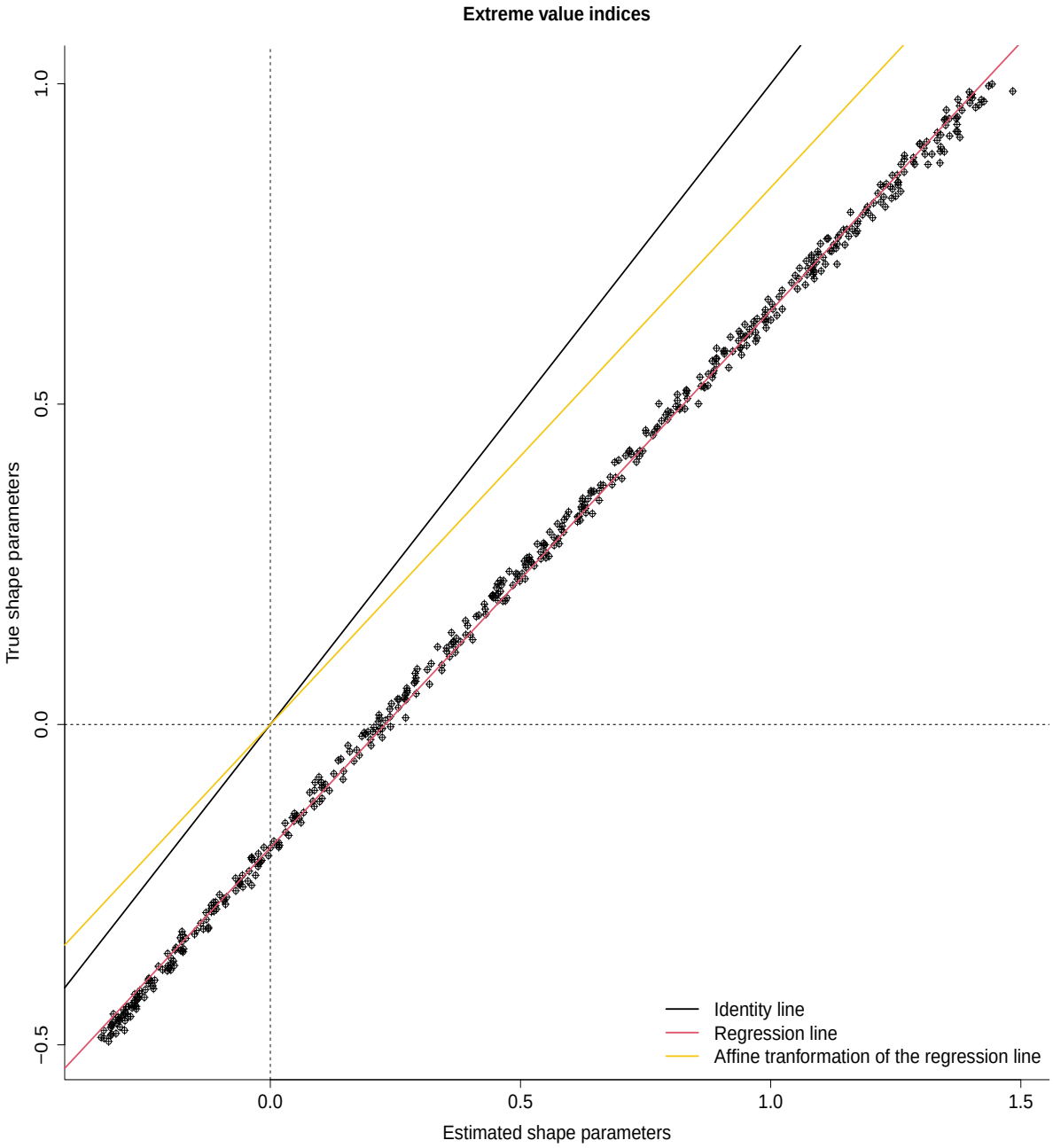


Figure 3.1: Plot of the simple linear regression model between the true GEV model shape parameter γ and the estimated one $\widetilde{\gamma}$ resulting from Procedure 3.3 in which **Stage 5-a)** is considered in the **Routine procedure**.

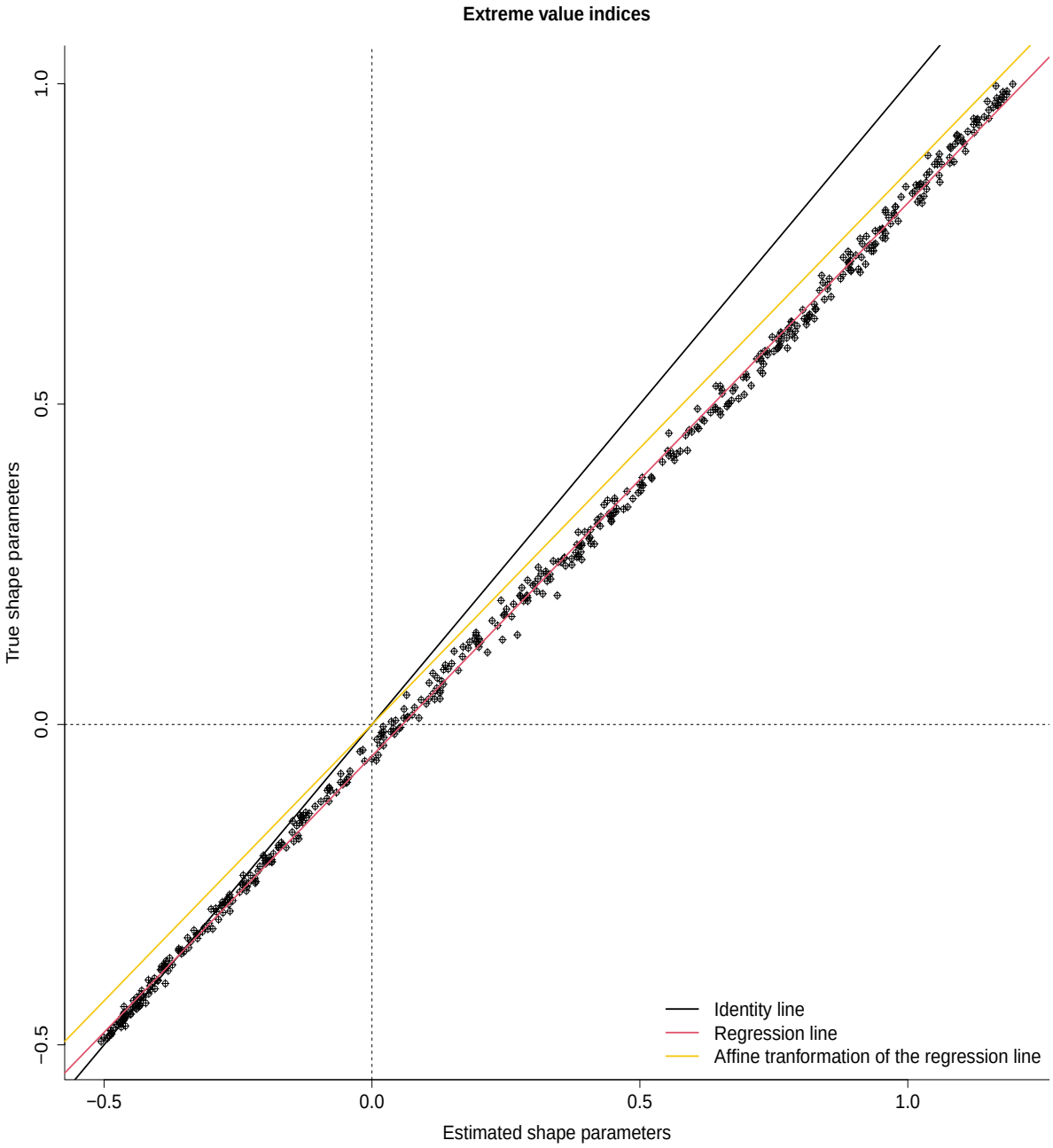


Figure 3.2: Plot of the simple linear regression model between the true GEV model shape parameter γ and the estimated one $\widetilde{\gamma}$ resulting from Procedure 3.3 in which **Stage 5-b)** is considered in the **Routine procedure**.

Variables	Beta	95% CI[1]	p-value
(Intercept)	-0.192	-0.194, -0.19	**<0.001**
Estimated shape	0.838	0.835, 0.84	**<0.001**

[1] CI = Confidence Interval

R^2 = 0.998786; Adjusted R^2 = 0.9987836; Sigma = 0.016; Statistic = 409,719; p-value = <0.001; df = 1; Log-likelihood = 1,370; AIC = -2,733; BIC = -2,720; Deviance = 0.122; Residual df = 498; No. Obs. = 500

Figure 3.3: Summary of the simple linear regression model between the true GEV model shape parameter γ and the estimated one $\widetilde{\gamma}$ from the result of Procedure 3.3 in which **Stage 5-a)** is considered in the **Routine procedure**, namely $\gamma = 0.838 \times \widetilde{\gamma} - 0.192 + \epsilon$, where ϵ is a normal random variable having the zero as mean and 0.016 as standard deviation parameters.

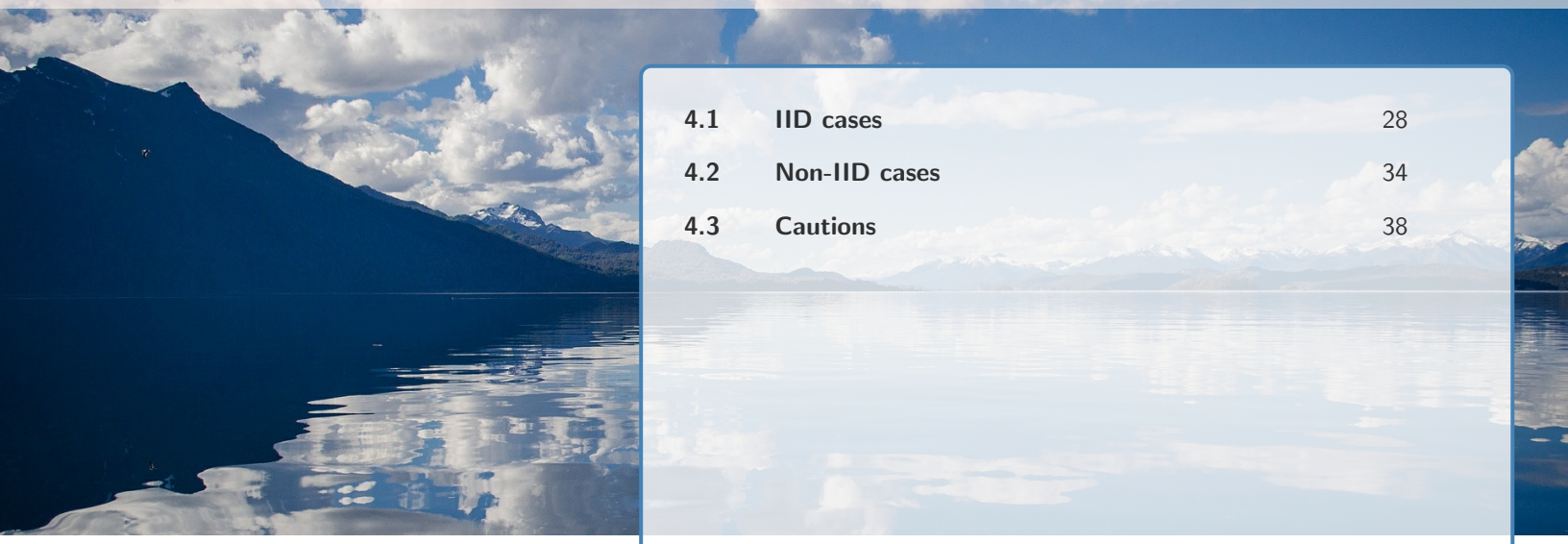
Variables	Beta	95% CI[1]	p-value
(Intercept)	-0.049	-0.051, -0.048	**<0.001**
Estimated shape	0.863	0.86, 0.865	**<0.001**

[1] CI = Confidence Interval

R^2 = 0.9989615; Adjusted R^2 = 0.9989594; Sigma = 0.015; Statistic = 479,038; p-value = <0.001; df = 1; Log-likelihood = 1,408; AIC = -2,809; BIC = -2,797; Deviance = 0.105; Residual df = 498; No. Obs. = 500

Figure 3.4: Summary of the simple linear regression model between the true GEV model shape parameter γ and the estimated one $\widetilde{\gamma}$ from the result of Procedure 3.3 in which **Stage 5-b)** is considered in the **Routine procedure**, namely $\gamma = 0.863 \times \widetilde{\gamma} - 0.049 + \epsilon$, where ϵ is a normal random variable having the zero as mean and 0.015 as standard deviation parameters.

4. Illustrations on univariate stochastic processes



4.1	IID cases	28
4.2	Non-IID cases	34
4.3	Cautions	38

Introduction

Consider the sequence $\mathcal{X} = (x_1, \ldots, x_N)$ of size $N = 10,000$ observations from a stationary process.

The goal is to estimate the upper bounds of return levels which are associated with return periods ranging from N to $10^5 N$.

We show how to properly use the strategy described in Chapter 3 with the sequence $\mathcal{X}$ in order to archive this goal.

4.1 IID cases

In this section, we consider the strategy described by the Procedure 3.2 using **Stage 5-b)** in the **Routine procedure**. We apply this strategy on simulated data of size $N = 10,000$ from the probability distributions given in the examples below. We take the number 123 as seed value for each simulation in the $\mathcal{R}$ statistical software. Numerical results are gathered in Table 4.1 where one can see that the estimated upper bounds for return levels of interest are greater than the true return levels. The graphical diagnostics provided in Figures 4.1–4.6 show that this conclusion is also true for return periods smaller than the sample size N.

Exemple 4.1 (Law1) The triangular distribution with parameters $a, b, c \in \mathbb{R}$ such that $a \leq c \leq b$ has probability density function defined as follows.

- For $a < x \leq c$
$$f(x; a, b, c) = \frac{2(x-a)}{(b-a)(c-a)}. \tag{4.1}$$
- for $c < x < b$
$$f(x; a, b, c) = \frac{2(b-x)}{(b-a)(b-c)}. \tag{4.2}$$

For illustration, we use the special case where $a = -1$, $b = +1$ and $c = 0$.

Exemple 4.2 (Law2) The normal distribution with location parameter $\mu \in \mathbb{R}$ and scale parameter $\sigma > 0$ has probability density function defined for $x \in \mathbb{R}$ by

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}. \tag{4.3}$$

For illustration, we use the special case where $\mu = 0$ and $\sigma = 1$.

Exemple 4.3 (Law3) The exponential distribution with rate parameter $\lambda > 0$ has probability density function defined for $x > 0$ by

$$f(x; \lambda) = \lambda \exp\{-\lambda x\}. \tag{4.4}$$

For illustration, we use the special case where $\lambda = 1$.

Exemple 4.4 (Law4) The generalized Pareto (GP) distribution with parameters $\mu \in \mathbb{R}$, $\sigma > 0$ and $\gamma \in \mathbb{R}$ has probability density function defined for $\gamma \neq 0$ and $x \in \mathbb{R}$ such that $1 + \gamma\left(\frac{x-\mu}{\sigma}\right) > 0$ by

$$f(x; \mu, \sigma, \gamma) = \frac{1}{\sigma}\left(1 + \gamma\,\frac{x-\mu}{\sigma}\right)^{-1-\frac{1}{\gamma}}. \tag{4.5}$$

For illustration, we use the special case where $\mu = 0$, $\sigma = 1$ and $\gamma = -0.2$.

Exemple 4.5 (Law5) For illustration, we consider the special case of the generalized Pareto (GP) distribution defined by the formula (4.5), where $\mu = 0$, $\sigma = 1$ and $\gamma = -0.5$.

Exemple 4.6 (Law6) The logistic distribution with location parameter $\mu \in \mathbb{R}$ and scale parameter $\sigma > 0$ has probability density function defined for $x \in \mathbb{R}$ by

$$f(x; \mu, \sigma) = \frac{1}{\sigma}\,\frac{\exp\left\{-\frac{x-\mu}{\sigma}\right\}}{\left(1 + \exp\left\{-\frac{x-\mu}{\sigma}\right\}\right)^2}. \tag{4.6}$$

For illustration, we use the special case where $\mu = 0$ and $\sigma = 1$.

Table 4.1: Table of estimated GEV distribution parameters obtained from the Procedure 3.2 when **Stage 5-b)** is considered in the **Routine procedure**. Upper bounds of return levels $\widehat{x}(T)$ associated with some return periods expressed in terms of numbers of observations are provided. The threshold $\widehat{u}$ is the quantile of observations whose order is $1-1/\widehat{i}$. The true values γ of shape parameters and the true values $x(T)$ of return levels for the considered probability distributions (Law1 to Law6) are also included.

	Law1	Law2	Law3	Law4	Law5	Law6
$\widetilde{\gamma}$	-0.5218	-0.1240	0.0923	-0.1722	-0.4659	0.0741
$\widetilde{\sigma}$	0.0831	0.6573	1.5631	0.5593	0.1258	1.7085
$\widetilde{\mu}$	0.8973	2.7595	5.8383	3.4388	1.8795	5.9212
$\widehat{i}$	207.0000	328.0000	349.3438	354.4688	290.3438	340.1562
$\widehat{u}$	0.8907	2.7872	5.8593	3.4177	1.8742	5.7311
$\widehat{\theta}$	0.9971	1.0000	0.9623	0.9234	0.9589	1.0000
$\widehat{\gamma}$	-0.4502	-0.1070	0.0796	-0.1486	-0.4020	0.0639
γ	-0.5000	0.0000	0.0000	-0.2000	-0.5000	0.0000
$\widehat{\sigma}$	0.0829	0.6573	1.5686	0.5517	0.1233	1.7085
$\widehat{\mu}$	0.8975	2.7595	5.8984	3.4830	1.8847	5.9212
$\widehat{x}(N)$	1.0500	4.6335	11.8930	4.9293	2.1171	12.3322
$x(N)$	0.9859	3.7190	9.2103	4.2076	1.9800	9.2102
$\widehat{x}(10N)$	1.0710	5.5713	17.1024	5.5899	2.1621	17.6244
$x(10N)$	0.9955	4.2649	11.5129	4.5000	1.9937	11.5129
$\widehat{x}\left(10^2N\right)$	1.0783	6.2995	23.3250	6.0554	2.1798	23.7215
$x\left(10^2N\right)$	0.9986	4.7534	13.8155	4.6845	1.9980	13.8155
$\widehat{x}\left(10^3N\right)$	1.0809	6.8683	30.7954	6.3858	2.1868	30.7809
$x\left(10^3N\right)$	0.9996	5.1993	16.1181	4.8009	1.9994	16.1181
$\widehat{x}\left(10^4N\right)$	1.0818	7.3129	39.7681	6.6205	2.1896	38.9586
$x\left(10^4N\right)$	0.9999	5.6120	18.4207	4.8744	1.9998	18.4207
$\widehat{x}\left(10^5N\right)$	1.0821	7.6605	50.5458	6.7871	2.1907	48.4323
$x\left(10^5N\right)$	1.0000	5.9978	20.7233	4.9208	1.9999	20.7233

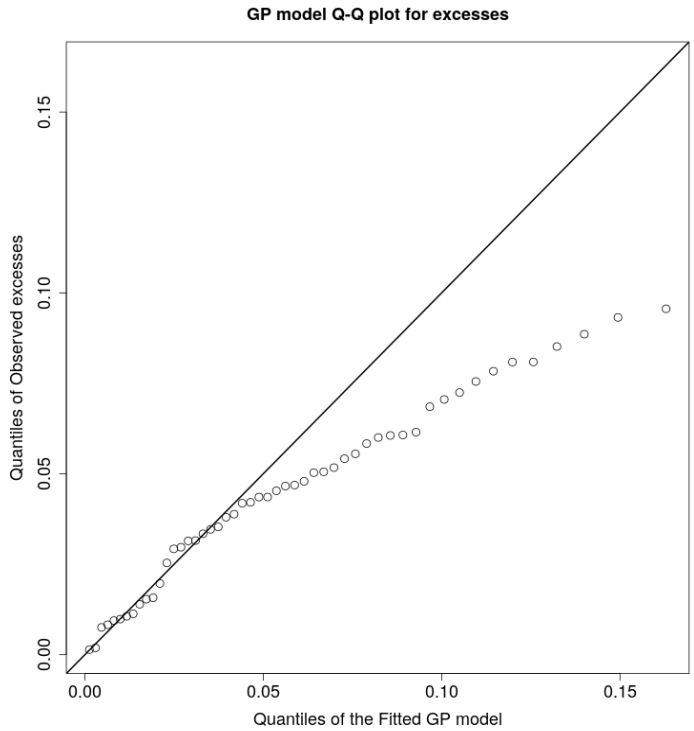


Figure 4.1: Diagnostics related to the GEV model for return levels upper bounds estimation of Law1.

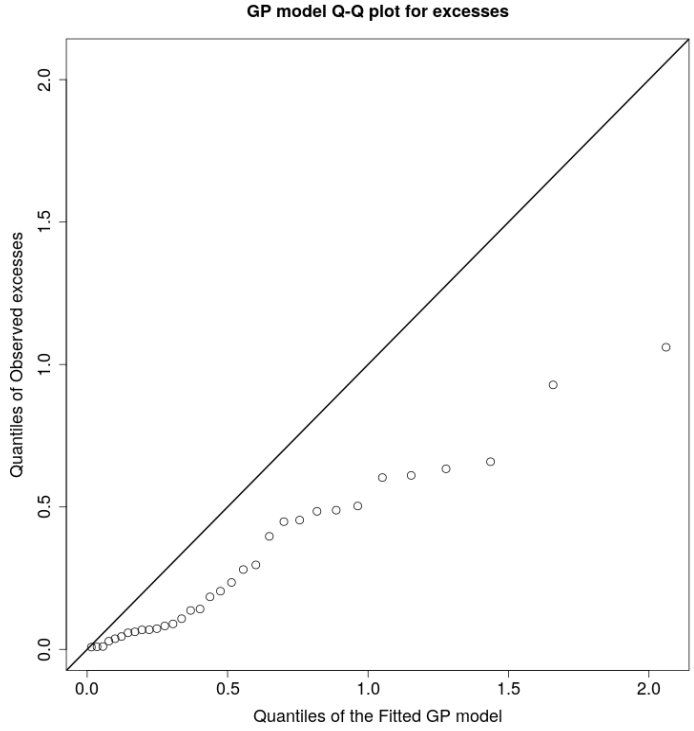


Figure 4.2: Diagnostics related to the GEV model for return levels upper bounds estimation of Law2.

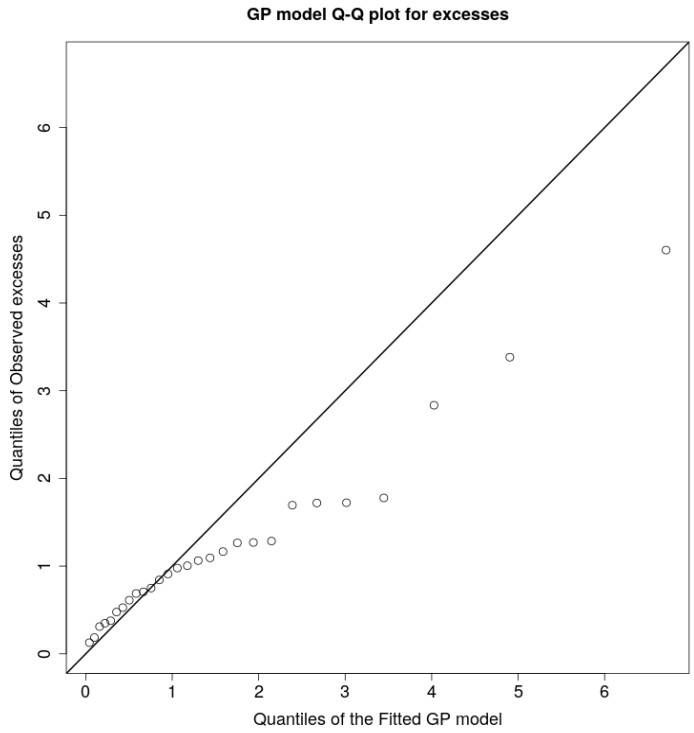


Figure 4.3: Diagnostics related to the GEV model for return levels upper bounds estimation of Law3.

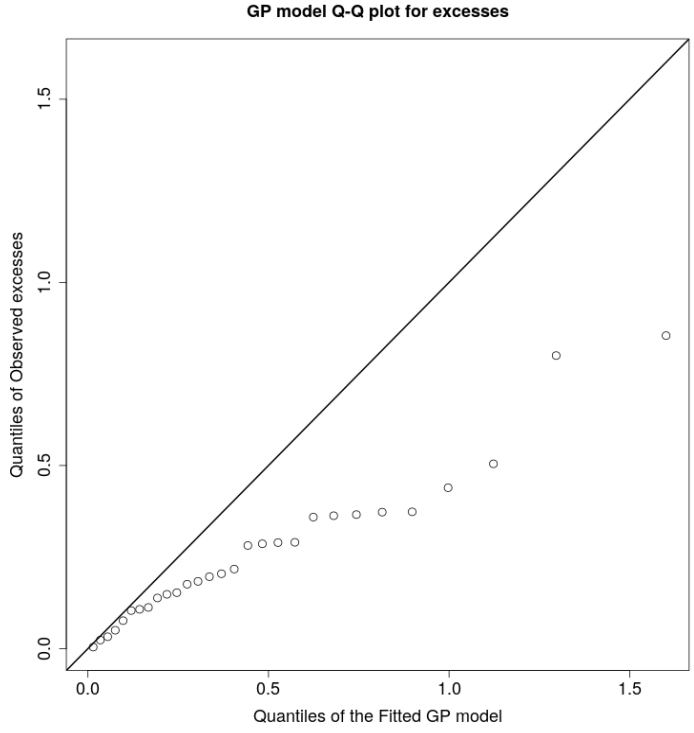


Figure 4.4: Diagnostics related to the GEV model for return levels upper bounds estimation of Law4.

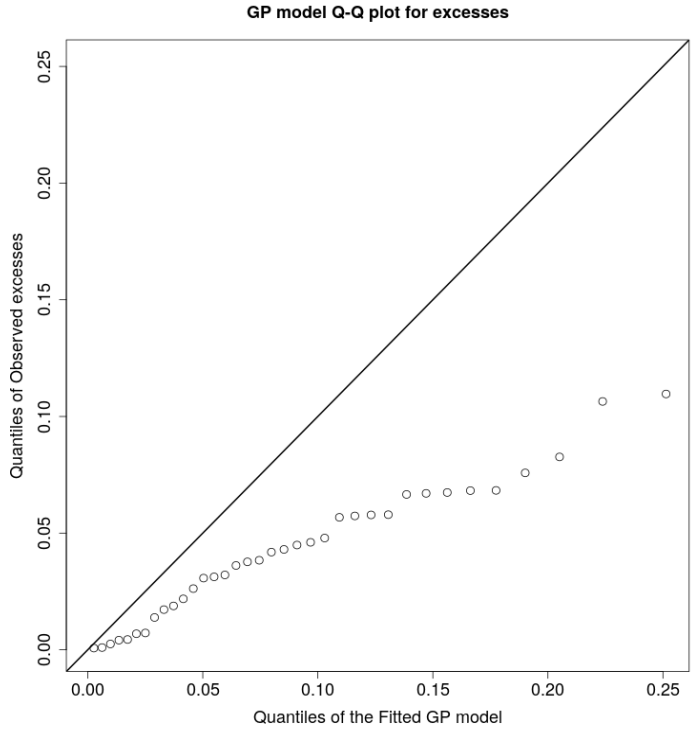


Figure 4.5: Diagnostics related to the GEV model for return levels upper bounds estimation of Law5.

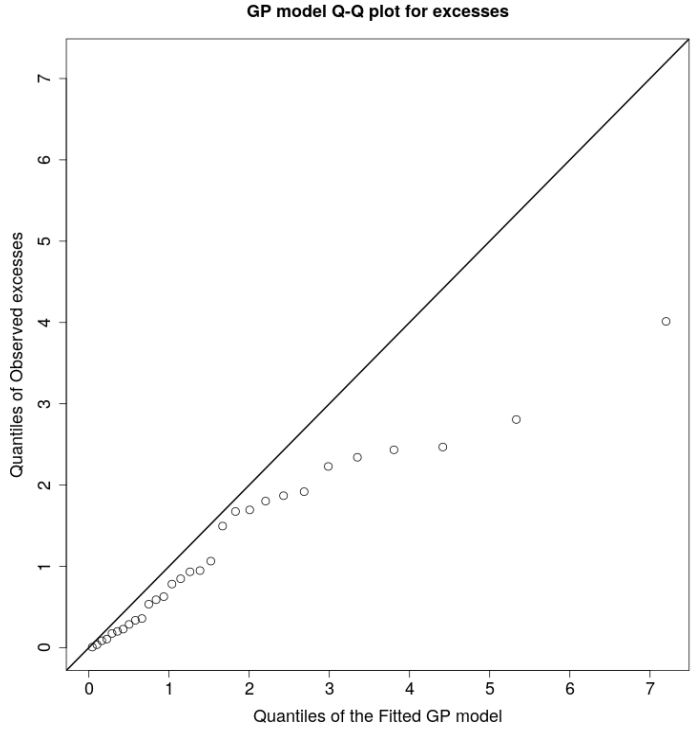


Figure 4.6: Diagnostics related to the GEV model for return levels upper bounds estimation of Law6.

4.2 Non-IID cases

In this section, we consider the strategy described by the Procedure 3.2 using **Stage 5-b)** in the **Routine procedure**. We apply this strategy on simulated data of size $N = 10,000$ from the stationary processes given in the examples below. We take the number 123 as seed value for each simulation in the $\mathcal{R}$ statistical software. Numerical results are gathered in Table 4.2 where one can see the estimated upper bounds for return levels of interest. The graphical diagnostics provided in Figures 4.7–4.10 show that the estimated upper bounds of return levels are greater than the true return levels for return periods smaller than the sample size N. It is therefore very likely that this conclusion remains true for longer return periods.

Exemple 4.7 (Process1) The autoregressive moving average process of first order, namely ARMA(1,1) with parameters ϕ and ψ is a time series X_t which satisfies

$$X_t - \phi\, X_{t-1} = \epsilon_t + \psi\, \epsilon_{t-1} \tag{4.7}$$

in which the random variables ϵ_t are independent and identically distributed. For illustration, we use the special case where $\phi = 0.25$, $\psi = 0.25$ and ϵ_t follows the normal distribution defined by (4.3) with parameters $\mu = 0$ and $\sigma = 1$.

Exemple 4.8 (Process2) The generalized autoregressive conditionally heteroscedastic process of first order, namely GARCH(1,1) with parameters α_0, α_1 and β is a time series X_t which satisfies

$$X_t = \sigma_{t|t-1}\, \epsilon_t \tag{4.8}$$

in which the random variables ϵ_t are independent and identically distributed, ϵ_t is independent of past x_{t-s}, $s = s1, 2, \cdots$, and

$$\sigma_{t|t-1} = \beta\, \sigma_{t-1|t-2} + \alpha_0 + \alpha_1\, x_{t-1}^2. \tag{4.9}$$

For illustration, we use the special case where $\alpha_0 = 0.25$, $\alpha_1 = 0.25$, $\beta = 0.25$ and ϵ_t follows the generalized Pareto distribution defined by (4.5) with parameters $\mu = 0$, $\sigma = 1$ and $\gamma = +0.2$.

Exemple 4.9 (Process3) For illustration, we consider the special case of the GARCH(1,1) process defined by the formula (4.8)-(4.9), where $\alpha_0 = 0.25$, $\alpha_1 = 0.25$, $\beta = 0.25$ and ϵ_t follows the Student distribution defined by (4.10) with parameters $\nu = 5$.

Exemple 4.10 (Process4) For illustration, we consider the special case of the GARCH(1,1) process defined by the formula (4.8)-(4.9), where $\alpha_0 = 0.25$, $\alpha_1 = 0.25$, $\beta = 0.25$ and ϵ_t follows the Student distribution defined by (4.10) with parameters $\nu = 2$.

Table 4.2: Table of estimated GEV distribution parameters obtained from the Procedure 3.2 when **Stage 5-b)** is considered in the **Routine procedure**. Upper bounds of return levels $\widehat{x}(T)$ associated with some return periods expressed in terms of numbers of observations are provided. The threshold $\widehat{u}$ is the quantile of observations whose order is $1-1/\widehat{i}$. The true values γ of shape parameters and the true values $x(T)$ of return levels for the considered stationary processes (Process1 to Process4) are not included.

	Process1	Process2	Process3	Process4
$\widetilde{\gamma}$	0.0096	0.5568	0.2861	0.9023
$\widetilde{\sigma}$	0.7183	9.9949	1.4868	24.0561
$\widetilde{\mu}$	3.0665	19.1773	3.7870	30.8539
$\widehat{i}$	341.0625	335.2188	333.7188	368.0000
$\widehat{u}$	3.0510	26.6459	4.3542	41.8074
$\widehat{\theta}$	0.9267	0.5210	0.6533	0.6416
$\widehat{\gamma}$	0.0083	0.4804	0.2469	0.7785
$\widehat{\sigma}$	0.7189	14.3693	1.6793	35.9047
$\widehat{\mu}$	3.1212	27.0331	4.4600	43.9850
$\widehat{x}(N)$	5.5714	148.7385	13.3388	592.2888
$\widehat{x}(10N)$	7.3016	458.8409	25.4470	3614.6998
$\widehat{x}\left(10^2N\right)$	9.0547	1393.9096	46.7373	21746.2584
$\widehat{x}\left(10^3N\right)$	10.8406	4219.9201	84.3097	130621.2637
$\widehat{x}\left(10^4N\right)$	12.6608	12762.7527	150.6398	784446.3008
$\widehat{x}\left(10^5N\right)$	14.5162	38587.7203	267.7429	4710885.9619

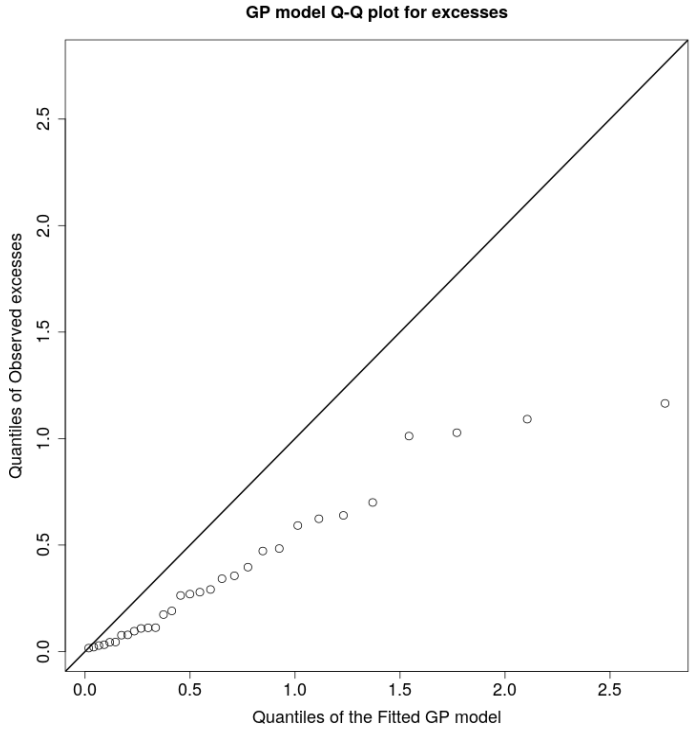


Figure 4.7: Diagnostics related to the GEV model for return levels upper bounds estimation of Process1.

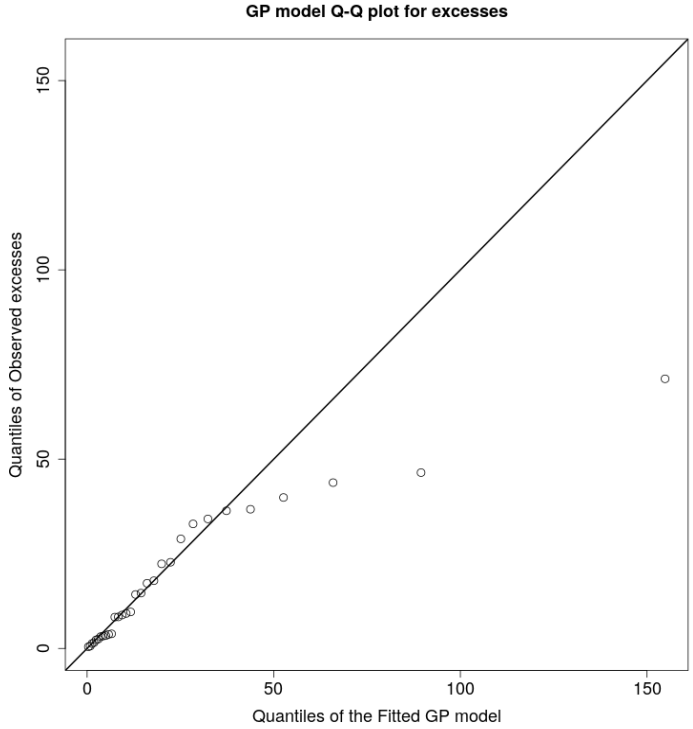


Figure 4.8: Diagnostics related to the GEV model for return levels upper bounds estimation of Process2.

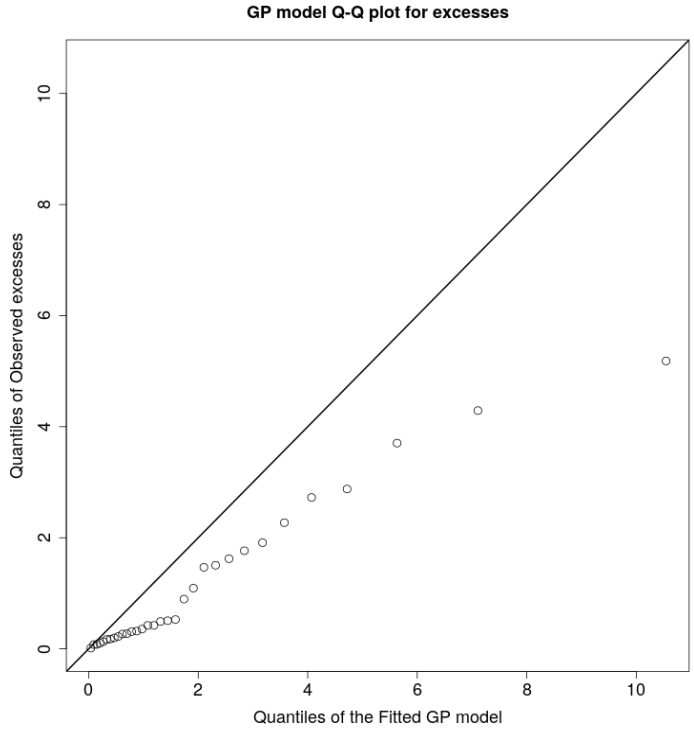


Figure 4.9: Diagnostics related to the GEV model for return levels upper bounds estimation of Process3.

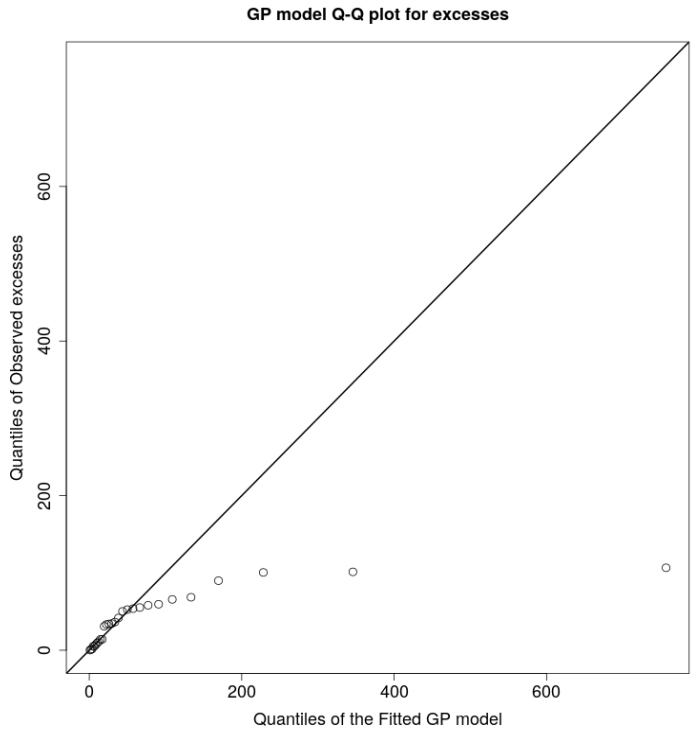


Figure 4.10: Diagnostics related to the GEV model for return levels upper bounds estimation of Process4.

4.3 Cautions

In this section, we consider the strategy described by the Procedure 3.2. We apply this strategy on simulated data of size $N = 10,000$ from the probability distributions given in the examples below. We take the number 123 as seed value for each simulation in the $\mathcal{R}$ statistical software. Numerical results gathered in Table 4.1 are obtained in the first particular case where **Stage 5-b)** of the **Routine procedure** is used. These results show that the estimated upper bounds for return levels of interest are not always greater than the true return levels. The similar numerical results provided in Table 4.4 are obtained in the second particular case where **Stage 5-a)** of the **Routine procedure** is used on the same data. It can be observed that the estimated upper bounds for return levels of interest are greater than the true return levels. We displayed the graphical diagnostics for return periods smaller than the sample size N in Figures 4.11–4.22. It follows that the estimated upper bounds for return levels of interest are not significantly greater than the true return levels in the first particular case, but are significantly greater than the true return levels in the second particular case. It is therefore necessary to always perform such diagnostics to check the reliability of estimators for the upper bounds of return levels resulting from the Procedure 3.2.

Exemple 4.11 (Law7) The Student distribution with degrees of freedom parameter $\nu > 0$ has probability density function defined for $x > 0$ by

$$f(x;\eta) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\,\pi}}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \tag{4.10}$$

where $\Gamma(\cdot)$ is the gamma function defined by

$$\Gamma(z) = \int_0^{+\infty} x^{z-1}\exp\{-x\}\,dx, \quad z > 0. \tag{4.11}$$

For illustration, we use the special case where $\nu = 5$.

Exemple 4.12 (Law8) For illustration, we consider the special case of the generalized Pareto (GP) distribution defined by the formula (4.5), where $\mu = 0$, $\sigma = 1$ and $\gamma = +0.2$.

Exemple 4.13 (Law9) For illustration, we consider the special case of the generalized Pareto (GP) distribution defined by the formula (4.5), where $\mu = 0$, $\sigma = 1$ and $\gamma = +0.5$.

Exemple 4.14 (Law10) The Laplace distribution with with location parameter $\mu \in \mathbb{R}$ and scale parameter $\sigma > 0$ has probability density function defined for $x \in \mathbb{R}$ by

$$f(x;\mu,\sigma) = \frac{1}{2\,\sigma}\exp\left\{-\frac{|x-\mu|}{\sigma}\right\}. \tag{4.12}$$

For illustration, we use the special case where $\mu = 0$ and $\sigma = 1$.

Exemple 4.15 (Law11) The log-gamma distribution with shape parameter $\alpha > 0$ and rate parameter $\beta > 0$ has probability density function defined for $x > 0$ by

$$f(x;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \frac{(\log x)^{\alpha-1}}{x^{\beta+1}}, \tag{4.13}$$

where $\Gamma(\cdot)$ is the gamma function defined by the formula (4.11). For illustration, we use the special case where $\alpha = 2$ and $\beta = 5$.

Exemple 4.16 (Law12) The log-logistic distribution with shape parameter $\alpha > 0$ and rate parameter $\beta > 0$ has probability density function defined for $x > 0$ by

$$f(x;\alpha,\beta) = \frac{\frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1}}{\left[1+\left(\frac{x}{\alpha}\right)^{\beta}\right]^2}. \tag{4.14}$$

For illustration, we use the special case where $\alpha = 2$ and $\beta = 5$.

Table 4.3: Table of estimated GEV distribution parameters obtained from the Procedure 3.2 when **Stage 5-b)** is considered in the **Routine procedure**. Upper bounds of return levels $\widehat{x}(T)$ associated with some return periods expressed in terms of numbers of observations are provided. The threshold $\widehat{u}$ is the quantile of observations whose order is $1-1/\widehat{i}$. The true values γ of shape parameters and the true values $x(T)$ of return levels for the considered probability distributions (Law7 to Law12) are also included.

	Law7	Law8	Law9	Law10	Law11	Law12
$\widetilde{\gamma}$	0.1789	0.2592	0.6005	0.1159	0.2250	0.2024
$\widetilde{\sigma}$	1.1883	3.6889	18.4688	1.1385	0.9442	1.1981
$\widetilde{\mu}$	4.5262	10.8452	36.1206	4.8954	4.5283	6.3007
$\widehat{i}$	306.4062	332.8750	362.2812	318.6250	253.5938	315.0312
$\widehat{u}$	4.5917	10.5808	34.0217	4.8447	4.4318	6.1480
$\widehat{\theta}$	0.9682	0.9296	0.9135	0.9978	1.0000	0.9379
$\widehat{\gamma}$	0.1543	0.2236	0.5181	0.1000	0.1941	0.1747
γ	0.2000	0.2000	0.5000	0.0000	0.2000	0.2000
$\widehat{\sigma}$	1.1951	3.7593	19.4997	1.1388	0.9442	1.2137
$\widehat{\mu}$	4.5647	11.1169	7.8373	4.8978	4.5283	6.3781
$\widehat{x}(N)$	10.0498	30.1492	208.1962	9.5580	9.5654	12.1054
$x(N)$	9.6776	26.5479	198.0000	8.5172	10.4989	12.6189
$\widehat{x}(10N)$	15.7351	54.4942	691.8872	13.7433	15.1797	18.4288
$x(10N)$	15.5469	45.0000	630.4555	10.8198	17.2416	20.0000
$\widehat{x}\left(10^2N\right)$	23.8110	95.0651	2282.6604	18.9869	23.9285	27.8422
$x\left(10^2N\right)$	24.7710	74.2447	1998.0000	13.1224	28.1539	31.6979
$\widehat{x}\left(10^3N\right)$	35.3275	162.9328	7526.2429	25.5856	37.6029	41.9100
$x\left(10^3N\right)$	39.3418	120.5943	6322.5553	15.4249	45.7852	50.2377
$\widehat{x}\left(10^4N\right)$	51.7568	276.5054	24814.2307	33.8932	58.9824	62.9418
$x\left(10^4N\right)$	62.4045	194.0536	19997.9999	17.7275	74.2292	79.6214
$\widehat{x}\left(10^5N\right)$	75.1954	466.5695	81813.6499	44.3526	92.4098	94.3858
$x\left(10^5N\right)$	98.9372	310.4787	63243.5541	20.0301	120.0545	126.1915

Table 4.4: Table of estimated GEV distribution parameters obtained from the Procedure 3.2 when **Stage 5-a)** is considered in the **Routine procedure**. Upper bounds of return levels $\widehat{x}(T)$ associated with some return periods expressed in terms of numbers of observations are provided. The threshold $\widehat{u}$ is the quantile of observations whose order is $1-1/\widehat{i}$. The true values γ of shape parameters and the true values $x(T)$ of return levels for the considered probability distributions (Law7 to Law12) are also included.

	Law7	Law8	Law9	Law10	Law11	Law12
$\widetilde{\gamma}$	0.3753	0.4531	0.7829	0.2732	0.4984	0.4513
$\widetilde{\sigma}$	1.6278	4.2128	21.9332	1.5411	1.3648	1.7179
$\widetilde{\mu}$	4.7236	11.2787	36.7877	5.0421	4.9624	6.5281
$\widehat{i}$	374.0000	380.2812	382.8125	379.0000	384.4688	382.5312
$\widehat{u}$	4.8475	11.0994	35.3485	4.9810	4.9448	6.4474
$\widehat{\theta}$	0.9952	0.8907	0.8926	1.0000	1.0000	0.8918
$\widehat{\gamma}$	0.3145	0.3797	0.6560	0.2289	0.4176	0.3781
γ	0.2000	0.2000	0.5000	0.0000	0.2000	0.2000
$\widehat{\sigma}$	1.6307	4.4397	23.9739	1.5411	1.3648	1.8091
$\widehat{\mu}$	4.7313	11.7794	39.3944	5.0421	4.9624	6.7301
$\widehat{x}(N)$	14.0337	40.2524	309.6191	12.4884	14.3332	18.2591
$x(N)$	9.6776	26.5479	198.0000	8.5172	10.4989	12.6189
$\widehat{x}(10N)$	29.5948	97.0075	1408.1295	22.4252	34.9974	41.1678
$x(10N)$	15.5469	45.0000	630.4555	10.8198	17.2416	20.0000
$\widehat{x}\left(10^2N\right)$	61.5682	232.5681	6373.8328	39.1797	88.8666	95.6852
$x\left(10^2N\right)$	24.7710	74.2447	1998.0000	13.1224	28.1539	31.6979
$\widehat{x}\left(10^3N\right)$	127.5021	557.4062	28856.9081	67.5508	229.7221	225.8481
$x\left(10^3N\right)$	39.3418	120.5943	6322.5553	15.4249	45.7852	50.2377
$\widehat{x}\left(10^4N\right)$	263.5156	1336.0503	130668.9829	115.6126	598.1366	536.7186
$x\left(10^4N\right)$	62.4045	194.0536	19997.9999	17.7275	74.2292	79.6214
$\widehat{x}\left(10^5N\right)$	544.1048	3202.5367	591720.7691	197.0350	1561.7716	1279.2000
$x\left(10^5N\right)$	98.9372	310.4787	63243.5541	20.0301	120.0545	126.1915

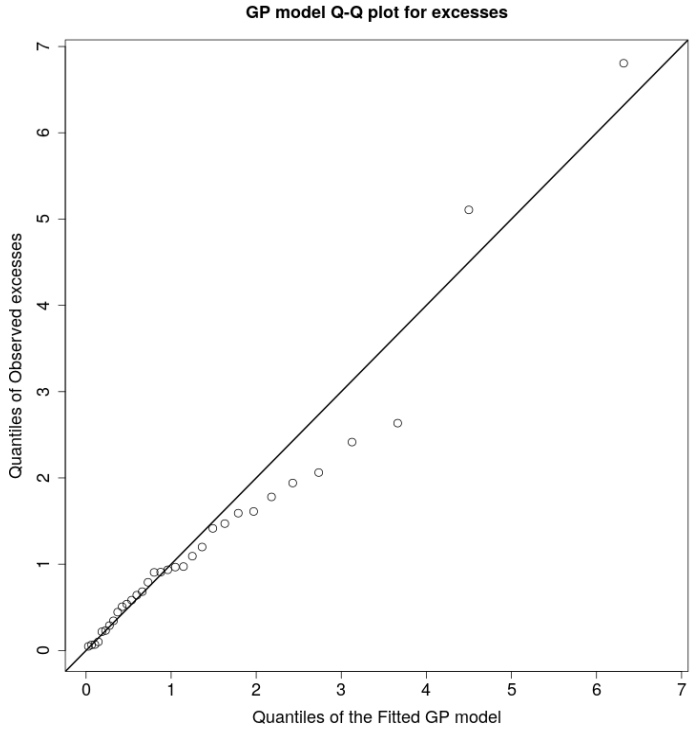


Figure 4.11: GEV model (Procedure 3.2 + **Stage 5-b)**) for return levels upper bounds estimation of Law7.

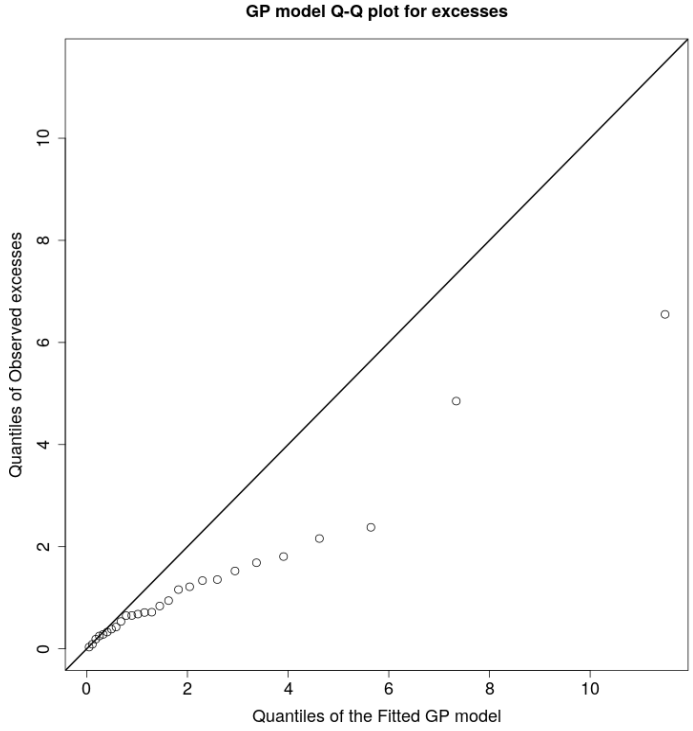


Figure 4.12: GEV model (Procedure 3.2 + **Stage 5-a)**) for return levels upper bounds estimation of Law7.

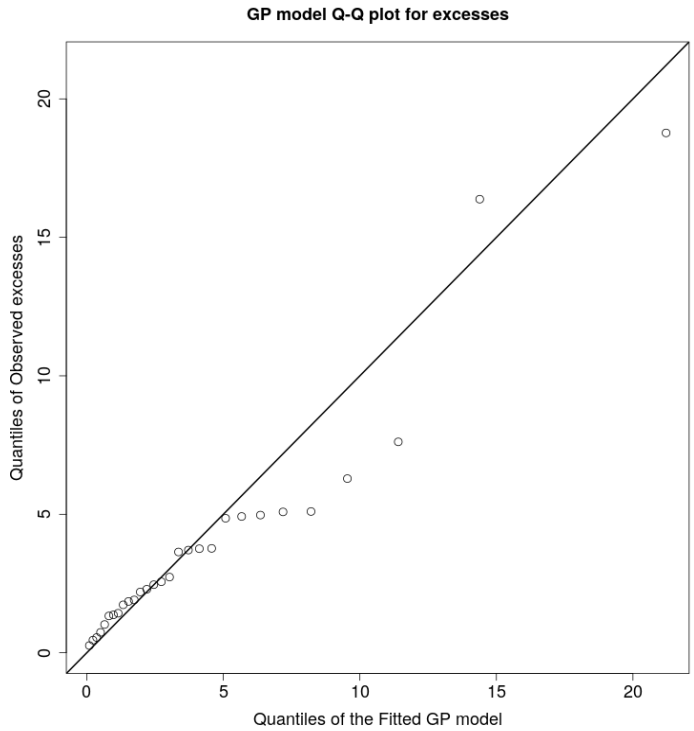


Figure 4.13: GEV model (Procedure 3.2 + **Stage 5-b)**) for return levels upper bounds estimation of Law8.

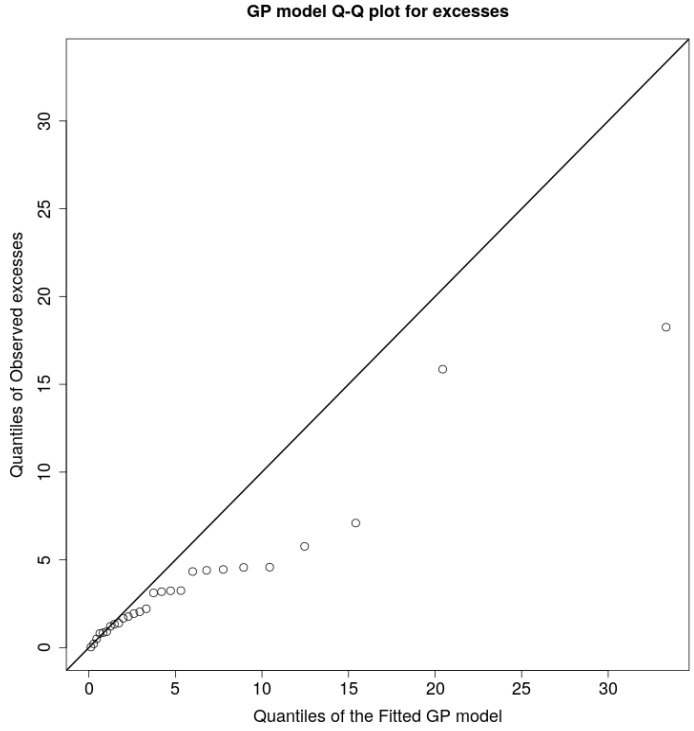


Figure 4.14: GEV model (Procedure 3.2 + **Stage 5-a)**) for return levels upper bounds estimation of Law8.

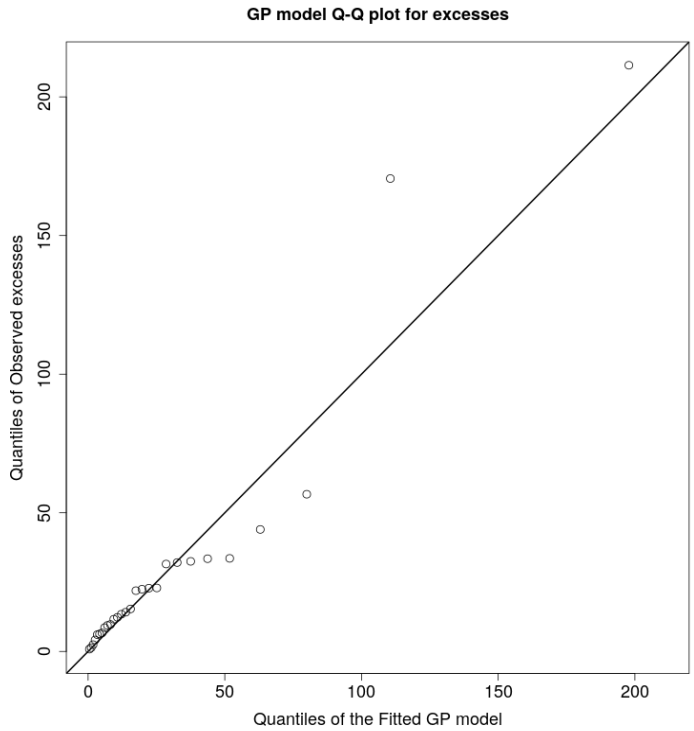


Figure 4.15: GEV model (Procedure 3.2 + **Stage 5-b)**) for return levels upper bounds estimation of Law9.

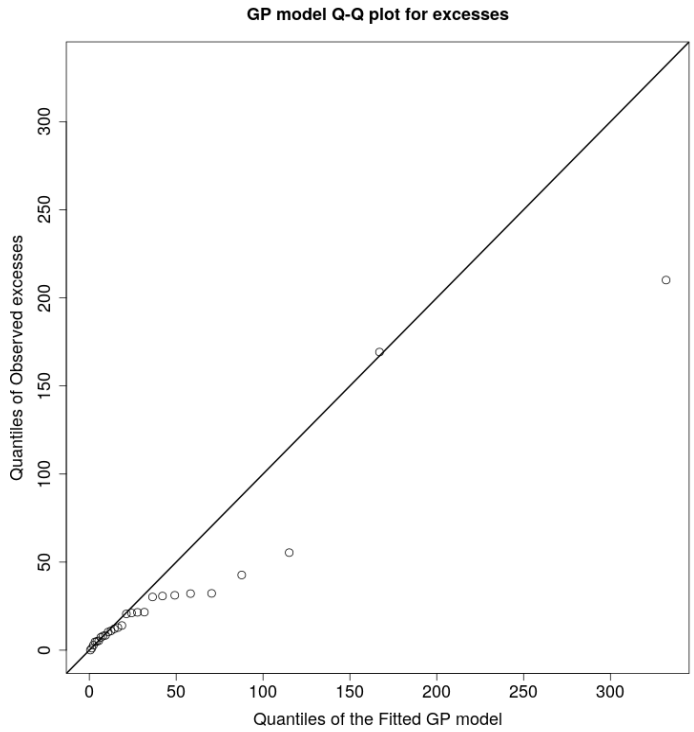


Figure 4.16: GEV model (Procedure 3.2 + **Stage 5-a)**) for return levels upper bounds estimation of Law9.

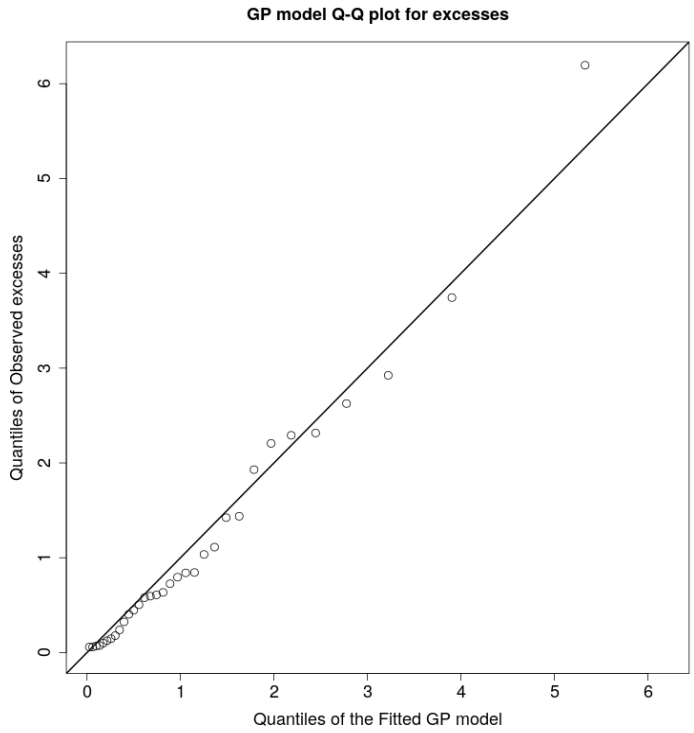


Figure 4.17: GEV model (Procedure 3.2 + **Stage 5-b)**) for return levels upper bounds estimation of Law10.

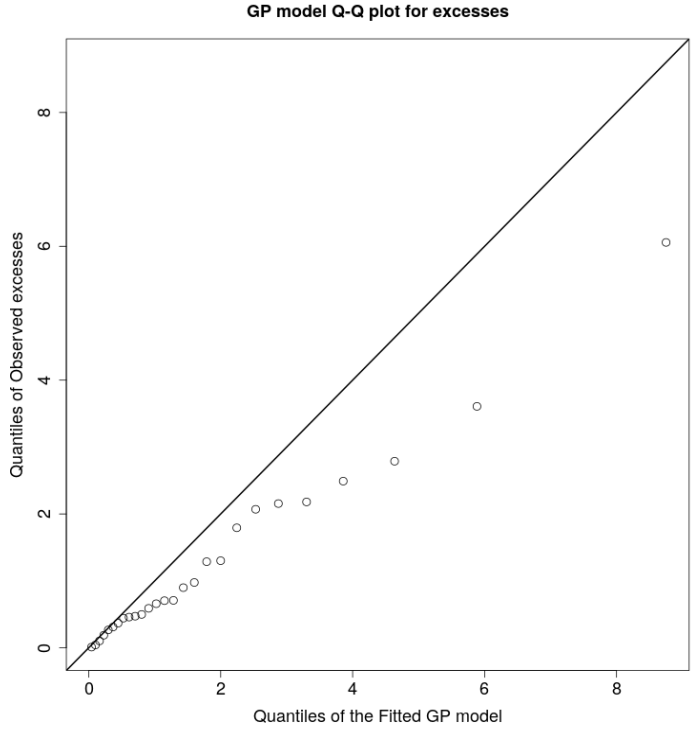


Figure 4.18: GEV model (Procedure 3.2 + **Stage 5-a)**) for return levels upper bounds estimation of Law10.

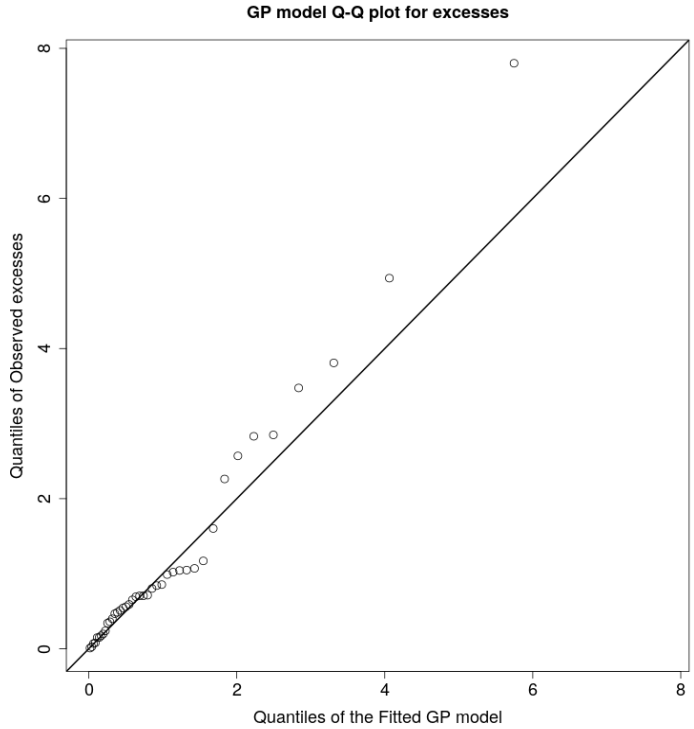


Figure 4.19: GEV model (Procedure 3.2 + **Stage 5-b)**) for return levels upper bounds estimation of Law11.

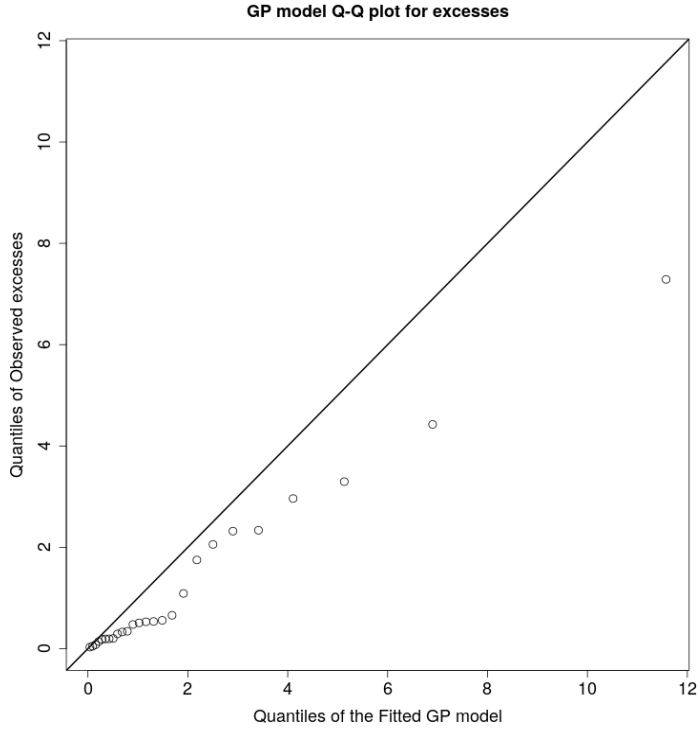


Figure 4.20: GEV model (Procedure 3.2 + **Stage 5-a)**) for return levels upper bounds estimation of Law11.

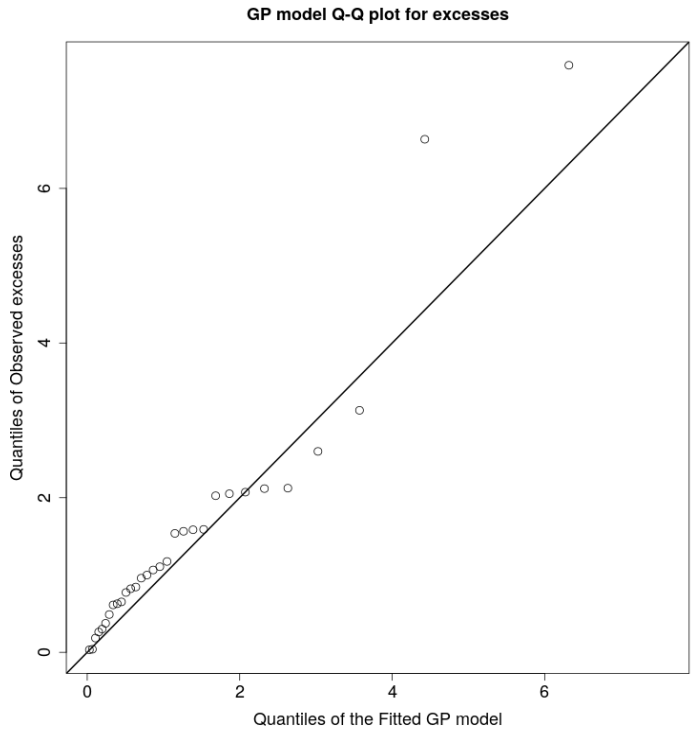


Figure 4.21: GEV model (Procedure 3.2 + **Stage 5-b)**) for return levels upper bounds estimation of Law12.

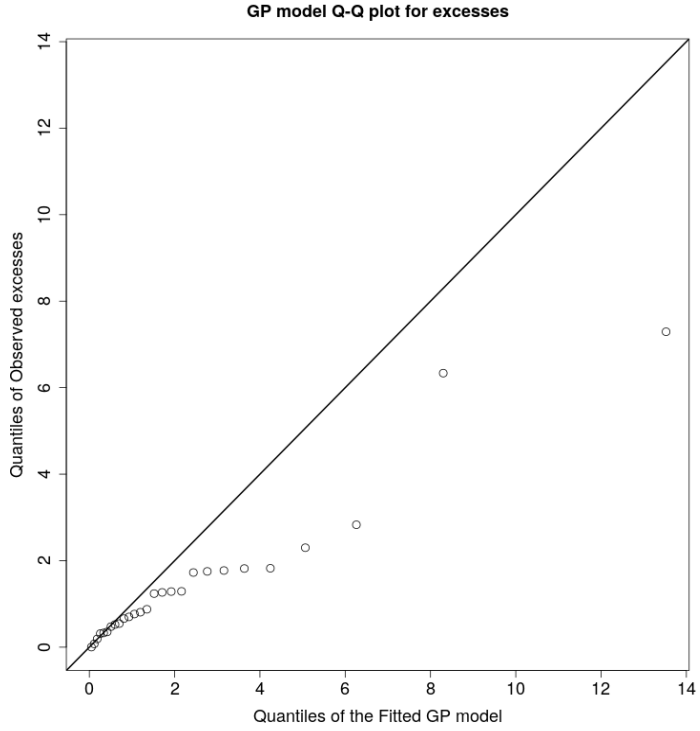
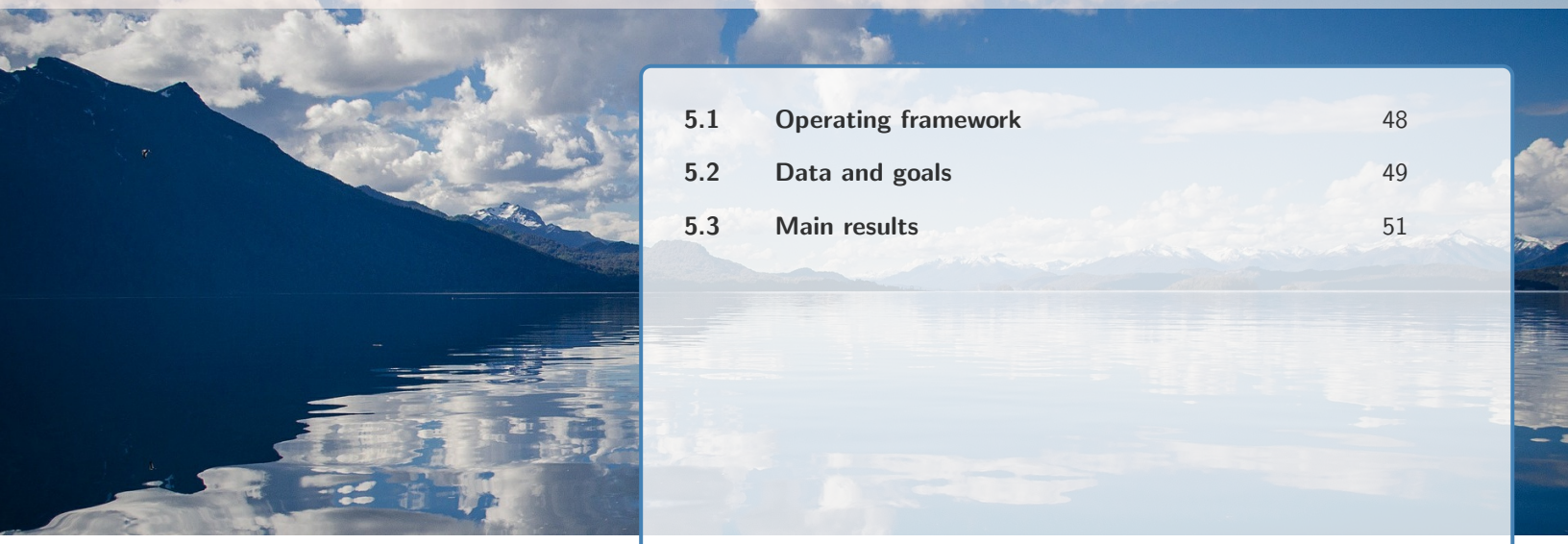


Figure 4.22: GEV model (Procedure 3.2 + **Stage 5-a)**) for return levels upper bounds estimation of Law12.

5. Illustrations on univariate time series in practice



5.1	**Operating framework**	48
5.2	**Data and goals**	49
5.3	**Main results**	51

5.1 Operating framework

We consider the pairs of longitudinal and lateral coordinates associated with the horizontal localization components of a vehicle. These errors are derived from the Ford autonomous vehicle dataset which can be freely downloaded at `https://avdata.ford.com/downloads/`. Data are collected by a fleet of Ford autonomous vehicles at different days and times during 2017-2018. The vehicles traversed an average route of 66 km in Michigan that included a mix of driving scenarios such as the Detroit Airport, freeways, city-centers, university campus and suburban neighbourhoods, etc. These experiments were carry out in various weather, lighting, construction and traffic conditions in dynamic urban environments. To get more details about the Ford autonomous vehicle datasets, visit the website `https://avdata.ford.com/`. A complete description of the collected datasets can be found in [1]. The seasonal variation consists of sunny, cloudy, fall and snow whereas the driving scenarios include freeway, residential, overpass, airport, bridge, tunnel, construction and vegetation. The raw dataset is divided into six parts or Logs, each part containing the driving weather and environmental conditions as shown in the Figure 5.1.

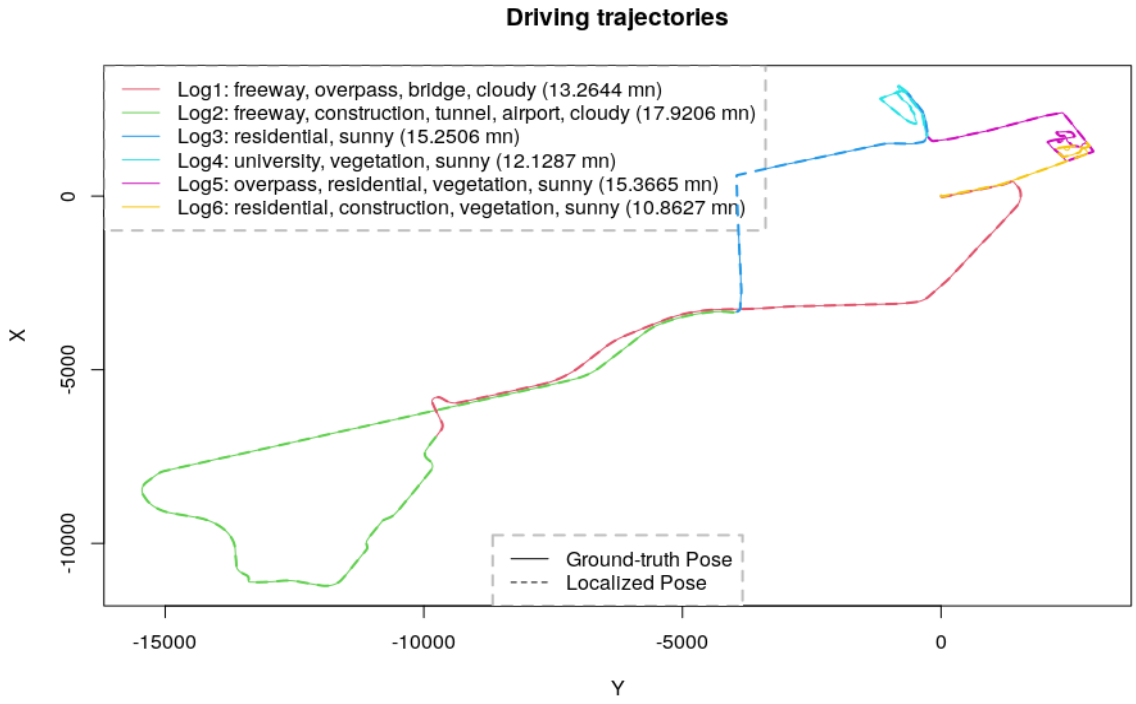


Figure 5.1: Overview of the driving scenarios.

5.2 Data and goals

Let E be the random variable associated with the **localization error** of a vehicle defined by

$$E = \sqrt{(X_G - X_L)^2 + (Y_G - Y_L)^2}, \tag{5.1}$$

where (X_G, Y_G) and (X_L, Y_L) are the random variables associated with coordinates of the true and predicted positions which are technically called **Ground-truth Pose** and **Localized Pose**, respectively. The box plots and the density plots of errors provided in Figure 5.2–5.3 give a complete view of how the observations of theses errors are distributed across the six driving contexts and the overall driving scenario designated as **Log0**. It is clear from these plots that the distributions of errors are asymmetric and right-tailed. In addition, the mean values of errors are greater in contexts Log1 and Log2 than in contexts Log4, Log5 and Log6. Numerical statistics which summarize these errors are gathered in Table 5.1. The goal is to estimate the upper bounds of the errors that can be obtained if the experiment in each context is repeated up to 10^5 times.

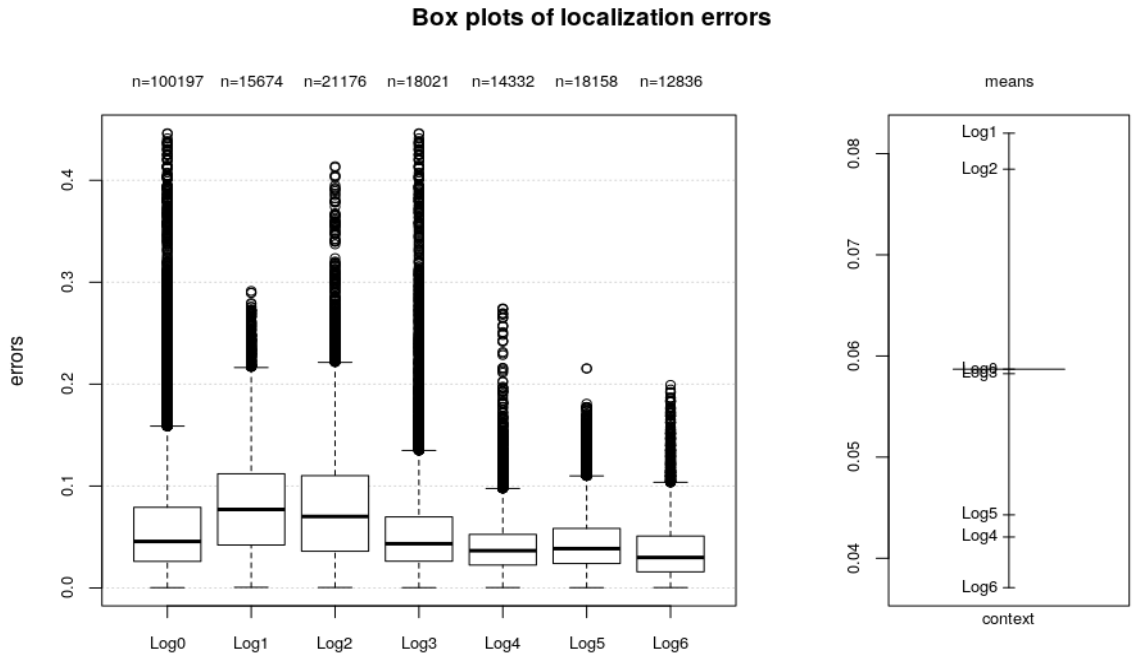


Figure 5.2: Box plots of errors E calculated by the formula (5.1) for each driving context or scenario.

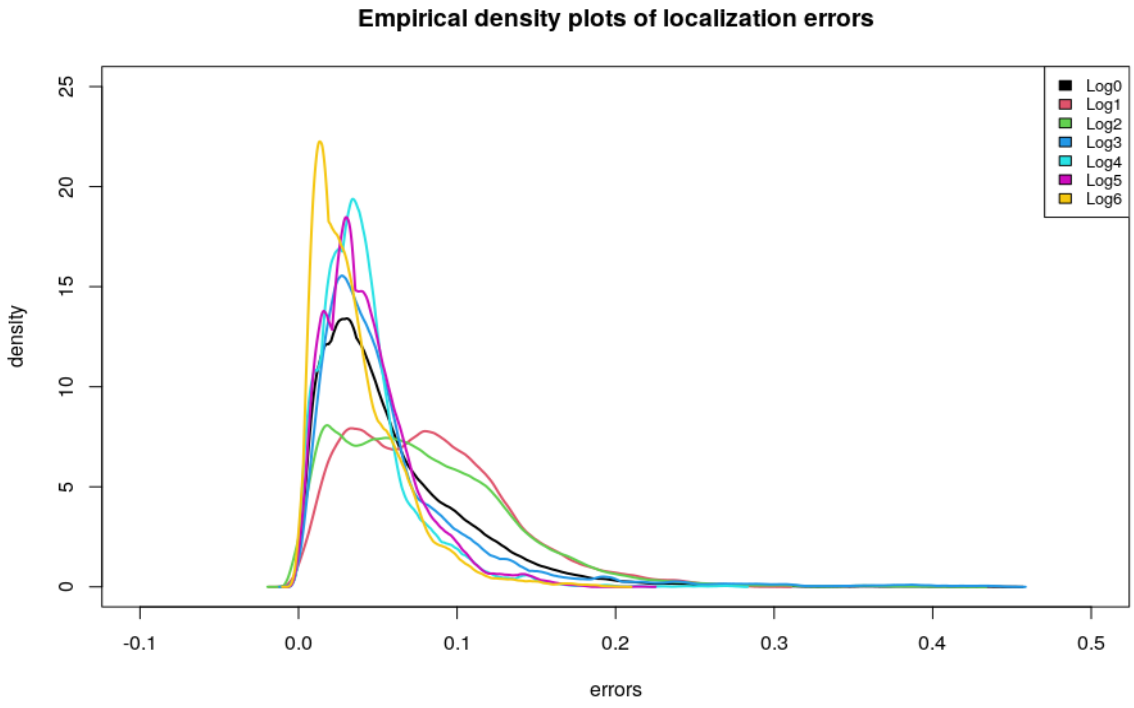


Figure 5.3: Density curves of errors E calculated by the formula (5.1) for each driving context scenario.

Table 5.1: Basic summary statistics of errors E calculated by the formula (5.1) for each driving context including those of the overall scenario **Log0**. The standard deviation (sd), the mean absolute deviation (mad) as well as the skewness (skew) of errors are included.

	Log0	Log1	Log2	Log3	Log4	Log5	Log6
n	100197	15674	21176	18021	14332	18158	12836
mean	0.0587	0.0820	0.0785	0.0582	0.0421	0.0443	0.0371
sd	0.0470	0.0494	0.0544	0.0540	0.0292	0.0288	0.0281
median	0.0457	0.0770	0.0701	0.0434	0.0366	0.0386	0.0299
mad	0.0355	0.0518	0.0545	0.0293	0.0222	0.0250	0.0240
min	0.0003	0.0006	0.0005	0.0003	0.0004	0.0003	0.0004
max	0.4462	0.2914	0.4137	0.4462	0.2742	0.2154	0.1992
range	0.4459	0.2908	0.4132	0.4459	0.2738	0.2151	0.1988
skew	1.9292	0.7937	1.1784	2.9410	1.9422	1.2521	1.5674
kurtosis	6.3444	0.5942	2.5278	12.0772	6.9300	2.0367	3.3751

5.3 Main results

To archived our main goals, we apply the strategy described in Chapter 3 to the sequence of error from each driving context. More precisely, this strategy consists of the Procedure 3.2 using **Stage 5-b)** in the **Routine procedure**. The estimated GEV model parameters along with the upper bounds for return levels of interest are provided in Table 5.2. It is worth noticing that for the overall driving scenario Log0, only the $10,000$ largest values of errors are fed to the considered strategy. These GEV distributions for return levels upper bounds estimation are validated by the diagnostic plots of Figures 5.4–5.10 in which theoretical and empirical quantiles of excesses over high thresholds are compared. At glance, one can observe the following results.

- The GEV distributions of errors are bounded in the contexts Log1 and Log2 since their shape parameters are smaller than zero.
- The GEV distributions of errors are heavy tailed in the contexts Log3 and Log4 since their shape parameters are positive and far from zero.
- The GEV distributions of errors are light tailed in the contexts Log5, Log6 and Log0 since their shape parameters are positive and close to zero.
- The dependency between large errors are strong in each driving context since the related extremal indexes are very close to zero.

It follows from the return levels provided in Table 5.2 that the amplitudes of errors depend on driving contexts. For instance, the errors which occur in average once every 10^5 times the experimentation period are at the scale of meter when driving in contexts such as Log1, Log2, Log4, Log5 and Log6 while these errors are at the scale of decameter when driving in contexts such as Log3. One can say that the results on Log3 are a bit exceptional compared to those on other driving contexts. In such a case, we suggest reapplying the procedure for over-estimating return levels on a time series of longer size. This obviously implies increasing (doubling, tripling, etc.) the duration of the experiments. It is worth doing this additional work in order to validate or reject the exceptional character of the results.

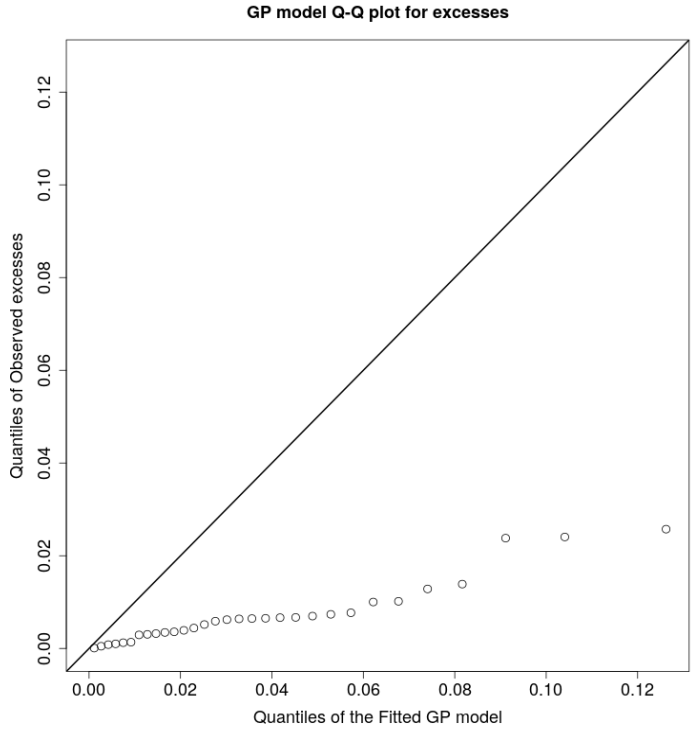


Figure 5.4: Diagnostics related to the GEV model for return levels upper bounds estimation in Log1.

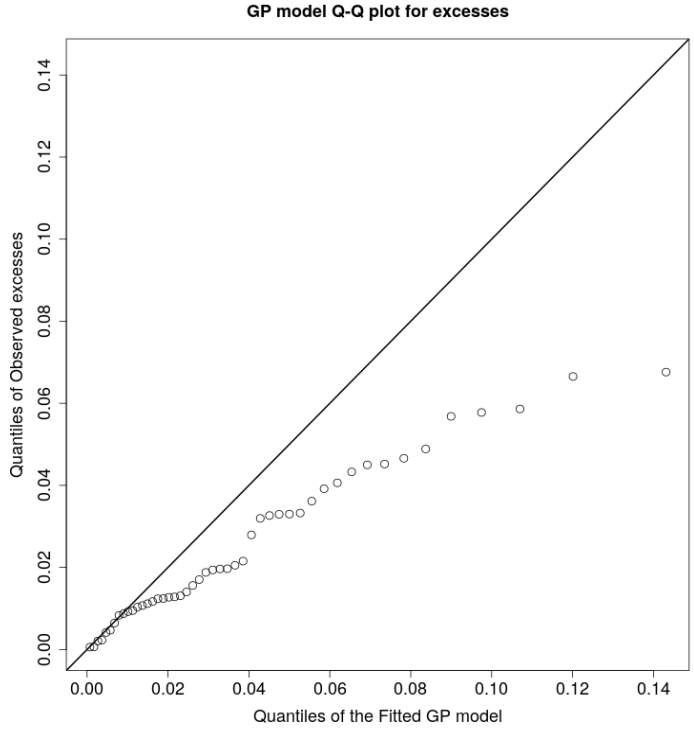


Figure 5.5: Diagnostics related to the GEV model for return levels upper bounds estimation in Log2.

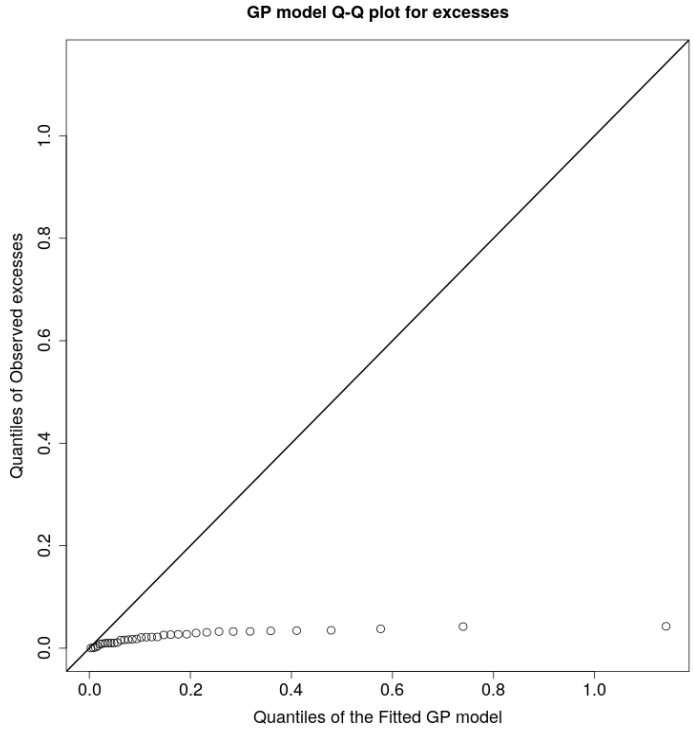


Figure 5.6: Diagnostics related to the GEV model for return levels upper bounds estimation in Log3.

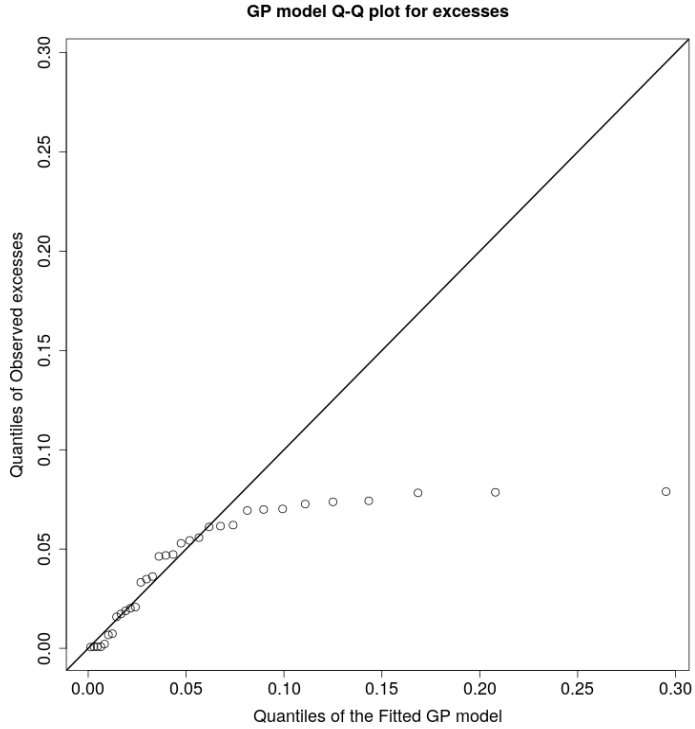


Figure 5.7: Diagnostics related to the GEV model for return levels upper bounds estimation in Log4.

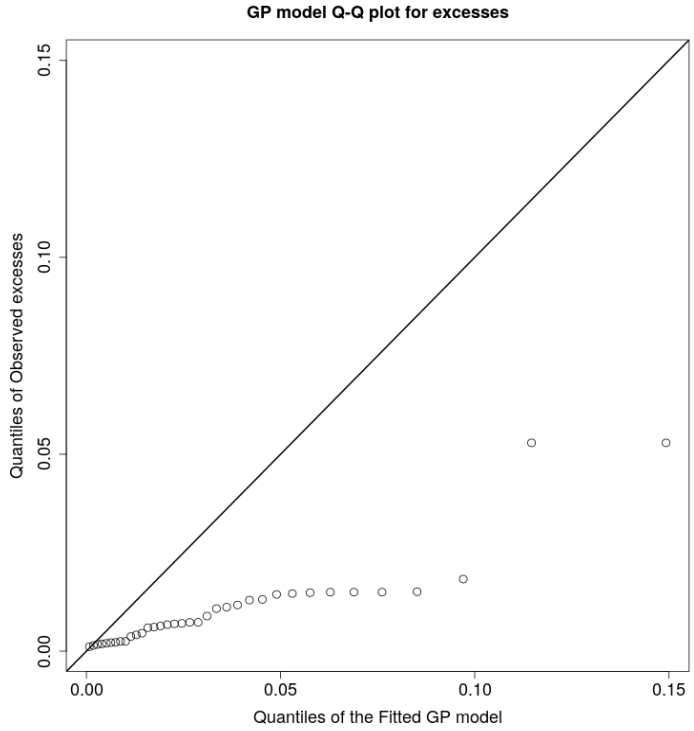


Figure 5.8: Diagnostics related to the GEV model for return levels upper bounds estimation in Log5.

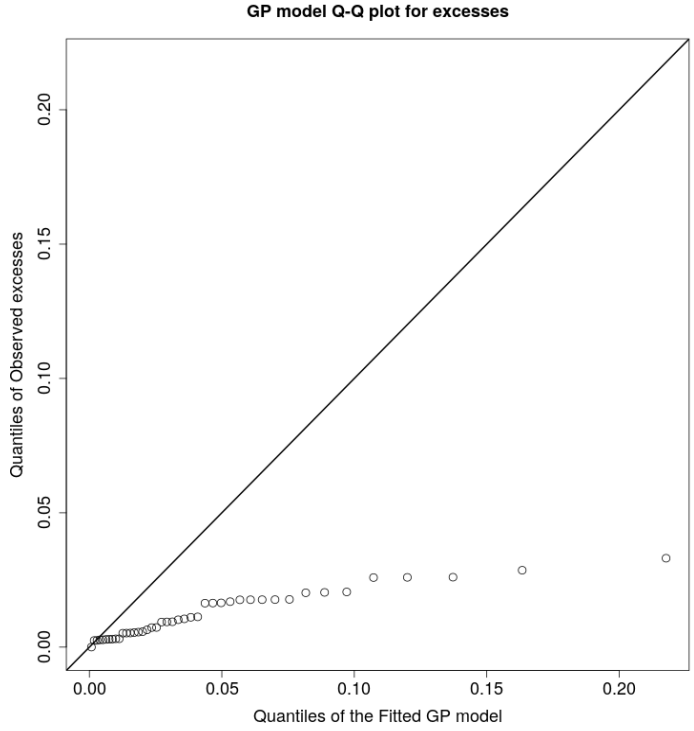


Figure 5.9: Diagnostics related to the GEV model for return levels upper bounds estimation in Log6.

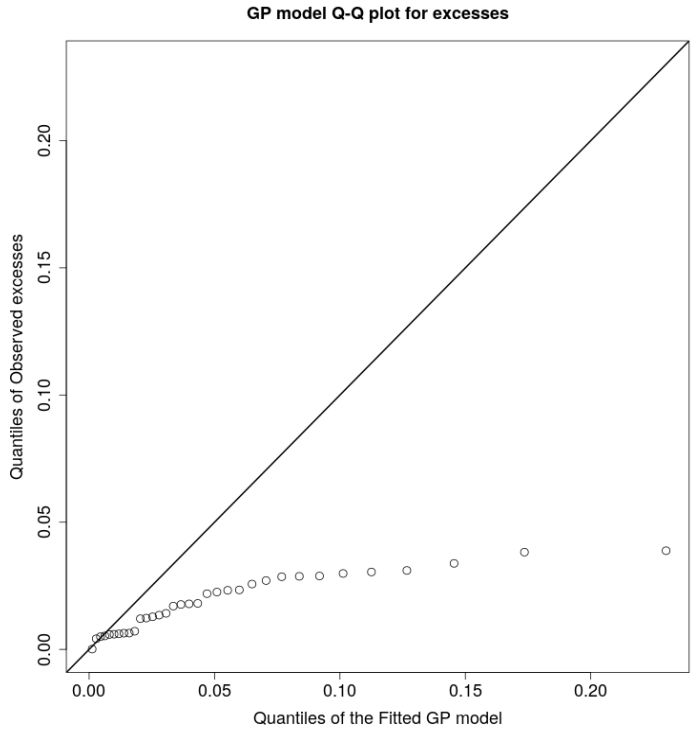
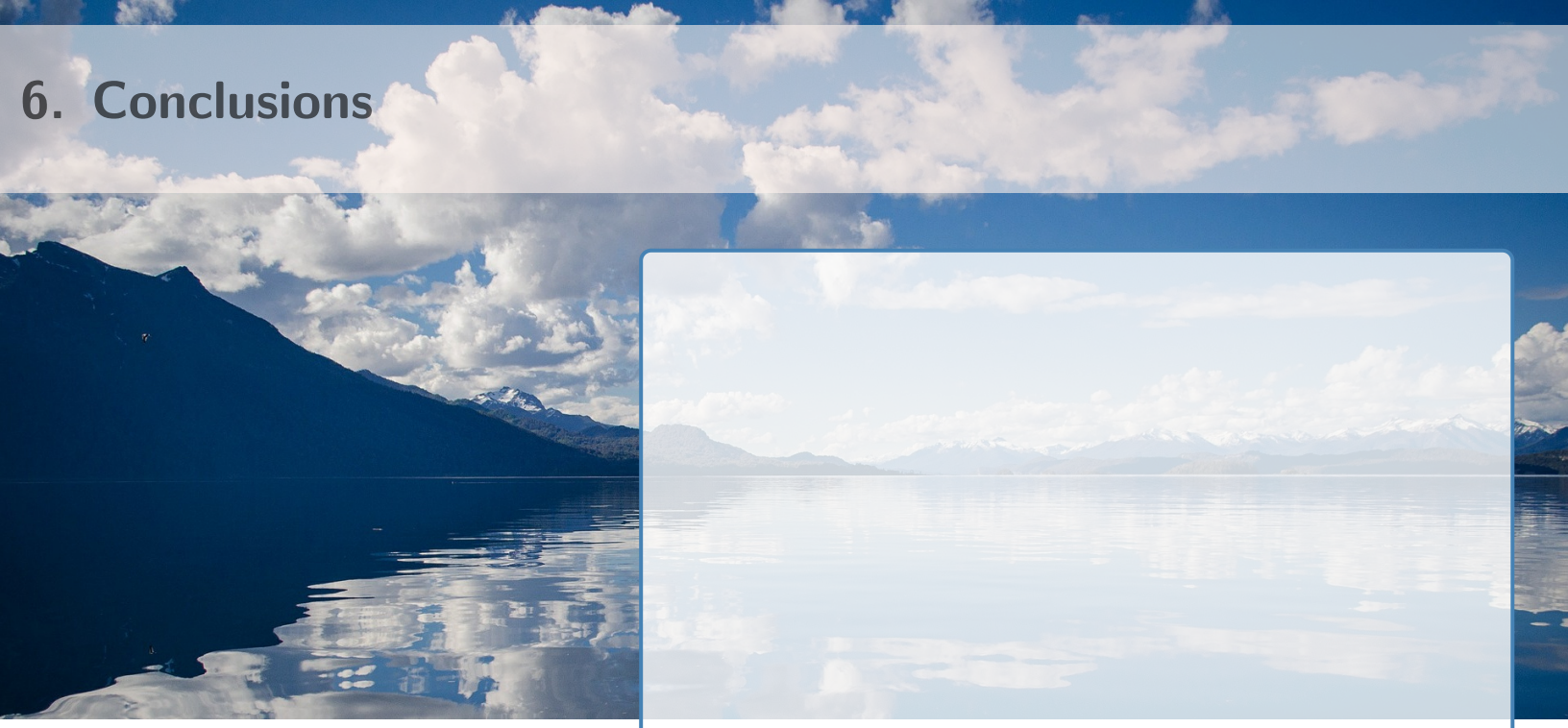


Figure 5.10: Diagnostics related to the GEV model for return levels upper bounds estimation in the overall driving scenario called **Log0**.

Table 5.2: Table of estimated GEV distribution parameters obtained from the Procedure 3.2 when **Stage 5-b)** is considered in the **Routine procedure**. Upper bounds of return levels $\widehat{x}$ associated with some return periods are provided. Here, the driving duration T is expressed in minutes and the threshold $\widehat{u}$ is the quantile of observations whose order is $1 - 1/\widehat{i}$..

	Log1	Log2	Log3	Log4	Log5	Log6	Log0
$\widetilde{\gamma}$	-0.2121	-0.1746	0.3721	0.2008	0.0196	0.1028	0.0524
$\widetilde{\sigma}$	0.0655	0.0768	0.0390	0.0314	0.0350	0.0329	0.0457
$\widetilde{\mu}$	0.1740	0.1681	0.0896	0.0761	0.0774	0.0599	0.2387
$\widehat{i}$	559.3125	574.3125	520.3750	467.0000	511.8438	361.4375	307.1875
$\widehat{u}$	0.2657	0.3461	0.4035	0.1953	0.1625	0.1661	0.4074
$\widehat{\theta}$	0.2272	0.0581	0.1578	0.1862	0.0730	0.0668	0.0778
$\widehat{\gamma}$	-0.1830	-0.1507	0.3211	0.1732	0.0169	0.0887	0.0452
$\widehat{\sigma}$	0.0478	0.0467	0.0776	0.0440	0.0368	0.0435	0.0523
$\widehat{\mu}$	0.2573	0.3404	0.1932	0.1388	0.1713	0.1627	0.3637
duration (T)	13.2644	17.9206	15.2506	12.1287	15.3665	10.8627	84.7935
$\widehat{x}(T)$	0.3762	0.4702	0.7022	0.3432	0.3063	0.3446	0.5600
$\widehat{x}(10T)$	0.4255	0.5233	1.5304	0.5696	0.3987	0.4979	0.7093
$\widehat{x}\left(10^2T\right)$	0.4576	0.5607	3.2597	0.9055	0.4942	0.6852	0.8742
$\widehat{x}\left(10^3T\right)$	0.4786	0.5871	6.8807	1.4059	0.5936	0.9148	1.0570
$\widehat{x}\left(10^4T\right)$	0.4924	0.6058	14.4646	2.1516	0.6968	1.1965	1.2598
$\widehat{x}\left(10^5T\right)$	0.5015	0.6190	30.3488	3.2628	0.8042	1.5421	1.4849

6. Conclusions



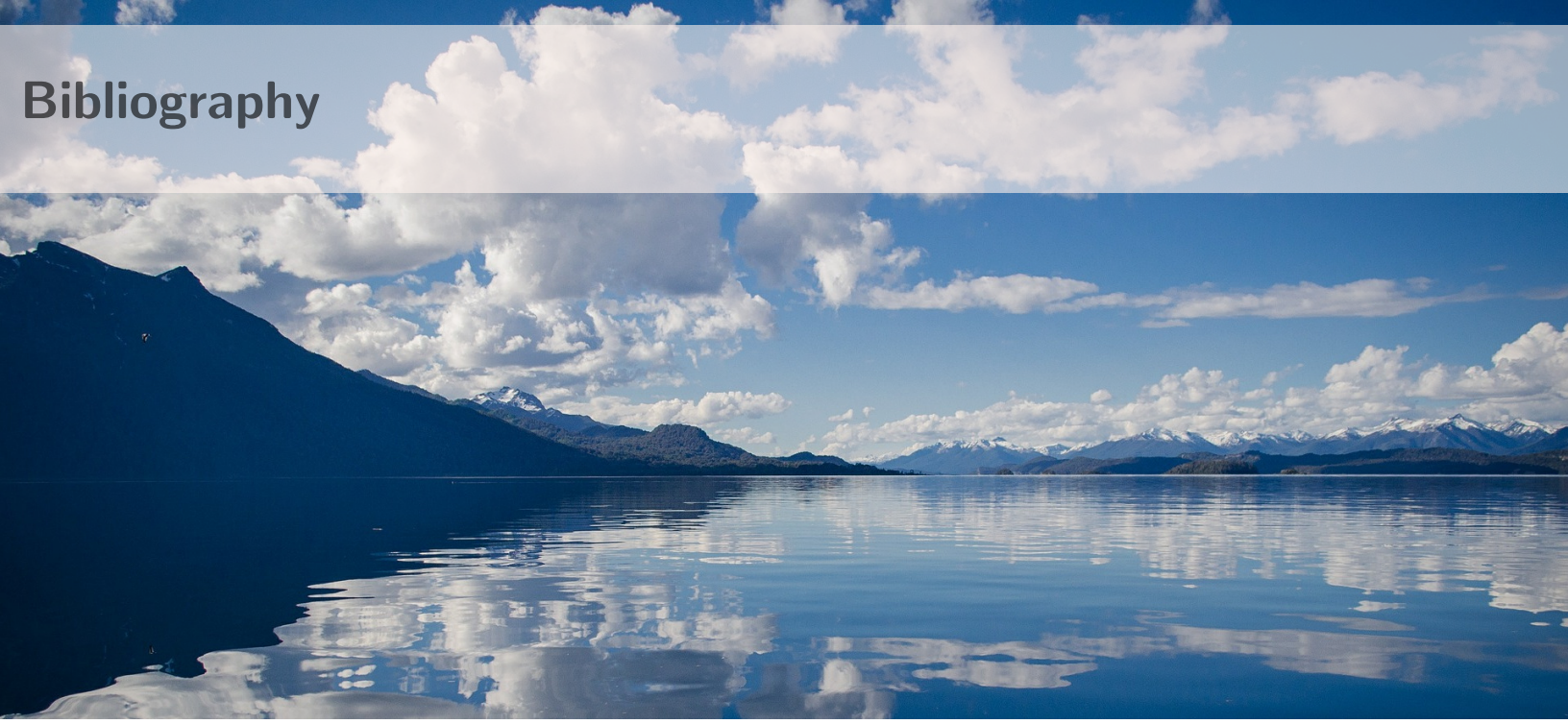
In this work, we have proposed an original strategy based on an algorithmic approach to construct estimators for upper bounds of return levels associated with continuous random variables and stationary processes.

We have demonstrated the theoretical validity of this strategy and we have proposed a diagnostic test that must be carried out to assess the quality of these estimators since the true stationary process from which the data come is generally unknown in practice.

We have automated all the stages of this strategy into an intuitive and interactive web application which is fast even for very long time series.

We plan to extend this work in the design of a similar approach suitable to construct estimators for the upper bounds of return levels associated with two dependent continuous random variables.

Bibliography



[1] S. Agarwal, A. Vora, and G. Pandey. "Ford Multi-AV Seasonal Dataset". In: *The International Journal of Robotics Research* 39.12 (2020), pp. 1367–1376. DOI: https://doi.org/10.1177/0278364920961451.

[2] J. Beirlant, Y. Goegebeur, and J. Segers. *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics, 2004.

[3] A. Bücher and J. Segers. "Inference for heavy tailed stationary time series based on sliding blocks". In: *Electronic Journal of Statistics* 12.1 (2018), pp. 1098–1125. DOI: https://doi.org/10.1214/18-EJS1415.

[4] Ö. Ciǵdem, E. Özge, and E. Esra. "A proposal method to select the optimal block size: An application to financial markets". In: *International Journal of Research in Technology and Management* 4.2 (2018), 5 pages.

[5] Ö. Ciǵdem, E. Özge, and H. Saygin. "A New Methodology for the Block Maxima Approach in Selecting the Optimal Block Size". In: *Tehnic ki vjesnik* 26.5 (2019), pp. 1292–1296.

[6] S. G. Coles. *An introduction to statistical modeling of extreme values*. Springer Series in Statistics, 2001.

[7] P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag, Berlin, 1997.

[8] M. Ferreira. "Heuristic tools for the estimation of the extremal index: a comparison of methods". In: *REVSTAT – Statistical Journal* 16.1 (2018), pp. 115–136. URL: http://hdl.handle.net/1822/50644.

[9] C. A. T. Ferro and J. Segers. "Inference for Clusters of Extreme Values". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 65.2 (2003), pp. 545–556. ISSN: 13697412, 14679868. URL: http://www.jstor.org/stable/3647520.

[10] R. A. Fisher and L. H. C. Tippett. "Limiting forms of the frequency distribution of the largest or smallest member of a sample". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 24.2 (1928), pp. 180–190.

[11] B. Gnedenko. "Sur La Distribution Limite Du Terme Maximum D'Une Série Aléatoire". In: *Annals of Mathematics* 44.3 (1943), pp. 423–453. ISSN: 0003486X.

[12] D. Prata Gomes and M. Manuela Neves. "Extremal index blocks estimator: the threshold and the block size choice". In: *Journal of Applied Statistics* 47.13-15 (2020), pp. 2846–2861. DOI: https://doi.org/10.1080/02664763.2020.1720626.

[13] M. R. Leadbetter, G. Lindgren, and H. Rootzén. *Extremes and related properties of random sequences and processes*. Springer-Verlag, New York Inc., 1983.

[14] A. Mefleh, R. Biard, and C. Dombry. "Permutation bootstrap and the block maxima method". In: *Communications in Statistics - Simulation and Computation* 50.1 (2021), pp. 295–311. DOI: https://doi.org/10.1080/03610918.2018.1563146.

[15] J. Oorschot and C. Zhou. *All Block Maxima method for estimating the extreme value index*. 2020. arXiv: 2010.15950 [math.ST].

[16] C. Y. Robert, J. Segers, and C. A.T. Ferro. "A sliding blocks estimator for the extremal index". In: *Electronic Journal of Statistics* 3 (2009), pp. 993–1020. DOI: http://dx.doi.org/10.1214/08-EJS345.

[17] C. Scarrott and A. MacDonald. "A review of extreme value threshold estimation and uncertainty quantification". In: *REVSTAT - Statistical Journal* 10.1 (2012), pp. 33–60.

[18] R. L. Smith and I. Weissman. "Estimating the Extremal Index". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 56.3 (1994), pp. 515–528.

[19] A. Tancredi, C. Anderson, and A. O'Hagan. "Accounting for threshold uncertainty in extreme value estimation". In: *Extremes* 9.2 (2006), pp. 87–106. DOI: https://doi.org/10.1007/s10687-006-0009-8.

[20] J. Wang, S. You, and Y. Wu. "A Method of Selecting the Block Size of BMM for Estimating Extreme Loads in Engineering Vehicles". In: *Mathematical Problems in Engineering* (2016), 9 pages. DOI: https://doi.org/10.1155/2016/6372197.

[21] I. Weissman and S. Y. Novak. "On blocks and runs estimators of the extremal index". In: *Journal of Statistical Planning and Inference* 66.2 (1998), pp. 281–288. ISSN: 0378-3758. DOI: https://doi.org/10.1016/S0378-3758(97)00095-5.

[22] G. Wu and G. Qiu. "Threshold Selection for POT Framework in the Extreme Vehicle Loads Analysis Based on Multiple Criteria". In: *Shock and Vibration* (2018), 9 pages. DOI: https://doi.org/10.1155/2018/4654659.

[23] X. Yang, J. Zhang, and W. X. Ren. "Threshold selection for extreme value estimation of vehicle load effect on bridges". In: *International Journal of Distributed Sensor Networks* 14.2 (2018), 12 pages. DOI: https://doi.org/10.1177%2F1550147718757698.