Extremal Index: estimation and resampling

Marta Ferreira

Centro de Matemática, Universidade do Minho

msferreira@math.uminho.pt

Abstract

The duration of extremes in time leads to a phenomenon known as clustering of high values, with a strong impact on risk assessment. The extremal index is a measure developed within Extreme Value Theory that quantifies the degree of clustering of high values. In this work we will consider the cycles estimator introduced in Ferreira & Ferreira (2018). A reduced bias estimator based on the Jackknife methodology will be presented. The bootstrap technique will also be considered in the inference and will allow to obtain confidence intervals. The performance will be analyzed based on simulation. We found our proposal effective in reducing bias and it compares favorably with some well-known methods. An application of the methods to real data will also be presented.

1 Introduction

The financial and (re)insurance industry is undergoing major challenges due to catastrophic losses increasingly demanding sophisticated risk management tools (Embrechts *et al.* [15], 1999). The climate change is amplifying the intensity and increasing the frequency of some extreme events such as heavy rains, floods, hurricanes and strong storms (Lee *et al.* [35], 2023). The theory of extreme values (EVT) allows us to infer about the possibility of observing values that are as extreme or more extreme than those ever seen. EVT thus proves to be the appropriate tool to analyze and predict catastrophic phenomena.

Consider $\{X_n^*\}_{n\geq 1}$ an independent and identically distributed (i.i.d.) sequence of random variables (r.v.) with marginal distribution function (d.f.) F. The main result of EVT states the possible limit family of d.f. for the univariate normalized maximum. We have that F belongs to the max-domain of attraction of G, if there exist real constants $a_n > 0$ and b_n, such that,

$$\lim_{n\to\infty} P(\max(X_1^*, ..., X_n^*) \le a_n x + b_n) = \lim_{n\to\infty} F^n(a_n x + b_n) = G(x)\,,$$

for all continuity points of

$$G(x) = \exp\left\{-\left(1+\xi\,\frac{x-\mu_G}{\sigma_G}\right)^{-1/\xi}\right\}\,,\;1+\xi\,\frac{x-\mu_G}{\sigma_G} > 0\,,$$

where $G(x) = \exp\{-\exp(-(x-\mu_G)/\sigma_G)\}$ if $\xi = 0$, and μ_G and σ_G are, respectively, the location and the scale parameters. The shape parameter ξ is called tail index and it characterizes the tail of the survival function: (reversed) Weibull ($\xi < 0$), Gumbel ($\xi = 0$) and Fréchet ($\xi > 0$). The (reversed) Weibull max-domain corresponds to d.f. with light-tails and finite right end-point, the Gumbel max-domain is the one of exponential-type tails and the Fréchet max-domain corresponds to the heavy-tail d.f. with infinite right end-point. Function G is usually called the Generalized Extreme Value (GEV) d.f.. There are a large number of results characterizing the max-domain of attractions. For a survey, see, e.g., de Haan and Ferreira ([10], 2006) and Beirlant *et al.* ([2], 2004). In particular, if a d.f. U belongs to the Fréchet max-domain of attraction with tail index $\xi > 0$, then $1 - U(x) = x^{-1/\xi} L_U(x)$, i.e., if $U(x)$ is heavy-tailed then $1-U(x)$ is a regularly varying function at ∞ with index $-1/\xi$, where $L_U(x)$ is a slowly varying function ($L(tx)/L(t) \to 1$, as $t \to \infty$). We can also state that $1 - U(x^{-1})$ and $L_U(x^{-1})$ are, respectively, regularly varying and slowly varying at zero.

The propensity for the occurrence of clusters of extreme values is a matter of concern, as the duration of abnormally high values over time can exacerbate the extreme phenomenon and make it more devastating (see, e.g., Faranda et al.[16], 2023). In EVT, the extremal index is a measure that assesses the degree of clustering. This parameter is usually denoted by θ, varying in the range $[0, 1]$, where, the closer to zero, the greater the propensity for the occurrence of clusters of extreme values. The null case $\theta = 0$ is abnormal (see, e.g., Leadbetter *et al.* [34] 1983) and will not be considered. The extremal index concept has already surpassed the borders of EVT, namely appearing in other areas such as Dynamical Systems (Moloney *et al.* [38] 2019, Freitas *et al.* [21] 2021, among others).

Figure 1 plots the daily maximum nitrogen dioxide (NO2) concentration in micrograms per cubic metre (ug/m^3) in the station Aotizhongxin of Beijing in China, where we can see large values occurring close together.

Along the paper, we will adopt notation $M_{i,j} = \max(X_{i+1}, .., X_j)$, $i \le j-1$, with $M_{i,j} = -\infty$ if $i > j-1$ and $M_{0,j} \equiv M_j$.

Let $\boldsymbol{X} = \{X_n\}_{n\geq 1}$ be a stationary sequence of r.v. with common marginal d.f. F. We say that $\boldsymbol{X}$ has extremal index θ if for each $\tau > 0$ there exists a sequence of normalized levels u_n, i.e., $n(1-F(u_n)) \to \tau$, as $n \to \infty$, such that $P(M_n \leq u_n) \to \exp(-\theta\tau)$. If $\theta = 1$ then the tail behavior of $\boldsymbol{X}$ resembles an i.i.d. sequence, where we have $P(M_n \leq u_n) \to \exp(-\tau)$. On the other hand, $\theta < 1$ leads to the occurrence of clusters of extreme values and thus to the existence of extremal local dependence.

In the stationary case, it is considered the dependence condition $D(u_n)$, that basically limits the long-range dependence at large values. More precisely, $\boldsymbol{X}$ satisfies condition $D(u_n)$ if for any integers $1 \leq i_1 < ... < i_q < j_1 < ... < j_{q'} \leq n$ for which $j_1 - i_q \geq l$, we have

$$\left| P\left(M_{i_1,i_q} \leq u_n, M_{j_1,j_{q'}} \leq u_n\right) - P\left(M_{i_1,i_q} \leq u_n\right) P\left(M_{j_1,j_{q'}} \leq u_n\right)\right| \leq \alpha_{n,l}, \tag{1}$$

with $\alpha_{n,l_n} \to 0$, as $n \to \infty$, for some sequence $l_n = o(n)$ and $l_n \to \infty$ (Leadbetter, [33] 1974). Therefore, for sufficiently separated time points i and j and large threshold u_n, the threshold exceedances, $X_i > u_n$ and $X_j > u_n$ are asymptotically independent.

If there exist normalizing real constants $a_n > 0$ and b_n such that $F^n(a_nx + b_n) \to G(x)$, if $\boldsymbol{X}$ satisfies $D(u_n)$ with $u_n = a_nx + b_n$ for each x such that $G(x) > 0$ and if $P(M_n \leq a_nx + b_n)$ converges for some x, then $P(M_n \leq a_nx + b_n) \to H(x) \equiv G^\theta(x)$. Now the location and scale parameters of the limiting GEV H are affected by θ. More precisely, we have $\mu_H = \mu_G + \sigma_G(\theta^\xi - 1)/\xi$ and $\sigma_H = \sigma_G\theta^\xi$. The shape parameter ξ remains untouchable meaning that H and G share the same max-domain of attraction. It is important to remark that ignoring θ may lead to incorrect tail inferences, e.g., underestimation (overestimation) of quantiles of F (H) if inference is based on H (F) from sample block-maxima (sample observations). Further details are found in, e.g., Beirlant *et al.* ([2], 2004).

Different characterizations of the extremal index inspired several estimation methods. For instance, under a suitable mixing condition, θ is the reciprocal of the limiting mean cluster size in the point process of exceedance times of a large threshold u_n, (Hsing *et al.* [30], 1988). O'Brien ([41], 1987) characterizes θ based on a conditional probability that quantifies to what extent extremes cluster together. In Ferro and Segers ([20], 2003), θ appears as a parameter of the limiting mixture d.f. of the interexceedance times of a high threshold, under a suitable mixing condition. The majority of estimators developed under these interpretations of θ are quite sensitive to the choice of the high threshold and the cluster identification. Estimation methods based on the relation between the two limiting GEV, H and G, of block-maxima exposed above involves the delicate selection of block size.

Thus, we can state these two major groups of estimators. One is based on clusters identification where we have to choose a clustering parameter and high threshold u_n, working under a suitable mixing condition. Here we include, among others, the following estimators: Nandagopalan ([39]1990), Runs and Blocks (Weissman & Novak, [53] 1998 and references there in), Intervals (only requiring the threshold; Ferro & Segers, [20] 2003), K-gaps (Süveges & Davison, [52] 2010), censored/truncated (Holěsovský & Fusek, [31, 32] 2020/22), cycles Estimator (Ferreira & Ferreira, [18] 2018). The second group is based on block maxima, where we only choose the block length for maxima, and which includes the estimators of Gomes ([24] 1993), Ancona-Navarrete & Tawn ([1] 2000), Northrop ([40] 2015) and Ferreira & Ferreira ([19] 2022).

In stationary models closely resembling an i.i.d. behavior at large values, i.e., having $\theta = 1$, usually estimators underestimate the value of θ. A discussion on this topic is promoted, e.g., in Ancona-Navarrete & Tawn ([1] 2000). Here we propose a pre-test in order to first evaluate if $\theta < 1$, since in inference it can be difficult to distinguish from the case $\theta = 1$ corresponding to the upper boundary point of the domain of θ values.

The asymptotic properties of the extremal index estimators allow approximations of the true confidence intervals for finite samples and the literature suggests that an exact confidence region may be better approximated by bootstrap resampling. See, e.g., Sebastião *et al.* ([47], 2013) and references therein.

This paper focuses on the cycles estimator of θ introduced in Ferreira & Ferreira ([18] 2018). A characterization of θ motivating this estimator is presented in Section 2. In Section 3 we introduce a reduced bias estimator based on the Jackknife methodology. This will be particularly useful to overcome the sensitivity of the cycles estimator to the threshold choice, as shall be illustrated in the simulation study of Section 5. We also consider a bootstrap version of the cycles estimator and derive confidence intervals. Under a dependent setup we must implement block bootstrap techniques in order to preserve the underlying dependence structure (see, e.g., Politis and Romano [44], 1994 and references therein). In Section 4 we present a test that will allow us to infer the case $\theta < 1$. We analyze the performance of the methods through a simulation study in Section 5. An illustration on an environmental dataset is presented in Section 6 and we conclude in Section 7.

2 Estimation of θ

The classical runs estimator is one of the most popular estimators of θ and is related to the O'Brien's ([41] 1987) limiting result,

$$\theta = \lim_{n\to\infty} P(M_{1,r_n} \le u_n | X_1 > u_n), \tag{2}$$

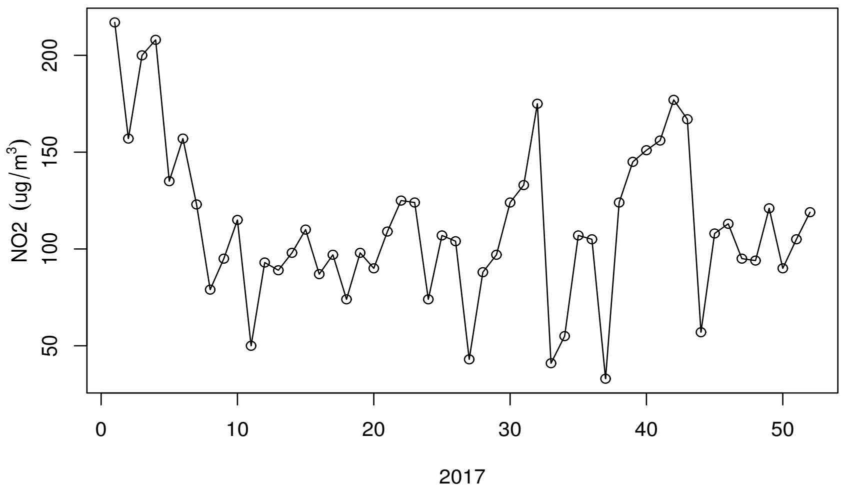


Figure 1: Daily maximum nitrogen dioxide (NO2) concentration in micrograms per cubic metre (ug/m^3) in the station Aotizhongxin of Beijing in China.

where $r_n = o(n)$. The runs estimator is just the empirical counterpart of (2):

$$\tilde{\theta}^{(R)} = \frac{\sum_{i=1}^{n-r+1} \mathbb{1}_{\{X_i > u, X_{i+1} \leq u, \ldots, X_{i+r-1} \leq u\}}}{N_u}, \tag{3}$$

where $\mathbb{1}_{\{.\}}$ is the indicator function and N_u denotes the number of exceedances of a high threshold u in the stationary sequence $\boldsymbol{X}$ (Hsing [29] 1993). We thus consider two different groups of exceedances of a threshold u as independent clusters if there are at least $r-1$ consecutive observations between them that are below u and r is the so called runs parameter.

In the work of Chernick *et al.* ([7] 1991) a similar O'Brien's result is stated, under a local mixing condition denoted $D^{(s)}(u_n)$. This condition basically states that within a cluster, an exceedance of a high threshold u_n is most likely to be followed by another exceedance within $s-1$ consecutive observations, under the validity of condition $D(u_n)$ in (1).

Formally, condition $D^{(s)}(u_n)$ holds for $\boldsymbol{X}$ if $D(u_n)$ also holds and if there exist $s > 0$ integer and sequences of integers r_n and l_n such that $r_n \to \infty$, $n\alpha_{n,l_n}/r_n \to 0$, $l_n/r_n \to 0$ and

$$\lim_{n\to\infty} nP\left(X_1 > u_n \geq M_{1,s}, M_{s,r_n} > u_n\right) = 0\,.$$

If $D^{(s)}(u_n)$ holds, then $D^{(s^*)}(u_n)$ also holds for all $s^* \geq$s.

Consider $u_n(\tau)$ such that, for $\tau > 0$,

$$\lim_{n\to\infty} nP\left(X_1 > u_n(\tau)\right) = \tau\,.$$

In Chernick *et al.* ([7] 1991, Corollary 1.3) it is established that if $\boldsymbol{X}$ satisfies $D^{(s)}(u_n)$, for some $s > 0$ and $u_n = u_n(\tau)$ for all $\tau > 0$, the extremal index θ of $\boldsymbol{X}$ exists if and only if

$$\lim_{n\to\infty} P\left(M_{1,s} \le u_n | X_1 > u_n\right) = \theta\,, \tag{4}$$

for all $\tau > 0$.

Therefore the runs estimator (3) is also an empirical counterpart of (4) if we take the runs parameter $r = s$. The Nandagopalan estimator corresponds to the case $r = 2$ and thus requires condition $D^{(2)}(u_n)$ to hold (Nandagopalan, [39] 1990). Its mathematical expression is just the ratio between the number of upcrossings (or downcrossings) and the number of exceedances of u_n. The simplicity of the Nadagopalan's estimator makes it a quite attractive estimation procedure. However it requires condition $D^{(2)}(u_n)$, which may be an unrealistic assumption in applications.

The cycles estimator presented in Ferreira and Ferreira ([18] 2018) is based on the Nadagopalan's estimator, but under the validation of a $D^{(s)}(u_n)$ for some positive s. This is not so restrictive and other estimators also demand this condition (see, e.g., Süveges & Davison, [52] 2010 and Holěsovský & Fusek, [31, 32] 2020/22). In the cycles estimator we generate the so called cycles process $\{Z_n = M_{(n-1)(s-1),n(s-1)}\}_{n\ge 1}$ from the stationary sequence $\boldsymbol{X}$ by taking the maximum of non overlapping blocks of successive r.v. X_i, $i = (n-1)(s-1)+1, ..., n(s-1)$, for each $n \ge 1$. It is proved that the stationary sequence $\{Z_n\}$ satyifies $D^{(2)}(u_n)$ (Ferreira and Ferreira, [18] 2018, Proposition 2.3). Thus the cycles estimator, denoted $\tilde{\theta}^{(C)}$, corresponds to the ratio between the number of upcrossings of threshold u within $\{Z_1, ..., Z_{[n/(s-1)]}\}$, denoted U_u^Z, and the number of exceedances N_u of $\boldsymbol{X}$, i.e.,

$$\tilde{\theta}^{(C)} = \frac{U_u^Z}{N_u}. \tag{5}$$

There are different methods in literature to check the validity of a $D^{(s)}(u_n)$ condition. The diagnostic plots of anti-$D^{(s)}(u_n)$ (Süveges, [51] 2007; Ferreira and Ferreira [18] 2018) are heuristic procedures. The proposal in Cai ([5], 2022) lies on a stability check of the runs estimator. The method presented in Fukutome *et al.* ([22, 23] 2014/2019) allows to select both s and u_n based on misspecification tests through the information matrix test (IMT) presented in Süveges and Davison ([52] 2010). The test is applied to all combinations of pairs (u, K) in admissible ranges of thresholds u and parameter K which is a runs parameter where exceedances lying closer together than K are considered to belong to the same cluster and declustering is based on choosing only the largest exceedance of each cluster. The pairs are tested for misspecification of the model, and the selected pair corresponds to the largest number of observations after declustering, within the pairs of low misspecification (IMT < 0.05) and with a number of exceedances above 80. We denote this automation procedure by Fukutome-Süveges-Davison, in short, FSD. Based on the FSD automation procedure for a given selected pair (u, K), we will assume the validity of condition $D^{(K+1)}(u)$ (see, e.g., Ferreira [17] 2018; Holěsovský & Fusek, [31, 32] 2020/22).

3 Jackknife Estimator

One of the purposes in applying the Jackknife methodology is the construction of estimators with lower bias and mean square error than those presented by the initial one.

In order to apply the Jackknife resampling technique to the cycles estimator, we will closely follow the methodology in Gomes *et al.* ([25], 2008) developed for the Nandagopalan's estimator. We recall that the cycles estimator is based on the Nandagopalan's idea and has a similar formulation, although it is more general and therefore can be applied to a greater diversity of processes.

Consider $X_{1;n} \leq X_{2:n} \leq ... \leq X_{n:n}$ the order statistics (o.s.) of the stationary sequence $\boldsymbol{X}$. Replacing the threshold u by the k-th upper o.s. in (5), we obtain the cycles estimator $\tilde{\theta}^{(C)}(k)$ as a function of the number k of o.s. larger than the threshold, i.e.,

$$\tilde{\theta}^{(C)}(k) = \frac{U_k^Z}{k}, \tag{6}$$

where U_k^Z denotes the number of upcrossings of $X_{k:n}$ within $\{Z_1, ..., Z_{[n/(s-1)]}\}$.

Gomes *et al.* (2008) assume that this type of estimators presents in its bias function mainly two dominant components of orders k/n and $1/k$. More precisely, if

$$k = k_n \to \infty,\, k = o(n), \;\text{ as } n \to \infty, \tag{7}$$

then the bias is given by

$$d_1(\theta)(k/n) + d_2(\theta)(1/k) + o(1/k) + o(k/n).$$

The same authors verified this assumption for i.i.d. and first order max auto-regressive Fréchet models satisfying the $D^{(2)}(u_n)$ condition. In accordance with the Generalized Jackknife (GJ) methodology, given the existence of two main terms of bias, we must consider three estimators of θ having the same type of bias. Let us consider $\widehat{\theta_n}^{(1)}$, $\widehat{\theta_n}^{(2)}$ and $\widehat{\theta_n}^{(3)}$, such that

$$E(\widehat{\theta_n}^{(i)} - \theta) = d_1(\theta)\phi_1^{(i)}(n) + d_2(\theta)\phi_2^{(i)}(n)\,,\; i = 1, 2, 3\,,$$

the second order GJ statistics is derived by $\widehat{\theta_n}^{(GJ)} := det(m)/det(m_1)$, where det denotes the determinant of a matrix and matrices m and m_1 are given by

$$m = \begin{bmatrix} \widehat{\theta_n}^{(1)} & \widehat{\theta_n}^{(2)} & \widehat{\theta_n}^{(3)} \\ \phi_1^{(1)} & \phi_1^{(2)} & \phi_1^{(3)} \\ \phi_2^{(1)} & \phi_2^{(2)} & \phi_2^{(3)} \end{bmatrix} \quad \text{and} \quad m_1 = \begin{bmatrix} 1 & 1 & 1 \\ \phi_1^{(1)} & \phi_1^{(2)} & \phi_1^{(3)} \\ \phi_2^{(1)} & \phi_2^{(2)} & \phi_2^{(3)} \end{bmatrix}.$$

It is proved that a GJ statistic obtained as described is unbiased (Gray and Schucany, [27] 1972).

Considering levels k, $\lfloor \delta k \rfloor + 1$ and $\lfloor \delta^2 k \rfloor + 1$, with $\lfloor x \rfloor$ denoting the integer part of x and δ a tuning parameter in $(0, 1)$, according to Gomes *et al.* (2008) and after some calculations, we obtain the reduced-bias GJ cycles estimator:

$$\tilde{\theta}^{(CGJ)}(k) = \frac{(\delta^2 + 1)\tilde{\theta}^{(C)}(\lfloor \delta k \rfloor + 1) - \delta\left(\tilde{\theta}^{(C)}(\lfloor \delta^2 k \rfloor + 1) + \tilde{\theta}^{(C)}(k)\right)}{(1-\delta)^2}. \tag{8}$$

4 Asymptotic tail independence

In i.i.d. sequences we have $\theta = 1$. However, a unit extremal index doesn't imply independence. Stationary Gaussian autorregressive processes are typical examples where $\theta = 1$, despite their time dependence. Observe that $\theta = 1$ corresponds to a boundary value in the domain of θ. This is usually a critical issue of the estimators performance of θ, as already discussed in Ancona-Navarrete ([1], 2000). Simulations of Section 5 will also illustrate this feature. In order to overcome this drawback, we propose to analyze in advance if $\theta < 1$, at least, in the cases were estimates of θ are close to one. In the following we describe our method.

In asymptotically tail independent sequences, the degree of dependence between exceedances of x decreases as x approaches x^F, and the sequence is becoming similar to an i.i.d. one at extremal levels. In order to account this feature, Ledford and Tawn ([36, 37], 1996/1997) proposed the following model:

$$P(U_2 > 1 - t | U_1 > 1 - t) = t^{1/\eta - 1} L(t), \text{as}, \; t \downarrow 0, \tag{9}$$

where $U_i = F(X_i))$, $i = 1, 2$, are standard uniform (transformed) marginals, $\eta \in (0, 1]$ and $L(t)$ is a slowly varying function at 0. We denote model (9) by L&T. If $\eta = 1$ and $L(t)$ converges to some positive constant c, then $P(U_2 > 1-t | U_1 > 1-t) \to c > 0$ leading to asymptotic tail dependence associated with $\theta < 1$. On the other hand, if $0 < \eta < 1$ or $L(t) \to 0$ as $t \downarrow 0$, we are in the asymptotic tail independent case linked to $\theta = 1$. Moreover, independence, positive association and negative association correspond, respectively, to $\eta = 1/2$ and $L(t) = 1$, $1/2 < \eta < 1$ and $0 < \eta < 1/2$. In (9) we are analyzing tail (in)dependence at lag-1, but we can also consider lag-m, for m positive integer, i.e.,

$$P(U_{1+m} > 1-t | U_1 > 1-t) = t^{1/\eta_m - 1} L_m(t), \text{as}, \; t \downarrow 0, \tag{10}$$

with $\eta_m \in (0,1]$ and $L_m(t)$ is a slowly varying function at 0 and where $\eta \equiv \eta_1$ and $L(t) \equiv L_1(t)$.

Observe that, if $\boldsymbol{X}$ satisfies $D^{(s)}(u_n)$, for some $s > 0$ and has extremal index θ, then $\theta < 1$ under lag-m tail dependence for some $m \in \{1, ..., s-1\}$, given (4).

There are different proposals to estimate η, for example, in the works of Draisma *et al.* ([13], 2004) and Chiapino *et al.* ([6], 2019), and references therein. However they all assume independence between random pairs, which is not our case. Therefore, we pursue the estimation of η in another direction.

In the L&T model, η is always positive and can be read as the tail index of r.v. $W = \min(U_1, U_2)$, since

$$P(W > 1-t) = P(U_2 > 1-t, U_1 > 1-t) = P(U_2 > 1-t | U_1 > 1-t) P(U_1 > 1-t) = t^{1/\eta} L(t^{-1}).$$

The tail index estimation is largely addressed in the literature of EVT. Some well-known estimators are the Hill ([28], 1975), Moments ([12], 1989), maximum likelihood (ML) of Smith ([48], 1987), Pickands ([43], 1975), among others. They are defined from the k upper order statistics, where $k \equiv k_n$ is an intermediate sequence as in (7). For instance, the Hill estimator is given by,

$$\widehat{\eta}_{k,n} := \frac{1}{k} \sum_{i=1}^{k} \log W_{n-i+1:n} - \log W_{n-k:n}, \tag{11}$$

where $W_{1:n} \leq ... \leq W_{n:n}$ are the o.s. of $W_i = \min(U_{i-1}, U_i)$, $i = 1, ..., n$. Observe that, for each value of k, i.e., for each chosen threshold $W_{n-k:n}$, we obtain an estimate of the tail index. In practice, k must be chosen not too high (large variance) neither too short (large bias). Asymptotic properties of a class of tail index estimators under a dependence framework, which include Hill, Moments, ML and Pickands estimators, were deduced in Drees ([14], 2003). In particular, in this later work we can find consistent estimators, $\widehat{\sigma^2}$, of the asymptotic variance of the tail index estimators, which can be used to test the hypotheses $H_0 : \eta = 1$ vs. $H_1 : \eta < 1$. If we do not reject H_0, i.e., if $\hat{\eta} > 1 + z_{0.05}\widehat{\sigma}$, where $z_{0.05}$ is the 5% quantile of the standard Gaussian d.f., then we accept $\theta < 1$. On the other hand, if we reject H_0, we proceed on testing $H_0 : \eta_m = 1$ vs. $H_1 : \eta_m < 1$ until we do not reject H_0 provided $m \leq s-1$, leading to the acceptance of $\theta < 1$. Otherwise, we cannot conclude about the value of θ.

5 Simulation study

In this section we use simulation to analyze the performance of the cycles estimator (6) in comparison with its reduced-bias GJ version in (8), where we take $\delta = 1/4,\ 1/2,\ 3/4$. We also compare with the well-known estimators of Ferro and Segers ([20], 2003) and Süveges and Davison ([52], 2010). We apply the block-bootstrap method to the cycles estimator in order to derive confidence intervals (CI). We recall that in the context of time series, we must apply bootstrap resampling based on blocks to maintain the process dependence structure. This is an important issue and several contributions on this topic are found in literature. For a survey, see, e.g., Gomes and Neves ([26], 2015) and references therein. To this end we need to choose an adequate block-length in order to perform block resampling. We consider the automatic block-length selection method in Politis and White ([45], 2004) and Paton *et al.* ([42], 2009) implemented in software R ([46], 2020), in package *blocklength* ([50], 2022). We derive bootstrap CI corresponding to the normal, the basic and the percent methods presented in Davison and Hinkley ([9], 1997; Chapter 5) and available in R package *boot*. We also apply the FSD method described in Section 2 to validate condition $D^{(s)}(u_n)$ for some positive integer s, as required by the cycles estimator.

The models below are used in simulations:

- 1st order max auto-regressive (MAR), $X_i = \max(\phi X_{i-1}, \epsilon_i)$, $i \geq 1$, $X_0 = \epsilon_1/(1-\phi)$, $\{\epsilon_i\}$ i.i.d. with standard Fréchet marginals and auto-regressive parameter $\phi = 0.5$ (Davis and Resnick [8], 1989), for which $D^{(2)}(u_n)$ holds and $\theta = 0.5$;

- moving maxima $X_i = \max_{j=0,\ldots,d} a_j Z_{i-j}$ with $\{Z_i\}$ i.i.d. standard Fréchet (MMFrec) (Deheuvels [11], 1983), where $d = 8$ and parameters $\alpha_i = 1/5$, $i \in \{1, 2, 3, 7, 8\}$, and $\alpha_i = 0$ otherwise, for which $D^{(5)}(u_n)$ holds and $\theta = 0.2$ (Ferreira and Ferreira [18] 2018);

- Markov chain (MC) with standard Gumbel marginals and bivariate logistic dependence, $P(X_i \leq x, X_{i+1} \leq y) = \exp(-(x^{-1/\alpha} + y^{-1/\alpha})^{\alpha})$, with $\alpha = 0.5$, having $\theta = 0.328$ (Smith [49], 1992);

- ARCH(1) process, $X_i = (\beta + \alpha X_{i-1}^2)^{1/2}\epsilon_i$, with i.i.d. Gaussian innovations $\{\epsilon_i\}$, parameters $\alpha = 0.7$ and $\beta = 2 \cdot 10^{-5}$, for which $\theta = 0.721$ (Cai, [5] 2022);

- AR(1) with Cauchy standard marginals (ARCau), $X_i = \rho X_{i-1} + \epsilon_i$, $\{\epsilon_i\}$ i.i.d. having Cauchy d.f. with mean 0 and scale $1 - |\rho|$ with $\rho = -0.6$, satisfying condition $D^{(3)}(u_n)$ and $\theta = 0.64$ (Chernick *et al.* [7], 1991);

- AR(1) process, $X_i = \phi X_{i-1} + \epsilon_i$, $i \geq 1$, $\{\epsilon_i\}$ i.i.d. $N(0,1)$, $X_0 \frown N(0, 1/(1-\phi^2))$, with parameter $\phi = 0.5$, where $D^{(1)}(u_n)$ holds and $\theta = 1$;

- an i.i.d. sequence of Fréchet r.v. and thus $\theta = 1$.

At our knowledge, there is no theoretical analysis on condition $D^{(s)}(u_n)$ for the MC model (an empirical essay was conducted in Ferreira and Ferreira [18] 2018). Cai ([5] 2022) proved that $D^{(s)}(u_n)$ condition doesn't hold for any s in the ARCH model given above.

Our simulation study is based on 1000 replicas of each given model with size $n = 1000$. We computed the absolute bias (abias) and the root mean squared error (rmse). The respective plots are in Figures 2 and 3. The AR model results are in Figure 4.

We can see that the GJ estimator reduces the bias and presents a more stable sample path of estimates in a large number of thresholds, when compared to the cycles estimator (full line), more incisively if $\delta = 1/4$. Looking at the rmse, the GJ estimator still provides lower values in several cases, particularly within $\delta = 1/4$, yet not so expressive as in the bias. We can also note that the cases of $\theta = 1$, i.e., the model AR with Gaussian marginals and the i.i.d. Fréchet sequence, do not behave so well in estimates. The cycles estimator shows an increasingly sharp bias and rmse, as well as the estimator of Süveges and Davison. In the GJ method, both bias and rmse do not grow as sharply but still remain relevant. Exception made for the i.i.d. Fréchet model where the GJ estimator with $\delta = 1/4$ and the Ferro and Segers one perform both well. We remark that $\theta = 1$ corresponds to a boundary value of θ's domain, which causes difficulties for inference in general. So, one possibility is to infer in advance if $\theta < 1$. We apply the test procedure described in Section 4, based on the L&T asymptotic tail (in)dependence parameter η. In Table 1 we present the acceptance percentage of $\theta < 1$, where in estimating η we consider the 50^{th} upper o.s. (i.e., $k = 50$ in (11)).

Table 1: Acceptance (%) of $\theta < 1$.

MAR	MMFrec	MC	ARCH	ARCau	AR	FrecInd
94	1	97.9	64.1	0	23.9	0

The results are quite good for models MAR, MMFrec, MC and FrecInd. In the AR and ARCH models they are also not bad, although they can be improved, e.g., using alternative estimators to Hill or analyzing more adequate choices of the threshold involved in the estimation. For a survey on the tail index estimation, see, e.g., Beirlant *et al.* ([3], 2012) and references therein.

In ARCau model, by its defining formula we have that a very large observation will be followed by a very small one and vice versa, typical of negatively associated consecutive pairs. So if $X_j > x$, then the values $X_{j+2}, X_{j+4}, ...$ are most likely to exceed x, whilst $X_{j+1}, X_{j+3}, ...$ are expected to fall below. As remarked in Bortot and Tawn ([4], 1998), most consecutive pairs will have only one large component which compromises the estimation of model (9). Indeed, we have obtained $\widehat{\eta} = 0.367$ and 95% CI $[0.265, 0.468]$ and thus, well far away from the expected value 1. The same authors suggest to model the lag-2 probabilities, i.e., by taking $m = 2$ in (10). According to our proposal in Section 4, and since ARCau model satisfies condition $D^{(3)}(u_n)$, we proceed on testing $H_0: \eta_m = 1$ vs. $H_1: \eta_m < 1$, for $m = 2$. Thus choosing the lag-2 in ARCau, we now obtain $\widehat{\eta}_2 = 0.865$ with 95% CI $[0.591, 1]$ and the acceptance of $\theta < 1$ is 82.6%, which are quite better results.

The bootstrap methodology can also be an alternative approach to conduct inference on the extremal index, namely, in the construction of CI that better estimate θ in the case of assuming a value close to or equal to 1. In order to evaluate this issue, we have also applied the block-bootstrap method as described above. The absolute bias and rmse of bootstrap estimates of the cycles estimator, denoted cycles-bootstrap, are plotted in Figure 5. Here we considered the threshold $X_{n-k:n}$ with $k = 100$. We can see that the cycles-bootstrap estimator decreases the bias, except in the AR model, but in the rmse this only happens with models ARCau, MMFrec, MAR and AR. For the estimated bootstrap 95% CI, we computed the proportion of intervals in the simulations that included the true value of θ (coverage) and the coverage divided by the mean range width (coverage/range). See Figure 6. Observe that the percentile CI has an overall better performance when compared to the Normal CI and basic CI. Observe also that none of the three CI considered (Normal, basic and percentile) can capture the value $\theta = 1$ in the AR model. They also present a very low coverage in the independent FrecInd process where θ is 1 too. In the ARCH model the coverage does not reach 50%, which also reveals little accuracy in the case of a high value of θ. These results reinforce the strategy proposed in this work of previously evaluating if $\theta < 1$.

6 Application

The analyzed data corresponds to the daily maximum of nitrogen dioxide (NO2) concentration in micrograms per cubic metre (ug/m^3) in the station Aotizhongxin from the Beijing Municipal Environmental Monitoring Center in China, in the period between March 1st 2013 and February 28th 2017. The data is available from `https://archive.ics.uci.edu/ml/index.php`. The observations are plotted in Figure 7, where successive high values occur. See also the year 2017 represented in Figure (1).

We start by analyzing if $\theta < 1$. We recall the testing procedure described in Section 4 based on the L&T model, more precisely, based on coefficient η estimated by (11). The estimates are plotted in Figure 8, using $1 < k < n$ upper o.s. (full line) along with the upper and lower 95% confidence bands (dotted lines). We can see that $\eta = 1$ is a plausible value for a large number of upper o.s. (approximately $k = 270$), and thus we infer that $\theta < 1$. We now move on to the estimation of the extremal index. The estimates can be seen in Figure 9. The black, red, blue and green lines are the cycles, GJ, Süveges and Davison and Ferro and Segers estimates, respectively. The bootstrap 95% CI correspond to the blue dashed lines (percentile), the orange dotted lines (Normal) and the brown dot-dashed lines (basic). Our guess is that possible values for θ range around 0.4, although the Süveges and Davison estimator points towards 0.5.

7 Conclusion

In extreme values inference of time series, the extremal index θ assumes a relevant role. If omitted it can lead to erroneous extrapolations, as is the case of extreme quantiles above the observed maximum. In addition, it is also associated with the clustering effect of extreme values, measuring the tendency of data for such behavior. Data series with a greater propensity for the appearance of clusters of high values exhibit a duration of extremes in time, which can worsen the damage, in the case, for example, of financial losses or natural catastrophes (see, e.g., Cai [5] 2022 and references therein). The importance of the extremal index is also attested by the various scientific contributions that appear in the literature (Ancona-Navarrete and Tawn [1] 2000, Beirlant *et al.* [2] 2004, Ferreira and Ferreira [18] 2018, Ferro and Segers [20] 2003, Nandagopalan [39] 1990, Northrop [40], Süveges and Davidson [52] 2010, among others). This work is based on an existing estimator, on which Jackknife and bootstrap resampling techniques were applied for bias reduction and interval estimation. Both methods contributed to a noticeable reduction in bias and some decrease in rmse was also recorded. However, the critical situation of estimating the boundary case $\theta = 1$ of its domain was not properly resolved, even with the use of bootstrap interval estimation. A test was proposed to evaluate if it is plausible to have $\theta < 1$ with promising results. This is a topic that deserves a deeper analysis and that we intend to improve. Given the potential revealed by the resampling methods and the progressive computational advances, refinements of the described methods and new applications extended to other estimators will follow.

Disclosure Statement

The authors report there are no competing interests to declare.

Acknowledgement

The author thanks the reviewers for their relevant comments and valuable suggestions that contributed to improving this work. The research of the author as partially financed by Portuguese Funds through FCT (Fundação para a Ciência e a Tecnologia) within the Projects UIDB/00013/2020 and UIDP/00013/2020.

References

[1] Ancona-Navarrete, M.A., Tawn, J.A. (2000). A comparison of methods for estimating the extremal index. Extremes 3, 5–38.

[2] Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J. (2004). Statistics of Extremes: Theory and Applications. Wiley, Chichester.

[3] Beirlant, J., Caeiro, F., Gomes , M.I. (2012). An Overview and Open Research Topics in Statistics of Univariate Extremes. REVSTAT-Statistical Journal 10(1), 1–31.

[4] Bortot, P., Tawn, J.A. (1998). Models for the extremes of Markov chains. Biometrika 85(4), 851–867.

[5] Cai, J.J. (2022). Statistical inference on $D^{(d)}(u_n)$ condition and estimation of the Extremal Index. arXiv:1911.06674

[6] Chiapino, M., Sabourin, A., Segers, J. (2019). Identifying groups of variables with the potential of being large simultaneously. Extremes 22, 193–222.

[7] Chernick M.R.. Hsing, T., McCormick, W.P. (1991). Calculating the extremal index for a class of stationary sequences. Advances in Applied Probability 23, 835–850.

[8] Davis, R., Resnick, S. (1989). Basic properties and prediction of max-ARMA processes. Advances in Applied Probability 21, 781–803.

[9] Davison, A.C., Hinkley, D.V. (1997). Bootstrap Methods and Their Application. Cambridge University Press

[10] de Haan, L., Ferreira, A. (2006). Extreme Value Theory: An Introduction. Springer, New York.

[11] Deheuvels, P. (1983). Point processes and multivariate extreme values. Journal of Multivariate Analysis 13, 257–272.

[12] Dekkers, A.L.M., Einmahl, J.H.J., de Haan, L. (1989). A moment estimator for the index of an extreme-value distribution. Annals of Statististics 17, 1833–1855.

[13] Draisma, G., Drees, H., Ferreira, A., de Haan, L. (2004). Bivariate tail estimation: dependence in asymptotic independence. Bernoulli 10, 251–280.

[14] Drees, H.. (2003). Extreme quantile estimation for dependent data, with applications to finance. Bernoulli 9(4), 617–657.

[15] Embrechts, P., Resnick, S., Samorodnitsky, G. 1999. Extreme Value Theory as a Risk Management Tool. North American Actuarial Journal 3(2), 30–41.

[16] Faranda, D., Messori, G., Yiou, P., Thao, S., Pons, F., Dubrulle, B. (2023). Dynamical footprints of hurricanes in the tropical dynamics. Chaos 33(1), 013101.

[17] Ferreira, M. (2018). Analysis of estimation methods for the extremal index. Electronic Journal of Applied Statistical Analysis 11(1), 296–306

[18] Ferreira, H., Ferreira, M. (2018). Estimating the extremal index through local dependence. Annales de l'Institut Henri Poincaré - Probabilités et Statistiques 54(2), 587–605.

[19] Ferreira, H., Ferreira, M. (2022). A new blocks estimator for the extremal index. Communications in Statistics - Theory and Methods. (in press)

[20] Ferro, C.A.T., Segers, J. (2003). Inference for clusters of extreme values. Journal of the Royal Statistical Society: Series B 65, 545–556.

[21] Freitas, A.C.M., Freitas, J.M., Soares, J.V. (2021). Rare events for product fractal sets. Journal of Physics A: Mathematical and Theoretical 54 (34), 34520.

[22] Fukutome. S.. Liniger M.A.. Süveges M. (2014). Automatic threshold and run parameter selection: a climatology for extreme hourly precipitation in Switzerland. Theoretical and Applied Climatology 120(3). 403–416.

[23] Fukutome, S., Liniger, M.A., Süveges, M. (2019). Correction to: Automatic threshold and run parameter selection: a climatology for extreme hourly precipitation in Switzerland. Theoretical and Applied Climatology 137(3-4), 3215

[24] Gomes, M. (1993). On the estimation of parameters of rare events in environmental time series. In V. Barnett and K. Turkman (Eds.), Statistics for the environment 2: Water Related Issues, pp. 225–241, Wiley.

[25] Gomes, M.I., Hall, A., Miranda, M.C. (2008). Subsampling techniques and the Jackknife methodology in the estimation of the extremal index. Comput. Stat. Data Anal. 52(4), 2022–2041.

[26] Gomes, D.P., Neves, M.M. (2015). Bootstrap and Other Resampling Methodologies in Statistics of Extremes. Communications in Statistics - Simulation and Computation 44(10), 2592–2607.

[27] Gray, H.L., Schucany, W.R. (1972). The Generalized Jackknife Statistic. Marcel Dekker.

[28] Hill, B. (1975). A simple general approach to inference about the tail of a distribution. Ann. Statist. 3, 1163–1174.

[29] Hsing, T. (1993). Extremal index estimation for a weakly dependent stationary sequence. Annals of Statistics 21, 2043–2071.

[30] Hsing, T., Hüsler, J., Leadbetter, M.R. (1988). On the exceedance point process for a stationary sequence. Probability Theory and Related Fields 78. 97–112.

[31] Holěsovský, J., Fusek, M. (2020). Estimation of the extremal index using censored distributions. Extremes 23(2), 197–213.

[32] Holěsovský, J., Fusek, M. (2022). Improved interexceedance-times-based estimator of the extremal index using truncated distribution. Extremes (in press).

[33] Leadbetter, M.R. (1974). On extreme values in stationary sequences. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete. 28(4), 289–303.

[34] Leadbetter, M.R., Lindgren, G., Rootzén, H. (1983). Extremes and Related Properties of Random Sequences and Processes. New York: Springer.

[35] Lee, H., K. Calvin, D. Dasgupta, G. Krinner, A. Mukherji, P. Thorne, C. Trisos, J. Romero, P. Aldunce, K. Barrett, G. Blanco, W.W.L. Cheung, S.L. Connors, F. Denton, A. Diongue-Niang, D. Dodman, M. Garschagen, O. Geden, B. Hayward, C. Jones, F. Jotzo, T. Krug, R. Lasco, J.-Y. Lee, V. Masson-Delmotte, M. Meinshausen, K. Mintenbeck, A. Mokssit, F.E.L. Otto, M. Pathak, A. Pirani, E. Poloczanska, H.-O. Pörtner, A. Revi, D.C. Roberts, J. Roy, A.C. Ruane, J. Skea, P.R. Shukla, R. Slade, A. Slangen, Y. Sokona, A.A. Sörensson, M. Tignor, D. van Vuuren, Y.-M. Wei, H. Winkler, P. Zhai, Z. Zommers (2023). Synthesis Report of the IPCC Sixth Assessment Report (AR6): Summary for Policymakers. Intergovernmental Panel on Climate Change, accepted.

[36] Ledford, A.W., Tawn, J.A. (1996). Statistics for near independence in multivariate extreme values. Biometrika 83, 169–187.

[37] Ledford, A.W., Tawn, J.A. (1997). Modelling dependence within joint tail regions. J. R. Statist. Soc. B 59, 475–99.

[38] Moloney, N.R., Faranda, D., Sato, Y. (2019). An overview of the extremal index. Chaos 29, 022101.

[39] Nandagopalan S. (1990). Multivariate extremes and estimation of the extremal index. Ph.D. Thesis. University of Nth Carolina. Chapel Hill.

[40] Northrop. P.J. (2015). An efficient semiparametric maxima estimator of the extremal index. Extremes 18(4). 585–603.

[41] O'Brien, G.L. (1987). Extreme values for stationary and markov sequences. The Annals of Probability 15, 281–291.

[42] Patton, A., Politis, D.N., White, H. (2009). Correction to "Automatic Block-Length Selection for the Dependent Bootstrap" by D. Politis and H. White. Econometric Review 28(4), 372–375.

[43] Pickands III, J. (1975). Statistical inference using extreme order statistics. Ann. Statist. 3, 119–131.

[44] Politis, D.N., Romano, J.P. (1994). The stationary bootstrap. Journal of the American Statistical Association 89, 1303–1313.

[45] Politis, D.N., White, H. (2004). Automatic Block-Length Selection for the Dependent Bootstrap. Econometric Reviews 23(1), 53–70.

[46] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[47] Sebastião J.R., Martins A.P., Ferreira H., Pereira L. (2013). Estimating the upcrossings index. TEST 22(4). 549–579.

[48] Smith, R.L. (1987). Estimating tails of probability distributions. Ann. Statist. 15, 1174–1207.

[49] Smith. R. (1992). The Extremal Index for a Markov Chain. Journal of Applied Probability 29(1), 37–45.

[50] Stashevsky, A. (2022). blocklength: Select an Optimal Block-Length to Bootstrap Dependent Data (Block Bootstrap). R package version 0.1.5. https://CRAN.R-project.org/package=blocklength

[51] Süveges, M. (2007). Likelihood estimation of the extremal index. Extremes 10, 41–55.

[52] Süveges, M., Davison A.C. (2010). Model misspecification in peaks over threshold analysis. Annals of Applied Statistics 4. 203–221.

[53] Weissman, I., Novak, S.Y. (1998). On blocks and runs estimators of the extremal index. Journal of Statistical Planning and Inference 66(2), 281–288.

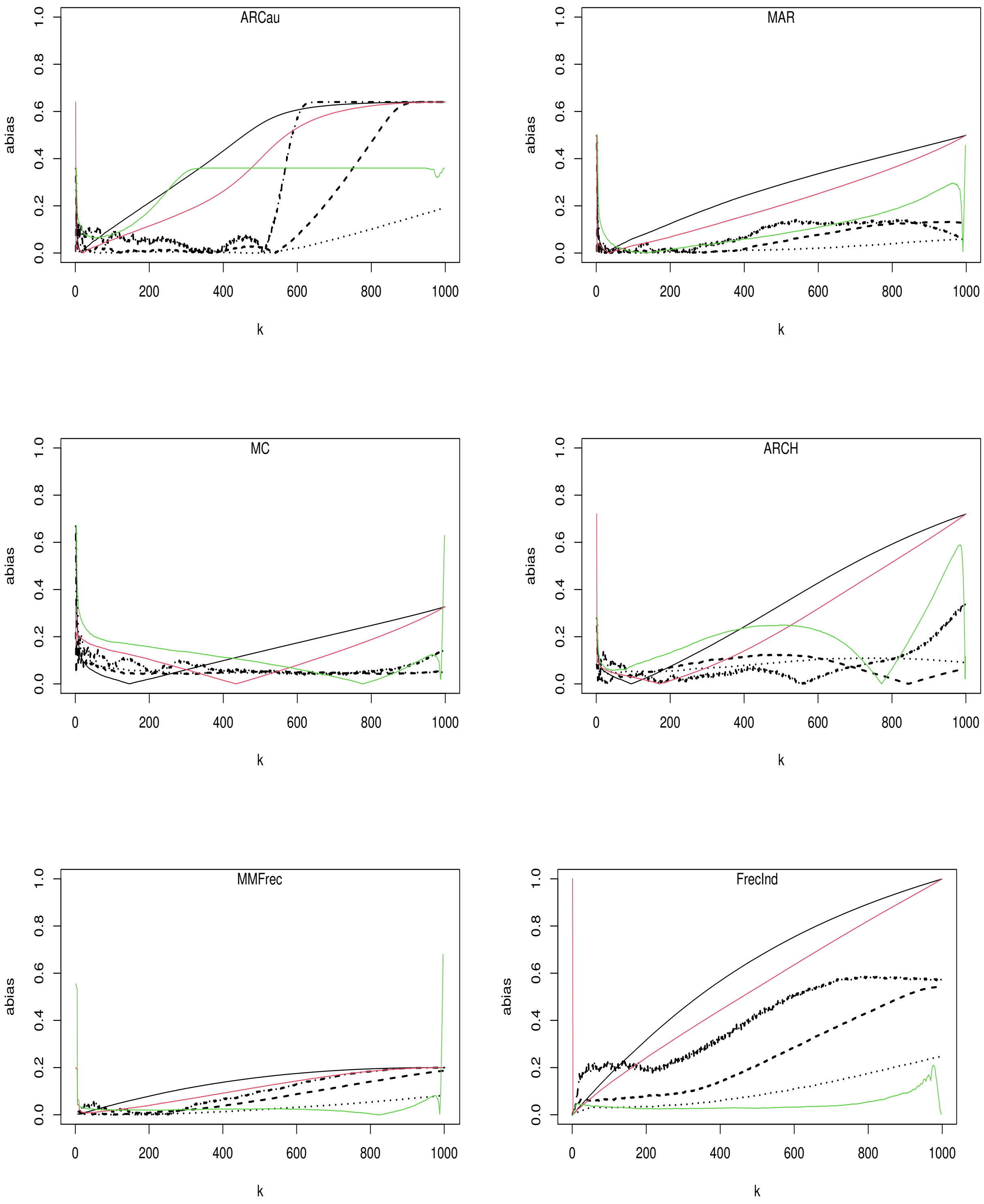


Figure 2: Absolute bias (abias) of the cycles estimator $\tilde{\theta}^{(C)}(k)$ in (6) (full line), the GJ cycles estimator $\tilde{\theta}^{(CGJ)}(k)$ in (8) with $\delta = 1/4$ (dotted line), $\delta = 1/2$ (dashed line), $\delta = 3/4$ (dashed-dotted line), the Ferro and Segers estimator (full green line) and the Süveges and Davison estimator (full red line), $1 < k < n$, for models: max auto-regressive (MAR), moving maxima (MMFrec), Markov chain (MC), ARCH(1), auto-regressive Cauchy (ARCau) and i.i.d. Fréchet (FrecInd).

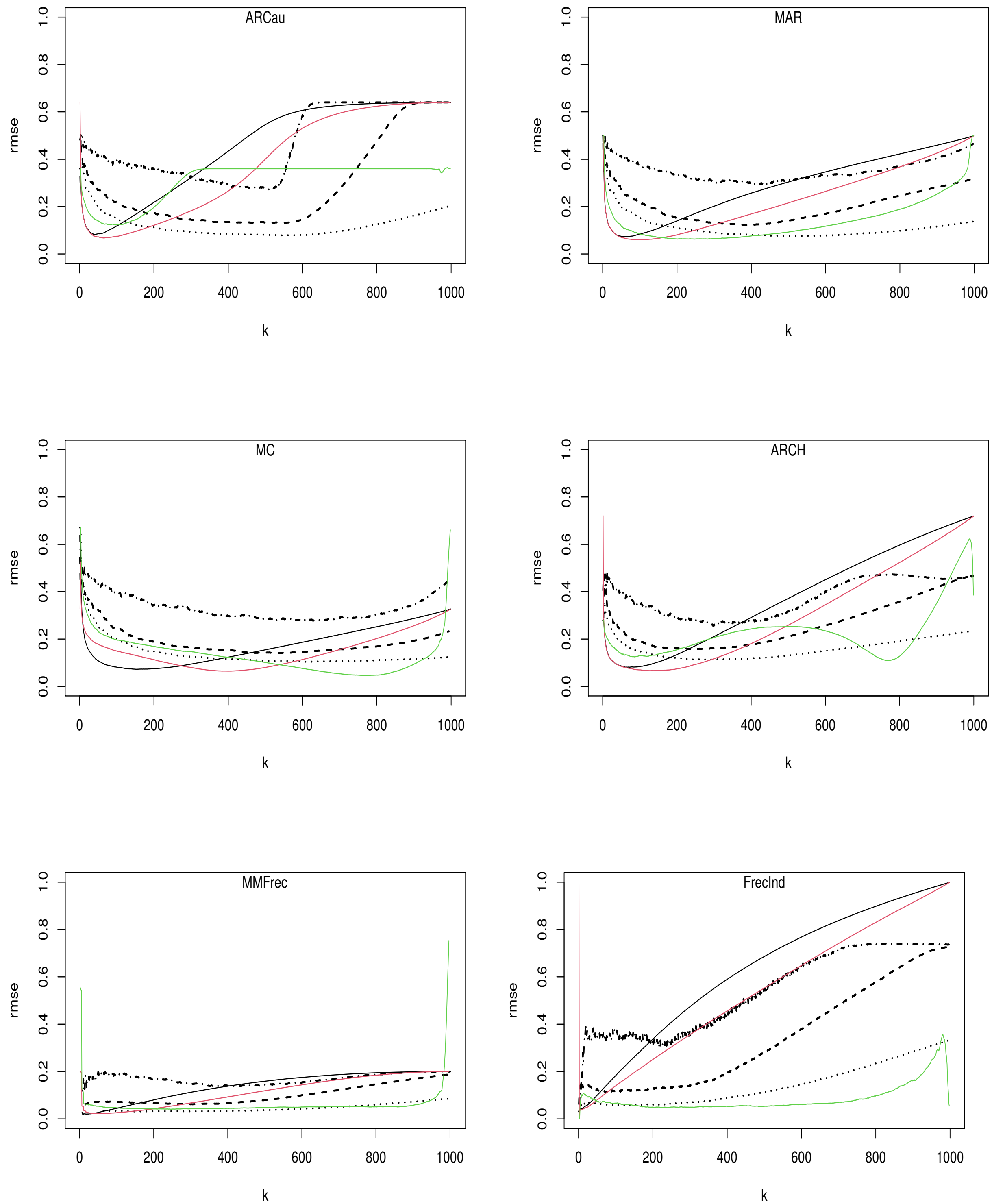


Figure 3: Root mean squared error (rmse) of the cycles estimator $\tilde{\theta}^{(C)}(k)$ in (6) (full line), the GJ cycles estimator $\tilde{\theta}^{(CGJ)}(k)$ in (8) with $\delta = 1/4$ (dotted line), $\delta = 1/2$ (dashed line), $\delta = 3/4$ (dashed-dotted line), the Ferro and Segers estimator (full green line) and the Süveges and Davison estimator (full red line), $1 < k < n$, for models: max auto-regressive (MAR), moving maxima (MMFrec), Markov chain (MC), ARCH(1), auto-regressive Cauchy (ARCau) and i.i.d. Fréchet (FrecInd).

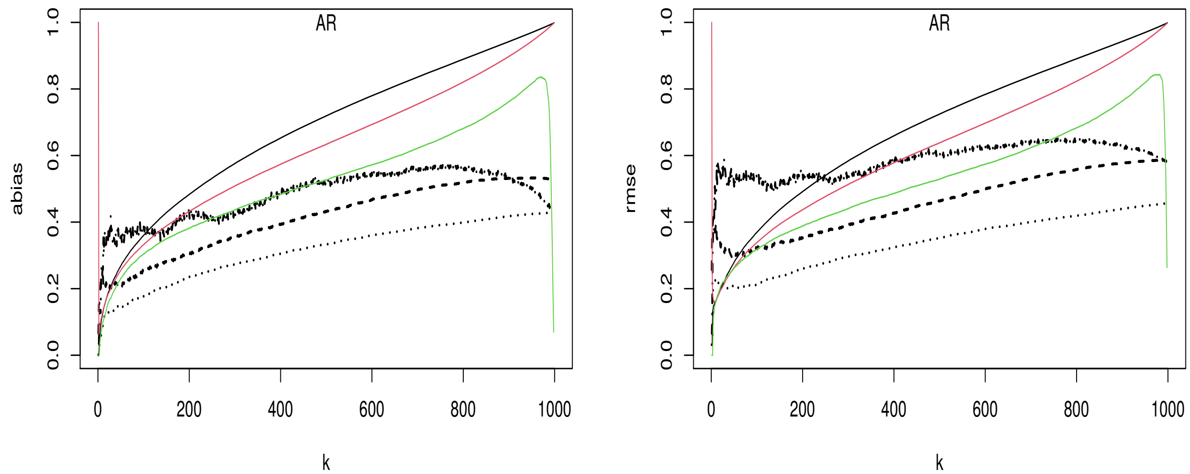


Figure 4: Absolute bias (left) and root mean squared error (right) of the cycles estimator $\tilde{\theta}^{(C)}(k)$ in (6) (full line), the GJ cycles estimator $\tilde{\theta}^{(CGJ)}(k)$ in (8) with $\delta = 1/4$ (dotted line), $\delta = 1/2$ (dashed line), $\delta = 3/4$ (dashed-dotted line), the Ferro and Segers estimator (full green line) and the Süveges and Davison estimator (full red line), for $1 < k < n$ in the AR(1) Gaussian model.

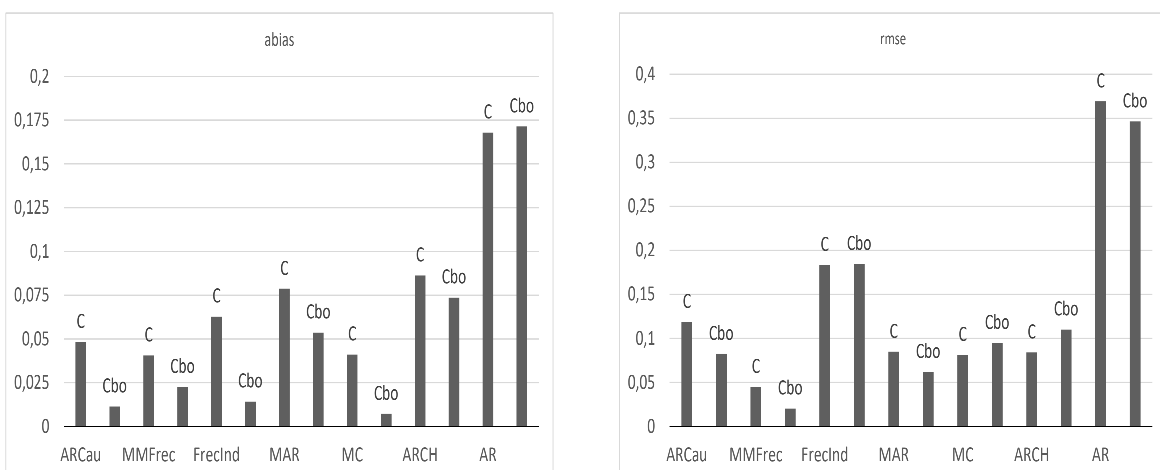
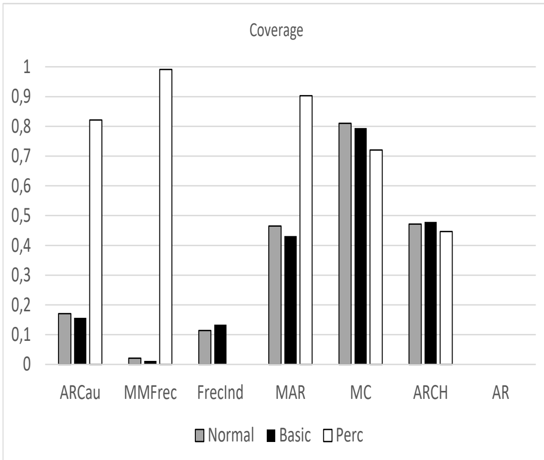


Figure 5: Absolute bias (left) and root mean squared error (right) of the cycles (C) and cycles-bootstrap (Cbo) estimators, using threshold $X_{n-k:n}$ with $k = 100$, for models: max auto-regressive (MAR), moving maxima (MMFrec), Markov chain (MC), ARCH(1), auto-regressive Cauchy (ARCau), auto-regressive Gaussian (AR) and i.i.d. Fréchet (FrecInd).



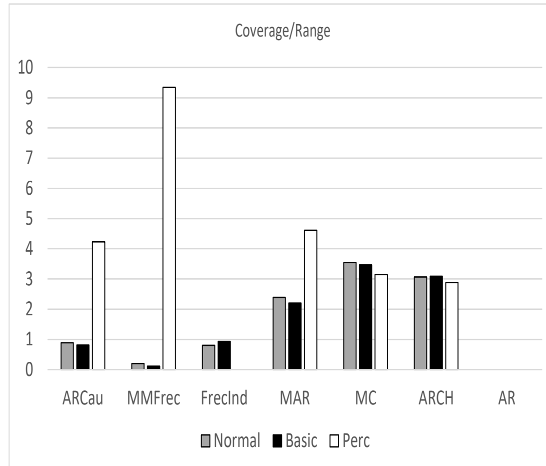


Figure 6: Fraction of the estimated bootstrap 95% CI that included the true value of θ (coverage) on the left and the rate coverage over the mean intervals range width (coverage/range) on the right, for models: max auto-regressive (MAR), moving maxima (MMFrec), Markov chain (MC), ARCH(1), auto-regressive Cauchy (ARCau), auto-regressive Gaussian (AR) and i.i.d. Fréchet (FrecInd).

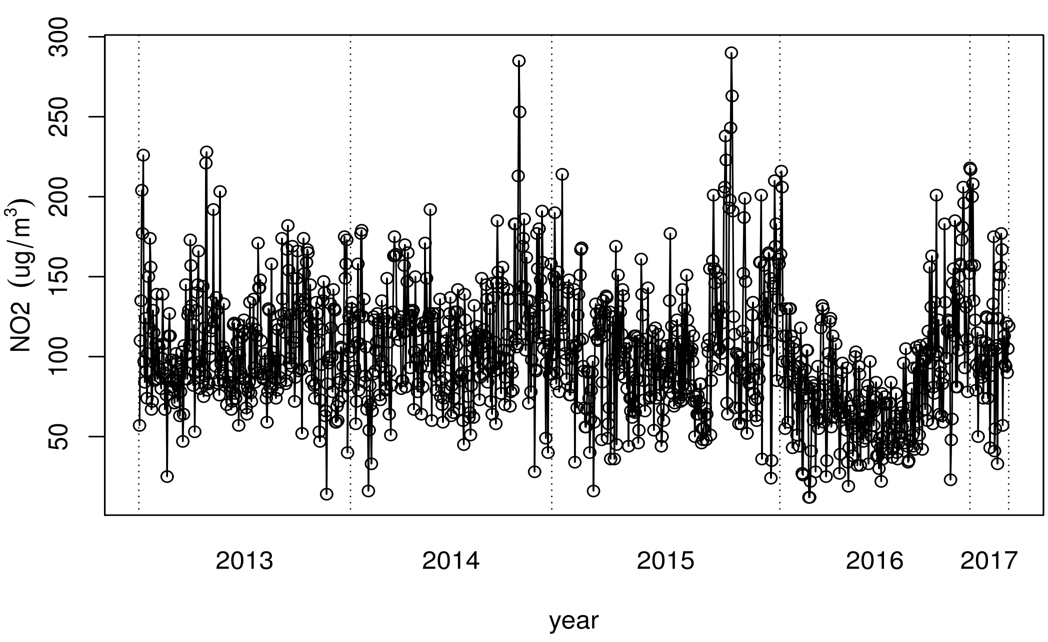


Figure 7: Daily maximum of nitrogen dioxide (NO2) air concentration (in ug/m^3) from March 2013 to February 2017 at station Aotizhongxin (Beijing-China). The vertical dotted bars separate the years.

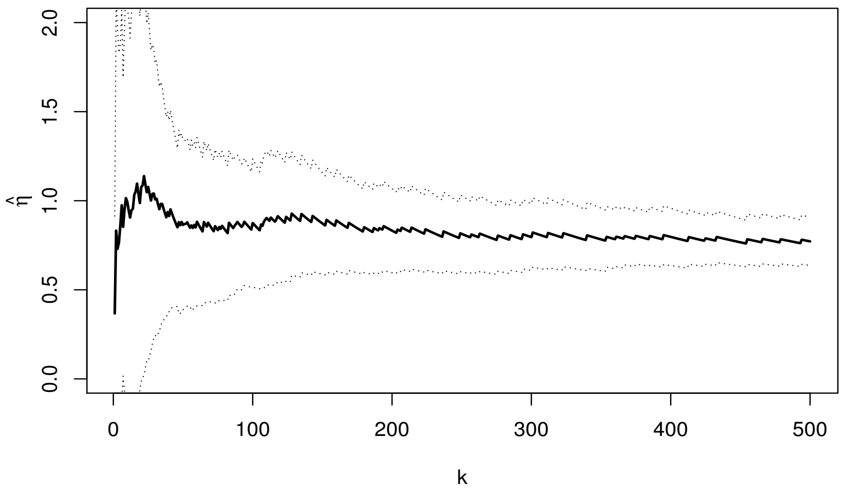


Figure 8: Estimates of η in daily maximum NO2 air concentration data, obtained with estimator (11), using $1 < k < n$ upper o.s. (full line). The upper and lower dotted lines are, respectively, the upper and lower 95% confidence limits.

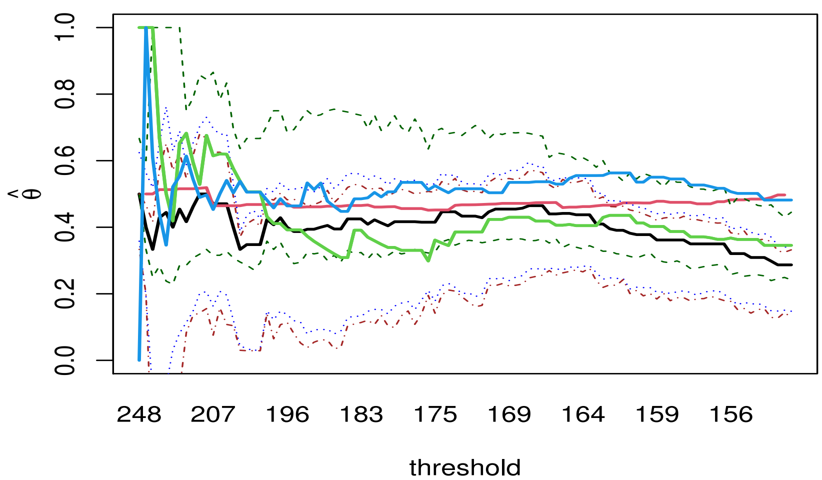


Figure 9: Estimates of θ in daily maximum NO2 air concentration data: cycles (full black line), GJ (full red line), Ferro and Segers (full green line) and Süveges and Davison (full blue line). The bootstrap 95% CI correspond to the green dashed lines (percentile), the blue dotted lines (Normal) and the brown dot-dashed lines (basic).