AIP Conference Proceedings

Comparing two block estimation procedures for the extremal index: An application

Cite as: AIP Conference Proceedings **2425**, 320004 (2022); https://doi.org/10.1063/5.0081320
Published Online: 06 April 2022

Dora Prata Gomes and Manuela Neves

View Online Export Citation

ARTICLES YOU MAY BE INTERESTED IN

Preface of the "Methods of Nonlinear Mathematical Physics"
AIP Conference Proceedings **2425**, 340001 (2022); https://doi.org/10.1063/5.0081616

Lock-in Amplifiers up to 600 MHz
Zurich Instruments
Watch

AIP Publishing

© 2022 Author(s).

Comparing Two Block Estimation Procedures for the Extremal Index: An Application

Dora Prata Gomes[1,2,a),b)] and Manuela Neves[3,4,c)]

[1] *Centro de Matemática e Aplicações (CMA), FCT, Universidade Nova de Lisboa. Portugal.*
[2] *Departamento de Matemática, FCT, Universidade Nova de Lisboa. Portugal.*
[3] *Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL). Portugal.*
[4] *Instituto Superior de Agronomia, Universidade de Lisboa. Portugal.*

[a)]Corresponding author: dsrp@fct.unl.pt
[b)]dsrp@fct.unl.pt
[c)]manela@isa.ulisboa.pt

Abstract. When extending the analysis of the limiting behaviour of the extreme values from independent and identically distributed sequences to stationary sequences a key parameter appears, the extremal index θ, whose accurate estimation is not easy and is not completely solved. Here we focus on the estimation of θ using blocks estimators, that can be constructed by using disjoint or sliding blocks. Both blocks construction require the choice of a threshold and a block length. The main objective of this work is to revisit another block estimation procedure that only depends on the block length, although some conditions on the underlying process need to be verified. An application will be presented for illustrating the proposed procedure.

INTRODUCTION

In many real situations a pronounced temporal clustering of the extreme values can be seen, indicating the presence of local dependence in the extremes. This motivate a search for reliable tools to describe these features because quantifying the nature of the dependence structure as well as the duration of extreme events becomes an essential part of the understanding of these time series data. The *extremal index* (EI) is the main parameter that describes and quantifies the clustering characteristics of the extreme values in many stationary time series. Its formal definition is given next.

Definition 1 ([1]) Suppose that $\{X_n\}_{n\geq 1}$ is a strictly stationary sequence of random variables with marginal distribution function (d.f.) F. This sequence is said to have an EI $\theta \in [0,1]$ if, for each $\tau > 0$, there exists a sequence of levels $u_n \equiv u_n(\tau)$, such that

$$n\Big(1 - F(u_n(\tau))\Big) \to \tau \qquad \text{and} \qquad \mathbb{P}\Big\{M_{1,n} \leq u_n(\tau)\Big\} \to \exp(-\theta\tau), \tag{1}$$

as $n \to \infty$ where $M_{1,n} = \max\{X_1, \ldots, X_n\}$.

An informal interpretation of θ is given in [1], namely θ being approximately the reciprocal of the mean cluster size. The *extremal index* takes values in the interval $[0,1]$. . A value close to 0 indicates a very strong short range extremal dependence, while a value close to 1 a rather weak dependence. The case $\theta = 0$ appears in pathological situations. For almost all cases of interest we have $\theta > 0$, the situation here considered.

Dependence in stationary sequences can take different forms, and it is impossible to develop a general characterization of the behavior of extremes unless some constraints are imposed. It is usual to assume a condition that limits the extend of long-range dependence at extreme levels, so that the events $X_i > u$ and $X_j > u$ are approximately independent, provided that u is high enough, and time points i and j have a large separation. This condition is denominated $D(u_n)$ *condition*, see [1].

This paper will be useful for the analysis of blocks estimation procedures for θ. To compare the two blocks estimation procedures, we first apply both to a stationary model. Conditions under which the second procedure holds are also verified. An application to a daily mean flow discharge rate time series and some comments are also presented.

International Conference of Numerical Analysis and Applied Mathematics ICNAAM 2020
AIP Conf. Proc. 2425, 320004-1–320004-4; https://doi.org/10.1063/5.0081320
Published by AIP Publishing. 978-0-7354-4182-8/$30.00

CLASSICAL BLOCKS ESTIMATOR *versus* ANOTHER APPROACH

The concept of *extremal index* given by interpreting θ^{-1} as the limiting mean cluster size of the exceedances yields the blocks method. This method consists of partitioning the n observations into consecutive $k_n = [n/r_n]$ contiguous blocks of a certain length, $r_n = o(n)$. In each block, the number of exceedances over a certain high threshold u_n are counted, and the blocks estimator is then defined as the reciprocal of the average number of exceedances per block among blocks with at least one exceedance. The blocks estimator, in [2], is given by

$$\widehat{\theta}_n^B(u_n) := \frac{\sum_{j=1}^{k_n} I\left(M_{(i-1)r_n, ir_n} > u_n\right)}{\sum_{i=1}^{n} I\left(X_i > u_n\right)}. \tag{2}$$

Blocks estimators can be constructed considering continuous blocks or sliding blocks. However, for both procedures the blocks estimator requires the choice of a threshold, u_n, and a block size, r_n. But, the behaviour of the estimates depend strongly of r_n and u_n. Some recent works trying to deal with that situation can be mentioned, such as [3, 4, 5, 6]. Let us consider a Max-Autoregressive Process model ([7]) to illustrate how the estimates depend on r_n and u_n. Let $\{Y_n\}_{n\geq 1}$ be a sequence of independent, standard Gumbel distributed random variables. For fixed α define

$$X_n = \max\left\{X_{n-1} - \alpha, Y_n + \log\left(1 - \exp(-\alpha)\right)\right\}, \qquad n \geq 1. \tag{3}$$

The EI of this process is given by $\theta = 1 - \exp(-\alpha)$, see [7]. Given the sample $(X_1, \ldots, X_n)$ and the associated ascending order statistics, $X_{1:n} \leq \ldots \leq X_{n:n}$, we shall consider the level u_n, in (2) substituted by the stochastic one, $X_{n-k:n}$.

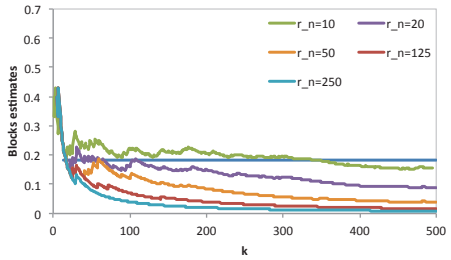

FIGURE 1. Estimates of $\widehat{\theta}_n^B$ plotted against k ($u_n = X_{n-k:n}$), of a sequence of length $n = 1000$, with block lengths $r_n = 10, 20, 50, 125, 250$, for the Max-Autoregressive Process with $\alpha = 0.2 \equiv (\theta = 0.1813)$

It seems difficult to decide what r_n should be chosen, there is some block size for which the path estimates do not cross the true value of the parameter. On the other hand the region of *extremal index* estimates that shows some stability around the true value of the parameter depends on r_n and even for a given r_n, it is not obvious how to choose the threshold appropriately.

In this section, we introduce another method, see [8], for estimating θ that not depends on threshold choice because the threshold is defined inside each block. The validity of this estimator can be extended to dependent processes satisfying the long-range approximate independence condition of [9], called $D(u_n)$ *condition*, and the $D^2(u_n)$ *condition* of [10].

Definition 2 ([1]) The $D(u_n)$ *condition* holds for a stationary sequence if for every integers p, q and $i_1 < i_2 < \cdots < i_p < j_i < j_2 < \cdots < j_q < n$ such that $j_1 - i_p > \ell \equiv \ell_n$, we have

$$\left|F_{i_1,i_2,\cdots,i_p,j_1,j_2,\cdots,j_q}(u_n, u_n, \cdots, u_n) - F_{i_1,i_2,\cdots,i_p}(u_n, u_n, \cdots, u_n)F_{j_1 j_2,\cdots,j_q}(u_n, u_n, \cdots, u_n)\right| \leq \alpha_{n,\ell}, \tag{4}$$

where $\lim_{n\to\infty} \alpha_{n,\ell_n} = 0$ for some sequence $\{\ell_n = o(n)\}$.

Definition 3 ([10]) Let $\{X_n\}_{n\geq 1}$ be a stationary sequence of random variables. $D^2(u_n)$ *condition* is said to be satisfied if

$$n\mathbb{P}\left\{X_j > u_n, X_{j+1} \leq u_n, M_{j+2,r_n} > u_n\right\} \to 0 \quad \text{as} \quad n \to \infty, \tag{5}$$

with u_n verifying the $D(u_n)$ *condition* and a sequence r_n of block sizes such that $n/r_n \to \infty$ and $r_n = o(n)$.

The proposed estimator was defined in the following way: let k_n denote the number of blocks, and r_n the respective block size. Let v_{ni} be a sequence of levels such that $r_n\mathbb{P}\{X_1 \leq v_{ni} < X_2\} \to 1$ as $n \to \infty$. Denoting $N_i(r_n, v_{ni})$ as the number of up-crossing of v_{ni} in *ith* block, the estimator is defined by

$$\widetilde{\theta}_n^B(r_n) := \frac{k_n}{\sum_{i=1}^{k_n} N_i(r_n, v_{ni})}. \tag{6}$$

The estimator in (6) depends on the validity of $D^{(2)}(u_n)$ *condition*, that can be checked by calculating the proportion of the anti-$D^{(2)}(u_n)$ events $\{X_{j+1} \leq u_n, M_{j+2,r_n} > u_n | X_j > u_n\}$ among the exceedances for a range of thresholds and block sizes, given u_n and r_n. By [11], the proportion of the anti-$D^{(2)}(u_n)$ is calculated by

$$p(u_n, r_n) = \frac{\sum_{j=1}^{n} I(X_j > u_n, X_{j+1} \leq u_n, M_{j+2,r_n} > u_n)}{\sum_{j=1}^{n} I(X_j > u_n)}, \tag{7}$$

for the observed sequence $\{X_1, \ldots, X_n\}$.

Under the validity of the $D^{(2)}(u_n)$ *condition*, it seems reasonable to substitute v_{ni}, in each block, by adequate levels such that the number of up-crossings is equal to 1, but low enough to identify exceedances. More precisely, we can define, $V_{ni} = \inf\{u : N_i(r_n, u) = 1\}$ with $i = 1, \ldots, k_n$.

For Max-Autoregressive Process with several values of θ ($\theta = 0.1813,\ 0.5034,\ 0.9093$), a sample of size $n = 1000$ was generated and the estimator in (6) was applied. Figure 2 display the proportion of the anti-$D^{(2)}(u_n)$ events and the estimates obtained for several values of block length with ($\alpha = 0.2,\ 0.7,\ 2.4$) $\equiv$ ($\theta = 0.1813,\ 0.5034,\ 0.9093$).

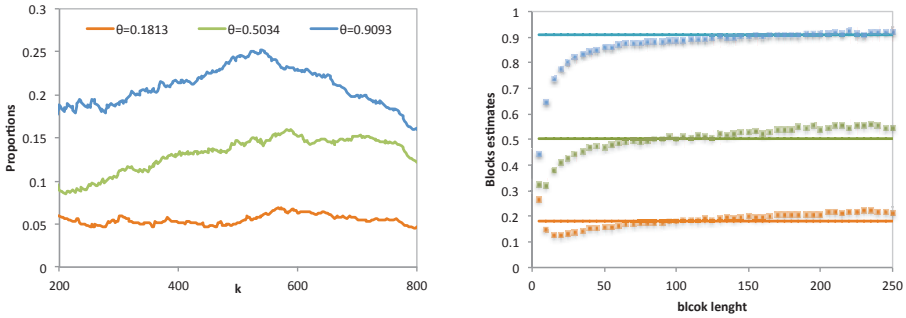

FIGURE 2. Observed proportions of $p(u_n, r_n)$ plotted against k ($u_n = X_{n-k:n}$) (left) and estimates of $\widetilde{\theta}_n^B$ plotted against different block lengths (right), from the Max-Autoregressive Process with several values of α ($\alpha = 0.2,\ 0.7,\ 2.4$) which correspond to values of θ ($\theta = 0.1813,\ 0.5034,\ 0.9093$), respectively, and $r_n = 100$.

As we can see, the procedure presents very good results, a large stability region, very close to the true value of the parameter θ. The observed proportions of $p(u_n, r_n)$ depend on the θ value, showing higher values for high values of θ and small values for small values of θ.

In [12] we applied a path stability algorithm, see [13] and [14] now adapted to the choice of r_n and to obtain a θ estimate who conducted to quite nice results for extreme value parameters estimation.

CASE STUDY

We conclude with an application of the blocks estimator and the other blocks procedure aforementioned to a time series of daily mean flow discharge rate (m^3/s) from 1 October, 1946 to 30 April, 2012 ("SNIRH: Sistema Nacional de Informação dos Recursos Hídricos"). The stationarity of the data can be assumed from November until April ($n = 11947$). We shall estimate the extremal index. In Figure 3, we depict estimates of the extremal index as a function of the block length parameter, ranging from $r_n = 120$ to $r_n = 1500$. We also checked the proportion of the anti-$D^{(2)}(u_n)$ events in our data and we verify very low proportion of anti-$D^{(2)}(u_n)$ values. The difficulties of choosing the block size, r_n, as well as of choosing the adequate level k, are clear on the left plot of this figure.

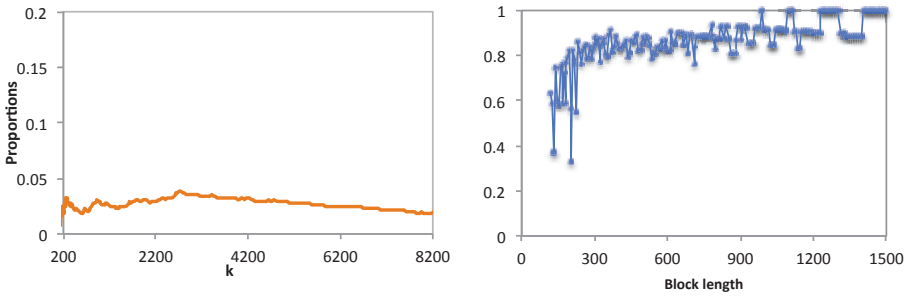

FIGURE 3. Observed proportions of $p(u_n, r_n)$ with $r_n = 100$ and estimates of $\widetilde{\theta}_n^{\beta}$ plotted against block length (right), for the daily mean flow discharge rate values.

A FEW COMMENTS

The estimation of the *extremal index* governs the clustering of the extremes of a univariate observational series. This work apply block estimation procedures to estimate this parameter, one of which may not depend of threshold choice. The comparison of the two above procedures needs some more research, mainly regarding to the $D^2(u_n)$ *condition*.

ACKNOWLEDGMENTS

This work was partially supported by the Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through the projects UIDB/00297/2020 (Centro de Matemática e Aplicações) and UIDB/00006/2020 (Centro de Estatística e Aplicações/UL).

REFERENCES

[1] M. R. Leadbetter, G. Lindgren, and H. Rootzén, *Extremes and related properties of random sequences and processes* (Springer, 1983).

[2] R. L. Smith and I. Weissman, Estimating the Extremal Index, *Journal of the Royal Statistical Society. Series B (Methodological)* **56**, 515–528 (1994).

[3] B. Berghaus and A. Bücher, Weak convergence of a pseudo maximum likelihood estimator for the extremal index, *Annals of Statistics* **46**, 2307–2335 (2018).

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

[5] H. Drees, Bias correction for estimators of the extremal index, 2011, arXiv:1107.0935 [stat.ME] .

[6] P. Northrop, An efficient semiparametric maxima estimator of the extremal index, *Extremes* **18**, 585–603 (2015).

[7] S. Richard L., The Extremal Index for a Markov Chain, *Journal of Applied Probability* **29**, 37–45 (1992).

[8] L. Canto e Castro, "Estudo de um método de estimação do indice extremal," in *I Congresso Ibero-Americano de Estadística e Investigación Operativa* (1992).

[9] C. A. T. Ferro and J. Segers, Inference for Clusters of Extreme Values, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **65**, 545–556 (2003).

[10] M. R. Leadbetter and S. Nandagopalan, "On exceedance point processes for stationary sequences under mild oscillation restrictions," in *Extreme Value Theory*, edited by J. Hüsler and R.-D. Reiss (Springer New York, New York, NY, 1989), pp. 69–80.

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

[12] D. P. Gomes and M. M. Neves, Extremal index blocks estimator: the threshold and the block size choice, Journal of Applied Statistics **47**, 2846–2861 (2020).

[13] F. Caeiro and M. I. Gomes, "Threshold selection in extreme value analysis," in *Extreme Value Modeling and Risk Analysis* (Chapman and Hall/CRC 2007, 2016), pp. 69 – 86.

[14] M. M. Neves, M. I. Gomes, F. Figueiredo, and D. Prata Gomes, Modeling Extreme Events: Sample Fraction Adaptive Choice in Parameter Estimation, *Journal of Statistical Theory and Practice* **9**, 184–199 (2015).