Measuring extremal clustering in time series

Marta Ferreira

Centro de Matemática, Universidade do Minho,
Portugal
msferreira@math.uminho.pt

Abstract. The propensity of data to cluster at extreme values is important for risk assessment. For example, heavy rains that last over time lead to catastrophic floods. The extremal index is a measure of Extreme Values Theory that allows measuring the degree of high values clustering in a time series. Inference about the extremal index requires the prior choice of values for tuning parameters which impacts the efficiency of existing estimators. In this work we propose an algorithm that avoids these constraints. Performance will be evaluated based on simulation. We also illustrate with real data.

Keywords: extreme values theory, stationary sequences, extremal index

1 Introduction

The occurrence of extreme values can lead to risky situations. Climate change and the global economic and financial crisis resulting from the COVID19 pandemic situation and the war in Ukraine have contributed to a continuous growing attention from analysts, namely, to the risk of extreme phenomena. The duration of extreme values in time means the generation of clusters, whose extension can aggravate the phenomenon. Extreme Values Theory (EVT) presents a set of adequate tools in this context. The extremal index is a measure of serial dependence assessing the propensity of data for the occurrence of clusters of extreme values. Figure 1 shows the maximum of sea-surge heights, where clusters of high values are visible.

More precisely, considering $\mathbf{X} = \{X_n\}_{n\geq 1}$ a stationary sequence of random variables (r.v.) with common marginal distribution function (d.f.) F and denoting $M_n = \max(X_1, ..., X_n)$, then $\mathbf{X}$ has extremal index $\theta \in (0, 1]$ if for each real $\tau > 0$ there exists a sequence of normalized levels u_n, i.e., satisfying $n(1 - F(u_n)) \to \tau$, as $n \to \infty$, such that $P(M_n \leq u_n) \to \exp(-\theta\tau)$. In the independent and identically distributed (i.i.d.) case, we have $P(M_n \leq u_n) \to \exp(-\tau)$ and thus $\theta = 1$. On the other hand, if $\theta = 1$ then the tail behavior of $\mathbf{X}$ resembles an i.i.d. sequence. Clustering of extreme values takes place whenever $\theta < 1$ and the smaller the θ the larger the propensity for clusters to appear. Under some dependence conditions, θ is stated as the arithmetic inverse of the mean cluster size (Hsing *et al.* [15] 1988).

Assuming F continuous, we have $U_i = F(X_i)$, $i = 1, ..., n$ standard uniform r.v. and $P(-n\log(F(M_n)) \geq \tau) \approx P(n(1 - F(M_n)) \geq \tau) = P(M_n \leq$

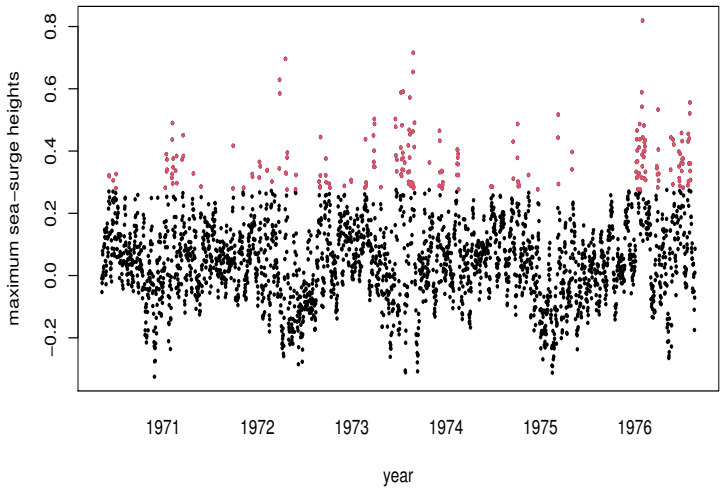

Fig. 1. Hourly maximum sea-surge heights in years 1971-1976 at the Newlyn coast, Cornwall, UK.

$u_n) \to \exp(-\theta\tau)$, with $F(M_n) = \max(U_1, ..., U_n)$. Thus, $Y_n = -n\log(F(M_n))$ and $Z_n = n(1 - F(M_n))$ follow asymptotically an exponential distribution with parameter θ. The maximum likelihood estimator was considered in Northrop ([19] 2015) based on Y_n. More precisely, dividing the time series, $X_1, ..., X_n$, into k_n blocks of length b_n, with $n = b_n k_n$ and considering $M_{ni} = M_{((i-1)b_n+1):(ib_n)} = \max(X_{(i-1)b_n+1}, ..., X_{ib_n})$, $i = 1, ..., k_n$, the maximum of the i-th block in the disjoint blocks case and $M_{ni} = M_{((i-1)):(i+b_n-1)} = \max(X_{i-1}, ..., X_{i+b_n-1})$, $i = 1, ..., n - b_n + 1$ the maximum of the i-th block in the sliding blocks case, the Northrop estimator is given by

$$\widetilde{\theta}^N = \left(\frac{1}{t_n} \sum_{i=1}^{t_n} \hat{Y}_{ni} \right)^{-1}, \tag{1}$$

where $\hat{Y}_{ni} = -b_n \log(\hat{F}(M_{ni}))$ and $\hat{F}$ denotes the empirical d.f. estimating the usually unknown F, with $t_n = k_n$ or $t_n = n - b_n + 1$, whether we are using disjoint or sliding blocks, respectively. In Berghaus and Bücher ([2] 2018) it was considered

$$\widetilde{\theta}^B = \left(\frac{1}{t_n} \sum_{i=1}^{t_n} \hat{Z}_{ni} \right)^{-1}, \tag{2}$$

with $Z_{ni} = b_n(1 - \hat{F}(M_{ni}))$, a more amenable formulation to derive the asymptotic properties. Here we consider the Berghaus and Bücher estimator with bias

adjustment given by

$$\widehat{\theta} = \widetilde{\theta}^{B} - 1/b_n. \tag{3}$$

We also consider the sliding blocks version since it usually performs better (Northrop [19] 2015, Berghaus and Bücher [2] 2018).

Observe that the estimators above only depend on a tuning parameter: the block length $b \equiv b_n$. This is an advantage of these methods since most estimators of θ presented in literature have two sources of uncertainty and thus two parameters to be defined in advance: the clustering generation of high values and the choice of a high threshold above which the clusters occur. To mention the best known ones, there is the Nandagopalan ([17] 1990), Runs and Blocks (Weissman & Novak, [22] 1998 and references there in), K-gaps (Süveges & Davison, [16] 2010), censored/truncated (Holěsovský & Fusek, [13, 14] 2020/22), cycles Estimator (Ferreira & Ferreira, [7] 2018). We also refer other estimators that require a single tuning parameter, like the Intervals estimator which needs to fix a high threshold Ferro & Segers, [9] 2003), and similar to the Northop estimator above, where we only choose the block length for maxima, we cite Gomes ([11] 1993), Ancona-Navarrete & Tawn ([1] 2000) and Ferreira & Ferreira ([8] 2022).

As already highlighted in the literature, there is no simple optimal methodology for the best choice of block length and a single estimate for θ. In EVT we have a typical bias-variance trade-off observed in sample paths estimates of rare events parameters. For blocks estimators, the bias decreases with b while the variance increases. A recurrent method is to plot the estimates obtained for successive block size values and visually identify case-by-case plateau zones of these estimates. The stability around a value is an indicator of a reasonable estimate, and this stability region, in general, should not be neither in too small nor in too large values of b, due to the trade-off between bias and variance already mentioned. In Figure 2, it is plotted a trajectory of estimates (full line) along with 95% confidence intervals (CI) (dashed line) obtained for each block length b from 1 to 100, in a random sample of dimension 1000 generated from a moving maximum model with standard Fréchet margins. We can see a plateau region in the estimates around the true value (horizontal line) $\theta = 0.5$ for the block sizes between 25 and 45. Observe the large variability occurring for large values of b and the higher bias for small values of b.

Some methods have been proposed in the literature to help in the choice of tuning parameters based on the stability regions of the estimates graph. See, e.g., Frahm *et al.* ([10] 2005), Gomes & Neves ([12] 2020) and their references. In particular, the algorithm proposed in Frahm *et al.* ([10] 2005) was implemented in the context of estimating the bivariate tail dependence and in Ferreira ([6] 2018) it was applied to extremal index estimators requiring the choice of a high threshold. In this work, our objective is to propose an adaptation of the algorithm developed in Frahm *et al.* ([10] 2005) applied to estimator (3) in order to find a suitable plateau of estimates taking into account the bias-variance trade-off. As a by product, this will allow us to circumvent the unique tuning parameter selection corresponding to the block size of where the sequence of maximums

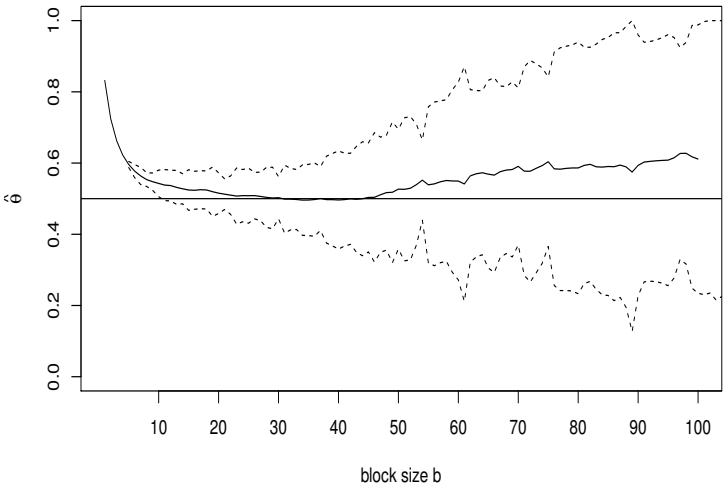

Fig. 2. Estimates of $\widehat{\theta}$ given in (3) for successive values of block size $b = 1, ..., 100$ (full line) obtained for a sample simulated from a moving maxima Fréchet model with $\theta = 0.5$ (horizontal line). The dashed lines correspond to 95% CI.

will be extracted, as described above. The method will be detailed in Section 2 and analyzed through simulation in Section 3. We end with an application to real data.

2 Estimation method

Our proposal estimation of θ is based on the bias corrected estimator $\widehat{\theta}$ in (3) by considering sliding blocks and on the heuristic plateau-finding algorithm of Frahm *et al.* ([10] 2005).

The algorithm is described in the following steps:

Step 1. Calculate estimates $\widehat{\theta}_b$ from estimator (3), for $1 \leq b \leq t < n$;

Step 2. Smooth the results of the previous step by taking means of $2w+1$ successive estimates; we have considered bandwidth $w = \lfloor 0.02t \rfloor$;

Step 3. Define plateaus of length $m = \lfloor \sqrt{t - 2w} \rfloor$, i.e., $p_j = \left(\overline{\widehat{\theta}}_j, ..., \overline{\widehat{\theta}}_{j+m-1}\right)$, $j = 1, ..., t - 2w - m + 1$;

Step 4. Compute the standard deviation s of $\overline{\widehat{\theta}}_1, ..., \overline{\widehat{\theta}}_{t-2w}$ and choose the first plateau p_j satisfying $\sum_{i=j+1}^{j+m-1} \left|\overline{\widehat{\theta}}_i - \overline{\widehat{\theta}}_j\right| \leq 2s$;

Step 5. The extremal index is estimated through $\frac{1}{m}\sum_{i=1}^{m}\overline{\widehat{\theta}}_{j+i-1}$, i.e., taking the average of the estimates that constitute the plateau chosen in the previous step. This will be denoted plateau estimator.

The estimators (1), (2) and (3) are already implemented in package *exdex* of software R (Northrop & Christodoulides [20] 2019) with respective CI. We use package *exdex* to compute estimator (3) under sliding blocks and respective upper and lower 95% CI bounds. We also apply steps 1, 2 and 3 to the lower and upper bounds of CI. Once the plateau of *theta* estimates is chosen in step 4, we pick the corresponding plateau in the CI limits and in step 5 we apply the average to the plateau values of the lower limit of the CI as well as the average of the plateau values of the upper limit of the CI.

We are going to analyze the estimation method described above through simulation. The models that will be used are the following:

- 1st order auto-regressive model with Cauchy standard marginals (ARC), $X_i = \rho X_{i-1} + \epsilon_i$, $\{\epsilon_i\}$ i.i.d. having Cauchy d.f. with mean 0 and scale $1-|\rho|$ and $\theta = 1-\rho$ if $\rho > 0$ (Chernick *et al.* [4], 1991); we consider $\rho = 0.9$ and $\theta = 0.1$;
- m-dependent model (MMU), $X_i = \max(U_i, U_{i+1}, ..., U_{i+m-1})$, $i \geq 1$, where $\{U_i\}$ is an i.i.d. sequence of r.v. (Newell [18] 1964), with $\theta = 1/m$; we consider U_i, $i \geq 1$, standard uniform r.v., $m = 3$ and thus $\theta = 1/3$;
- moving maxima Fréchet model (MMF), $X_i = \max_{j=0,...,d} a_j Z_{i-j}$ with $a_j \geq 0$, $\sum_{j=0}^{d} a_j = 1$ and $\{Z_i\}$ i.i.d. standard Fréchet where $\theta = \max_{j=0,...,d} a_j$ (Weissman & Cohen [21] 1995); we consider $d = 2$ and parameters $a_0 = 1/3$, $a_1 = 1/6$, $a_2 = 1/2$, and thus $\theta = 1/2$;
- ARCH(1) process, $X_i = (\beta + \alpha X_{i-1}^2)^{1/2}\epsilon_i$, with i.i.d. Gaussian innovations $\{\epsilon_i\}$, $\alpha = 0.7$ and $\beta = 2 \cdot 10^{-5}$, where $\theta = 0.721$ (Cai, [3] 2019);
- 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 $\theta = 1-\phi$ (Davis and Resnick [5] 1989); we consider $\phi = 0.1$, $\theta = 0.9$;
- an i.i.d. sequence (Ind) of Fréchet r.v. where $\theta = 1$.

3 Simulation study and application

The simulation study is based on random generation of samples with size 1000 replicated 1000 times, within each of the models described above. We consider different models with different values of θ. We apply the estimation plateau method of Section 2 both to estimate θ and respective 95% CI lower and upper bounds. In Table 1 are the estimation global results of the plateau method. See also the simulation results of $\widehat{\theta}$ given in (3) for each block size b in Figure 3 as well as the results of plateau method. We can observe in each model that the plateau estimate (thicker gray horizontal full line) is located in a plateau zone of the sample path of estimates plotted as a function of block size b (full black line), as expected. We can also see that the plateau estimate is close to the true

value (blue horizontal full line). In all cases it is verified that the limits of 95% CI estimated by the plateau method (thicker gray horizontal dotted-dashed lines) include the true value of θ. In the ARCH case, the estimates closest to the true value of θ occur for large values of b where the variability is very high, which makes it difficult to apply the plateau methodology. Even so, the root mean squared error (rmse) of 0.1126 is not very expressive. The independent model (Ind) has $\theta = 1$ and, therefore, constitutes a frontier value of the parameter support, which typically leads to difficulties in statistical estimation. Still, the plateau method showed relatively low bias and rmse. Observe also that in the MAR model we have $\theta = 0.9$, quite near to boundary value 1, and the plateau method does a very good job.

Table 1. Simulation results of plateau method: average of θ estimates (mean), average of lower and upper 95% CI bounds estimates, bias, root mean squared error (rmse) and standard deviation of θ estimates (sd).

	mean	lower	upper	bias	rmse	sd
ARC ($\theta = 0.1$)	0.1106	0.0841	0.1372	0.0106	0.0218	0.0190
MMU ($\theta = 1/3$)	0.3587	0.3042	0.4139	0.0254	0.0494	0.0424
MMF ($\theta = 0.5$)	0.5160	0.4379	0.5940	0.0160	0.0636	0.0616
ARCH ($\theta = 0.721$)	0.7634	0.6267	0.8920	0.0424	0.1126	0.1044
MAR ($\theta = 0.9$)	0.9017	0.7779	0.9763	0.0017	0.0827	0.0827
Ind ($\theta = 1$)	0.9709	0.8756	0.9969	-0.0291	0.0643	0.0573

3.1 Application to real data

We illustrate the method with the real data *newlyn* available in R package *exdex*, consisting of 2894 sea-surge heights measured at the coast at Newlyn, Cornwall, UK, over years 1971-1976. The observations correspond to the maximum hourly surge heights during periods of 15 hours. See the left plot in Figure 4. Previous analysis of this data can be seen in Northrop ([19] 2015) and references therein. The sample path of estimates from (3) as a function of block size b and respective 95% confidence limits are plotted on the right graph of Figure 4. The horizontal full line corresponds to the plateau estimate of θ given by 0.2577 and the horizontal dotted-dashed lines to the plateau 95% CI estimate $(0.2206, 0.2948)$.

4 Conclusion

This work addresses the estimation of the extremal index θ. This is an important measure in time series, namely in assessing risky phenomena, as it measures the

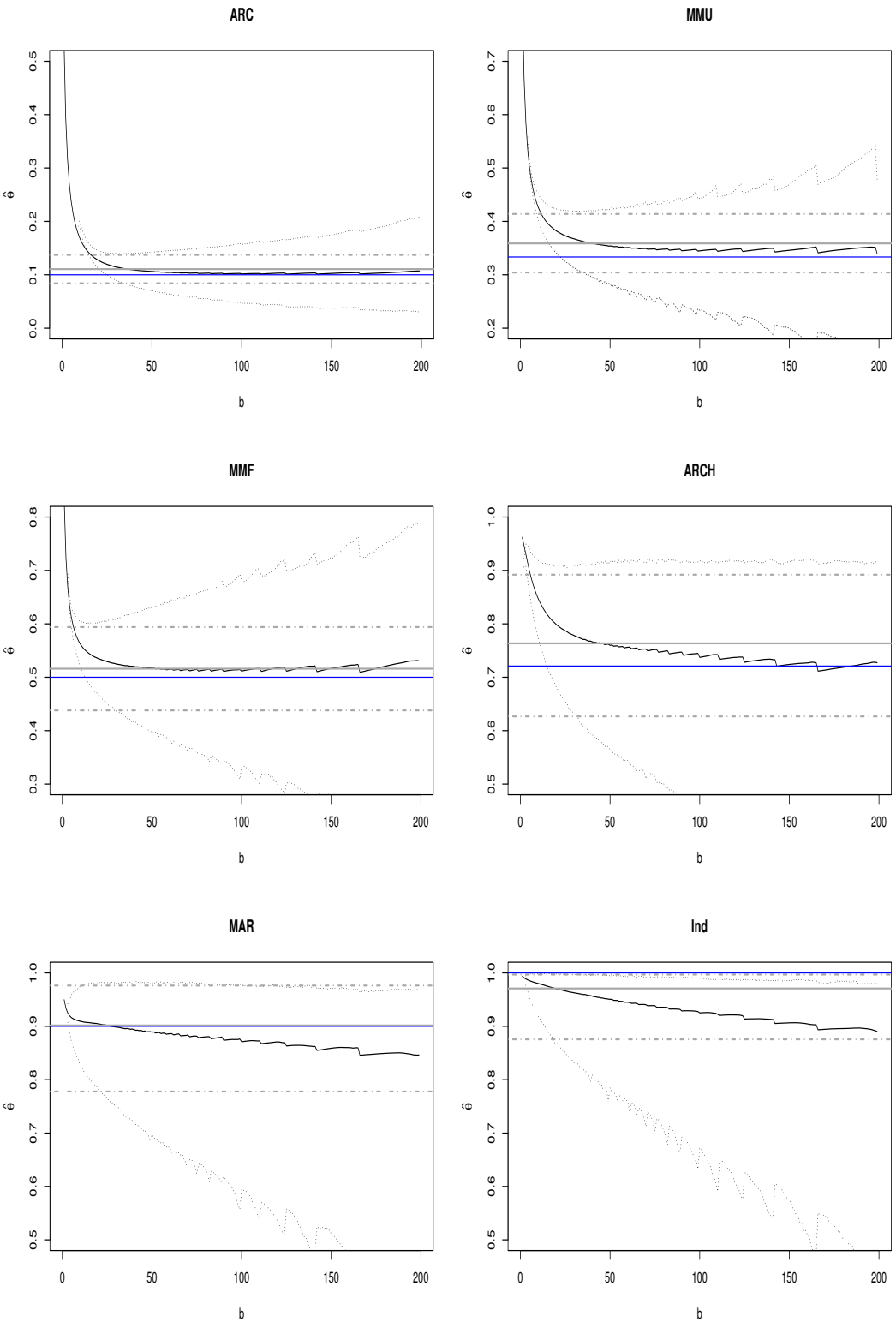

Fig. 3. Simulation results: average of estimates of θ for each block size $b = 2, ..., 200$ using $\widehat{\theta}$ in (3) (full black line) and average of respective 95% CI upper and lower bounds (dotted lines); plateau estimation of θ (thicker gray horizontal full line) and respective plateau estimates of 95% CI upper and lower bounds (thicker gray horizontal dotted-dashed lines). The true value of θ corresponds to the blue horizontal full line.

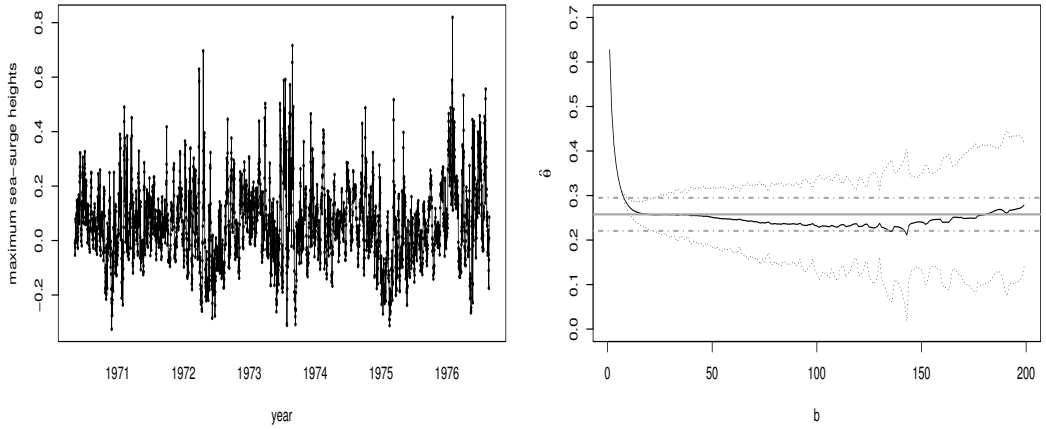

Fig. 4. Left: Maximum hourly (within successive 15 hours periods) surge heights time series at Newlyn cost, Cornwall, UK in years 1971-1976; Right: Sample path estimates obtained from estimator in (3) (full line) and respective 95% CI limits (dotted lines) for successive values of block size b, plateau estimate of θ (horizontal full line) and respective 95% CI plateau estimate limits (horizontal dotted-dashed lines).

propensity for occurrence of clusters of extreme values. The estimation of θ requires the prior setting of tuning parameters values that impact the precision of estimates. In this work we present an algorithm that allows to estimate θ free of tuning parameters. We applied this methodology to diverse models and the results are encouraging in several cases. In the future it is intended to continue the study of this methodology and develop it in order to improve its applicability to different types of models.

Acknowledgments

The author was financed by Portuguese Funds through FCT - Fundação para a Ciência e a Tecnologia within the Projects UIDB/00013/2020 and UIDP/00013/2020 of Centre of Mathematics of the University of Minho.

References

1. Ancona-Navarrete, M.A., Tawn, J.A.: A comparison of methods for estimating the extremal index. Extremes 3, 5–38 (2000)
2. Berghaus, B., Bücher, A.: Weak convergence of a pseudo maximum likelihood estimator for the extremal index. The Annals of Statistics 46(5) 2307–2335 (2018)

3. Cai, J.J.: Statistical inference on $D^{(d)}(u_n)$ condition and estimation of the Extremal Index. arXiv:1911.06674 (2019)
4. Chernick M.R.. Hsing, T., McCormick, W.P.: Calculating the extremal index for a class of stationary sequences. Advances in Applied Probability 23, 835–850 (1991)
5. Davis, R., Resnick, S.: Basic properties and prediction of max-ARMA processes. Advances in Applied Probability 21, 781–803 (1989)
6. Ferreira , M.: Heuristic Tools for the Estimation of The Extremal Index: A Comparison of Methods. REVSTAT-Statistical Journal 16(1), 115–136 (2018)
7. Ferreira, H., Ferreira, M.: Estimating the extremal index through local dependence. Annales de l'Institut Henri Poincaré - Probabilités et Statistiques 54(2), 587–605 (2018).
8. Ferreira, H., Ferreira, M.. A new blocks estimator for the extremal index. Communications in Statistics - Theory and Methods (2022) (in press)
9. Ferro, C.A.T., Segers, J.: Inference for clusters of extreme values. Journal of the Royal Statistical Society: Series B 65, 545–556 (2003)
10. Frahm, G., Junker, M., Schmidt, R.: Estimating the tail-dependence coefficient: Properties and pitfalls. Insurance Mathematics & Economics 37, 80–100 (2005)
11. Gomes, M.: 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 (1993)
12. Gomes, D.P., Neves, M.M.: Extremal index blocks estimator: the threshold and the block size choice. Journal of Applied Statistics 47(13-15), 2846–2861 (2020)
13. Holěsovský, J., Fusek, M.: Estimation of the extremal index using censored distributions. Extremes 23(2), 197–213 (2020)
14. Holěsovský, J., Fusek, M.: Improved interexceedance-times-based estimator of the extremal index using truncated distribution. Extremes 25, 695–720 (2022).
15. Hsing, T., Hüsler, J., Leadbetter, M. R.: On the exceedance point process for a stationary sequence. Probab. Theory Related Fields 78 97–112 (1988)
16. Süveges. M.. Davison A.C.: Model misspecification in peaks over threshold analysis. Annals of Applied Statistics 4. 203–221 (2010)
17. Nandagopalan S.: Multivariate extremes and estimation of the extremal index. Ph.D. Thesis. University of Nth Carolina. Chapel Hill (1990)
18. Newell, G.F.: Asymptotic Extremes for m-Dependent Random Variables. Annals of Mathematical Statistics, 35, 1322–1325 (1964)
19. Northrop, P.J.: An efficient semiparametric maxima estimator of the extremal index. Extremes 18(4). 585–603 (2015)
20. Northrop, P.J., Christodoulides, C. exdex: Estimation of the Extremal Index. R package version 1.0.1. (2019) `https://CRAN.R-project.org/package=exdex`
21. Weissman, I., Cohen, U.: The extremal index and clustering of high values for derived stationary sequences. J. Appl. Prob. 32, 972–981 (1995).
22. Weissman, I., Novak, S.Y.: On blocks and runs estimators of the extremal index. Journal of Statistical Planning and Inference 66(2), 281–288 (1998)