New Approaches to Extremal Index Estimation

M. CRISTINA MIRANDA[1,2,3], MANUELA SOUTO DE MIRANDA[2], M. IVETTE GOMES[3]
[1]ISCA,[2] CIDMA
University of Aveiro
Campus Santiago, 3810-193 Aveiro
[3]CEAUL
University od Lisbon
PORTUGAL

Abstract: The extremal index is a parameter associated with the extreme value distributions of dependent stationary sequences. Under certain local dependence conditions, exceedances above a specified threshold tend to occur in isolated clusters. The reciprocal of the extremal index can be interpreted as the limiting size of these clusters. Accurately estimating the size of such clusters is crucial for analyzing real data and can significantly influence decision making processes that impact population well being. The paper presents a recent method for the estimation of the extremal index which starts by the estimation of the parameter itself and, only then, to use that estimate in the cluster mean size estimation. The procedure starts with the estimation of a specific proportion by the corresponding relative frequency. Thus, it is very simple, intuitive, it has good statistical properties, and it does not depend on the method used for the mean cluster estimation. The interpretation of the extremal index as a proportion is known, but it has not been used directly as an estimation method. In recent years, various authors have proposed different estimators for the extremal index. This paper applies some of the latest estimation methods for the extremal index to real data and analyses their performance using training and test samples. The results are compared with other well known estimators, for which *R* packages are available. The results show a better performance of the Proportion estimator, followed by the Gaps estimator, when compared to the other considered index estimators.

Key-Words: Extremal Value Theory, Extremal index, Robust estimation, Proportion estimator, Negative Binomial, Stationary sequences, Clusters of exceedances

Received: March 15, 2024. Revised: August 6, 2024. Accepted: September 7, 2024. Published: October 22, 2024.

1 Introduction

Extreme value theory (EVT) addresses the study of extreme value distributions. The occurrence of rare events in real life, such as floods, strong winds, extreme temperatures, epidemic peaks, or stock market crashes, often has catastrophic consequences for populations. This reality spurred the development of EVT around the 1950´s, particularly in the Netherlands, due to its geographic characteristics and the need to survive in a territory below the sea level. In the 1980's, the classical independent and identically distributed (IID) approach was extended to stationary sequences, [1]. Under suitable local dependence conditions, it is also possible to obtain the limit distribution for the maximum of stationary sequences. In fact, this limit distribution presents the same shape parameter but the location and scale are affected by an extra parameter, the Extremal Index (EI), hereby denoted by θ.

In an IID scenario, extreme values typically manifest as rare and isolated occurrences. However, in dependent sequences, the dynamics change significantly. The structure of dependence leads to clusters of extreme values rather than isolated events. As the strength of dependence increases, so does the size of these clusters. The EI can be interpreted as the reciprocal of the mean size of clusters of exceedances above a specified threshold. Therefore, EI estimation is crucial as it provides a measure of the duration of periods with consistently high or low values in the sequence under study.
A different approach is to consider the sequence as a compound Poisson Process. Extreme events occur as rare phenomena as in a Poisson Process, [2]. Under suitable local dependence conditions, when the sequence is stationary but not I.I.D., the process of exceedances converges to a compound Poisson Process with intensity related to the EI, [3]. When analysed in the dynamical systems framework, the EI may be seen as a measure about the dynamics of the underlying systems, [4].

Since the first proposed estimators, in the 90's, estimating the EI is an issue that has been addressed with increased attention. Some of the existing proposals are based on the size of the clusters and how these clusters are identified. This is the case with the runs estimator and the blocks estimator which require additional parameters in turn, [5], [6], [7]. The intervals estimator, [8], is one of the most

popular in the literature and is based on the interexceedances times of overpassing a high threshold. The intervals estimator motivated the work of Süveges, who proposed a maximum likelihood estimator, the so called Kgaps estimator, [9], which also depends of a K parameter. A detailed synthesis of the EI estimation options may be found at the book by [10], or in the recent paper of [11]. Variations of the initial blocks and runs estimators include the disjoint blocks estimator and the sliding blocks estimator, [12]. Based on the maxima method, [13], proposed a semiparametric approach producing more efficient estimates. Relevant asymptotic results and a moments EI estimator is presented in the paper from [14]. To reduce bias, [15], applies the Jackknife methodology. A key problem with EI estimation is the choice of the threshold to be considered, originating the process of exceedances. Some of the existent estimators are too sensitive to the choice of that parameter, [16]. Recent research addresses to this problem, proposing a new way to select the threshold in a context of non parametric estimation of the extremal index of stochastic processes, [17]. When academics are called to apply their results to real data, it is frequent not to encounter the conditions assumed to validate the results. In that sense, there is the recognized need for robust procedures which are not so vulnerable in the absence of those conditions. The estimator proposed in [18], is a robust proposal to estimate the EI.

More recently, an interesting proposal uses artificial censoring of interexceedances times with some evidence of stability improvement and less sensitivity to the choice of the parameter evolved, D, in this case, and to the choice of the high threshold, [19]. In the present paper, two different proposals are considered to estimate the EI: a robust estimator, based on the Negative Binomial (NB) model, and a proportion estimator, also based in the interexceedances times.

Following the introduction, **Section** 2 refers to the basic principles of EVT in the case of IID samples and for specific dependent sequences. In **Section** 3 and 4, two recently proposed EI estimators are presented. **Section** 5 presents other known estimators to be considered in the data analysis. **Section** 6 contains the results of applying both robust and non robust estimators to a real data set. The paper concludes in **Section** 7 with comments and conclusions.

2 EVT classic theory

2.1 IID case

The theory of extreme values is highly developed when dealing with IID samples. In this case, there exists the well known theorem in EVT that allows us to obtain the distribution of the limit of the maximum when appropriately normalized, and whenever such a limit exists. This theorem provides a convenient way to understand the behaviour of extreme values in large samples drawn independently and identically distributed, $(Y_1, Y_2, ..., Y_n)$, with some unknown common cumulative distribution function (CDF), $F(y)$. The Fisher, Tippett, Gnedenko extremal types theorem determines the limiting CDF of the maximum as a member of the general extreme value (EV) distribution family. If $Y_{n:n} = \max\{Y_1, Y_2, ..., Y_n\}$ and if there are constants $a_n > 0$ and $b_n \in \mathbb{R}$ such that the standardized maximum $Z = (Y_{n:n} - b_n)/a_n$ converges for some non degenerate distribution, then, such a limit CDF is of the type of

$$EV_\xi(y) = \begin{cases} \exp(-(1+\xi y)^{-1/\xi}), 1+\xi y > 0, \\ \qquad\qquad\qquad \text{if } \xi \neq 0 \\ \exp(-\exp(-y)), y \in \mathbb{R}, \\ \qquad\qquad\qquad \text{if } \xi = 0. \end{cases} \tag{1}$$

Several advancements have expanded the applicability of EVT to dependent sequences, just as sketched in **Section** 2.2. This is especially relevant in the analysis of time series, where dependencies among observations are typical. In such cases, recent developments in EVT have adapted to account for these dependencies, offering insights into the distribution of extremes and enhancing our understanding of their behaviour over.

2.2 EVT for dependent structures

When we have an IID sequence, extreme values tend to be rare events, usually far apart from each other in time. But when some dependence structure is present, this affects their behaviour, originating clusters of high values close to each other. Dependent observations tend to exhibit similar patterns. When an extreme value is observed, the next observation is likely to be close to the previous one. The presence of autocorrelation induces proximity between observations, leading to clusters of high values rather than isolated extremes.

In some sequences, those clusters appear separated in time. So in limit, the clusters ocuurrence times tend to be independent. This paper addresses to the problem of estimating a parameter that measures the interdependence among sequences

of random variables. For sequences that verify certain stationary and mixing conditions, the classic EVT remains applicable with few changes, [1].

Definition 2.1. $D(u_n)$ The dependent sequence $\{X_n\}_{n\geq 1}$, with marginal CDF F, verifies the mixing condition $D(u_n)$ of [1], if, for each $i_1, ..., i_p, j_1, ..., j_q$ such that

$$1 \leq i_1 < ... < i_p < j_1 < ... < j_q \leq n, \quad j_1 - i_p \geq l,$$

and the notation

$$F_{k_1 \ldots k_r}(u_n) := \mathbb{P}(X_{k_1} \leq u_n, ..., X_{k_r} \leq u_n),$$

$$\left|F_{i_1 \ldots i_p j_1 \ldots j_q}(u_n) - F_{i_1 \ldots i_p}(u_n) F_{j_1 \ldots j_q}(u_n)\right| < \alpha_{n,l}$$

where $\alpha_{n,l_n} \to 0$ as $n \to \infty$ for some sequence l_n with $l_n = o(n)$.

If the D condition in **Definition** (2.1) is assumed for a stationary sequence $\{X_n\}_{n\geq 1}$, then the limit distribution of the maximum will be one of the three types, Fréchet ($\xi > 0$), Gumbel ($\xi = 0$) or Weibull ($\xi < 0$), depending on the ξ value, in (1).

Other local dependence conditions are usually necessary in the framework of the EI estimation so that the asymptotic distribution of the sample maximum can be calculated. These may be formulated in terms of the D^k condition, [20], [21], which may be stated as follows:

Definition 2.2. $D^k(u_n)$ Suppose that a sequence $\{X_n\}_{n\geq 1}$ verifies the D Leadbetter condition, given above in **Definition** (2.1). The $D^k(u_n)$ condition is said to hold if for some $\{k_n\}_{n\geq 1}$ such that

$$k_n \to \infty, \quad k_n \alpha_{n,l_n} \to 0, \quad k_n l_n / n \to 0$$

as $n \to \infty$, we have

$$nP\left(X_1 > u_n, M_{1,k} \leq u_n < M_{k,r_n}\right) \underset{n\to\infty}{\to} 0$$

with $\{r_n = [n/k_n]\}_{n\geq 1}$, $[x]$ denoting the integer part of x and $M_{i,j} = \max\{X_i, ..., X_j\}$, $(j > i)$.

Understanding the joint distribution of k consecutive terms enables us to derive the limiting distribution of the maximum for these sequences.

2.3 The extremal index

Consider a strictly stationary process $\{X_n\}_{n\geq 1}$ with CDF F, with finite or infinite right endpoint. The process is said to have an extremal index $\theta \in [0, 1]$ if, for each $\tau > 0$, there is a sequence $\{u_n\}_{n\geq 1}$ such that, as $n \to \infty$:

a. $n\bar{F}(u_n) \to \tau$ and

b. $\mathbb{P}[M_n \leq u_n] \to \exp(-\tau\theta)$,

where $\bar{F} := 1 - F$ and $M_n := M_{0,n}$ denotes the maximum of n consecutive observed variables, defined by $M_{k,l} := \max\{X_i : i = k+1, \ldots, l\}$.

Let $\{Y_n\}_{n\geq 1}$ be an IID sequence, and suppose $\{X_n\}_{n\geq 1}$ is a stationary sequence with the same marginal F. Then, if there exist real constants a_n and b_n, such that

$$F_{Y_{n:n}}(a_n y + b_n) \to H(y),$$

where $H(y)$ needs to be of the type of the EV distribution, in (1), then, for the stationary sequence $\{X_n\}_{n\geq 1}$, with an extreme value index, θ,

$$F_{X_{n:n}}(a_n y + b_n) \to H^\theta(y).$$

So, the EI value has an effect of clustering and shrinking the extreme values when we compare the two types of samples. The EI is a parameter that may be interpreted as a measure of the dimension of the clusters of exceedances over some high threshold. In fact, it may be interpreted as the limit of the reciprocal of the cluster mean size, [22]. For independent sequences or asymptotically independent sequences, $\theta = 1$. This means that exceedances do not form clusters as it happens with dependent samples. The value of θ relates to the dimension of the clusters: the closest to zero, the greater the clusters size will be (in average). Figure 1 illustrates the difference between sequences for which $\theta = 1$ and an associated dependent sample with a small value of θ.

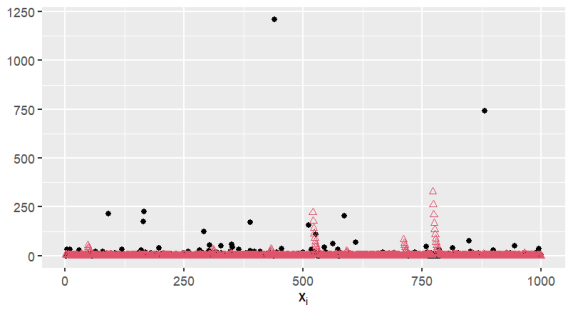

Figure 1: Circles represent a sample with $\theta = 1$; triangles denote a dependent sample, with $\theta = 0.2$.

3 Robust estimation based on the NB

A recent proposed method also based in the limit mean size estimation of the clusters uses the uniparametric NB model and robust estimation of its parameter. The model was suggested by [23], upon the NB2 reparametrization of the model, [24], which turns easier robust estimation in the framework of

the generalized linear regression. The process was investigated aiming to improve robustness of the runs estimator in the sense that we are interested in modelling the number of exceedances that occur before a non exceedance and to bypass difficulties caused by the presence of too many zeros and distributional overdispersion, [18].

Assume a GLM with independent response variables Y_i, $i = 1, \ldots, n$, such that Y_i follow a NB distribution. With the NB2 reparametrization, the probability function NB can be expressed in terms of μ_i and σ, where σ is a shape parameter and μ_i is the conditional mean, being given by

$$f(y_i; \mu_i, \sigma) = \frac{\Gamma(y_{i^+}\sigma^{-1})}{\Gamma(y_i+1)\Gamma(\sigma^{-1})}\left(\frac{1}{1+\sigma\mu_i}\right)^{\sigma^{-1}}\left(\frac{\sigma\mu_i}{1+\sigma\mu_i}\right)^{y_i}, \quad (2)$$

where $y_i = 0, 1, ..., n$, $\Gamma(.)$ is the Gamma function and $\sigma > 0$ is the overdispersion parameter of the Y_i distribution. Under that parametrization (and denoting by $\mathbf{x}_i$ an observation of a potential k dimensional regressor), the conditional variance is:

$$\mu_i = \mathbb{E}[Y_i|\boldsymbol{x}_i], \;\; V(\mu_i) = V(Y_i|\boldsymbol{x}_i) = \mu_i + \sigma\mu_i^2. \quad (3)$$

Expression (3) shows that when $\sigma \to 0$, $V(\mu_i)$ converges to the variance of a Poisson distribution. Thus, we will work on the GLM setting

$$Y_i \sim NB(\mu_i, \sigma), \; i = 1, \ldots, n,$$

$$g(\mu_i) = g(\mathbb{E}[Y_i|\boldsymbol{x}_i]) = \boldsymbol{x}_i^T\boldsymbol{\beta}x_i^T + \epsilon_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon_i, \quad (4)$$

with the link function $g(.) = \log(.)$ and where $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$ is the unknown parameter. Robust estimators for $\boldsymbol{\beta}$ and σ are based on maximum likelihood estimators (MLE), substituting estimating equations by bounded functions. They are already disposable from *R* package **robNB** for the model parameter that we are interested in. Particularly, we only need to estimate the intercept, thus without regressors. So, from (4), the cluster mean size estimate will be $\hat{\mu}_i = \exp\hat{\beta}_0$ and the EI estimate will be its reciprocal, i.e.,

$$\hat{\theta}_{NB} = 1/\hat{\mu}_i = \exp\{-\hat{\beta}_0\}.$$

4 Proportion estimator

The EI proportion estimator appeared recently, [25], on the basis of a completely different estimation approach. While other methods are developed aiming to find the best estimator for the limiting clusters mean size, and then to take the reciprocal of that estimate as the EI estimate, the proportion estimator goes directly to the estimation of the EI as a distributional parameter by its own. The procedure emerged taking a particular attention to the distribution of the interexceedances times, like it is presented in [8]. Those authors proved that the distribution of the interexceedances times is the mixture:

$$(1-\theta)\epsilon_0 + \theta f_{D(\theta)}, \quad (5)$$

where ϵ_0 is the degenerate probability distribution at 0 and $f_{D(\theta)}$ represents an exponential distribution with mean θ^{-1}. Thus, as referred in that paper, the EI is the proportion of strictly positive interexceedances times.

More formally, let $\{X_n\}_{n\geq 1}$ be a strictly stationary sequence of random variables with marginal cumulative function $F(.)$ and consider a chosen high threshold u. Suppose the exceedances occur at times $j_1, ..., j_N$, N being the number of exceedances, *i.e.*,

$$N = \sum_{i=1}^{n} \mathbb{I}_{\{X_n > u\}}, \quad (6)$$

Consider also S_i ($i = 1, \ldots, N, \;\; N \leq n$), the times of isolated exceedances or of the first element of a cluster, and

$$T_i^* = S_{i+1} - S_i \text{ with } i = 1, \ldots, N-1.. \quad (7)$$

In the limit distribution, clusters occur at the singularities of the compound Poisson process. Sequential exceedances will belong to the same cluster with null interexceedances times and they represent the multiplicity of each singularity of the compound Poisson process. So, in the limit distribution, the T_i^* $i = 1, \ldots, N-1$) correspond to the realizations of a random variable T_θ, whose distribution is characterized by (5).

The simplest way and also the most intuitive method for estimating a proportion is to use the relative frequency. Besides, the estimator has good properties, as strong convergence results are known, without requirements of differentiability or even continuity, for the EI parameter space is the interval $[0, 1]$ and the estimates will take values in $(0, 1)$.

Particularly, they will take values in $(0, 1)$ and they are defined by

$$\hat{\theta}_P = \frac{\#\{T_i^* > 0\}}{N-1}, \quad (8)$$

where N ($N < n$) is the number of exceedances to the fixed threshold observed in the sample.

In what concerns robustness, things are not so clear. The most commonly used criteria for evaluating robustness are B robustness and the breakdown point (BP) value. B robustness ensures that a robust estimator has a bounded influence function, which, in simple terms, measures the sensitivity of the estimator to changes in individual observed values. The breakdown point, on the other hand, is related to the estimator's resistance, where a strictly positive BP indicates that the estimator can handle a certain proportion of altered observations without leading to a breakdown in the estimate. Actually, the proportion estimator is not B robust. Nevertheless, it has a positive breakdown point, thus it is robust in the last sense.

Perhaps even more interesting is the fact that the type of dependence structure is ignored in the estimation process. The proportion estimator seems not to be very sensitive to the type of dependence within clusters of exceedances, like it happens with real data sequences. This is because the proportion of strictly positive interexceedance times is unaffected by the numerical values of the exceedances. In fact, what truly matters are the times between clusters of exceedances, rather than the times within different clusters.

There are only a few technical details: the convergence is based on the number of exceedances in the sample and not directly in the sample size, as usual. Different samples with the same n dimension may have different numbers of exceedances. So, with an unique sample there is only one observation of the relative frequency. It is impossible to interpret the EI as the binomial parameter (exception for the Bernoulli model). Fortunately, the nature of extreme values assure that exceedances are rare events. Most real data sets have the size enough for splitting the original sample into several subsamples with constant number of exceedances N. In fact, some big data collections contain too many zero values and they are already cyclic. Thus, some subsamples will turn possible to define a particular N value. The properties of the relative frequency (consistency and positive BP) will assure good results or, at least, reasonable results even in the worst cases. In simulations studies the last question is overpassed using only samples with the same number of exceedances (which are random realizations of a Poisson distribution), eventually discarding other generated samples. Another technical suggestion deals with the choice of N , which must cover the dependence structure in such a large part as possible. That is an issue in current investigation by the authors.

5 Other EI estimators under consideration

In this paper the authors choose to give particular attention to the EI estimators based on the interexceedance times initially proposed by [8]. In previous simulation studies they included the Blocks and the Runs estimators, [26], [7], also. In this paper the option was to analyse the EI estimators that are based on the interexceedance times only. To compare the performance of the author's proposals referred above, namely, $\hat{\theta}_{NB}$ and $\hat{\theta}_{Prop}$, the following were also computed at the same high thresholds:

- the Intervals estimator, [8], hereby denoted by $\hat{\theta}_{Int}$;
- the K generalized Gap estimator, $\hat{\theta}_{KGaps}$, [9];
- the DGaps estimator, $\hat{\theta}_{DGaps}$, [19].

5.1 Intervals estimator

The authors in [8], pointed out the nature of the asymptotic distribution of the interexceedance times $T_i = j_{i+1} - j_i$ as part of a parametric family of distributions indexed by the extremal index. Under the same notation used in **Section** 4, with a preliminary estimator of θ, on the basis of

$$\hat{\theta}_n(u) = \frac{2\left(\sum_{i=1}^{N-1} T_i\right)^2}{(N-1)\sum_{i=1}^{N-1} T_i^2} \tag{9}$$

and

$$\hat{\theta}^*_n(u) = \frac{2\{\sum_{i=1}^{N-1}(T_i-1)\}^2}{(N-1)\sum_{i=1}^{N-1}(T_i-1)(T_i-2)}, \tag{10}$$

the Intervals estimator is $\hat{\theta}_{Int}(u)$ is defined by

$$\hat{\theta}_{Int}(u) = \begin{cases} \min\{1, \hat{\theta}_n(u)\} & \text{if } \max T_i : \\ & 1 \le i \le N-1 \le 2 \\ \min\{1, \hat{\theta}^*_n(u)\} & \text{if } \max T_i : \\ & 1 \le i \le N-1 > 2 \end{cases}. \tag{11}$$

5.2 K Gaps estimator

The $\hat{\theta}_{KGaps}$ estimator eliminates small interexceedance times by setting them to zero. This approach appears reasonable at a first glance: if a period of high sales is briefly interrupted by an external event, the total count of high sales days shouldn't be markedly impacted. This interruption can be viewed as an outlier within the larger cluster

of high sales days. Therefore, a more appropriate strategy might involve disregarding such values and treating the entire set of observations as a single cluster. The K Gaps estimator is a maximum likelihood estimator, turned possible due to the inclusion of a new variable in the likelihood function, namely, the K Gap, $Su^K(u_n) = \max\{T(u_n) - K; 0\}$, [9], [19], with $T(u_n) = \min\{j \geq 1 : X_{j+1} > u | X_1 > u\}$.

5.3 D Gaps estimator

The $\hat{\theta}_{DGaps}$ estimator is presented by [19], as a generalization of the $\hat{\theta}_{KGaps}$ estimator. An artificial scheme of censoring is introduced where the small interexceedance times are censored. This procedure requires a new time parameter D to be conveniently chosen. Consider the order statistics of the interexceedance times as defined in **Section**(5.1). The authors in [19], apply a type I left censoring after choosing a time censor $D \geq 0$ and consider the loglikelihood function of the censored sample. Their estimator is the value which minimizes that function.

6 Application to real data

Recent years brought new challenges to international markets in several fields like electronics and pharmaceutical products, either due to COVID or justified by the constraints of war. In times of COVID, pharmaceutical laboratory suffered extra stress and this is still presently pointed occasionally at news. High values of drug sales frequently happen in clusters. This can be explained by some epidemiology or by some other phenomena, not so easy to identify. Being able to estimate the size of the clusters allows for better planning the production and, on the other side, may help to study eventual hidden causes for the peaks of sales. High levels of air pollution tend to be associated with high levels of antihistaminic sales. Available since the 40's, this type of drugs represent an option of treatment that can improve life quality to a vast number of people, [27]. The sales numbers of this type of drugs show clusters of high values which may be relevant, either to pharma business or to the medical research. The data set used in the present work was obtained from Kaggle [28], and consists of weekly pharma sales data from antihistamines for systemic use (R06, accordingly to Anatomical Therapeutic Chemical (ATC) Classification System), since 2014 to 2019. As shown in Figure 2 , this data set presents clusters of high values separated from each other. We'll assume the validity of the conditions of local dependence.

Following the *holdout* procedure, which is typical in time series analysis, [29], the sample was divided

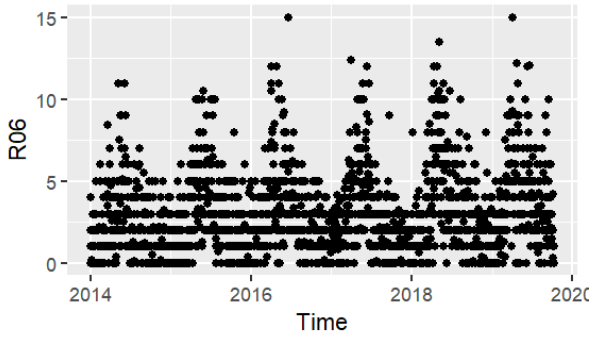

Figure 2: Sales data from antihistamines for systemic use (R06), since 2014 to 2019

into a training sub sample containing the first 60% of observations and a test sample, consisting of the latest 40% observations (Initially, 70% of the first observations were considered for the training sample. However, the decision to use 60% instead was made, to ensure that the test sample visually preserved the behaviour of the entire series more accurately). The resulting samples are illustrated in Figure 3. To obtain the results presented in this paper the R packages *extRemes*, *fExtremes* and *exdex* were used.

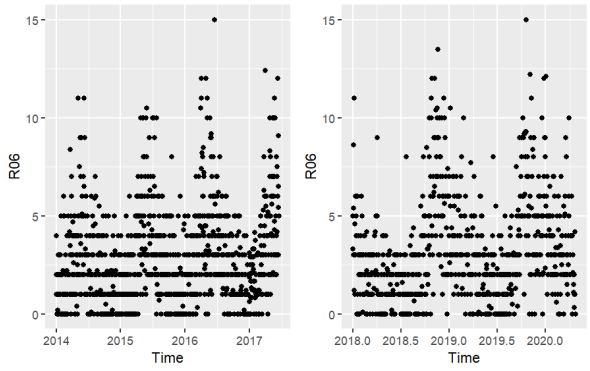

Figure 3: Sales data from antihistamines for systemic use (R06): training sample (left) an test sample (right).

Different estimates were computed using the training sample and then tested by comparing these estimates with the reciprocal of the average size of the observed clusters of exceedances in the test sample. These clusters were defined as groups of exceedances above the specified high thresholds after the first one observed. A set of thresholds corresponding to the highest quantiles, from 0.6 to .98 was considered. For the $KGaps$ and $DGaps$ estimators, the suitable combination of the pair (u, K) (of the threshold and the tuning parameter) was selected, accordingly to the graphical

diagnostics based on the information matrix test, [9]. Figure 4 shows the behaviour of the obtained estimates for the data. Apart from the intervals estimate, it is possible to identify two main tendencies: one group with higher values includes the KGaps and the Proportion estimators, and another group with lower values, with the NB and the DGaps estimators. In the present case, it seems that the former would perform better if used to preview the future θ value as the corresponding lines are closer to the observed value of θ.

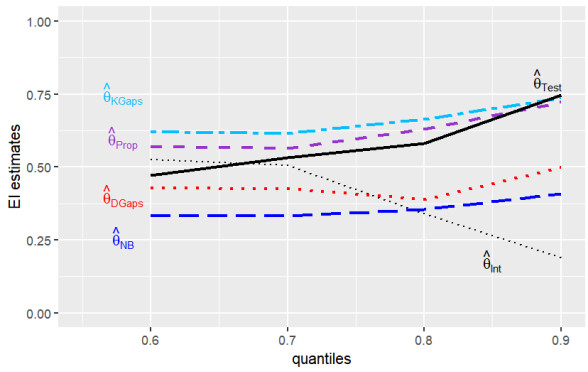

Figure 4: EI estimates for the antihistamines (R06) data: estimates computed with the training sample and the reciprocal of the mean size of the observed clusters for the test sample (θ_{Test}).($K = 1$ and $D = 4$) for the KGaps and DGaps, respectively.)

Considering the former group corresponding to better performances the best results are provided by the Proportion estimator herein suggested. Thus, while ignoring completely the type of dependence and being very simple to interpret, the well known properties of the relative frequency assure a good performance. The approach by the NB estimator does not show advantages, though the trajectory is similar to the θ_{Test} but with an almost constant considerable bias. It is also relevant to notice that the Intervals estimator seems to produce poor estimates.

7 Conclusions

This paper gives a contribution to the estimation of the EI. Among different existing proposals in the literature, two recent new approaches were applied to a real data set: a robust estimator based on the uniparametric NB model (3), and an estimator based on the interexceedance times (4). The resulting estimates were then compared for different high thresholds showing a better performance from the Proportion estimator. Assuming the $\hat{\theta}_{Test}$ is the most close to the real value, the $\hat{\theta}_{Prop}$ estimator seems to have better performance since their paths are the closest to the $\hat{\theta}_{test}$ path.

The simplicity, the properties, and the performance of the Proportion estimator make it well suited for application in more practical fields, particularly when the dependence structure of the clusters is unknown, which is often the case. However, some disadvantages arise in simulation studies, potentially due to the fact that the choice of N does not depend on the sample size, as well as limitations in the application of resampling methods.

Local dependence conditions might be difficult to prove. So, other alternative robust forms must be found. The $\hat{\theta}_{NB}$ seems to be the one that shows more stability concerning the threshold level.
The main conclusion after this paper is that there is still much work to do for the academic community to provide more accurate tools for the decision makers. For a parameter that takes values between 0 and 1, the results obtained by different approaches show a range af values that is considerably high, like it happens with the present data.

Future research will focus on identifying the optimal number of exceedances for analysis (related to the threshold choice), and conducting a deeper study on robustness, particularly by considering simulated contaminated samples. Additionally, further investigations will explore the application of the method in outlier detection and comparisons of robust estimators. The variation in estimates produced by different methods suggests the need for developing confidence intervals for the EI.

Acknowledgment:
Research partially supported through the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia), by the Center for Research and Development in Mathematics and Applications (CIDMA),UIDB/04106/2020, doi.org/10.54499/UIDB/04106/2020, doi.org/10.54499/UIDP/04106/2020,and Centre of Statistics and its Applications (CEAUL), within project UIDB/00006/2020 (https://doi.org/10.54499/UIDB/00006/2020).

References:

[1] M. R. Leadbetter, G. Lindgren, and H. Rootzén. *Extremes and Related Properties of Random Sequences and Processes*. Springer-Verlag, New-York, 1983.

[2] N. Moloney, D. Faranda, Y. Sato, and N. R. Moloney. An overview of the extremal index.

Chaos: An Interdisciplinary Journal of Nonlinear Science, 29(2), 2019.

[3] N. Chenavier, A. Darwiche, and A. Rousselle. Compound Poisson approximation for simple transient random walks in random sceneries. *Am. J. Probab. Math. Stat*, 21:293–306, 2024.

[4] T. Caby, D Faranda, S. Vaienti, P. Yiou, S Vaienti, and P Yiou. On the Computation of the Extremal Index for Time Series. *Journal of Statistical Physics*, 179:5–6, 2020.

[5] S Nandagopalan. *Multivariate Extremes and Estimation of the Extremal Index*. PhD thesis, University of North Carolina, 1990.

[6] T. Hsing. Extremal Index Estimation for a Weakly Dependent Stationary Sequence. *The Annals of Statistics*, 21:2043–2071, 1993.

[7] R. L. Smith and I. Weissman. Estimating the Extremal Index on JSTOR. *Journal of the Royal Statistical Society. Series B*, 56(3):515–528, 1994.

[8] C. A. T. Ferro and J. Segers. Inference for Clusters of Extreme Values. *J. R. Statist. Soc. B*, 65(2):545–556, 2003.

[9] M. Süveges and A. C. Davison. Model misspecification in peaks over threshold analysis. *The Annals of Applied Statistics.*, 4(1):203–221, 2010.

[10] J. Beirlant, Y. Goegebeur, J. Teugels, and J. Segers. *Statistics of Extremes*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd, Chichester, UK, aug 2004.

[11] G. Buriticá, N. Meyer, T. Mikosch, and O. Wintenberger. Some variations on the extremal index. *Journal of Mathematical Sciences (United States)*, 273(5):687–704, jun 2021.

[12] C. Y. Robert, J. Segers, and C. A. T. Ferro. A Sliding Blocks Estimator for the Extremal Index. *Eletronic Journal of Statistics*, 3:993–1020, dec 2009.

[13] P. J. Northrop. An efficient semiparametric maxima estimator of the extremal index. *Extremes*, 18:585–603, 2015.

[14] A. Bücher and T. Jennessen. Method of moments estimators for the extremal index of a stationary time series. *https://doi.org/10.1214/20-EJS1734*, 14(2):3103–3156, jan 2020.

[15] M. Ferreira. Extremal index: estimation and resampling. *Computational Statistics*, 39(5):2703–2720, jul 2024.

[16] D. Prata Gomes and M. M. Neves. Extremal index blocks estimator: the threshold and the block size choice. *Journal of Applied Statistics*, 47(13-15):2846, nov 2020.

[17] N. Markovich. The Discrepancy Method for Extremal Index Estimation. *Springer Proceedings in Mathematics and Statistics*, 339:341–355, 2020.

[18] M. C. Miranda, M. Souto de Miranda, and M. I. Gomes. A new proposal for robust estimation of the extremal index. *Journal of Statistical Computation and Simulation*, pages 1–18, jan 2024.

[19] J. Holešovský and M. Fusek. Estimation of the extremal index using censored distributions. *Extremes*, 23:197–213, 2020.

[20] M. R. Chernick, T. Hsing, and W. P. McCormick. Calculating the extremal index for a class of stationary sequences. *Advances in Applied Probability*, 23:835–850, 12 1991.

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

[22] T. Hsing, J. Hüsler, and M.R. Leadbetter. On the Excedance of Point Process for a Stationary Sequence. *Probability Theory and Related Fields*, 78:97–112, 1988.

[23] W. H. Aeberhard, E. Cantoni, and S. Heritier. Robust inference in the negative binomial regression model with an application to falls data. *Biometrics*, 70(4):920–931, dec 2014.

[24] J. M. Hilbe. *Negative Binomial Regression*. Cambridge University Press, Cambridge, 2nd edition, jan 2011.

[25] M. Souto de Miranda, M. . Miranda, and M. I. Gomes. A direct approach in extremal index estimation. In *New Frontiers in Statistics and Data Science (SPE 2023, Guimarães, Portugal), in press*. Springer, 2024.

[26] T. Hsing. Estimating Parameters of Rare Events. *Stochastic Processes and their Applications*, 37(1):117–139, 1991.

[27] M. Grundström, Åslög Dahl, Tinghai Ou, Deliang Chen, and Håkan Pleijel. The relationship between birch pollen, air pollution and weather types and their effect on antihistamine purchase in two Swedish cities. *Aerobiologia*, 33(4):457–471, dec 2017.

[28] M. Zdravković. Pharma sales data. Kaggle, url:https://www.kaggle.com/ds/466126, 2020. Accessed: 2024-05-13.

[29] V. Cerqueira, L. Torgo, and I. Mozetič. Evaluating time series forecasting models: an empirical study on performance estimation methods. *Machine Learning*, 109(11):1997–2028, nov 2020.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

All the authors contributed to the conceptualization as well as to the the final writing.
M. Cristina Miranda was responsible for chosing the data and for the code to produce and visualise the results.
Manuela Souto de Miranda made a strong contribution in the robustness area.
M. Cristina Miranda and M. Ivette Gomes gave contributions in the EVT framework.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

Research partially supported through the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia)

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.