Discussiones Mathematicae
Probability and Statistics 30 (2010) 21–33

EXTREMAL BEHAVIOUR OF STATIONARY PROCESSES: THE CALIBRATION TECHNIQUE IN THE EXTREMAL INDEX ESTIMATION

D. Prata Gomes

CMA and Mathematics Department, Faculty of Science and Technology
New University of Lisbon
Monte de Caparica 2829–516 Caparica, Portugal

e-mail: dsrp@fct.unl.pt

AND

Maria Manuela Neves

CEAUL and Mathematics Department, Instituto Superior de Agronomia,
Technical University of Lisbon
Tapada da Ajuda, 1349–017, Lisboa, Portugal

e-mail: manela@isa.utl.pt

Abstract

Classical extreme value methods were derived when the underlying process is assumed to be a sequence of independent random variables. However when observations are taken along the time and/or the space the independence is an unrealistic assumption. A parameter that arises in this situation, characterizing the degree of local dependence in the extremes of a stationary series, is the extremal index, θ. In several areas such as hydrology, telecommunications, finance and environment, for example, the dependence between successive observations is observed so large values tend to occur in clusters. The extremal index is a quantity which, in an intuitive way, allows one to characterise the relationship between the dependence structure of the data and their extremal behaviour. Several estimators have been studied in the literature, but they endure a problem that usually appears in semiparametric estimators - a strong dependence on the high level u_n, with an increasing bias and a decreasing variance as the threshold decreases.

The calibration technique (Scheffé, 1973) is here considered as a procedure of controlling the bias of an estimator. It also leads to the construction of confidence intervals for the extremal index. A simulation study was performed for a stationary sequence and two sets of stationary data are under study for applying this technique.

Keywords: extreme value, stationary sequences, extremal index, estimation, calibration technique.

2000 Mathematics Subject Classification: 62G32, 62G09, 62G05, 62J05.

1. Introduction and motivation

Extreme Value Analysis deals with events that are more extreme than any that have already been observed. Many studies deal with independent and identically distributed (i.i.d.) observations but in several situations the independence between consecutive observations is an unrealistic assumption. Extreme conditions often persist along several consecutive observations. In fact, most environmental datasets have a complex structure: they show a time-dependent variation and a short-term clustering, which are typical behaviour for extreme value data.

As an illustration of this situation let us consider two data sets:

Example 1. The data plotted in Figure 1 are the daily minimum temperatures, recorded to the nearest degrees Fahrenheit at Wooster, Ohio, during the years from 1983 to 1987. These data are freely available at http://cdiac.ornl.gov/epubs/ndp/ushcn/newushcn.html.

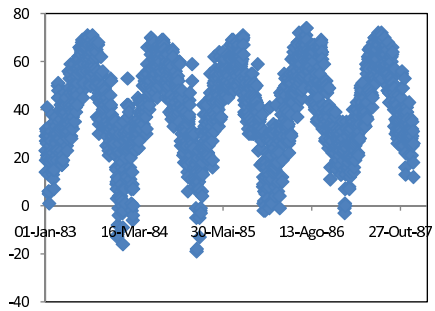

Figure 1. Daily minimum Wooster temperatures from 1983 to 1987.

From Figure 1 it is clear that:

- Large positive observations correspond to extreme cold conditions;
- There is a strong annual cycle in the data;
- An exceptionally cold winter day has quite different characteristics from an exceptionally cold summer day;
- A tendency for extreme values to occur close to one another is also evident.

There is evidence for a quadratic trend in the Wooster series (Coles *et al*, 1994). The series is approximately stationary over the winter (December to February months) during which all the observed annual minimum temperatures have occurred. We focus only on the winter months and present results under the assumption of stationarity throughout this season and over years, see Figure 2.

Example 2. Daily mean river levels from hydrometric station at Fraga, during the years from 1946/47 to 1996/97. Stationarity was achieved by considering only the data from November to February, according to what was also used in Example 1, see Figure 2.

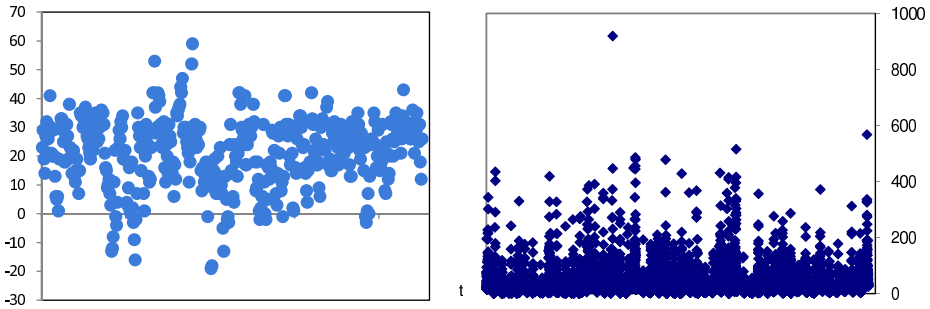

Figure 2. Daily minimum temperatures in December, January and February from 1983 to 1987 (left); Daily mean levels in November, December, January and February from 1946/47 to 1996/97 (right).

The classical extreme value theory gives conditions for the existence of normalizing sequences $\{a_n > 0\}$ and $\{b_n\}$ such that, for $u_n = a_n x + b_n$,

$$P\{M_n \leq u_n\} \to G(x)$$

as $n \to \infty$, where $G(\cdot)$ is a non-degenerate distribution function that belongs to one of the Gumbel, Fréchet and Weibull families that are usually termed as the *extreme value distributions*. The results were derived under the hypothesis of i.i.d random variables.

But, as can be seen in Figure 1 and 2, extreme events in the real world are often synonymous with clusters of large values. So, for a dependent structure, the exceedances over a high level tend to occur in clusters instead of isolated. This motivated the modification of the standard methods and the characterization of the extremes of stationary processes, the most natural generalization of a sequence of i.i.d random variables.

To study the extremal properties that occur in almost all series that appeared in applications, we consider only processes with any form of short range dependence for which, at long lags, the extremes are independent, i.e., processes that satisfy the $D(u_n)$ *condition* of Leadbetter *et al.* (1983).

A new parameter, θ, named the extremal index, appears now. It is roughly interpreted as the inverse of the mean of the cluster size. Now the limiting distributions for the independent and for the stationary sequences are not the same, unless $\theta = 1$.

Leadbetter *et al.* (1983) established the following result:

– Let $X_1, X_2, \ldots, X_n$ be a stationary process and $X_1^*, X_2^*, \ldots, X_n^*$ a sequence of independent variables with the same marginal distribution. Define

$$M_n = \max(X_1, X_2, \ldots, X_n) \qquad \text{and} \qquad M_n^* = \max(X_1^*, X_2^*, \ldots, X_n^*).$$

If the $D(u_n)$ *condition* holds with $u_n = a_n x + b_n$ for each x

$$P[(M_n^* - b_n)/a_n \leq x] \to G_1(x),$$

as $n \to \infty$ for normalizing sequences $\{a_n > 0\}$ and $\{b_n\}$ where G_1 is a non-degenerate distribution function, if and only if

$$P[(M_n - b_n)/a_n \leq x] \to G_2(x)$$

where $G_2(x) = G_1^\theta(x)$, for a constant θ such that $0 < \theta \leq 1$.

θ is the extremal index and G_2 is an extreme value distribution but with parameters different from those of G_1. If (μ, σ, γ) are the parameters of G_2 and $(\mu^*, \sigma^*, \gamma^*)$ are the parameters of G_1, their relationship is

$$\gamma = \gamma^*, \qquad \mu = \mu^* - \sigma^* \frac{1 - \theta^\gamma}{\gamma}, \qquad \sigma = \sigma^* \theta^\gamma.$$

The estimation of θ is then very important not only by its own importance but also because its influence in the other parameters.

Several estimators have appeared in literature motivated by different probabilistic interpretations of θ. Those estimators show a strong dependence on the high level u_n used in the exceedances definition. When the level u_n decreases the variance decreases but the bias increases.

The objective of this study is to show that the calibration technique can be used as a tool for reducing the bias of an estimator as well as for providing confidence intervals for the parameter. This is a preliminary study; some simulation results already obtained are encouraging, but more work is needed.

2. EXTREMAL INDEX ESTIMATION

One way of interpreting the extremal index of a stationary sequence is in terms of the tendency of the process to cluster at extreme levels. A rough interpretation of θ is

$$\theta = (\text{limiting mean cluster size})^{-1},$$

where the limiting is in the sense of clusters of exceedances of increasingly high thresholds.

The clusters of exceedances may be identified asymptotically as runs of consecutive exceedances and cluster sizes as run lengths. Under regularity conditions the conditional expected run length is approximately equal to $1/\theta$ (Nandagopalan, 1990). A suggestion was then to estimate θ by the reciprocal of the sample average run length.

Given a sequence of r.v.'s observations, $X_1, X_2, \ldots, X_n$, from a process which satisfies the $D(u_n)$ *condition*, where n is large and u_n is a high threshold, the most basic form of cluster identification (that does not require any knowledge of clustering characteristics of the process), led to a naive non-parametric estimator of θ, the *up-crossing estimator*, $\widehat{\theta}_n^{UC}(u_n)$, defined as:

$$\widehat{\Theta}_n^{UC} := \frac{\sum_{i=1}^{n-1} I\left(X_i \leq u_n < X_{i+1}\right)}{\sum_{i=1}^{n} I(X_i > u_n)}$$

(Nandagopalan, 1990 and Gomes, 1990, 1992, 1993).

The asymptotic properties of the *up-crossing* estimator were established in Nandagopalan (1990), Hsing (1993), Smith and Weissman (1994) and Weissman and Novak (1998), under several different conditions. Nandagopalan (1990) showed that, for random levels u_n, $\widehat{\Theta}_n^{UC}(u_n)$ is a weakly consistent estimator.

The asymptotic normality of $\widehat{\Theta}_n^{UC}(u_n)$ was derived in Hsing (1993) and Weissman and Novak (1998). The first moments of the estimator $\widehat{\Theta}_n^{UC}(u_n)$, the variance and the bias were derived in Hsing (1993).

Figure 3 shows simple path of the estimates obtained for both real data sets (Example 1 and 2). Since u_n is unknown, the corresponding order statistics is considered, $u_n := X_{k:n}$, where $X_{k:n}$ denotes the kth descending order statistics associated to the sample $(X_1, X_2, \ldots, X_n)$, $X_{1:n} \geq X_{2:n} \geq \ldots \geq X_{n:n}$.

$\widehat{\theta}_n^{UC}(k)$ is plotted for a range of thresholds chosen up to 20% of the sample length, where $u_n = X_{k:n}$, $\quad (5 \leq k \leq 0.2 \times n)$.

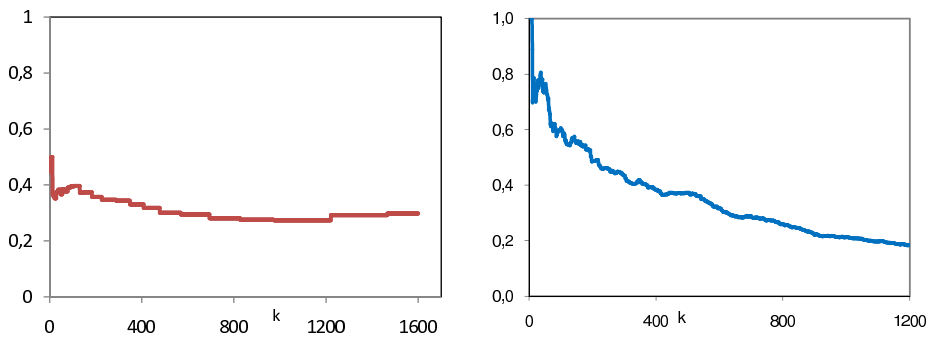

Figure 3. A simple path for the up-crossing estimates of the extremal index for several values of k: Daily minimum temperatures (left) and Daily mean river levels (right).

A problem that arises is how to choose the level u_n or k for obtaining the estimates. Intensive computational methods such as *Bootstrap*, *Jackknife* and *subsampling* have been considered to help in estimating a value for the level.

In this work the *a calibration technique* will be applied for obtaining estimates and/or confidence intervals for the extremal index.

3. The calibration technique and extremal index estimation

Calibration aims at estimating the values of a variable from values of a related variable. We have linear calibration when we assume there is a linear relationship between both variables. We then shall have

$$\widehat{\Theta}^{UC} = \beta_1 + \beta_2\theta,$$

where we measure the value of $\widehat{\Theta}^{UC}$ in order to estimate the values of θ. In the general case we would have

$$\widehat{\Theta}^{UC} = g(\theta),$$

with g known. To carry calibration we obtain values of $\widehat{\Theta}^{UC}$, $\widehat{\theta}^{UC}$, for given values of θ and adjust the function g.

In the case of linear calibration we are led to adjust linear regression of $\widehat{\Theta}^{UC}$ on θ, (see Andrews, 1970; Williams, 1969 and Scheffé, 1973).

In our case θ is the extremal index, $\widehat{\Theta}^{UC}$ the up-crossing estimator and we obtain values $\widehat{\theta}^{UC}$ at know values of θ, e.g., $\theta_1 = 0.1, \theta_2 = 0.2, \ldots, \theta_{n_\theta} = 0.9$ for each value of k ($k : \ u_n := X_{k:n}, \ \ X_{1:n} \geq X_{2:n} \geq \ldots \geq X_{n:n}$) to adjust the linear regression,

$$\widehat{\theta}^{UC} = \widehat{\beta}_1(k) + \widehat{\beta}_2(k)\theta, \tag{1}$$

where $\widehat{\beta}_1(k)$ and $\widehat{\beta}_2(k)$ are the least squares estimates for the coefficients.

Besides adjusting the linear regression we can obtain the corresponding confidence band, see Figure 4.

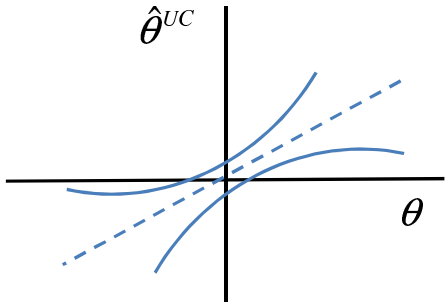

Figure 4. Confidence band.

The α level confidence band is bounded by

$$\widehat{\beta}_1(k) + \widehat{\beta}_2(k)\theta(-1)^h\widehat{\sigma}\left(c_1 + c_2\left(n_\theta^{-1} + k(\theta - \overline{\theta})^2\right)^{1/2}\right), \tag{2}$$

where $h = 1$ (lower), 2 (upper) and $\widehat{\sigma}$ is the estimate for the variance error. Constants c_1 and c_2 are calculated as follows:

– Let define

$$S_1 = n_\theta^{-1/2} \qquad \text{and} \qquad S_2 = \left(n_\theta^{-1} + kM^2\right)^{1/2},$$

where

$$M = \max\left\{\overline{\theta} - \theta^{(1)}, \theta^{(2)} - \overline{\theta}\right\}, \quad k = 1/\sum_{i=1}^{n}(\theta_{n_i} - \overline{\theta})^2, \quad \overline{\theta} = \sum_{i=1}^{n_\theta}\theta_{n_i}/n_\theta,$$

where $\theta^{(1)}$ and $\theta^{(2)}$ are the minimum and the maximum of θ_i, respectively.

After c has been obtained by entering Tables (see Scheffé, 1973) with $s_1 = S_1/z_\alpha$ and $s_2 = S_2/z_\alpha$, where z_α is the upper $\alpha/2$-point of the standard normal distribution, c_1 and c_2 are given by

$$c_1 = cz_\alpha\nu^{1/2}\left(\chi^2_{1-\delta}\right)^{-1/2} \qquad , \qquad c_2 = c\left(p\,\chi_\delta^{F_{p,\nu}}\right)^{1/2},$$

with $p = 2$, where $\chi_\delta^{F_{p,\nu}}$ is the upper δ-point of the F-distribution with p and ν df and $\chi^2_{1-\delta}$ is the lower δ-point of the chi-square distribution with ν df.

Figure 5 represents a graphical explanation of the calibration procedure for obtaining confidence bands.

We can now invert the equation (1)

$$
\begin{aligned}
\theta &= \left(\widehat{\theta}^{UC} - \widehat{\beta}_1(k)\right) / \widehat{\beta}_2(k) \\
&= a(k)\, \widehat{\theta}^{UC} + b(k)
\end{aligned} \tag{3}
$$

and the limits (2) are obtained as

$$
\theta_{UP} = \overline{\theta} + C^{-1} \left(\widehat{\beta}_2(k) D_1 + \widehat{\sigma} c_2 \left(n_\theta^{-1} C + k D_1^2 \right)^{1/2} \right), \tag{4}
$$

$$
\theta_{LOW} = \overline{\theta} + C^{-1} \left(\widehat{\beta}_2(k) D_2 - \widehat{\sigma} c_2 \left(n_\theta^{-1} C + k D_2^2 \right)^{1/2} \right), \tag{5}
$$

with

$$
\begin{aligned}
C &= \widehat{\beta_2}^2 - (\widehat{\sigma} c_2)^2 k, \\
D_1 &= D_1(\widehat{\theta}^{UC}) = \widehat{\theta}^{UC} - \widehat{\beta}_1(k) - \widehat{\beta}_2(k)\overline{\theta} + \widehat{\sigma} c_1, \\
D_2 &= D_2(\widehat{\theta}^{UC}) = \widehat{\theta}^{UC} - \widehat{\beta}_1(k) - \widehat{\beta}_2(k)\overline{\theta} - \widehat{\sigma} c_1.
\end{aligned} \tag{6}
$$

Expressions above give the bands of θ for the α level, once $\widehat{\theta}^{UC}$ is obtained.

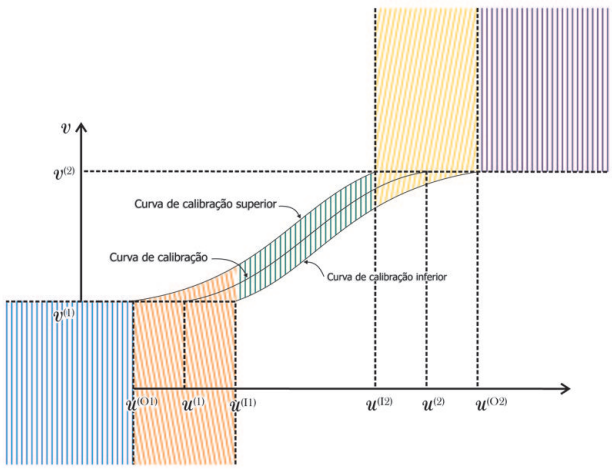

Figure 5. Schematic diagram of calibration chart ($u \equiv \widehat{\theta}^{UC}$ and $v \equiv \theta$).

To use equations (3), (4) and (5) we need the endpoints of the three calibration intervals:

– For $v = 1, 2$, $\widehat{\theta}^{UC(v)} = \beta_1 + \beta_2 \theta^{(v)}$;

– $\widehat{\theta}^{UC(I1)}(\widehat{\theta}^{UC(I2)})$ is found by putting $\theta = \theta^{(1)}$ $(\theta^{(2)})$ in (2) with $h = 2(1)$;

– $\widehat{\theta}^{UC(01)}(\widehat{\theta}^{UC(02)})$ was found by putting $\theta = \theta^{(1)}$ $(\theta^{(2)})$ in (2) with $h = 1(2)$.

Once $\widehat{\theta}^{UC}$ is obtained:

– For $\widehat{\theta}^{UC(1)} \leq \widehat{\theta}^{UC} \leq \widehat{\theta}^{UC(2)}$; the point estimate of θ is given by putting $\widehat{\theta}^{UC}$ in (3);

– For $\widehat{\theta}^{UC(01)} \leq \widehat{\theta}^{UC} \leq \widehat{\theta}^{UC(I2)}$ $(\widehat{\theta}^{UC(I1)} \leq \widehat{\theta}^{UC} \leq \widehat{\theta}^{UC(02)})$, the upper (lower) endpoint of the interval estimate for θ is given by putting $\widehat{\theta}^{UC}$ in (4) and (5).

4. Simulation study

In Prata Gomes (2008) several stationary processes were considered and the extremal index was obtained. For those models a simulation study applying the calibration procedure for estimating θ was carried out.

Here we are going to present the moving-maximum process, Deheuvels (1983), of order q, in short denoted by MMM(q), defined by

$$X_t = \max_{0 \leq i \leq q} Z_{t-i}, \quad t > q,$$

where Z_i are independent standard Fréchet random variables.

The extremal index exists and is $\theta = 1/(1+q)$.

For several values of q, and obviously θ, a sample of size $n = 1000$ is obtained from that model. A set of k (number of upper order statistics) values for which the simple path of $\widehat{\theta}^{UC}$ shows some stability was chosen.

For n_θ pairs $(\theta, \widehat{\theta}^{UC})$ the calibration technique was applied and the results (only for 3 values of k) are shown in Table 1.

Table 1. Real values and confidence intervals (CI) for θ.

MMM(q)						
θ	CI for $k = 103$		CI for $k = 104$		CI for $k = 105$	
0.1	0	0.158898	0	0.15864	0	0.157282
0.1111	0	0.170274	0	0.17	0	0.168642
0.125	0	0.184648	0	0.184352	0	0.182989
0.1429	0.100246	0.203397	0.100651	0.203071	0.101763	0.201694
0.1667	0.127416	0.228792	0.127762	0.22842	0.128815	0.227012
0.2	0.164797	0.265251	0.165067	0.264806	0.166048	0.263331
0.25	0.219072	0.321707	0.219245	0.321139	0.220158	0.319536
0.3333	0.304702	0.418379	0.30476	0.417593	0.305688	0.415743
0.5	0.468591	1	0.468491	1	0.469676	1

Given a data set and once fitted a model for which there exists the extremal index, the construction of a table based on the model, gives the possibility of obtaining an estimate of θ as well as obtaining a confidence interval.

5. Conclusions

As it was said this is a preliminary study on using calibration technique as an auxiliary tool for correcting an estimator from bias. We are now developing a computational procedure in R that can:

- fit a stationary model, for which the extremal index does exist, to a given data set;
- consider the inclusion of other estimators;
- consider possible non-linear calibration models.

Acknowledgments

The authors wish to thank Prof. J. Tiago Mexia for his helpful suggestion on using the calibration technique as a tool for improving the extremal index estimation, perhaps aiming to overcome some drawbacks that the classical estimators for that parameter reveal.

References

[1] F. Andrews, *Calibration and statistical inference*, J. Ann. Statist. Assoc. **65** (1970), 1233–1242.

[2] S.G. Coles, J.A. Tawn and R.L. Smith, *A sazonal Markov model for extremely low temperatures*, Environmetrics **5** (1994), 221–339.

[3] P. Deheuvels, *Point processes and multivariate extreme values*, J. Multivariate Analysis **13** (1983), 257–272.

[4] M.I. Gomes, *Statistical inference in an extremal markovian model*, COMPSTAT (1990), 257–262.

[5] M.I. Gomes, *Modelos extremais em esquemas de dependência*, I Congresso Ibero-Americano de Esdadistica e Investigacion Operativa (1992), 209–220.

[6] M.I. Gomes, *On the estimation of parameters of rare events in environmental time series*, *Statistics for the Environment* (1993), 226–241.

[7] T. Hsing, *Extremal index estimation for weakly dependent stationary sequence*, Ann. Statist **21** (1993), 2043–2071.

[8] M.R. Leadbetter, G. Lindgren and H. Rootzen, *Extremes and related properties of random sequences and series*, Springer Verlag. New York 1983.

[9] S. Nandagopalan, *Multivariate Extremes and Estimation of the Extremal Index*, Ph.D. Thesis. Techn. Report 315, Center for Stochastic Processes, Univ. North-Caroline 1990.

[10] D. Prata Gomes, Métodos computacionais na estimação pontual e intervalar do índice extremal. Tese de Doutoramento, Universidade Nova de Lisboa, Faculdade de Ciẹncias e Tecnologia 2008.

[11] H. Scheffé, *A statistical theory of calibration*, *Ann. Statist* **1** (1973), 1–37.

[12] R. Smith and I. Weissman, *Estimating the extremal index*, J. R. Statist. Soc. B, **56** (1994), 515–528.

[13] I. Weissman and S. Novak, *On blocks and runs estimators of the extremal index*, J. Statist. Plann. Inf. **66** (1998), 281–288.

[14] E.J. Williams, *Regression methods in calibration problems*, Proc. 37th Session, Bull. Int. Statist. Inst. **43** (1) (1969), 17–28.

Received 27 January 2010