On Clustering of High Values in Statistically Stationary Series

M.R. Leadbetter, I. Weissman
L. de Haan, H. Rootzen

Technical Report # 155

SIAM INSTITUTE FOR MATHEMATICS AND SOCIETY

Rand

HEALTH SCIENCES PROGRAM
RAND CORPORATION

DEPARTMENT OF APPLIED EARTH SCIENCES
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY

DEPARTMENT OF STATISTICS
UNIVERSITY OF BRITISH COLUMBIA
On Clustering of High Values in Statistically Stationary Series

by

M.R. Leadbetter, I. Weissman, L. de Haan, H. Rootzen

Technical Report 155

April 1990

Prepared in part under support to SIMS from the United States Environmental Protection Agency.

Department of Statistics, University of British Columbia, 2021 West Mall, Vancouver, British Columbia, CANADA V6T 1W5.
This is one of a series of technical reports produced under a Cooperative Agreement, "Study on Statistics and Environmental Factors in Health (Acid Deposition, Human Exposure, Quality Assurance, and Global Climate Change)" sponsored by the U.S. Environmental Protection Agency. The study is supported by a grant to SIMS (Societal Institute of the Mathematical Sciences) from the EPA. Cooperator members of the study are Rand Corporation, Stanford University (Departments of Statistics and Applied Earth Sciences), University of British Columbia (Department of Statistics), University of Washington (Department of Statistics), and SIMS.

SIMS is a nonprofit corporation founded in 1973 to "foster the application of mathematics to societal problems". To achieve this objective, SIMS has organized appropriate research and education studies and conferences, with mathematics and statistics playing a principal role.

Support for SIMS' programs has been forthcoming from private foundations, industry, and government agencies. Governance of SIMS resides with a Board of Directors from academe, government/industry, and the private sector. SIMS' research results have been recorded in a series of Technical Reports, which have been distributed to those requesting them. Most of the Technical Reports have been published in scientific journals.

SIMS Technical Reports are available from:

SIMS Publications
97 Parish Road South
New Canaan, Connecticut
06840

PRINCIPAL INVESTIGATORS

Donald L. Thomsen, Jr.
Societal Institute of the Mathematical Sciences

Naihua Duan
The Rand Corporation

Peter Guttorp
U. of Washington, Statistics

A. John Petkau,
U. of British Columbia, Statistics

Paul Switzer
Stanford University, Statistics

SIMS BOARD OF DIRECTORS

Joel E. Cohen
Rockefeller University
William T. Golden
New York, New York
Ettore F. Infante
University of Minnesota
William H. Kruskal
University of Chicago
Brockway McMillan
Bell Telephone Laboratories
Emanuel R. Piore
Rockefeller University

Howard Raiffa
Harvard University
Mina S. Rees
City University of New York
Fred S. Roberts
Rutgers University
T. Richard Savage
Yale University
Donald L. Thomsen, Jr.
Societal Institute of the Mathematical Sciences
Arthur C. Upton
New York University Medical Center
On Clustering of High Values in Statistically Stationary Series

M.R. Leadbetter*, I. Weissman, L. de Haan, H. Rootzen

SUMMARY

When dealing with concentrations of certain pollutants - extremes rather than averages are compared with air quality standards. Extreme value theory is therefore very important for the analysis of environmental data. For independent data - extreme concentrations are isolated events. But for dependent data, as in most environmental phenomena, extreme concentrations occur in clusters. So, for practical reasons, we are interested not just in the distribution of very large concentrations and how often they occur, but also, when they occur, how long they last. For the latter - the extremal index is very informative, because its reciprocal is the mean value of the cluster size for high values.

The present paper is the first paper that deals with the estimation of the extremal index. The results are applied to two data sets. One set is sulphate data from station 65 (Penn.State, PA). The sulphate measurements are part of a major study on acid rain in the north eastern part of the United States, from 1976 to 1982 (MAP35/PCN, Gentleman, Zidek and Olsen, 1985). The second data set consists of heights of tides measured at Den--Helder, Holland in winter periods from 1932 to 1985.

* The first author presented the paper in the Fourth International Meeting on Statistical Climatology, New Zealand, 1989.
On Clustering of High Values in Statistically Stationary Series

by

M.R. Leadbetter¹, Ishay Weissman², Laurens de Haan, Holger Rootzen

Abstract: For statistically dependent observed series as in many meteorological contexts, high values may tend to occur in clusters whose statistical properties are of interest e.g. in prediction of storm damage. This paper focuses on estimation of the extremal index for the observed series (defined as the inverse of the mean cluster size for high levels.) The importance of this parameter lies in both its role in the cluster structure, and in that it completely determines the modifications necessary to apply classical extreme value theory to statistically dependent (stationary) series. Two estimation procedures are described and applied to (a) acid precipitation data in rain episodes in Pennsylvania USA, (b) high tide records in Den Helder, Holland. Theory of T. Hsing [3] and S. Nandagopalan ([8]) is used for discussion of asymptotic properties of the estimators.

1 Introduction

The statistical properties of high values of observed (e.g. meteorological) series have traditionally been discussed by application of classical extreme value theory, assuming that the observations are well modeled as a sequence of statistically independent and identically distributed (i.i.d.) values. One finds that under such a model very high values tend to occur singly - i.e. are typically preceded and followed by "moderate" values. On the other hand when observations are serially dependent a high value will tend to induce further high values, so that these high values occur in clusters (cf. Figure 1).

![Figure 1](image)

1 Research supported by the Air Force Office of Scientific Research Contract No. F49620 85C 0144, and U.S.-Israel Binational Science Foundation Grant No. 86-00285.
2 Partially supported by the US Environmental Protection Agency through a co-operative research agreement with SIAM's Institute for Mathematics and Society (SIMS).
A study of the cluster structure has two important purposes:

(i) To model statistically the duration of "episodes" of high values of series of interest (such as wind speed, leading to storms; rainfall, leading to floods, etc.)

(ii) To provide appropriate modifications to the classical theory of extreme values to deal with dependence among the observations in discussing the largest and other high values in long time periods.

The question of precisely defining a "cluster" of high values will be taken up in Section 3. For the moment however we simply regard cluster of "exceedances" as the groups of consecutive observed values above a high level u (cf. Figure 1).

The statistical properties of the lengths of the clusters are of course basic to the discussion of cluster structure. In particular it turns out that under wide conditions, the mean cluster length (i.e. the average cluster size above a high level in a long time period) is an important parameter. We denote this parameter by $1/\theta$ and refer to $\theta$ as the "extremal index" of the observed series (cf. [5] [6]).

It may be shown ([6]) that the value of $\theta$ lies between zero and one. In the "classical" case for i.i.d. observations the high values tend to be isolated as noted so that clusters have size 1 and correspondingly $\theta = 1$. This is true also under serial dependence in a number of cases (e.g. many Gaussian models). On the other hand for highly dependent series the mean cluster size will exceed one and correspondingly one then has $\theta < 1$. Thus in understanding the clustering phenomenon it is important to estimate the value of $\theta$ - a topic considered in Section 3 using theory of [3] and [8], illustrated by practical examples in Section 4.

In addition to its relevance to cluster structure, the extremal index $\theta$ has important bearing on the extreme value theory for dependent observations. Specifically if the values observed are $X_1, X_2, \ldots, X_n$ and

$$M_n = \max(X_1, X_2, \ldots, X_n)$$

then the classical theory discusses the possible limit laws for $M_n$ of the form

$$(1.1) \quad P\{a_n(M_n - b_n) \leq z\} \to G(z) \quad \text{as } n \to \infty$$

under the assumption that the terms $X_i$, are i.i.d. In particular the theory shows that $G$ must (aside from linear normalizations) then be one of the three so-called "extreme value types" $G(z) = \exp(-e^{-z})$, $G(z) = \exp(-z^{-\alpha})$ $\alpha > 0$, $z > 0$, or $G(z) = \exp(-(-z)^\alpha)\alpha > 0$, $z < 0$.

If on the other hand observations $X_i$ are allowed to be statistically dependent but keep the same (marginal) distribution function, more recent theory (cf. [6]) shows that then under wide conditions the only change to (1.1) is the replacement of $G$ by $G^\theta$ . It is readily shown that $G^\theta$ is of the same extreme value type as $G$ itself, so that the classical criteria for the type of limit may be used, just as if the observations were i.i.d.

Thus knowledge of the value of the extremal index is all that is necessary in modifying the classical limiting distribution of the maximum to deal with a wide class of (stationary)
series of dependent observations. For other "order statistics" (e.g. second, third largest values etc.) the modification depends more intimately on the detailed statistical properties of the cluster sizes but $\theta$ still plays an important role.

Thus estimation of $\theta$ is also important for extreme value theory applications. It is perhaps of interest to note that the classical i.i.d. based theory has been traditionally used in applications where the data is clearly dependent. In fact this would perhaps be so for most naturally occurring phenomena, such as meteorological series. The success of extreme value theory in such cases stems from the fact noted above that the type of the limit law is unchanged by the introduction of dependence, though the constants may change. Since the constants are estimated, the dependence in the data is not necessarily apparent from the estimation process. Of course, an estimation of $\theta$ is really involved implicitly in such an analysis even if not carried out explicitly.

In Section 2 the specific theory surrounding the extremal index $\theta$ is briefly indicated (and may be omitted by a reader less interested in this background material), and Section 3 discusses estimation procedures for $\theta$. This is illustrated in Section 4 from measurements of acid levels in periods of rain, obtained from the acid deposition study [2], and by tide level series from the port of Den Helder, Holland.

2 The extremal index

A detailed theory regarding the clustering of high values may be found in [6] [4] and here we indicate only some main features.

If $X_1, X_2, \ldots$ are i.i.d. random variables with common distribution function (d.f.) $F$, and $\tau > 0$, let $u_n = u_n(\tau)$ be a family of "levels" such that

$$n(1 - F(u_n)) \to \tau$$

(2.1)

It is easily shown (since the d.f. for $M_n$ is $F^n$) that (2.1) is entirely equivalent to

$$P\{M_n \leq u_n\} \to e^{-\tau}.$$  

(2.2)

This simple equivalence underlies much of classical extreme value theory – for example (1.1) may be put in the form (2.2) by writing $u_n = x/a_n + b_n$, $G(x) = e^{-x}$. Further, use of the equivalence yields so-called "domain of attraction" criteria in terms of the tail 1-F(x) to determine which type of extreme value limit occurs in particular cases.

Typically $u_n(\tau)$ increases as $n$ increases and "high values" of the observations $X_i$ are interpreted as being those which exceed the level $u_n(\tau)$. As $n$ increases these "exceedances" become rarer and (after change of time scale) occur at points of time which become like a Poisson Process. The points of this Poisson Process are of course isolated and do not occur in clusters.

Assume now that $X_1, X_2, \ldots$ can be dependent, forming a statistically stationary sequence. Then under a mild dependence decay condition at large separations (cf. [6]) it may be shown that if (2.1) holds then

$$P(M_n \leq u_n) \to e^{-\theta \tau}$$

(2.3)

for a parameter $\theta$, $0 \leq \theta \leq 1$ which does not depend on $\tau$. The parameter $\theta$ is in fact the extremal index and it follows simply that
(2.4) \[ P\{a_n(M_n - b_n) \leq z\} \to (G(z))^\theta \]

if (1.1) holds when there is no dependence between the \( X_i \).

For the dependent sequence, exceedances of the level \( u_n(\tau) \) now occur in clusters (Figure 1) which, after time scale change, are situated at (approximately) Poisson points when the level is high. The exceedances then approximate more and more closely \( z \) so-called "Compound Poisson Process" where the Poisson positions of the clusters (occurring at the rate \( \theta \tau \)) are "compounded" or "marked" by the cluster sizes. As noted above the mean cluster size tends to 1/\( \theta \) as the level \( u_n \) increases.

Finally in this sketch of basic concepts we note that the distribution of the \( r \)th largest value \( M^{(r)}_n \) (\( r \)th "order statistic") may be obtained from that for \( M_n(= M^{(1)}_n) \) in either i.i.d. or dependent contexts. This is done by a simple consideration of the limiting Poisson (or Compound Poisson) nature of the exceedances by noting that the event \( \{M^{(r)}_n \leq u_n\} \) occurs precisely when there are less than \( r \) exceedances of \( u_n \) among \( X_1, X_2, ..., X_n \). In the i.i.d. case this leads to the limiting distribution

\[
(2.5) \quad P\{a_n(M^{(r)}_n - b_n) \leq z\} \to G(x) \sum_{s=0}^{r-1} (-\log G(x))^s/s! \]

when \( M_n \) has limiting distribution given by (1.1). In the dependent case this must be modified to read

\[
(2.6) \quad P\{a_n(M^{(r)}_n - b_n) \leq z\} \to G^\theta(x) \sum_{s=0}^{r-1} \sum_{i=0}^{s-1} (-\log G^\theta(x)^s/s!) \pi_{s,i} \]

where \( \pi_{s,i} \) are constants - specifically being for each \( s \), the \( s \)th convolution probabilities of the limiting cluster size distribution.

A comparison of (2.5) and (2.6) highlights the fact that the introduction of dependence makes much more of a difference for \( r \)th largest values (for \( r=2,3,... \)) than for \( r=1 \) when \( G \) is simply replaced by \( G^\theta \). This is especially important to note in applications where e.g. the second (or \( r \)th for \( r \geq 2 \)) largest values are periodically reported, rather than the maximum itself.

3 Estimation of \( \theta \)

In view of the interpretation of \( \theta^{-1} \) as limiting expected size of an exceedance cluster it is natural to estimate \( \theta \) simply as the reciprocal of the average cluster size in the given observations. That is one estimates \( \theta \) by

\[
\hat{\theta} = Z/\sum_{i=1}^{Z} Y_i
\]

where \( Y_i \) is the size of the \( i \)th cluster and \( Z \) is the total number of clusters. But the sum of cluster sizes is simply the total number \( N \) of exceedances of the level \( u_n \) so that

\[
(3.1) \quad \hat{\theta} = Z/N = \text{(number of clusters) / (number of exceedances)}.\]

Two practical questions arise, the first concerning the precise definition of a cluster. A natural definition would be that any group of consecutive exceedances forms a cluster. This "exceedance run" definition is simple but difficult to deal with theoretically and
has the disadvantage that two or more nearby but separated groups should perhaps be regarded as just one cluster. Another definition is obtained by dividing the $n$ observation times into smaller blocks of specified size and regarding all exceedances (if any) in any one block as forming a cluster (cf. [4]). Of course this can split single clusters and combine two or more into one. However it is simple to deal with theoretically, does not depend critically on the block size, and is asymptotically equivalent to the first definition under general conditions ([7]). In Section 4 we use both definitions, and vary the block sizes for the latter "block based" definition.

The second practical question is philosophically more serious but provides guidance concerning the proper procedure to be used. When the level $u_n$ satisfies (2.1) there are insufficient exceedances to give statistical "consistency" for the estimate $\hat{\theta}$. That is as $n$ increases the value of $\hat{\theta}$ does not necessarily converge appropriately to the value $\theta$. It has, however, been shown ([3] for the block-based estimator and [8] for the runs estimator) that consistency can be achieved by the use of somewhat lower levels - replacing (2.1) by

$$1 - F(u_n) = \gamma_n \to 0, \quad n\gamma_n \to \infty.$$  

F is typically unknown but $\gamma_n$ is approximately the proportion of observations exceeding $u_n$. Thus (3.2) suggests selecting a small value for $\gamma_n$ and choosing $u_n$ so that the proportion of observations exceeding $u_n$ is $\gamma_n$ (i.e. $u_n$ is the $(1- \gamma_n)$ percentile of the data). If possible $\gamma_n$ should be selected so that $n\gamma_n$ (approximately the number of values exceeding $u_n$) is large. For the first application in the next section we have (for $n = 504$) taken $\gamma_n = .05, .075$ and 0.1 so that $n\gamma_n = 25, 38$ and 50. Of course consistency is concerned with increasing values of $n$ rather than fixed sample sizes. However the guidance provided in the choice of $u_n$ does seem valuable.

It has also been shown in [3] under appropriate conditions, when $u_n$ is chosen by (3.2), that a block-based estimator $\hat{\theta}$ is approximately normal with mean $\theta$ and variance estimated by

$$\hat{\theta}(\hat{\theta}^2V - 1)/(n\gamma_n)$$

where (using the previous notation)

$$V = \sum^Z_i Y_i^2/Z = \text{average of squares of cluster sizes}.$$  

This applies when $n$ is large and $0 < \theta < 1$. Similar properties for the runs estimator are considered in [8].

This result enables one to provide confidence intervals or a hypothesis test regarding $\theta$, i.e. to determine the extent of clustering of high values in the observed data.

4 Applications

Our first example comes from acid levels ([2]) in periods of rain measured in Pennsylvania, USA. In the study one measurement of each quantity of interest was made in each of a series of 504 consecutive "rain events" and these are here taken as the basic data - focussing specifically on sulfate (sulfur VI). An excerpt from the data is shown in Table 1 in which the observed values greater than 93 $\mu$ mole/L are reported together with the numbers (in time sequence) of the rain periods in which they occurred.
Three choices were made for the level $u_a$ (viz. 93, 99, 110 $\mu$ mole/L) corresponding approximately to the 90, 92½, and 95 percentiles of the data, the numbers of exceedance values being respectively $N = 50, 34, 22$. The observed distributions of cluster sizes are given in Table 2 along with the estimate $\hat{\theta}$ and its estimated standard deviation (s.d.) $\hat{\sigma}$.

<table>
<thead>
<tr>
<th>Rain Period Sulfate</th>
<th>55</th>
<th>60</th>
<th>64</th>
<th>65</th>
<th>73</th>
<th>74</th>
<th>75</th>
<th>77</th>
<th>83</th>
<th>85</th>
<th>102</th>
<th>129</th>
<th>150</th>
</tr>
</thead>
<tbody>
<tr>
<td>140</td>
<td>95</td>
<td>150</td>
<td>110</td>
<td>99</td>
<td>130</td>
<td>120</td>
<td>110</td>
<td>110</td>
<td>120</td>
<td>130</td>
<td>98</td>
<td>110</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rain Period Sulfate</th>
<th>160</th>
<th>168</th>
<th>176</th>
<th>177</th>
<th>184</th>
<th>187</th>
<th>188</th>
<th>228</th>
<th>229</th>
<th>237</th>
<th>242</th>
<th>247</th>
<th>253</th>
</tr>
</thead>
<tbody>
<tr>
<td>95</td>
<td>110</td>
<td>130</td>
<td>130</td>
<td>150</td>
<td>110</td>
<td>110</td>
<td>280</td>
<td>160</td>
<td>96</td>
<td>95</td>
<td>98</td>
<td>95</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rain Period Sulfate</th>
<th>315</th>
<th>317</th>
<th>324</th>
<th>334</th>
<th>353</th>
<th>374</th>
<th>375</th>
<th>376</th>
<th>377</th>
<th>397</th>
<th>398</th>
<th>402</th>
<th>404</th>
</tr>
</thead>
<tbody>
<tr>
<td>94</td>
<td>110</td>
<td>200</td>
<td>330</td>
<td>100</td>
<td>110</td>
<td>110</td>
<td>95</td>
<td>99</td>
<td>340</td>
<td>99</td>
<td>140</td>
<td>95</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rain Period Sulfate</th>
<th>405</th>
<th>415</th>
<th>439</th>
<th>452</th>
<th>453</th>
<th>455</th>
<th>469</th>
<th>470</th>
<th>476</th>
<th>480</th>
<th>485</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>150</td>
<td>100</td>
<td>100</td>
<td>190</td>
<td>130</td>
<td>130</td>
<td>105</td>
<td>110</td>
<td>150</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1

Rain periods and sulfate values ($\mu$ mole/L) exceeding 93 $\mu$ mole/L

In the above table runs occur where two or more rain periods with consecutive index values appear. These are indicated by underlining.

<table>
<thead>
<tr>
<th>Cluster Definition</th>
<th>Frequency of cluster size</th>
<th>$u_a$</th>
<th>$N$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>$Z$</th>
<th>$\hat{\theta}$</th>
<th>$V$</th>
<th>s.d.$\hat{\sigma}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Runs</td>
<td></td>
<td>93</td>
<td>50</td>
<td>27</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>37</td>
<td>.74</td>
<td>2.27</td>
<td>.060</td>
</tr>
<tr>
<td>&quot;</td>
<td></td>
<td>99</td>
<td>38</td>
<td>22</td>
<td>8</td>
<td>-</td>
<td>-</td>
<td>30</td>
<td>.79</td>
<td>1.80</td>
<td>.051</td>
</tr>
<tr>
<td>&quot;</td>
<td></td>
<td>110</td>
<td>22</td>
<td>14</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>18</td>
<td>.82</td>
<td>1.67</td>
<td>.067</td>
</tr>
<tr>
<td>Block (5)*</td>
<td></td>
<td>93</td>
<td>50</td>
<td>21</td>
<td>10</td>
<td>3</td>
<td>-</td>
<td>34</td>
<td>.68</td>
<td>2.59</td>
<td>.052</td>
</tr>
<tr>
<td>&quot;</td>
<td></td>
<td>99</td>
<td>38</td>
<td>15</td>
<td>10</td>
<td>1</td>
<td>-</td>
<td>26</td>
<td>.68</td>
<td>2.46</td>
<td>.050</td>
</tr>
<tr>
<td>&quot;</td>
<td></td>
<td>110</td>
<td>22</td>
<td>12</td>
<td>5</td>
<td>-</td>
<td>-</td>
<td>17</td>
<td>.77</td>
<td>1.88</td>
<td>.064</td>
</tr>
<tr>
<td>Block (10)*</td>
<td></td>
<td>93</td>
<td>50</td>
<td>11</td>
<td>11</td>
<td>3</td>
<td>2</td>
<td>27</td>
<td>.54</td>
<td>4.22</td>
<td>.050</td>
</tr>
<tr>
<td>&quot;</td>
<td></td>
<td>99</td>
<td>38</td>
<td>13</td>
<td>8</td>
<td>3</td>
<td>-</td>
<td>24</td>
<td>.63</td>
<td>3.00</td>
<td>.056</td>
</tr>
<tr>
<td>&quot;</td>
<td></td>
<td>110</td>
<td>22</td>
<td>12</td>
<td>5</td>
<td>-</td>
<td>-</td>
<td>17</td>
<td>.77</td>
<td>1.88</td>
<td>.064</td>
</tr>
</tbody>
</table>

* Figure in parenthesis indicates block length.

Table 2

Cluster size and estimates ($\hat{\theta}$) for the extremal index ($\theta$) with estimated standard deviation ($\hat{\sigma}$).
Confidence intervals for \( \theta \) are complicated by the dependence of the variance of \( \hat{\theta} \) on \( \theta \) and the fact that \( \hat{\theta} \) is only approximately unbiased. However, calculations show that e.g. the classical 95\% interval \((\hat{\theta} \pm 2\sigma)\) is likely to give a reasonable approximation to the true interval (and will be conservative in cases where the bias is small). For example for the runs estimate based on the \( u_n = 99 \) level, this 95\% confidence interval for \( \theta \) is (.68, .90).

Such elementary comparisons of differences between the estimated values with standard deviations show no statistically significant differences between the different procedures but all estimates \( \hat{\theta} \) are significantly less than one, so that it may be safely concluded that \( \theta < 1 \) i.e. clustering is present. This is confirmed by a test of the null hypothesis that the data is i.i.d. (which implies \( \theta = 1 \)) based on the number of runs one then expects to find in the data. This avoids the bias problem referred to above. Specifically under the hypothesis of i.i.d. data, conditionally on \( N \) one has (\cite{11})

\[
\mathcal{E}\hat{\theta} = \mathcal{E} \frac{Z}{N} = \left( \frac{n - N + 1}{n} \right)
\]

\[
\sigma^2 = \text{var} \hat{\theta} = \frac{1}{2} \frac{(n - N)[2N(n - N) - n]}{(n^3 N)}.
\]

For the three levels considered the estimated values, means and variances are:

<table>
<thead>
<tr>
<th>( u_n )</th>
<th>( \hat{\theta} )</th>
<th>Expected value under s.d. under independence</th>
<th>independence ( \mathcal{E}\hat{\theta} )</th>
<th>( \sigma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>93</td>
<td>.74</td>
<td>.90</td>
<td>.040</td>
<td></td>
</tr>
<tr>
<td>99</td>
<td>.79</td>
<td>.93</td>
<td>.040</td>
<td></td>
</tr>
<tr>
<td>110</td>
<td>.82</td>
<td>.96</td>
<td>.042</td>
<td></td>
</tr>
</tbody>
</table>

Since the estimated values differ from the expected values for i.i.d. values by approximately four standard deviations one can clearly reject the i.i.d. hypothesis — again providing strong evidence that \( \theta < 1 \), i.e. clustering is present.

One cautionary (and instructive) note concerns the increase of the estimates \( \hat{\theta} \) as the level increases which is apparent (even if not statistically significant) in all three procedures. For dependent, but non-clustering data (\( \theta = 1 \)) one would certainly expect the clusters of size greater than one to become “thinned out” more and more as the level increases, leading to higher values of \( \hat{\theta} \). Hence some modest caution needs to be attached to our (statistically highly significant) conclusion \( \theta < 1 \). The observed consistent increase in \( \hat{\theta} \) with level suggests that one’s procedure in practice should be to try higher levels until the value of \( \hat{\theta} \) either stabilizes at some value less than one, or else tends to one, or until the level is so high that exceedances are very few, and the estimated standard deviation \( \hat{\sigma} \) of \( \hat{\theta} \) thus shows a substantial increase. That point appears to have been reached in the above table with an increase in \( \hat{\sigma} \) of between 11\% and 28\% from the 93 to the 110 level.

The final point to be made regarding this data is that the block based estimates tend to be somewhat smaller than for runs, and the larger block size gives the smaller estimates (as would be expected). For “smoothly varying” observations the runs definition seems preferable but a choice should be made from a visual inspection of the data. As noted the estimators are asymptotically equivalent (\cite{8}) so that their closeness to each other is
another indication as to whether the level being used is sufficiently high for the given data set.

The second study concerns the heights of tides measured at Den Helder, Holland in winter periods from 1932-1985.

Four levels were selected - corresponding to the 90, 92\(\frac{1}{4}\), 95, 97\(\frac{1}{4}\) percentiles (actual levels 106, 114, 128, 149 cm.). The estimates of \(\theta\) and their standard deviations are given in Table 3 for estimators based on runs and on blocks of length 4,5,10.

<table>
<thead>
<tr>
<th>Cluster Definition</th>
<th>(u_n) (cm.)</th>
<th>N</th>
<th>(Z)</th>
<th>(\hat{\theta})</th>
<th>s.d.(\hat{\theta})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Runs</td>
<td>106</td>
<td>947</td>
<td>602</td>
<td>.64</td>
<td>.028</td>
</tr>
<tr>
<td></td>
<td>114</td>
<td>722</td>
<td>465</td>
<td>.64</td>
<td>.032</td>
</tr>
<tr>
<td></td>
<td>128</td>
<td>470</td>
<td>324</td>
<td>.69</td>
<td>.036</td>
</tr>
<tr>
<td></td>
<td>149</td>
<td>233</td>
<td>167</td>
<td>.72</td>
<td>.046</td>
</tr>
<tr>
<td>Block (4)*</td>
<td>106</td>
<td>947</td>
<td>616</td>
<td>.65</td>
<td>.025</td>
</tr>
<tr>
<td></td>
<td>114</td>
<td>722</td>
<td>483</td>
<td>.67</td>
<td>.028</td>
</tr>
<tr>
<td></td>
<td>128</td>
<td>470</td>
<td>335</td>
<td>.71</td>
<td>.033</td>
</tr>
<tr>
<td></td>
<td>149</td>
<td>233</td>
<td>175</td>
<td>.75</td>
<td>.043</td>
</tr>
<tr>
<td>Block (5)*</td>
<td>106</td>
<td>947</td>
<td>572</td>
<td>.60</td>
<td>.027</td>
</tr>
<tr>
<td></td>
<td>114</td>
<td>722</td>
<td>456</td>
<td>.63</td>
<td>.030</td>
</tr>
<tr>
<td></td>
<td>128</td>
<td>470</td>
<td>317</td>
<td>.67</td>
<td>.035</td>
</tr>
<tr>
<td></td>
<td>149</td>
<td>233</td>
<td>173</td>
<td>.74</td>
<td>.044</td>
</tr>
<tr>
<td>Block (10)*</td>
<td>106</td>
<td>947</td>
<td>440</td>
<td>.46</td>
<td>.032</td>
</tr>
<tr>
<td></td>
<td>114</td>
<td>722</td>
<td>361</td>
<td>.50</td>
<td>.036</td>
</tr>
<tr>
<td></td>
<td>128</td>
<td>470</td>
<td>258</td>
<td>.55</td>
<td>.041</td>
</tr>
<tr>
<td></td>
<td>149</td>
<td>233</td>
<td>147</td>
<td>.63</td>
<td>.055</td>
</tr>
</tbody>
</table>

*Figure in parenthesis indicates block length.

Table 3

Estimates \(\hat{\theta}\) for the extremal index \((\theta)\) with estimated standard deviation \((\hat{\sigma})\)

Again in this data there is clear evidence that \(\theta < 1\), indeed some values of \(\hat{\theta}\) differ from 1 by as much as fourteen standard deviations. Again also the estimates \(\hat{\theta}\) are increasing with increasing level and do not show signs of stabilizing with the levels used. However the more than 60\% increase in standard deviation for \(\hat{\theta}\) from low to high levels suggests that one may not wish to use yet higher levels in view of the attrition of exceedances. In this application the shortest block size used actually gives estimates which are larger than (and quite close to) those for runs. Again, unless the observations tend to oscillate rapidly when above a high level, the use of the runs estimator seems preferable in avoiding the ambiguity of block size choices.

In summary, the two data sets analyzed both have extremal index estimates which are significantly less than one, demonstrating the presence of clustering of high values in each
case. The differences between the "runs" and "block based" estimators are not statistically significant; the former may be preferred in practice because of its unambiguous definition, especially for "smoothly varying" observed series, though the latter is theoretically more tractable. Finally we note that it is possible in principle to estimate the limiting cluster size distribution by similar means, but this requires yet larger amounts of data to obtain satisfactory statistical properties for the estimates.

Acknowledgement.

We are very grateful to S. Nandagopalan and Neela Srinivasan for their help in analysis of the tide data.

References


