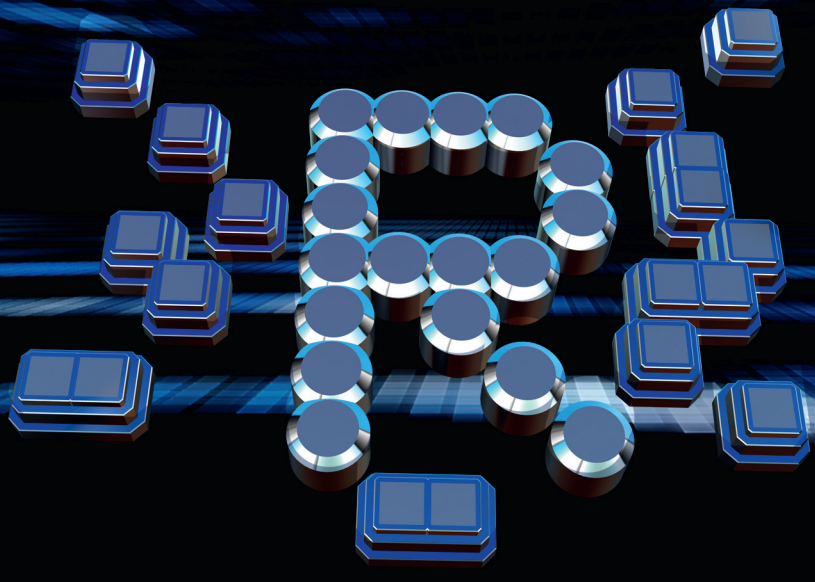
WILEY

Extreme Values

Statistical Analysis Using R



Lee Fawcett • David Walshaw

WILEY SERIES IN PROBABILITY AND STATISTICS

3.4 Generalisation to the r–largest order statistics

Given a sequence of IID random variables $X_1, X_2, \ldots$ we have shown how the generalised extreme value distribution (expression 6) can be used to model the set of normalised annual maxima. This approach is highly inefficient since all but the maximum in each year (or block) are discarded – other observations which could be considered extreme are simply thrown away because they are not as extreme as the maximum value in that year. A generalisation of the result in expression (6) attempts to overcome this, by incorporating the largest r order statistics from each year, where r is any positive integer. If we denote the r largest order statistics in an IID sample by $M^{(1)} \geq M^{(2)} \geq \ldots \geq M^{(r)}$, for $r \geq 1$, then the technique here is to obtain the limiting joint distribution of

$$\left(\frac{M_n^{(1)} - b_n}{a_n}, \frac{M_n^{(2)} - b_n}{a_n}, \ldots, \frac{M_n^{(r)} - b_n}{a_n} \right).$$

It can be shown that the complete class of limiting non–degenerate joint distributions is in fact given by the probability density function

$$\begin{aligned} f(x_1, x_2, \ldots, x_r; \mu, \sigma, \xi) &= \sigma^{-r} \exp \left\{ - \left[1 + \xi \left(\frac{x^{(r)} - \mu}{\sigma} \right) \right]_+^{-1/\xi} \right. \\ &\quad \left. - \left(1 + \frac{1}{\xi} \right) \sum_{j=1}^{r} \log \left[1 + \xi \left(\frac{x^{(j)} - \mu}{\sigma} \right) \right]_+ \right\} \end{aligned} \tag{14}$$

for $j = 1, \ldots, r$. As before, the case $\xi = 0$ is taken as the limit as $\xi \to 0$ in (14), to give

$$\begin{aligned} f(x_1, x_2, \ldots, x_r; \mu, \sigma, \xi) &= \sigma^{-r} \exp \left\{ - \exp \left[- \left(\frac{x^{(r)} - \mu}{\sigma} \right) \right] \right. \\ &\quad \left. - \sum_{j=1}^{r} \left(\frac{x^{(j)} - \mu}{\sigma} \right) \right\}. \end{aligned}$$

Obviously, the case where $r = 1$ is equivalent to the annual maxima approach, for which the GEV holds as the limiting distribution. The increased precision over the traditional annual maxima approach (due to more extremes being incorporated into the analysis) has obvious appeal; however, Smith (1986) shows that as r increases, the rate of convergence to the limiting distribution decreases rapidly, and so the number of order statistics to include must be considered carefully. Such methods must also take account of serial correlation, and are vulnerable to the effects of seasonal variation. Papers in the Journal of Hydrology by Smith (1986) and Tawn (1988b) illustrate the use of r–largest methods; in Section 3.5.6, we apply the technique to the 10 largest sea levels observed each year in Venice.

4 The basic model for threshold exceedances: the Generalised Pareto distribution

4.1 History and theoretical motivation

Generalisation of the classical annual maxima approach for modelling extreme values to the r–largest order statistics method was discussed in Section 3.4, the main advantage being the inclusion of more extreme data in the analysis, leading to more precise inferences on the extremal behaviour of the process under study. However, only the r largest values within each year (or block) are used, and any other extremes discarded. In the present chapter, we discuss an approach which aims to include *all* extreme values in the analysis, extreme in the sense that they exceed some pre–determined high level, or *threshold*.

Threshold methods developed rapidly during the 1980s, culminating in the Davison and Smith (1990) paper which applied these techniques to an environmental data set that displayed short–term serial dependence and seasonal variation (see Chapter 5 for more detail on such modelling issues). Since then, threshold methods have become the standard tool for many practitioners involved in modelling extreme values. Relative to the traditional annual maxima and r–largest approaches, threshold methods attempt to maximise efficiency by using *all* extreme values in their analysis; however, as we shall discuss in Chapter 5 (and Section 4.1.2 below), the fact that we use *all* extremes can itself create problems (though, as Davison and Smith (1990) illustrate, pragmatic solutions to these problems can be found).

4.1.1 The generalised Pareto distribution

Ignoring, for now, the practical implications of using *all* our extreme data, again consider a sequence of IID random variables $X_1, X_2, \ldots, X_n$ with common distribution function F. Then for a sufficiently large threshold u, the distribution of $(X - u)$, conditional on $X > u$, is approximately

$$G(y; \bar{\sigma}, \xi) \;=\; 1 - \left(1 + \frac{\xi y}{\bar{\sigma}}\right)_+^{-1/\xi}, \tag{15}$$

where $\bar{\sigma}$ (> 0) and ξ $(-\infty < \xi < \infty)$ are scale and shape parameters respectively. The shape parameter ξ in the GPD takes exactly the same value as that for the corresponding GEV distribution; the scale parameter $\bar{\sigma}$ is equal to $\sigma + \xi(u - \mu)$, where σ and μ are the scale and location parameters (respectively) in the corresponding GEV distribution (see expression 6). Specifically, G is defined on $0 < y < \infty$ if $\xi > 0$, and $0 < y < -\bar{\sigma}/\xi$ if $\xi \leq 0$. The case $\xi = 0$ is interpreted as the limit $\xi \to 0$, and is the exponential distribution with rate $1/\bar{\sigma}$. This is known as the *generalised Pareto distribution*, or GPD. The GPD is a limiting distribution for excesses over thresholds if, and only if, the parent distribution lies in the domain of attraction of one of the three extreme value distributions (see Theorem 3.1). However, since the limiting distribution of sample maxima follows one of the distributions given in Theorem 3.1 no matter what the parent distribution, the GPD is the *only* non–degenerate limiting distribution for excesses over thresholds of IID sequences. Until this point we have used the notation $\bar{\sigma}$ to denote the scale parameter for the GPD, so as to distinguish it from the corresponding parameter of the GEV distribution. For notational

convenience we now drop this distinction, using σ to denote the scale parameter within either family.

The GPD yields several important properties. One first of these, known as the 'threshold stability property', is that if $(X - u_0)$ follows a generalised Pareto distribution (conditional on $X > u_0$), then $(X - u)$ also follows a generalised Pareto distribution for any $u > u_0$. In other words, once a suitably high enough threshold has been found such that the GPD may be assumed a valid model for excesses over that threshold, then the GPD holds for excesses over any higher threshold too. Another property unique to the GPD is that if $N \sim$ Poisson, and $X_1, \ldots, X_N$ are IID random variables following a GPD, then $\max\{X_1, \ldots, X_N\}$ has the GEV distribution. As will be demonstrated, the threshold stability property can be exploited in graphical procedures for threshold selection and assessing the fit of the GPD. The second property suggests that, if the exceedances of u occur as a Poisson process with threshold excesses which are IID and generalised Pareto distributed, then the maximum value over any block size has a generalised extreme value distribution. Thus, if we assume extreme events occur over time as a Poisson process, models which fit the GEV to sets of block maxima are consistent with models which fit the GPD to sets of threshold excesses.

As with fitting the GEV, most practitioners use numerical maximum likelihood estimation to fit (15) to a set of threshold excesses. For $-1 < \xi \leq -0.5$, maximum likelihood estimators exist (in large enough samples), though in general (as before) do not possess all of the standard asymptotic properties; when $\xi \leq -1$, maximum likelihood estimators do not, in general, exist. For $\xi > -0.5$, maximum likelihood estimators are asymptotically normal and efficient (Smith, 1985). Luckily, in most environmental applications, values of $\xi \leq -0.5$ are rare, but do occur from time to time. Again, the Bayesian methodology provides a framework within which this problem can be avoided. However, in a Bayesian setting, use of the GPD(σ, ξ) model may be restrictive since the scale parameter σ is dependent on the choice of threshold level u; an uninformative prior for σ then becomes informative at higher thresholds. To overcome this, the GPD can be reparameterised with scale and shape parameters $\tilde{\sigma}$ and ξ (respectively, ξ remaining unchanged), where $\tilde{\sigma} = \sigma - \xi u$. Under this parameterisation, both parameters are threshold–independent. As demonstrated in Chapter 3 for the GEV, estimates of extreme quantiles can be obtained through inversion of (15).

4.1.2 When should we model threshold exceedances?

Since extremes are – by their very nature – scarce, any modelling approach that increases precision has obvious appeal. The r–largest approach attempts to increase precision through the inclusion of more data, but could still be considered rather wasteful. The aim of threshold methods is to maximise precision by analysing *all* extremes. However, we might not have access to the entire dataset; it might be the case that we only have the set of derived block maxima – in which case the most appropriate method of analysis would be to fit the GEV to our data directly.

As we shall discuss in Section 4.3 of this Chapter, and Chapter 5, there might also be various modelling issues to address arising as a direct consequence of using threshold exceedances, not least the problem of short term temporal dependence. Shown in Figure

9 is the autocorrelation function, and partial autocorrelation function, for the entire series of daily rainfall measurements discussed in Chapter 3 (recall that we analysed the set of rainfall annual maxima in Chapter 3). These plots were produced using the commands:

```
> acf(rain)
> pacf(rain)
```

where, as in Chapter 3, the series of daily rainfall measurements are stored in the vector `rain`.

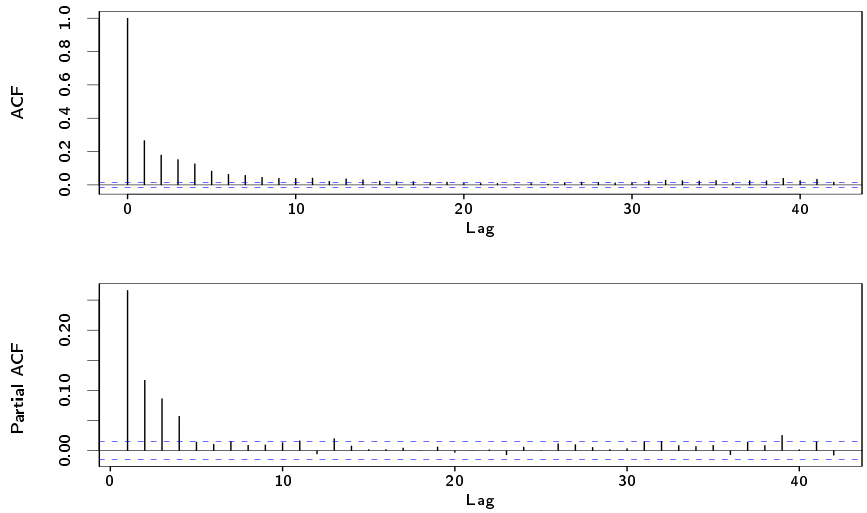


Figure 9: Autocorrelation function, and partial autocorrelation function, for the rainfall data

Both plots in Figure 9 indicate the presence of short term autocorrelation in the series, particularly the plot of partial autocorrelations which eliminates the effect of intermediate autocorrelations. For example, a process that is truly first–order Markov might have significant autocorrelations at lag 2 and beyond owing wholly to the temporal effects induced by the strength of dependence between successive observations. However, the partial autocorrelation function would provide a better indication of the order of dependence, showing only the first partial autocorrelation as significant. For the rainfall data, we see significance in the partial autocorrelation function up to lag 4, suggesting genuine short–term temporal dependence. The GPD in the form it is given in equation (15) assumes our series is IID, which is clearly not the case here. Though (as we shall see in Chapter 5) various techniques have been developed to circumvent the problem of temporal dependence, it is not always obvious how to implement these and parameter estimates can be sensitive to the technique chosen. When temporal dependence and other modelling issues might arise as a direct consequence of using all threshold exceedances, it might be considered preferable to work with a set of block maxima (or perhaps a set of "cluster maxima" – see Section 4.3.3).

4.1.3 How is the GPD used?

Once we have identified our set of threshold exceedances, a typical application would fit the model in (15) via maximum likelihood (perhaps) to obtain estimates of the scale

and shape parameters σ and ξ (the problems associated with maximum likelihood for particular values of ξ, for example, are discussed in Section 3.1.4 and also apply here). We can then use our estimates of σ and ξ to obtain estimates of return levels by inversion of (15). For example, suppose that a GPD with parameters σ and ξ is a suitable model for exceedances of a threshold u by a variable X, i.e. for $x > u$,

$$\Pr(X > x|X > u) = \left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]_+^{-1/\xi}.$$

Then

$$\Pr(X > x) = \lambda_u \left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]_+^{-1/\xi},$$

where $\lambda_u = \Pr(X > u)$. Hence, the level x_t that is exceeded on average once every t observations is the solution of

$$\lambda_u \left[1 + \xi\left(\frac{x_t-u}{\sigma}\right)\right]_+^{-1/\xi} = \frac{1}{t}.$$

Rearranging, we get

$$x_t = u + \frac{\sigma}{\xi}[(t\lambda_u)^\xi - 1],$$

provided t is sufficiently large to ensure that $x_t > u$, and $\xi \neq 0$. If $\xi = 0$, we have the exponential case, and so

$$x_t = u + \sigma\log(t\lambda_u),$$

again provided t is sufficiently large. By construction, x_t is the t–observation return level; however, it is often more convenient to give return levels on an annual scale, so that the *r–year* return level is the level expected to be exceeded once every r *years*. If there are n_y observations per year, this corresponds to the t–observation return level with $t = r \times n_y$. Hence, the r–year return level q_r is defined by

$$q_r = u + \frac{\sigma}{\xi}\left[(rn_y\lambda_u)^\xi - 1\right], \tag{16}$$

unless $\xi = 0$, in which case

$$q_r = u + \sigma\log(rn_y\lambda_u). \tag{17}$$

We can then estimate the r–year return level q_r by substituting our estimates of σ and ξ (say $\hat{\sigma}$ and $\hat{\xi}$) into equation 16 (or equation 17 when $\xi = 0$); λ_u can be estimated empirically as the proportion of threshold exceedances, and u is the threshold chosen to identify extremes (see Section 4.2.1). The usual approach for estimating standard errors for the GPD parameters can be used (i.e. inversion of the expected information matrix) and, as in Chapter 3, the delta method can be used to obtain estimated standard errors for return levels. We can also use profile likelihood to estimate more appropriate confidence intervals for return levels.

4.2 Simple case study

In this Section, we demonstrate a simple application of the GPD to a set of threshold exceedances. We illustrate this by using a set of threshold exceedances from the full rainfall series used in Chapter 3. Recall that this can be loaded into **R** by first installing the ismev package:

```
> library(ismev)
```

and then typing:

```
> data(rain)
```

Typing

```
> help(rain)
```

gives a description of the rainfall series. In this Section, we will

- exploit the threshold stability property of the GPD to produce a graphical tool for identifying a suitable threshold u;
- maximise the log–likelihood function for the GPD in **R**;
- use **R** to obtain the expected information matrix, and then invert this to obtain the estimated variance–covariance matrix for $(\hat{\lambda}_u, \hat{\sigma}, \hat{\xi})^T$;
- use the fitted values of the GPD parameters to estimate some return levels q_r, along with standard errors for these;
- use **R** to plot the profile log–likelihood for some return levels and obtain profile likelihood confidence intervals;
- use **R** to check the goodness–of–fit of the GPD to our series of threshold exceedances,

4.2.1 Identifying a suitable threshold: the mean residual life plot

In a mean residual life (MRL) plot we make use of the fact that if the GPD is the correct model for all exceedances x_i above some high threshold u_0, then the *mean excess*, i.e. the mean value of $(x_i - u)$, plotted against $u > u_0$, should give a linear plot. This is because $E[X_i - u_0]$ is a linear function of $u : u > u_0$. By producing such a plot for values of u starting at zero, we can select reasonable candidate values for u_0. In **R**, the following code sets up a vector of possible thresholds, starting at zero and going up to the maximum value in our dataset in steps of 0.1:

```
> u<-seq(0,max(rain),0.1)
```

The vector `x` will now be set up to take the corresponding values for the mean excess over each value in `u`:

```
> x<-vector('numeric', length(u))
```

Then the following code computes the mean excess for each value in `u` and stores it in `x`:

```
> for(i in 1:length(x))
  {
    threshold.exceedances<-rain[rain>u[i]]
    x[i]<-mean(threshold.exceedances-u[i])
  }
```

The MRL plot is then produced using the following code, giving the plot in Figure 10:

```
> plot(x~u,type='l', main='MRL plot',ylab='mean excess')
```

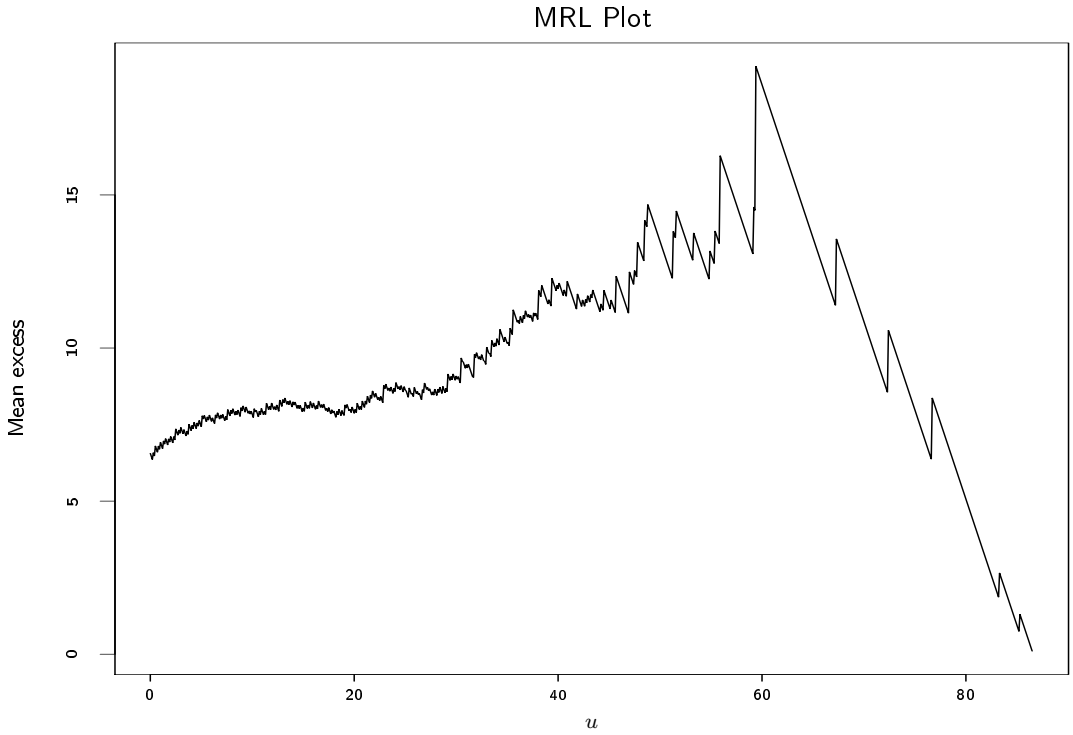


Figure 10: Mean residual life plot for the rainfall data

Though interpretation of these plots can be subjective, linearity in Figure 10 might be suggested at about $u_0 = 30$mm (information in the far right–hand–side of these plots is unreliable; here, variability is high due to the limited amount of data above such high thresholds). Using $u_0 = 30$ as our threshold for identifying extremes, we can then obtain our set of threshold exceedances for modelling with the generalised Pareto distribution:

```
> above.threshold<-rain[rain>30]
> threshold.exceedances<-above.threshold-30
```

We can look at our set of threshold exceedances by typing:

```
> threshold.exceedances
  [1]  1.8  2.5  1.8 14.5  0.5 13.2  5.6  8.1  2.0  1.8  3.0  9.1  0.5  1.8  2.3
 [16]  3.0  0.5  2.5 18.5  5.3 10.6  0.5  4.3  2.8  0.5 15.7  1.8  3.5  3.5  1.8
 [31]  4.8  5.3  7.8 46.7  2.3  4.0  3.8  6.6  0.5 15.7 56.6  5.6 17.8 17.5  4.3
 [46] 18.5  0.7 13.4 29.4  5.1 23.3  3.5  0.5  0.2 10.9 12.7 53.3 24.9 29.2  1.8
```

```
 [61]  7.3  2.5  4.0 37.3  1.2  0.2  6.1  6.8  8.4  1.0  3.3 17.0  2.0  3.0  8.1
 [76]  0.5 42.4  4.3  7.1  3.0 10.9  9.9 17.0  6.3  0.5  0.5 25.9  1.8 21.3 55.3
 [91] 11.9  0.5  3.0  5.6 25.9 14.2  8.1  4.3  1.8  2.0  1.8  5.6 15.2  0.5  9.4
[106]  0.2 14.5  1.8  3.8 21.6  5.3 29.4  3.5  5.3  0.5  6.8 17.8 12.9  7.6 25.4
[121]  5.3 12.4  3.0  3.0 10.1  4.8  8.1  9.4  4.0  5.6  4.3  3.5  1.0  6.6  6.3
[136]  8.4  8.1 17.0  1.0  0.5  1.2  5.6 18.8 11.9  1.7  1.2 21.3  3.5  7.6  9.4
[151]  9.4 15.7
```

Thus, we have identified 152 observations as being extreme.

4.2.2 Fitting the GPD

The GPD log–likelihood function can be derived in the same way that the log–likelihood for the GEV was derived in Section 3.2.1; this is left as an exercise for the reader, but can be shown to be:

$$\ell(\sigma, \xi; \boldsymbol{y}) \quad = \quad -152\log\sigma - (1 + 1/\xi) \sum_{i=1}^{152} \log_e \left(1 + \frac{\xi y_i}{\sigma}\right)_+, \tag{18}$$

where $\boldsymbol{y} = (y_1, \ldots, y_{152})$ are the set of exceedances above threshold $u_0 = 30$. For the case $\xi = 0$, interpreted as $\xi \to 0$, we have the log–likelihood for an exponential distribution with rate $1/\sigma$. We thus define the GPD log–likelihood in **R** in the following way: