Chapter 4

Dependent extremes

4.1 Introduction

The threshold–based approach to modelling extremes, as discussed in Chapter 3, has obvious practical advantages over the more traditional 'block maxima' approach of Chapter 2. Since extremes are, by their very nature, scarce, a modelling procedure which allows the inclusion of more data in the analysis has got to be a good thing. Indeed, the whole point of a threshold–based analysis is that we include *all* extremes in the analysis, extreme in the sense that the observations used have all exceeded some pre–determined threshold u. At a practical level, we might expect a threshold based analysis to result in estimates of return levels with much reduced standard errors and, correspondingly, narrower confidence intervals.

However, using information on *all* extremes, as natural and appealing as it may sound, brings with it it's own set of problems – not least, the issue of *temporal dependence*, sometimes referred to as *serial correlation*. For example, if we are analysing extremes of temperature, and we have hourly records, then we might expect dependence from hour to the next. Such dependence is also likely to be present in the extremes of the process. For example, in a threshold analysis, we are likely to include all hourly observations from a heatwave, and these extreme temperatures are also very likely to depend on one another from one hour to the next. This, of course, seems obvious, and is common to the extremes of most environmental time series (e.g. hourly wind speeds during a storm, sea–surge measurements during a hurricane...). And this is what makes such datasets interesting! However, such temporal dependence violates one of the most basic modelling assumptions we have made in this course so far – the assumption that our series of extremes are independent. In fact, this is the first "I" of "IID" – we will look at the "ID" part in the next chapter!

4.2 Example: Hourly wind speeds at High Bradfield

Hourly maximum wind gusts (in knots) were collected at High Bradfield, in the Peak District, over a period of 10 years from January 1st 2003 to December 31st 2012, giving total of 87,672 observations (including any missing values; 81,835 after the missing values have been removed). Suppose these hourly observations are in the vector `brad` in R; the series without the missing values has been stored in the vector `brad2`.

A plot of the partial autocorrelation function can be useful in ascertaining the extent of any temporal dependence in a series. Similarly, a plot of each observation against it's neighbour can be produced to check for the presence of serial correlation. In R:

```
> pacf(brad2)
> plot(brad[1:87671],brad[2:87672]
```

produces the plots shown in Figure 4.1. The partial autocorrelation function shows significant autocorrelation between observations one and two hours apart (lags 1 and 2); however, this plot doesn't focus on the extremes of the series. Neither does the plot of the time series against the version at lag 1 (right hand side) – however, it is clear from this plot that there is dependence between consecutive extremes even above a high threshold (horizontal and vertical lines in the plot). This could cast doubt on the validity of the GPD when fitted to all exceedances of this threshold.

Figure 4.2 delves a bit deeper; here, we see a plot of a portion of 150 observations from the wind speed series at Bradfield. Again, a possible threshold of 50 knots has been superimposed. It is clear that extreme wind gusts tend to occur in *clusters*; this makes perfect sense, as a storm might last for several hours with sustained gusts. How can we proceed? Ignoring this dependence, and proceeding with a fit of the GPD to the set of all threshold exceedances is an option, and a reasonable fit to the data might still be achieved. However, there are now fewer independent observations in the data than the likelihood assumes, and there is a real chance that the standard errors attached to estimates of the GPD parameters, and associated return levels, will be under–estimated.

4.3 Maxima of dependent (but stationary) series

The book by Leadbetter *et al.* (1983) considers, in great detail, properties of extremes of dependent processes. A key result often used is 'Leadbetter's $D(u_n)$ condition', which ensures that long–range dependence is sufficiently weak so as not to affect the asymptotics of an extreme value analysis. This condition is stated more formally in the Definition below.

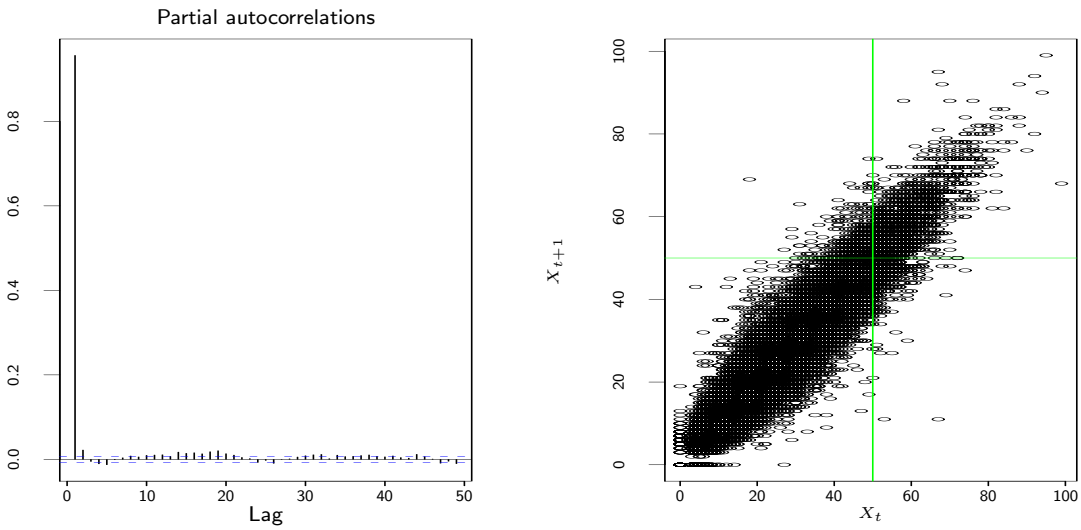


Figure 4.1: Partial autocorrelation function (left) and plot of the hourly wind gusts against the next value in the series (right).

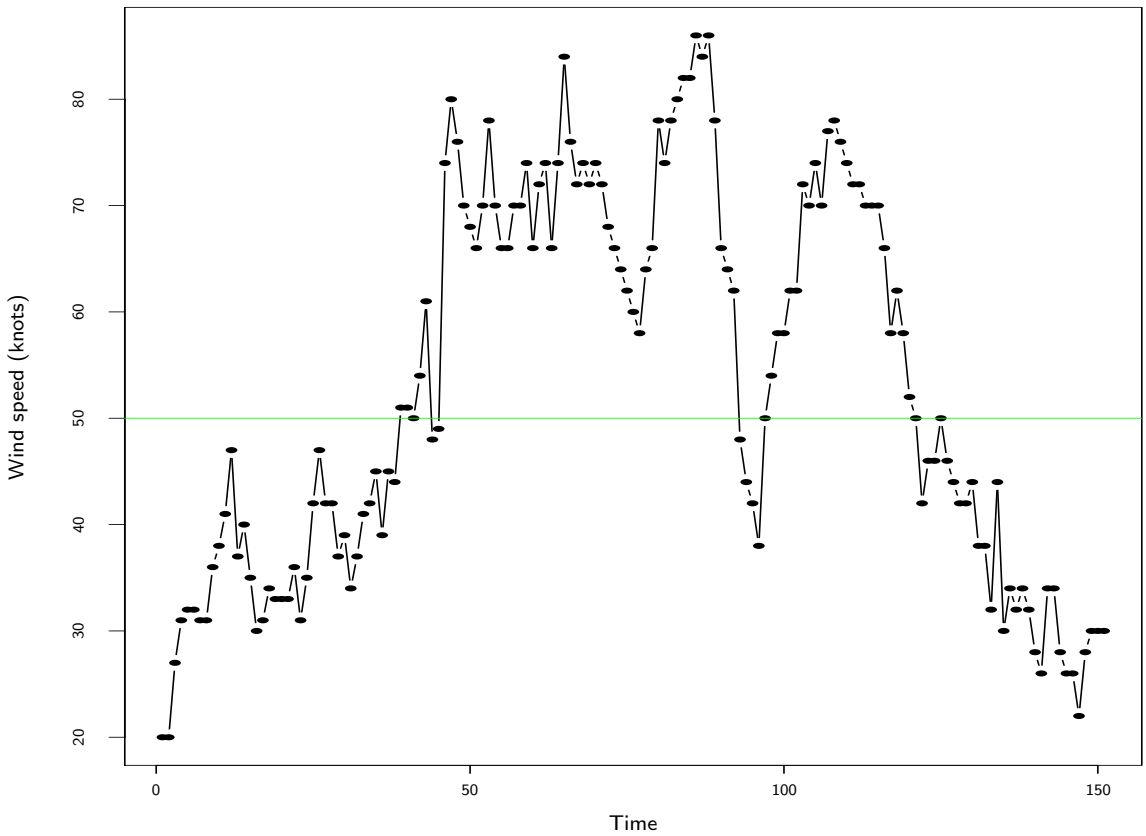


Figure 4.2: Time series plot of a small section of the Bradfield wind speed data, with threshold.

Definition *(Leadbetter's $D(u_n)$ condition)*

A stationary series $\tilde{X}_1, \tilde{X}_2, \ldots$ is said to satisfy the $D(u_n)$ condition if, for all $i_1 < \ldots < i_p < j_1 < \ldots < j_q$ with $j_1 - i_p > l$,

$$\begin{aligned} \Big| \Pr\left\{\tilde{X}_{i_1} \leq u_n, \ldots, \tilde{X}_{i_p} \leq u_n, \tilde{X}_{j_1} \leq u_n, \ldots, \tilde{X}_{j_q} \leq u_n\right\} \\ -\Pr\left\{\tilde{X}_{i_1} \leq u_n, \ldots, \tilde{X}_{i_p} \leq u_n\right\} \Pr\left\{\tilde{X}_{j_1} \leq u_n, \ldots, \tilde{X}_{j_q} \leq u_n\right\} \Big| \quad \leq \quad \alpha(n,l), \end{aligned} \tag{4.1}$$

where $\alpha(n,l) \to 0$ for some sequence l_n such that $l_n/n \to 0$ as $n \to \infty$.

For sequences of independent variables, the difference in probabilities in the above expression is exactly zero for *any* sequence u_n. More generally, we will require that the $D(u_n)$ condition holds only for a specific sequence of thresholds u_n that increases with n. For such a sequence, the $D(u_n)$ condition ensures that, for sets of variables that are far enough apart, the difference in probabilities expressed in (4.1), while not zero, is sufficiently close to zero to have no effect on the limit laws for extremes.

Theorem *(Extremes of dependent sequences)*

Let $\tilde{X}_1, \tilde{X}_2, \ldots$ be a stationary series satisfying Leadbetter's $D(u_n)$ condition, and let $\tilde{M}_n = \max\{\tilde{X}_1, \ldots, \tilde{X}_n\}$. Now let $X_1, X_2, \ldots$ be an *independent* series with X having the same distribution as $\tilde{X}$, and let $M_n = \max\{X_1, \ldots, X_n\}$. Then if M_n has a non–degenerate limit law given by $\Pr\{(M_n - b_n)/a_n \leq x\} \to G(x)$, it follows that

$$\Pr\left\{(\tilde{M}_n - b_n)/a_n \leq x\right\} \to G^{\theta}(x) \tag{4.2}$$

for some $0 \leq \theta \leq 1$.

The parameter θ is known as the *extremal index*, and quantifies the extent of extremal dependence: $\theta = 1$ for a completely independent process, and $\theta \to 0$ with increasing levels of (extremal) dependence. Since G in the above theorem is necessarily an extreme value distribution, and due to the *max–stability* property (see Leadbetter *et al.*, 1983), then the distribution of maxima in processes displaying short–range temporal dependence (characterised by the extremal index θ) is also a GEV distribution; the powering of the limit distribution by θ only affects the location and scale parameters of this distribution.

The above theorem implies that if maxima of a stationary series converge – which, from Chapter 2, we know they will do – then, provided an appropriate $D(u_n)$ condition is satisfied, the limit distribution is related to the limit distribution of an independent series. The effect of dependence, as seen in expression (4.2), is just a replacement of G as the limit distribution with G^{θ}. In fact, if G corresponds to the GEV distribution with

parameters (μ, σ, ξ), then

$$\begin{aligned} G^{\theta}(x) &= \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}^{\theta} \\ &= \exp\left\{-\left[1+\xi\left(\frac{x-\mu^*}{\sigma^*}\right)\right]^{-1/\xi}\right\}, \end{aligned}$$

where $\mu^* = \mu - \frac{\sigma}{\xi}\left(1-\theta^{-\xi}\right)$ and $\sigma^* = \sigma\theta^{\xi}$. Thus, if the (approximate) distribution of M_n is GEV with parameters (μ, σ, ξ), then the (approximate) distribution of $\tilde{M}_n$ is GEV with parameters (μ^*, σ^*, ξ).

4.3.1 Modelling block maxima

Provided long–range dependence is weak, we can proceed to model block maxima from series with short–range extremal dependence as outlined in Chapter 2, since the distribution of block maxima falls within the same family of distributions as would be appropriate if the series were truly independent. This is fantastic news! Short–range temporal dependence is a much more plausible assumption than complete independence, and our modelling approach is still valid! However, the main difference – excluding the change in parameters from (μ, σ, ξ) to (μ^*, σ^*, ξ) – is that our implied n (the number we are taking the maxima over) is now effectively reduced due to the dependence, so convergence of maxima to the limit distribution will be slower. And shouldn't we be using threshold methods anyway, which use information on *all* extremes and not just those that are the maximum within their block?

4.3.2 Modelling threshold exceedances

Though the modelling procedure for fitting the GEV to a set of annual maxima is unchanged for series which display short–term temporal dependence, some revision is needed of the threshold exceedance approach. If all threshold exceedances are used in our analysis, and the GPD fitted to the set of threshold excesses, the likelihoods we use will be incorrect since they assume independence of sample observations. In practice, several techniques have been developed to circumvent this problem, including:

1. filtering out an (approximately) independent set of threshold exceedances
2. fitting the GPD to *all* exceedances, ignoring dependence, but then appropriately adjusting the inference (usually an inflation of standard errors) to take into account the reduction in information
3. Explicitly modelling the temporal dependence in the process

Though the first approach above is by far the most widely–used, research has recently focussed on the relative merits of the other two approaches. The third approach makes

use of multivariate extreme value theory, and so we shall re–visit this idea in more detail in Chapter 6. For now, let us consider the first approach which is probably the most commonly–used.

Declustering

Since the mid–1990s, various methods for *declustering* a series of extremes, to extract a set of independent extremes, have been discussed in the literature. The most natural, commonly–used method of declustering is that of *runs declustering*. This is how it works:

1. Choose an auxiliary 'declustering parameter' (which we call κ)
2. A cluster of threshold excesses is then deemed to have terminated as soon as at least κ consecutive observations fall below the threshold
3. Go through the entire series identifying clusters in this way
4. The maximum (or 'peak') observation from each cluster is then extracted, and the GPD fitted to the set of cluster peak excesses.

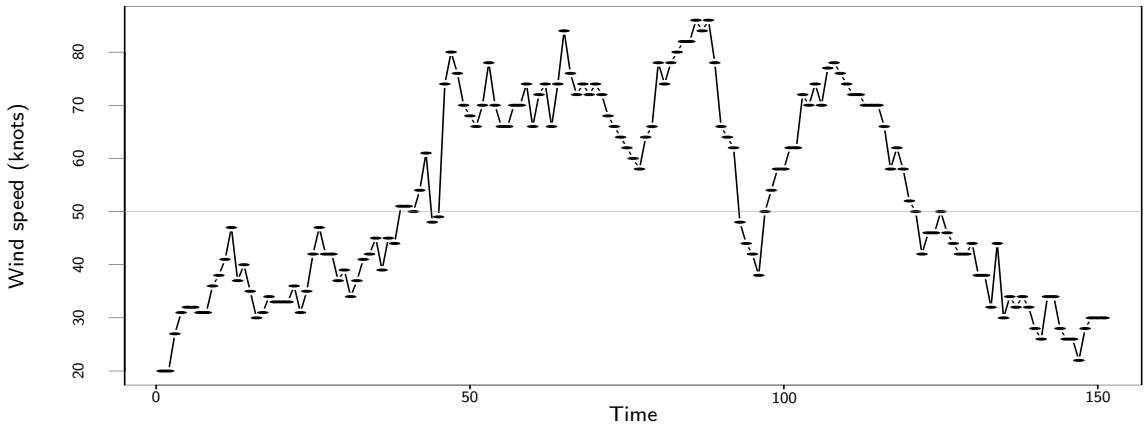
This approach is often referred to as the *peaks over threshold* approach (POT, Davison and Smith, 1990) and is widely accepted as the main pragmatic approach for dealing with clustered extremes. Although this approach is quite easy to implement, there are issues surrounding the choice of κ; if

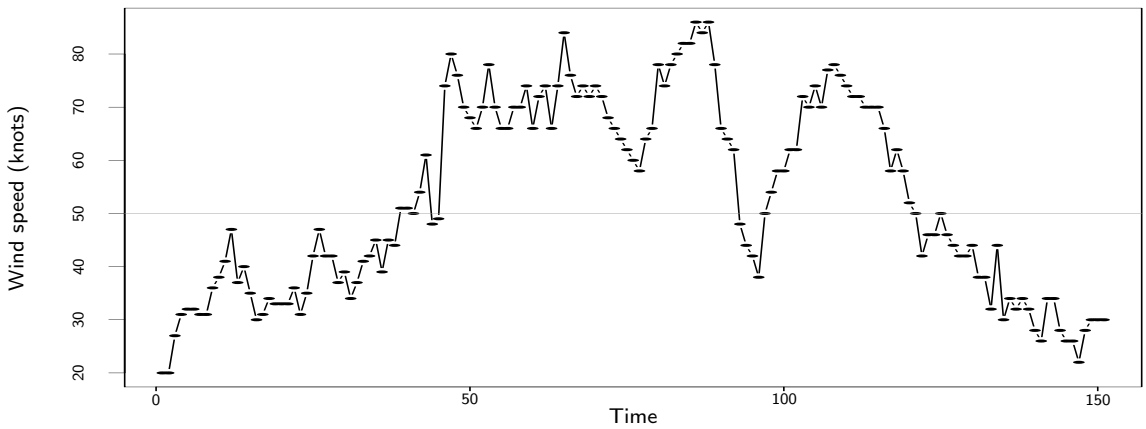
- κ is too small, the cluster peaks will not be far enough apart to safely assume independence
- κ is too large, there will be too few cluster exceedances on which to form our inference

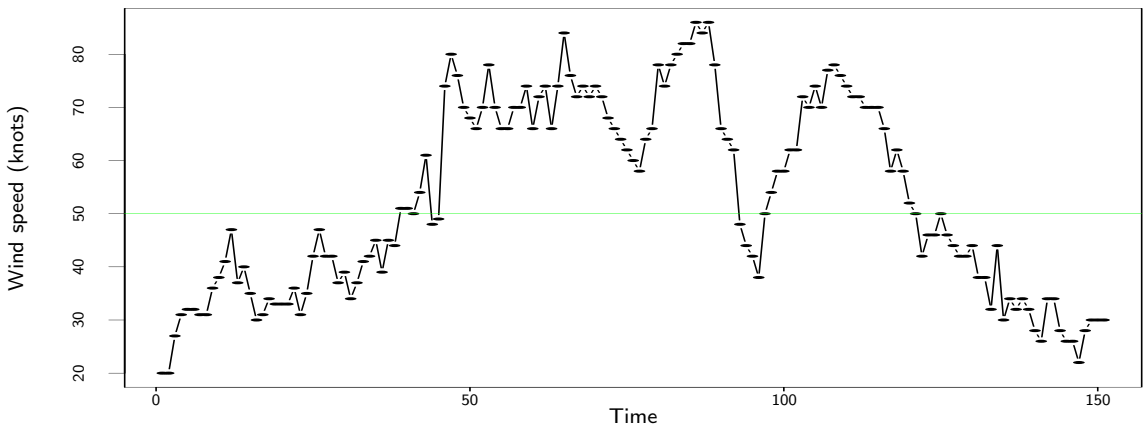
It has also been shown (Fawcett and Walshaw, 2012[1]) that parameter estimates can be sensitive to the choice of κ, and κ is all too often chosen arbitrarily.

[1]Estimating return levels from serially dependent extremes, *Environmetrics* **23**(3), pp 272–283

Let us return to the wind speeds at High Bradfield. Suppose we implement runs declustering with (a) $\kappa = 2$; (b) $\kappa = 4$ and (c) $\kappa = 10$. Identify the clusters associated with (a), (b) and (c) on the snapshots below (top to bottom).







Suppose we use $\kappa = 10$. Obviously, we wouldn't want to identify clusters by hand for the full Bradfield wind speed series (recall that we have 10 years of hourly observations!). Unfortunately, there is no function in ismev to perform runs declustering. I have written the following R code to perform this declustering, for $\kappa = 10$:

```
cluster10=function(dataset,threshold){
   x=list()
   z=list()
   j=1
    {
    for(i in (11):length(dataset)){
     if(dataset[i-10]>threshold
     & dataset[i-9]<=threshold & dataset[i-8]<=threshold
     & dataset[i-7]<=threshold & dataset[i-6]<=threshold
     & dataset[i-5]<=threshold & dataset[i-4]<=threshold
     & dataset[i-3]<=threshold & dataset[i-2]<=threshold
     & dataset[i-1]<=threshold
     & dataset[i]<=threshold){
        x=max(dataset[j:i])
        ifelse(i !=length(dataset), j<-i+1, NA)
        z=c(z,x)}}}
     return(z)}
```

Now, in R, typing

```
> cluster.peaks=as.numeric(cluster10(brad2,50))
```

stores the set of cluster peaks in the vector `cluster.peaks`.

Fitting the model

Now we can fit the GPD to this set of cluster peak excesses using the function `gpd.fit` in ismev:

```
A=gpd.fit(cluster.peaks,50)
$threshold
[1] 50

$nexc
[1] 299

$conv
[1] 0
```

```
$nllh
[1] 992.3284

$mle
[1] 11.9363320 -0.1607165

$rate
[1] 1

$se
[1] 0.8970203 0.0488930
```

Thus our GPD, fitted to the set of cluster peak excesses, has

$$\hat{\sigma} = 11.936(0.897) \quad \text{and} \quad \hat{\xi} = -0.161(0.049)$$

(standard errors in parentheses). Notice, however, that the function has returned an estimate of the threshold exceedance rate (which we called λ_u in the last chapter) of 1; this is because now, all of the data we are supplying `gpd.fit` with are above the threshold (recall that we are using the set of extracted cluster peaks!). Thus, we can put this 'right' by typing

```
> A$rate = length(cluster.peaks)/length(brad2)
```

giving

$$\hat{\lambda}_{u,\text{cp}} = 0.00365(0.0002),$$

again with the standard error in parentheses; the subscript notation on $\hat{\lambda} - u, \text{cp}$ – is to remind us that this is the rate of cluster peak exceedances above the threshold, rather than the full rate of threshold exceedances.

Return level estimation

We can now produce estimates of return levels in the usual way; that is, for the GPD we would use equation (3.9) from Chapter 3, but now the threshold exceedance rate $\hat{\lambda}_u$ must be replaced with the *cluster peak* exceedance rate $\hat{\lambda}_{u,\text{cp}}$ – which in this example is 0.00365, or 0.365%.

So, for example, an estimate of the 50 year gust speed at High Bradfield, would be

$$50 + \frac{11.936}{-0.161}\left[(50 \times \{365.25 \times 24\} \times 0.00365)^{-0.161} - 1\right] = 101.533 \text{ knots.}$$

We can also use profile likelihood to obtain a 95% confidence interval for this estimate – in the usual way, using the ismev function `gpd.prof`:

```
> gpd.prof(A,50,npy=365.25*24,xlow=92,xup=130)
> abline(v=101.533,lty=2)
```

This gives the plot shown in Figure 4.3; the corresponding 95% confidence interval is approximately (94.7 knots, 117 knots).

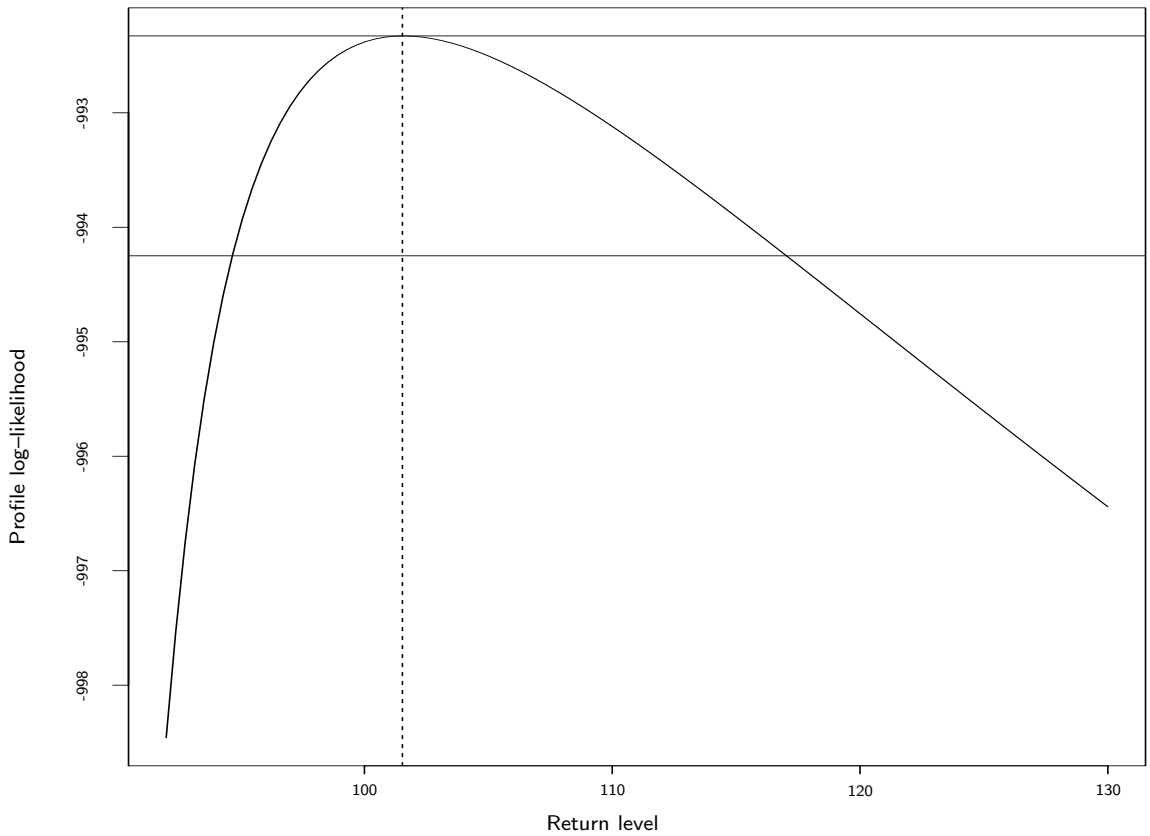


Figure 4.3: Profile log–likelihood as a result of a cluster peaks analysis of the wind speed data at High Bradfield.

4.4 Analyses of cluster peaks: words of warning