5

Non–stationary extremes

5.1 Introduction

In the context of environmental processes, it is common to observe non–stationarity – for example, due to different seasons having different climate patterns, or perhaps due to more long term trends owing to climate change. The models that were introduced in Chapters 2 and 3 assume that the observations used are independent and identically distributed. We examined the effects of dependence in Chapter 4, and we looked at how we might work around this at a practical level (for example, using a declustering technique to filter out a set of independent extremes). But what about extremes for which we cannot assume stationarity?

To date, no general theory for non–stationary extremes has been established (unlike the general theory we saw in Section 4.3 for dependent extremes). In practice, it is common to adopt pragmatic 'workarounds' based on the type of non–stationarity observed. For this reason, in this Chapter we will give some specific examples of how practitioners have dealt with non–stationarity in recent work and publications.

5.2 Annual maximum sea levels in Venice

Recall question 4 in problems sheet 2, which investigated the use of the r–largest order statistics model for extreme sea levels at Venice (for the years 1961–2011 inclusive). These data are available to download from the course webpage. Once saved, the data can be read into R using the command `read.table`:

```
> venice = read.table("venice.txt")
```

In this section, we will focus on modelling the set of annual maxima, that is, column 2 of `venice` – we can extract these by typing:

```
> venice.anmax = venice[, 2]
```

A time series plot of the annual maxima is shown in Figure 5.1. Your analysis of the Venice data in problems sheet 2 should have shown unsatisfactory fits for all values of r used, and Figure 5.1 gives us a clue as to why: there seems to be rather strong evidence for a positive trend over the years, and a substantial part of the variability in the data will probably be explained by the systematic variation in sea levels over time.

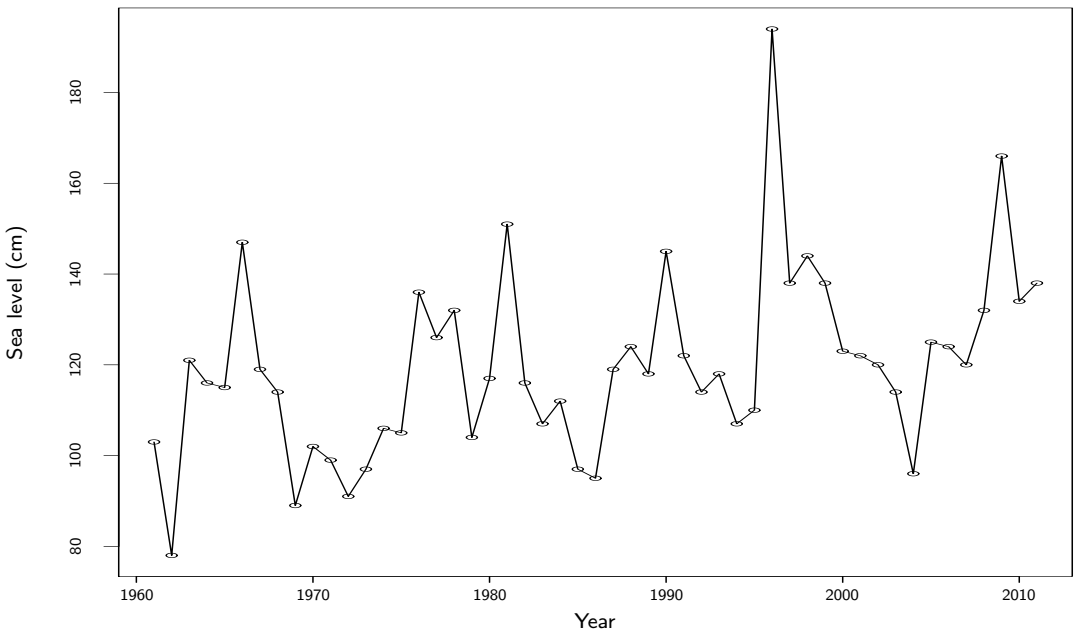

Figure 5.1: Time series plot showing the annual maximum sea levels (in cm above the average) observed at Venice, 1960–2011

One way of capturing this trend is by allowing the GEV location parameter μ to vary across time. From Figure 5.1, a simple linear trend in time seems plausible for our annual maximum sea levels X, and so we could use the model

$$X_t \sim \text{GEV}(\mu(t), \sigma, \xi),$$

where

$$\mu(t) = \beta_0 + \beta_1 t \tag{5.1}$$

and t is an indicator of year. In this way, variations over time are modelled as a linear trend in the location parameter of the GEV. As in a standard simple linear regression, β_1 represents the slope – in this case, the annual rate of change in sea–level maxima at Venice. The time–homogeneous model is a special case of this time dependent model, with $\beta_1 = 0$; since this is *nested* within the model which allows for a time dependence, the deviance statistic can be used to formally compare models.

5.2.1 Parameter estimation

Recall, from Section 2.1.4 in Chapter 2, the log–likelihood function for the GEV:

$$\ell(\mu, \sigma, \xi; \boldsymbol{x}) = -m\log\sigma - (1/\xi + 1)\sum_{i=1}^{m} \log\left[1 + \xi\left(\frac{x_i - \mu}{\sigma}\right)\right]_+ - \sum_{i=1}^{m}\left[1 + \xi\left(\frac{x_i - \mu}{\sigma}\right)\right]_+^{-1/\xi},$$

where m is the number of block maxima $x_1, x_2, \ldots, x_m$. We simply replace μ in the above expression with equation (5.1), giving

$$\begin{aligned}\ell(\beta_0, \beta_1, \sigma, \xi; \boldsymbol{x}, \boldsymbol{t}) \quad = \quad & -m\log\sigma - (1/\xi + 1)\sum_{i=1}^{m} \log\left[1 + \xi\left(\frac{x_i - (\beta_0 + \beta_1 t_i)}{\sigma}\right)\right]_+ \\ & - \sum_{i=1}^{m}\left[1 + \xi\left(\frac{x_i - (\beta_0 + \beta_1 t_i)}{\sigma}\right)\right]_+^{-1/\xi},\end{aligned}$$

with the usual replacement when $\xi = 0$. We could then maximise this log–likelihood in R from first principles – that is, by applying the `nlm` routine to a function which returns the *negative* log–likelihood. However, the ismev function `gev.fit` has an option to allow for non–stationary modelling of the GEV parameters. Within the function, we have:

```
> library(ismev)
> head(gev.fit)[1:4]
[1] function (xdat, ydat = NULL, mul = NULL, sigl = NULL, shl = NULL,
[2]     mulink = identity, siglink = identity, shlink = identity,
[3]     muinit = NULL, siginit = NULL, shinit = NULL, show = TRUE,
[4]     method = "Nelder-Mead", maxit = 10000, ...)
```

Obviously the first argument, `xdat`, corresponds to our series of annual maxima. The second argument, which is set to `NULL` as a default, corresponds to a matrix of covariates which can be used for non–stationary modelling of the GEV parameters; the argument `mul` tells R which column(s) of `ydat` to use as covariates for the linear modelling of the location parameter. Notice that, through the arguments `sigl` and `xil`, we can also allow the scale and shape parameters to depend on a covariate, or covariates, in `ydat`.

To allow for a linear trend in μ through time, `ydat` will need to be a matrix with just a single column, where the values in the column are a time counter from 1 to 51 (as we have 51 annual maxima). Thus:

```
> ti = matrix(ncol = 1, nrow = 51)
> ti[, 1] = seq(1, 51, 1)
```

Now to fit the GEV to allow for a linear trend in μ, we type:

```
> gev.fit(venice.anmax, ydat = ti, mul = 1)
$model
$model[[1]]
[1] 1

$model[[2]]
NULL

$model[[3]]
NULL


$link
[1] "c(identity, identity, identity)"

$conv
[1] 0

$nllh
[1] 216.0626

$mle
[1] 96.98579330  0.56414269 14.58435088 -0.02731421

$se
[1] 4.24930969 0.13948421 1.57840034 0.08270996
```

Thus, we have

$$\hat{\beta}_0 = 96.986(4.249) \quad \hat{\beta}_1 = 0.564(0.139) \quad \hat{\sigma} = 14.584(1.578) \quad \hat{\xi} = -0.027(0.083)$$

(with standard errors in parentheses). The value of the negative log–likelihood at these parameter estimates is 216.0626; `NULL` in `model[[2]]` and `model[[3]]` indicates that we have not asked for any modelling of the scale and shape parameters here – only the location parameter (hence `model[[1]]` = 1). Thus, we have

$$\hat{\mu}(t) = 96.986 + 0.564t,$$

giving an estimated increase in maximum sea levels at Venice of about 0.564cm per year. For example, the estimated value for μ in the year 2013 would be

$$\hat{\mu}(53) = 96.986 + 0.564 \times 53 = 126.878;$$

we could, of course, use the delta method to obtain the corresponding standard error. Using this simple linear model, we can estimate μ for $t = 1, 2, \ldots, 51$ to cover the years for which we have data (i.e. 1961, 1962, ..., 2011). Superimposing $\mu(t)$ on the original time series shown in Figure 5.1 summarises our fit, and the effect of the linear trend in μ; this updated plot can be seen in Figure 5.2.

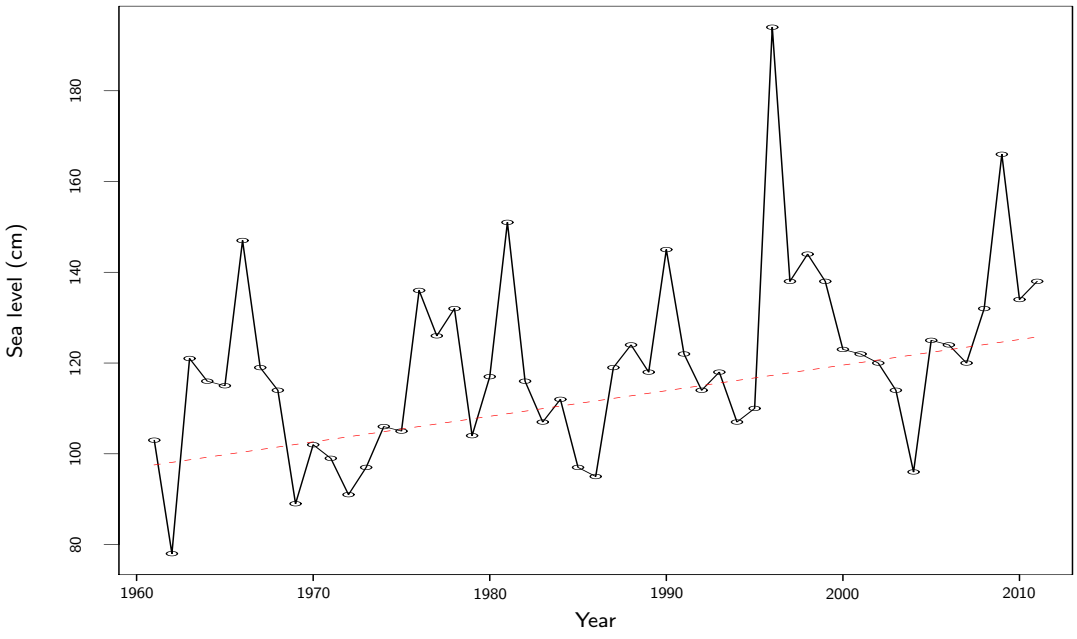

Figure 5.2: Time series plot showing the annual maximum sea levels (in cm above the average) with the trend for μ superimposed (dashed red line)

5.2.2 Model choice

Fitting a completely stationary model to the set of annual maxima, as you did in question 4(a) of problems sheet 2, gives:

```
> gev.fit(venice.anmax)
$conv
[1] 0

$nllh
[1] 222.7145

$mle
[1] 111.09925486  17.17548761  -0.07673265

$se
[1] 2.6280070 1.8033672 0.0735214
```

Is the non–stationary model worthwhile? Is the trend we observe in Figure 5.1 *significant*? In other words, does the non–stationary model provide an improvement in fit over the simpler model shown here? We can use a version of the result in Section 2.4 (page 43) to address this question. Generally, maximum likelihood estimation of nested models leads to a simple test procedure of one model against the other. With models $\mathcal{M}_0 \subset \mathcal{M}_1$, we define the deviance statistic as

$$D = 2\left\{\ell_1(\mathcal{M}_1) - \ell_0(\mathcal{M}_0)\right\},$$

where $\ell_1(\mathcal{M}_1)$ and $\ell_0(\mathcal{M}_0)$ are the maximised log–likelihood under models $\mathcal{M}_1$ and $\mathcal{M}_0$ respectively. The asymptotic distribution of D is given by the χ^2_k distribution with k degrees of freedom, where k is the difference in dimensionality of $\mathcal{M}_1$ and $\mathcal{M}_0$; thus, calculated values of D can be compared to critical values from χ^2_k, where large values of D suggest that model $\mathcal{M}_1$ explains substantially more of the variation in the data than $\mathcal{M}_0$.

We could, of course, use this method to check for a more complex association through time. For example, to check for a quadratic trend we might use a model $\mathcal{M}_2$ with the following form for μ:

$$\mu(t) = \beta_0 + \beta_1 t + \beta_2 t^2.$$

The matrix of covariates would now need two columns to include t and t^2, that is:

```
> ti2 = matrix(ncol = 2, nrow = 51)
> ti2[, 1] = seq(1, 51, 1)
> ti2[, 2] = ti2[, 1]^2
```

Then we could fit the model which allows for a quadratic trend in sea–levels in the following way (output not shown):

```
> gev.fit(venice.anmax, ydat = ti2, mul=c(1, 2), show = FALSE)
```

You could try this yourself – this gives a maximised log–likelihood of –216.0555. Comparing with model $\mathcal{M}_1$ we have

$$D = 2\{-216.0555 - (-216.0626)\} = 0.0142,$$

which is *small* compared to $\chi^2_1(0.05) = 3.841$. Thus, allowing for a quadratic dependence in time does not improve on our model which allows for a linear trend through time, and so we would reject model $\mathcal{M}_2$.

5.2.3 Model diagnostics

Before estimating return levels, we should check the goodness–of–fit of our model which allows for a linear trend in μ. The lack of homogeneity in the distributional assumptions for each observation, however, mean some modification of the standard procedures (e.g. probability plots and quantile plots) is required. For example, for the Venice annual maximum sea levels, we have

$$X_t \sim \text{GEV}(\mu(t), \sigma, \xi), \quad t = 1, 2, \ldots, 51,$$

giving a different GEV in each year indicated by t. What we need to do is standardise so that we can assume the X_t are IID across all years t; usually, the set of non–stationary annual maxima are transformed to a common Gumbel distribution with distribution function $F(y) = \exp\{-e^{-y}\}$. We can obtain the required transformation by equating $F(y_t)$ to $\text{GEV}(x_t; \hat{\mu}(t), \hat{\sigma}, \hat{\xi})$ and solving for y_t:

Now that we have transformed the original data, with a yearly–varying GEV, to a single common distribution, we can apply the standard graphical diagnostics. For example, we can compare the empirical probabilities and quantiles of y_t to their theoretical counterparts from the Gumbel distribution. Fortunately, the usual command in ismev can be used to produce these plots. For example, the following code results in the plots shown in Figure 5.3. It is clear from these plots that the model allowing for a linear trend in μ is adequate for our data.

```
> A = gev.fit(venice.anmax, ydat = ti, mul = 1, show = FALSE)
> gev.diag(A)
```

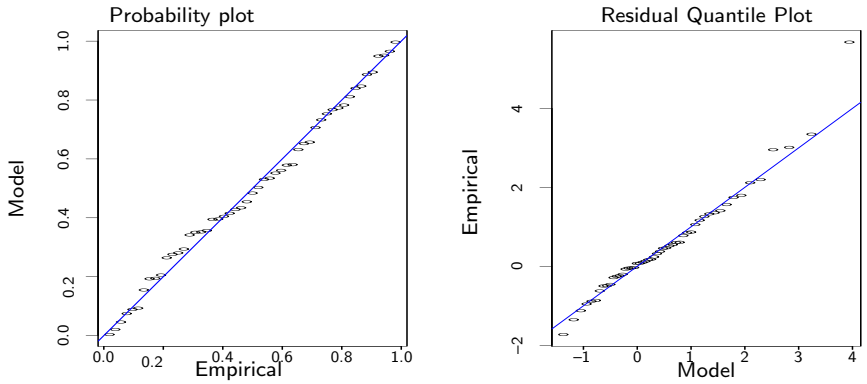

Figure 5.3: Diagnostic plots to assess the goodness–of–fit of the GEV model for the annual maximum sea–levels at Venice, allowing for a linear trend in μ

5.2.4 Return level estimation

Recall Equation (2.10) from Chapter 2 for estimating return levels from the GEV:

$$\hat{z}_r = \hat{\mu} + \frac{\hat{\sigma}}{\hat{\xi}}\left[\left(-\log\left(1 - r^{-1}\right)\right)^{-\hat{\xi}} - 1\right].$$

Since we have a time–varying location parameter $\hat{\mu}(t) = \hat{\beta}_0 + \hat{\beta}_1 t$, we will clearly have time–varying estimates of return levels $\hat{z}_r(t)$. For example, an estimate of the sea level we might expect to see in Venice once every 100 years is given by

$$\hat{z}_r(t) \quad = \quad (96.986 + 0.564t) - \frac{14.584}{0.027}\left[\left(-\log\left(1 - 100^{-1}\right)\right)^{0.027} - 1\right]. \qquad (5.2)$$

Figure 5.4 shows how we might expect this estimate to vary for $t = 52, 53, \ldots$, i.e. for the years 2012, 2013, ... We could treat these as forecasts of the 100–year return levels as we move through time; obviously, such forecasts will assume the linear trend for μ continues beyond the range of data we have observed and will, of course be subject to error (which we can estimate by constructing point–wise 95% confidence intervals using the profile log–likelihood, for example).

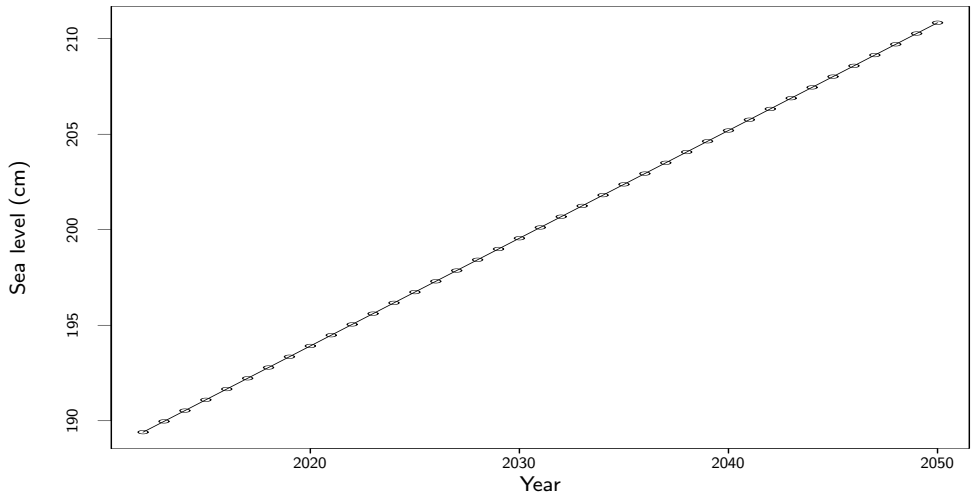

Figure 5.4: Time series plot showing the forecasted 100–year return levels at Venice, for the years 2012–2050

5.3 Sea levels and the Southern Oscillation Index

A different situation which could use the same approach as that in the previous section is where the extremal behaviour of a series is related to another variable, rather than time. For example, studies have revealed a link between annual maximum sea levels at Fremantle, Australia, and the mean value of the Southern Oscillation Index (SOI, an indicator of meteorological volatility due to effect such as El Niño); see Figure 5.5. Thus, the following model for X_t, the annual maximum sea level at Fremantle in year t, might be suitable:

$$X_t \sim \text{GEV}(\mu(t), \sigma, \xi),$$

where

$$\mu(t) \quad = \quad \beta_0 + \beta_1 \text{SOI}(t), \tag{5.3}$$

where $\text{SOI}(t)$ denotes the mean value of the SOI in year t. However, the plot in the right–hand–side of Figure 5.5 also reveals a possible trend in sea levels through time, suggesting

$$\mu(t) \quad = \quad \beta_0 + \beta_1 t, \tag{5.4}$$

where $t = 1, 2, \ldots$, as in Example 5.2. We can combine Equations (5.3) and (5.4) to allow for a dependence on time *and* SOI by letting

$$\mu(t) \quad = \quad \beta_0 + \beta_1 \text{SOI}(t) + \beta_2 t; \tag{5.5}$$

however, a technique of *forward selection* should be used to check whether or not any of Equations (5.3), (5.4) or (5.5) give significant improvement over the stationary model.

The sea level data for Fremantle are part of the ismev package; once ismev has been installed, the data can be loaded by typing

```
> data(fremantle)
```

We can then look at the data:

```
> head(fremantle)
  Year SeaLevel   SOI
1 1897     1.58 -0.67
2 1898     1.71  0.57
3 1899     1.40  0.16
4 1900     1.34 -0.65
5 1901     1.43  0.06
7 1903     1.19  0.47
```

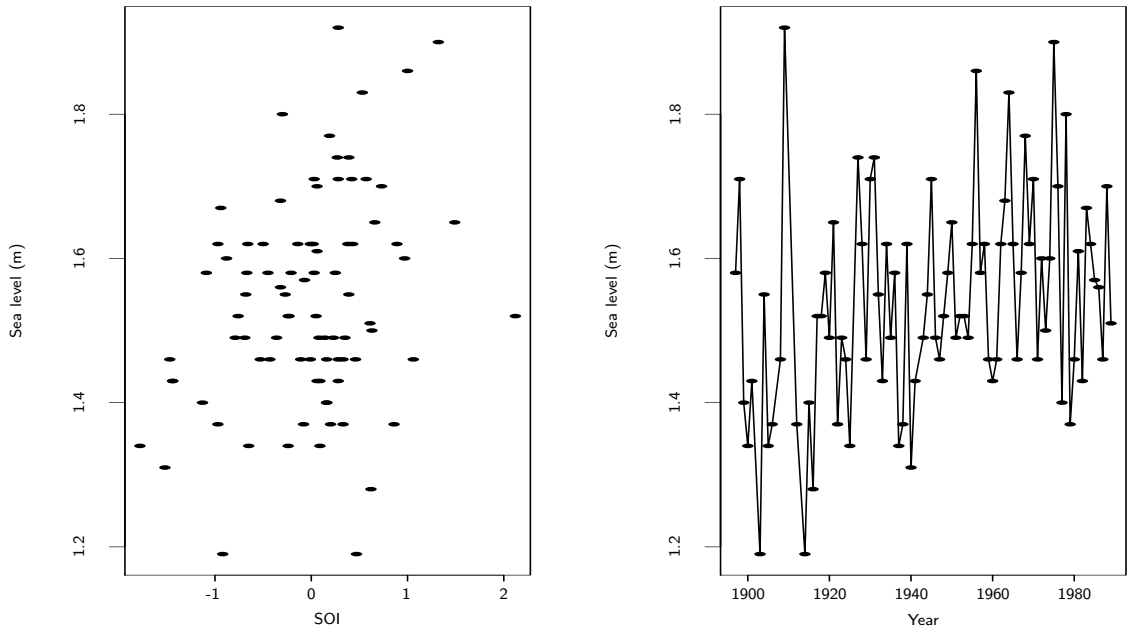

Figure 5.5: Sea levels at Fremantle, Western Australia, plotted against (1) mean Southern Oscillation Index (left), and (2) year (right).

Column 1 is an indicator of year; the annual maximum sea levels, and mean SOI values, are in columns 2 and 3 respectively. Although we have data from 1897–1989, data are missing for the years 1902, 1907, 1910–11, 1913, 1926 and 1942, giving us 86 years of data. Assuming our data have been continuously collected, we can then set up the year indicator t:

```
> ti = seq(1, 86, 1)
```

Then our matrix of covariates, including the mean SOI values and t, can be constructed:

```
> covar = matrix(ncol = 2, nrow = 86)
> covar[, 1] = fremantle[, 3]
> covar[, 2] = ti
```

Fitting the stationary model gives:

```
> gev.fit(fremantle[, 2])
$conv
[1] 0

$nllh
[1] -43.56663

$mle
[1]  1.4823409  0.1412671 -0.2174320

$se
[1] 0.01672502 0.01149461 0.06377394
```

Now we should try incorporating SOI and time, one variable at a time (via Equations (5.3) and (5.4) respectively), to see which (if any!) gives the most significant improvement over the stationary model. First, allowing for a dependence on the mean value of the Southern Oscillation Index gives the following output in R:

```
> A = gev.fit(fremantle[, 2], ydat = covar, mul = 1, show = FALSE)
> A$nllh
[1] -47.21114
> A$mle
[1]  1.48985338  0.06188902  0.13960518 -0.26848380
> A$se
[1] 0.01655406 0.02315637 0.01150991 0.06399288
```

Comparing to the stationary model, we have

$$D = 2\{47.21114 - 43.56663\} = 7.28902,$$

which is greater than $\chi^2_1(0.05) = 3.841$, suggesting a significant improvement over the stationary model. Allowing for a dependence in time, gives:

```
> A = gev.fit(fremantle[, 2], ydat = covar, mul = 2, show = FALSE)
> A$nllh
[1] -49.78972
> A$mle
[1]  1.387186155  0.002140832  0.124716473 -0.128545018
> A$se
[1] 0.0274796482 0.0005215259 0.0104146285 0.0679844086
```

Again, comparing to the stationary model, we have

$$D = 2\{49.78972 - 43.56663\} = 12.44618,$$

suggesting that including a dependence on time gives a more significant improvement than allowing for a dependence on SOI (since $12.44618 > 7.28902$). Thus, our current "best model" is

$$X_t \sim \text{GEV}(\mu(t), \sigma, \xi),$$

where $\mu(t) = \beta_0 + \beta_1 t$, and

$$\begin{aligned} \hat{\mu}(t) &= 1.387 + 0.002t \\ \hat{\sigma} &= 0.125 \\ \hat{\xi} &= -0.129 \end{aligned}$$

Now let's see if including mean SOI on top of this adds further improvement:

```
> A = gev.fit(fremantle[, 2], ydat = covar, mul = c(1, 2), show = FALSE)
> A$nllh
[1] -53.8257
> A$mle
[1]  1.389381297  0.055171074  0.002232467  0.121147089 -0.154480161
> A$se
[1] 0.0272538644 0.0197789753 0.0005178779 0.0100390306 0.0636920071
```

So we have

$$D = 2\{53.8257 - 49.78972\} = 8.07286,$$

which is significant when compared to $\chi^2_1(0.05) = 3.841$! So we should include *both* time and SOI as covariates, giving us our final model:

$$X_t \sim \text{GEV}(\mu(t), \sigma, \xi),$$

where

$$\mu(t) = \beta_0 + \beta_1 t + \beta_2 \text{SOI}(t),$$

and

$$\begin{aligned} \hat{\mu}(t) &= 1.389 + 0.055t + 0.002\text{SOI}(t) \\ \hat{\sigma} &= 0.121 \\ \hat{\xi} &= -0.154 \end{aligned}$$

Checks of model goodness–of–fit can be dome in the usual way, and return level estimates obtained given values of time and SOI.

5.4 Rainfall in New York

Recall Section 3.2 in which we modelled rainfall extremes in New York using a threshold–based approach. The Generalised Pareto distribution was applied to rainfall exceedances over a threshold of $u = 30$mm, giving estimates of the scale and shape as

$$\hat{\sigma} = 7.44(0.958) \quad \hat{\xi} = 0.184(0.101)$$

(respectively). Recall that the GPD(σ, ξ) arises from the GEV(μ, σ, ξ), where the GPD scale parameter is a function of the GEV location and shape parameters. Thus, attempting to model any trend in our threshold exceedances is usually done through linear modelling of the scale parameter σ (the GPD doesn't have a location parameter *per se*). Since the scale parameter σ must be positive, we might choose to model a trend through time as

$$\sigma(t) \quad = \quad \exp\{\beta_0 + \beta_1 t\}, \tag{5.6}$$

where t is, once again, an indicator of time. There are 17,531 observations in the rainfall series (stored in the vector `rain` in R); thus, to check for a dependence on time, we set up the following covariate matrix in R:

```
> ti = matrix(ncol = 1, nrow = 17531)
> ti[, 1] = seq(1, 17531, 1)
```

Then the code to fit the model which allows for a linear trend through time is similar to that used in the previous sections for the Venice and Fremantle sea level data, but now we must specify the exponential "link function" for the scale parameter, as specified by Equation (5.6):

```
> rain = scan("newyork.txt")
> library(ismev)
> gpd.fit(rain, threshold = 30, ydat = ti, sigl = 1, siglink = exp, show
= FALSE)
```

You can try this yourself; the fitted model gives

$$\hat{\sigma} = \exp\{1.804 + 0.00002t\} \quad \text{and} \quad \hat{\xi} = 0.198$$

with a maximised log–likelihood of –484.6017. Thus, comparing to the stationary model (results shown on page 56), we get

$$D = 2\{-484.6017 - (-485.0937)\} = 0.984,$$

which is small relative to $\chi^2_1(0.05) = 3.841$. Thus, there is no evidence of a (linear) trend in the log–scale parameter.

5.5 Generalisation

With reference to the examples discussed so far, we could model non–stationarity through *any* of the parameters in our extremal model (be that the GEV or the GPD). For example, take a non–stationary GEV model to describe the distribution of X_t, for $t = 1, 2, \ldots, m$:

$$X_t \sim \text{GEV}(\mu(t), \sigma(t), \xi(t)),$$

where each of the model parameters have an expression in terms of a parameter vector $\boldsymbol{\beta}$ and some covariates. The likelihood is then

$$L(x_t; \boldsymbol{\beta}) = \prod_{i=1}^{m} g(x_t; \mu(t), \sigma(t), \xi(t)),$$

where g is the GEV density function. From this, we can form the log–likelihood, and then maximise in the usual way (for example, using `nlm` in R).

In terms of threshold exceedances Y_t, $t = 1, 2, \ldots, k$, we could replace the GEV with the GPD:

$$Y_t \sim \text{GPD}(\sigma(t), \xi(t)),$$

with σ being defined as in Equation (5.5) to retain the positivity of the GPD scale.

5.6 Wind speed extremes at High Bradfield

Recall the wind speed data observed at High Bradfield, in the Peak District, first introduced in Section 4.2. These data were collected every hour, over a period of 10 years, from January 1st 2003 to December 31st 2012. A time series plot of the first four years of data (just over 35,000 observations) is shown below in Figure 5.6. We thought about how to deal with the dependence between successive observations in Chapter 4 (by declustering); however, it is clear from Figure 5.6 that wind speeds at this location also vary seasonally, suggesting a departure from the ideal of stationarity.

To investigate further, Figure 5.7 shows the wind speed distribution at High Bradfield by month. Although any seasonal changes are less obvious in the middle portion of the data, there is some evidence for seasonal variation in the extremes – this can be seen by marked differences in the upper quartile wind speeds, by month, and in the upper tails more generally. The green line cutting across the boxplots in Figure 5.7 corresponds to a monthly varying threshold used to identify wind speeds as extreme; twelve separate mean residual life plots (see Section 3.2.1) were applied to the wind speeds from each month to obtain these thresholds.

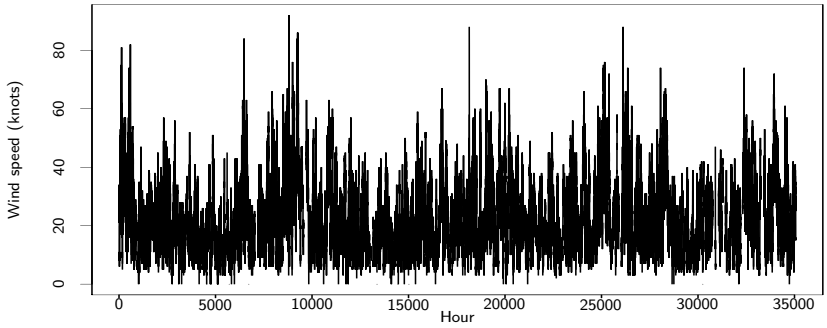

Figure 5.6: Time series plot showing the first four years of wind speed data at High Bradfield in the Peak District.

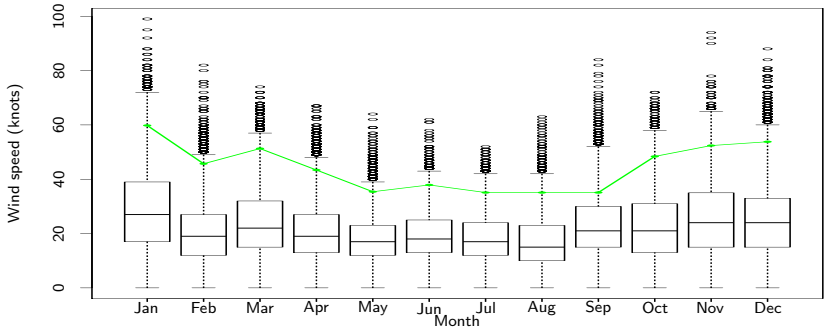

Figure 5.7: Boxplots of wind speed data at High Bradfield, by month.

In the literature, various methods have been considered to deal with non–stationarity arising from such seasonal variation, and we will now outline each of them.

5.6.1 Single season approach

Under the single season approach, an extremal model is fitted to the extremes of an environmental process from the season which gives rise to the 'most extreme' extremes. For example, with reference to the Bradfield wind speed data and Figures 5.6 and 5.7 above, January is clearly the windiest month. Thus, a single season approach would disregard data from all other months, and perhaps fit the GPD to threshold excesses from January only. This approach clearly has some appeal: it is easy to implement and it focuses on the largest (or smallest) extreme values. However, apart from assuming that extremes within the chosen month, or season, are stationary (and experience suggests that certain times in January are windier than other times in the same month), this is extremely wasteful of data – especially if we then decluster to filter out any dependence.

5.6.2 Seasonal piecewise approach

It is usual in strongly seasonal climates for the occurrence of extreme winds to be confined to a certain part of the yearly cycle. In the U.K., for example, it is very unusual for wind damage to occur outside the period October–March. The seasonal variation observed at Bradfield (Figures 5.6 and 5.7) might be expected, since prolonged, anticyclonic periods are more prevalent during June, July and August than in winter months. A model which takes account of seasonal variability will identify all gusts which are large *given the time of year* as extreme. There is only a point to modelling the extremes which occur in summer months, for example, if we believe that they can help us understand what happens in winter months, where genuinely large events *can* occur. For this to be realistic, we must assume that the same mechanism is responsible for the generation of large gusts throughout the year, and it is just the scale of this mechanism which changes. Indeed, in temperate climates (such as that of the U.K.), the same alternating sequence of anticyclones and depressions leads to most of the storms which occur throughout the year; the seasonal variability comes from the severity of these systems.

For wind speed data, there is no natural partition of the year into separate seasons. Here, we might take our seasonal unit to be one month; Fawcett and Walshaw (2006; 2007; 2008) argue that by dividing the year into twelve equal length seasons, we can strike a good balance between the two conflicting requirements of (a) reflecting reasonably accurately the continuous nature of seasonal changes in climate, and (b) retaining a substantial amount of data for analysis within each season.

Table 5.1 shows the results of fitting a separate GPD to excesses above the monthly varying thresholds shown in Figure 5.7. Temporal dependence has been accounted for by filtering out a set of independent threshold exceedances using runs declustering; after extensive discussions with a meteorologist, various values of cluster separation interval κ were used, depending on the month, to carefully identify clusters – where a cluster of extreme wind speeds was deemed to be a "storm".

The approach followed so far – known as the *seasonal piecewise approach* – avoids the problems of non–stationarity as a result of seasonal variability (provided we can safely assume stationarity *within* each seasonal unit). However, in terms of return level inference, it would not make practical sense to have monthly varying estimates of the r–year return level z_r. To include information from *all* months in our return level estimation procedure, we solve

$$\prod_{m=1}^{12} H(\hat{z}_r; \hat{\lambda}_{u_m}, \hat{\sigma}_m, \hat{\xi}_m) \quad = \quad 1 - \frac{1}{rn_y}$$

for $\hat{z}_r$, where H is the GPD distribution function and n_y is the (average) number of observations per year (see Section 3.2.5 for more details about return level estimation

Month (m)	u_m	n_m	$\hat{\sigma}_m$	$\hat{\xi}_m$
1	55.341	28	21.373 (5.358)	−0.420 (0.183)
2	41.531	24	15.130 (4.635)	−0.226 (0.233)
3	48.100	29	23.277 (6.316)	−0.894 (0.259)
4	39.910	29	14.853 (4.448)	−0.440 (0.249)
5	31.943	46	9.456 (1.990)	−0.158 (0.147)
6	35.670	35	12.329 (2.592)	−0.409 (0.143)
7	32.290	36	12.517 (2.609)	−0.605 (0.161)
8	32.639	34	10.199 (2.361)	−0.203 (0.159)
9	33.232	49	18.772 (3.668)	−0.255 (0.138)
10	44.914	34	11.669 (3.533)	−0.274 (0.254)
11	48.394	33	14.991 (3.381)	−0.225 (0.149)
12	49.341	35	18.681 (4.229)	−0.416 (0.166)

Table 5.1: Extreme wind speeds at Bradfield: separate months model fitted to cluster peak excesses.

using the GPD). Thus, we need to solve

$$\prod_{m=1}^{12} \left\{ 1 - \hat{\lambda}_u \left[1 + \hat{\xi} \left(\frac{\hat{z}_r - u_m}{\hat{\sigma}_m} \right) \right]_+^{-1/\hat{\xi}_m} \right\} = \frac{1}{rn_y} \tag{5.7}$$

for $\hat{z}_r$. This equation cannot be solved analytically; rather, a numerical procedure must be used. This is easy to do in R. Suppose we wanted to estimate the 10–year return level. First of all, define a function `f` which returns equation (5.7) above:

```
> f = function(z){
+       r = 10
+       ny = 365.25*24
+       sigma = c(21.373, 15.130, ... )
+       xi = c(-0.420, -0.226, ... )
+       u = c(55.341, 41.531, ... )
+       nobs = c(31*24*10, 28*24*7+29*24*3, ... )
+       lambda = c(28/nobs[1], 24/nobs[2], ... )
+       component = vector("numeric", 12)
+       inner = vector("numeric", 12)
+       for(m in 1:12){
+        inner[m] = max((1+xi[m]*((z - u[m])/sigma[m])), 0)
+        component[m] = 1-lambda[m]*((inner[m])^(-1/xi[m]))}
+       answer = prod(component)-(1-1/(r*ny))
+       return(answer)}
```

	Return period (r years)			
	10	50	200	1000
$\hat{z}_r$ (st. err.)	102.33 (3.970)	104.59 (15.951)	106.21 (23.793)	108.89 (44.865)

Table 5.2: Return level estimates for the Bradfield wind speed data (units are knots). Standard errors are shown in parentheses.

Then we apply the `uniroot` function to find the root of `answer`; this function requires a range of values to search within:

```
> uniroot(f, lower = 0, upper = 200)
$root
[1] 102.3291

$f.root
[1] -6.242917e-11

$iter
[1] 12

$init.it
[1] NA

$estim.prec
[1] 6.103516e-05
```

Thus, we have $\hat{z}_{10} = 102.3291$ – that is, we can expect to see a wind speed in excess of 102.3 knots about once every 10 years. Replacing r with 50, 200 and 1000 gives the return level estimates shown in Table 5.2 overleaf.

Table 5.2 shows standard errors for our return level estimates, obtained via the delta method (see Section 3.2.5). However, we now have

$$V = \begin{pmatrix} \frac{\hat{\lambda}_{u_1}(1-\hat{\lambda}_{u_1})}{N_1} & 0 & \dots & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots & \ddots & \vdots \\ \vdots & \ddots & \frac{\hat{\lambda}_{u_{12}}(1-\hat{\lambda}_{u_{12}})}{N_{12}} & 0 & \dots & 0 \\ 0 & \dots & 0 & v_{1,1} & \dots & v_{1,24} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & v_{24,1} & \dots & v_{24,24} \end{pmatrix},$$

where $v_{i,j}$ denotes the (i,j)–th term of the variance–covariance matrix of $\hat{\sigma}_m$ and $\hat{\xi}_m$,

$m = 1, \ldots, 12$. Hence, by the delta method,

$$\mathrm{Var}(\hat{z}_r) \approx \nabla z_r^T V \nabla z_r,$$

where

$$\nabla z_r^T = \left[\frac{\partial z_r}{\partial \lambda_{u_1}}, \ldots, \frac{\partial z_r}{\partial \lambda_{u_{12}}}, \frac{\partial z_r}{\partial \sigma_1}, \ldots, \frac{\partial z_r}{\partial \sigma_{12}}, \frac{\partial z_r}{\partial \xi_1}, \ldots, \frac{\partial z_r}{\partial \xi_{12}} \right],$$

evaluated at $(\hat{\lambda}_{u_1}, \hat{\sigma}_1, \hat{\xi}_1, \ldots, \hat{\lambda}_{u_{12}}, \hat{\sigma}_{12}, \hat{\xi}_{12})$. A modification of the usual procedure for obtaining confidence intervals based on the profile log–likelihood is also available (see, for example Fawcett (2005)), but this goes beyond the scope of MAS8306.

As an aside, notice how the estimate of the 50–year return level wind speed here differs to that when we assumed stationarity in Section 4.3.2: assuming stationarity gives $\hat{z}_{50} = 101.533$ knots, a slight under–estimation relative to the approach which uses a piecewise seasonal approach to inference. In fact, this under–estimation is a common observation when we fail to account for seasonal variability correctly.

5.6.3 Smoothly varying seasonal parameters

Various authors (e.g. Fawcett and Walshaw (2006)) have investigated the use of continuously varying parameters for the GPD when seasonal variation is present. For example, Fourier forms can be used to allow the GPD scale and shape to vary smoothly through time. However, such analyses for the Bradfield wind speed data yielded little, if any, improvement over the seasonal piecewise approach (in terms of model fit and precision of standard errors for return levels), and so the added computational burden was deemed unnecessary.