Tutorial Sheet 2

1 Part A: Seasonal and trend analysis.

Note

Several of these examples are similar in nature to exam questions, in that you are asked to comment on statements being made or on analysis already completed.

Task 1

The following excerpt is taken from the European Environment Agency web site, concerning one of their key indicators (CLIM008) for snow cover (accessed Jan 2011).
"Data from satellite monitoring (NESDIS-database at NOAA) from 1966 to 2005 show that monthly snow-cover extent in the northern hemisphere is decreasing by 1.3 % per decade (Figure 2.1), with the strongest retreat in spring and summer (UNEP, 2007)."
Discuss critically the statement concerning the decrease per decade, and the comment concerning the spring and summer effects. Figure 1 shows a 'trend' a 12 month running mean and the original values expressed as anomalies, compare and contrast the three representations.

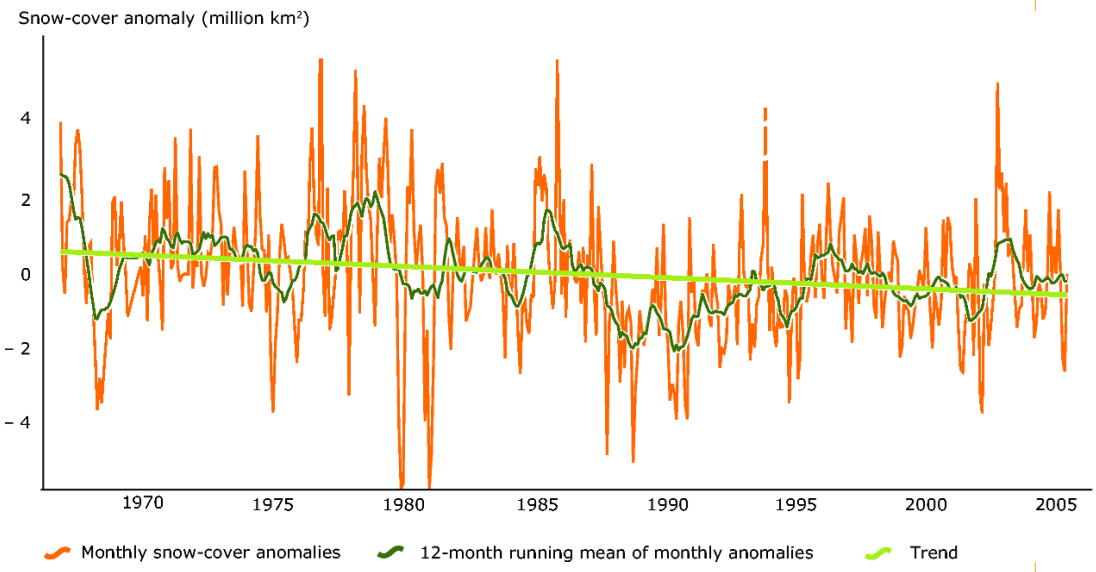


Figure 1: Figure 1: Trends in Snow Cover

Solution
The data plot (Fig 1) shows cycles and also a linear trend (with the slope estimated as % per decade). The green curve shows a 12 month running mean (so a smoothed curve), one could query the simplicity of the linear regression, since the running mean (moving average) does show fluctuations, although much smoothed in comparison to the original anomaly data. The comment that there is a strongest retreat in spring and summer suggest quite strongly that a simple straight line regression for the entire time period does not capture the 'heterogeneity' of rate of decline.

Task 2

It is of interest to look at temperature records in Loch Lomond and to consider what the trend might be in water temperature. The data shown here are for Cailness in the north basin of the Loch. Comment on the patterns over time in this record. What other plots might you look at and how might you model this data to investigate the changes in temperature over time. There are some missing data values which you might consider how to impute. The data are available to explore on Moodle if you want to look at this in detail.

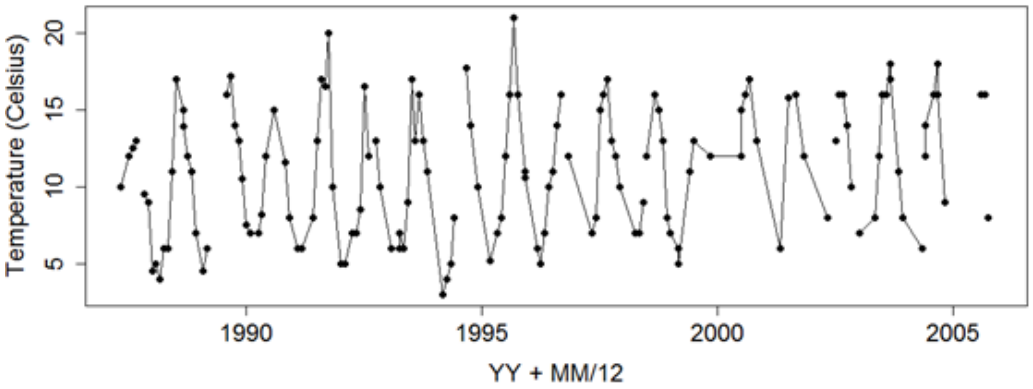


Figure 2: Figure 2: Cailness temperature over time

Solution

Temperature records- here the approach since we have approximately monthly data requires both a trend and seasonality term, since we expect that temperature will be seasonal. This is a very similar problem to what we will look at with the Central England temperature record, in the lab. I would expect to see a time series plot of these data (with time expressed as year.decimal part of the year), and a seasonality plot, where each temperature is plotted against position within the year. This latter plot will for sure show a seasonal cycle and we could easily imagine fitting some sort of sin/cosine curve to this, hence the idea to use a harmonic regression. However we also have a problem in that there is some missing data, and so as a discussion point we could consider how best to impute these missing values.

Imputation rules (this again would be more typical of an exam style question): Possible answers here could range from replacement by a simple average of the values surrounding the missing value, to a more complex approach which would also use the seasonal signal (e.g. if the value was missing in January 2006, you could replace the missing value by the average of the January 2005 and Jan 2007 values). You could also use the fitted sine/cosine seasonal curve as a tool for imputation.

You could be asked to look at the original equation for the harmonic regression (which is non-linear in the params) and show how it can be linearised (straight from the notes).

Task 3

The following excerpts are taken from the National Snow and Ice Data Centre (NSIDC) site, concerning arctic sea ice:

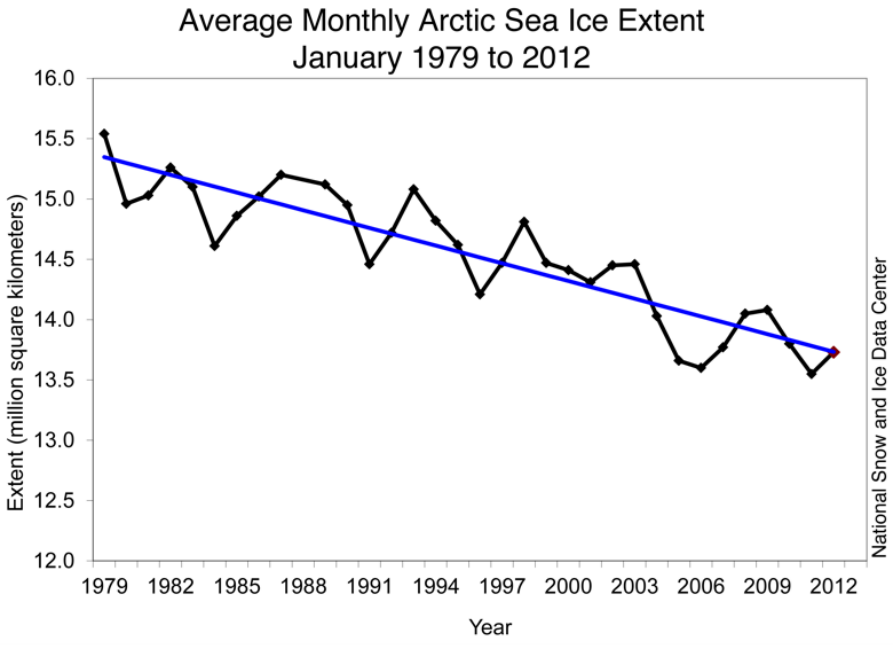
Statement 3a

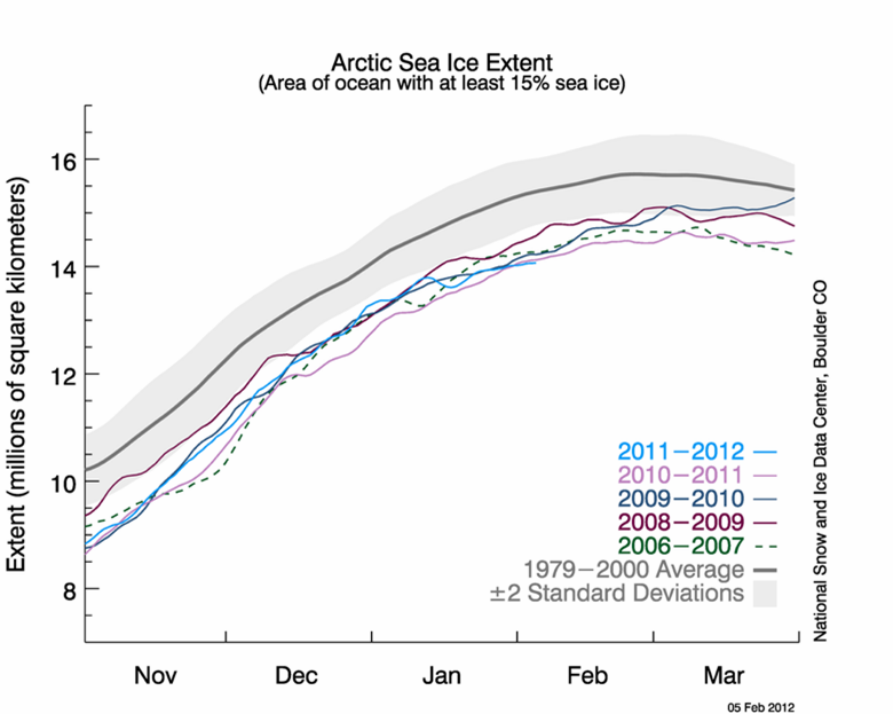
"Arctic sea ice extent in January 2012 averaged 13.73 million square kilometers (5.30

million square miles). This is the fourth-lowest January ice extent in the 1979 to 2012 satellite data record, 1.10 million square kilometers (425,000 square miles) below the 1979 to 2000 average extent. Including the year 2012, the linear rate of decline for January ice extent over the satellite record is 3.2% per decade. Based on the satellite record, before 2005, average January ice extent had never been lower than 14 million square kilometers (5.41 million square miles). January ice extent has now fallen below that mark six out of the last seven years."

Statement 3b

"The growth rate for Arctic sea ice in January was the slowest in the satellite record. After growing relatively quickly early in January, ice extent declined briefly in the middle of the month, and then grew more slowly than normal for the rest of the month. Overall, the Arctic gained 765,000 square kilometers (295,000 square miles) of ice during the month. This was 545,000 square kilometers (210,000 square miles) less than the average ice growth rate for January 1979 to 2000."





Discuss critically the two statements above concerning the 2012 sea-ice making reference to Figures 3.1 and 3.2, specifically the decrease per decade. Comment on the statistical methods and assumption which could be used to fit the models shown in the Figures. How would you interpret Figure 3.2 in terms of the comparison of the 5 curves shown to the 1979-2000 average curve?

Solution

Statement 3a, several aspects to comment on, the most straightforward concerns the linear rate of decline (3.2% per decade), no uncertainty mentioned. Fig 3.1, shows the straight line fit and the fluctuations around the line (why no uncertainty bands?). The second aspect to comment on really is about the rank information, fourth lowest, and January ice extent fallen below 14 million km^2 6 out of last 7 years (these latter comments would be picked up when we discuss extremes in the next section of the lectures). Figure 3.2 shows several curves, including a global average curve and its standard deviation band and then the curves for individual decades. We can see that several of the decade curves lie outside the grey bands, all are below the mean curve.

Task 4

Carbon dioxide concentrations are routinely measured at many places around the globe- one such data set is shown below with also some text describing how the plot and smooth curve was produced.

"To reduce noise in the determination of the global estimate due to atmospheric variability and measurement gaps, we fit a smooth curve to the weekly measurements.

To approximate the long-term trend and average seasonal cycle at a site, a function of the form

$$S(t) = \alpha_o + \alpha_1 t + \alpha_2 t^2 + \sum_{k=1:4} [b_{2k-1}\sin(2\pi kt) + b_{2k}\cos(2\pi kt)]$$

is fitted to the measurements [Thoning et al., 1989]. The above function includes 3 polynomial parameters (quadratic) and 8 harmonic parameters, associated with the sine and cosine terms which can be converted to amplitude and phase of each harmonic, if desired."

Figure 4 shows the smooth curve, $S(t)$, in red fitted to weekly background CO2 measurements from Ascension Island for 2000-2009.

Discuss the approach described in the statement concerning Figure 4 to quantify **trend**.

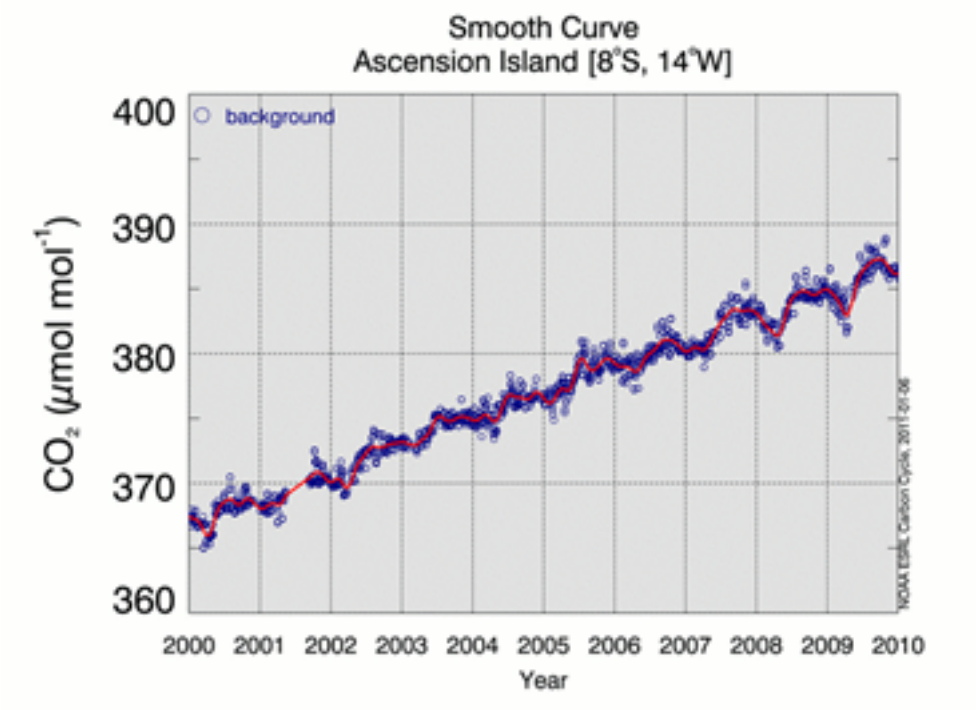


Figure 3: Figure 4: CO_2 data from Ascension Island

Solution

Critically, we have time series data, there are gaps, so a smooth curve is fitted, using a quadratic as in the equation. However there is also a need to include some cyclical terms, again introduced in the linear modelling framework using sin and cosine terms. Very standard regression modelling, but could and should also pick up on the fact there is no mention of the error terms or assumptions and since this is time series, we could expect that observations at a weekly scale are correlated, and what effect that might have on the estimates and any inference – i.e. less independent data so we might expect the standard estimates of our parameters to increase. An alternative approach may be to use a more flexible method (e.g. smoothing/GAM) to incorporate the seasonal pattern but with a linear term for trend. Models could be compared using an approximate F test. The advantage of the parametric approach adopted is the (relative) ease of interpretation.

Task 5

The following excerpt is taken from the European Environment Agency web site, concerning one of their key indicators (CSI012) for European temperature (accessed Jan 2010):

"The Earth has experienced considerable temperature increases in the last 100 years, especially in the most recent decades. These changes are unusual in terms of both magnitude and rate of change. The rate of change in the global average temperature is accelerating from 0.08°C per decade over the last 100 years, to 0.13°C per decade over the past 50 years up to 0.23°C per decade over the last 10 years (all values represent land & ocean area) (IPCC, 2007a). As such the indicative target (to keep climate change within 'safe' limits) of 0.2°C per decade has been exceeded in the recent years."

The figure below shows the time series plot of the CSI012 indicator, which has been developed to address specific policy issues concerning the trend and rate of change in the European annual and seasonal temperature.

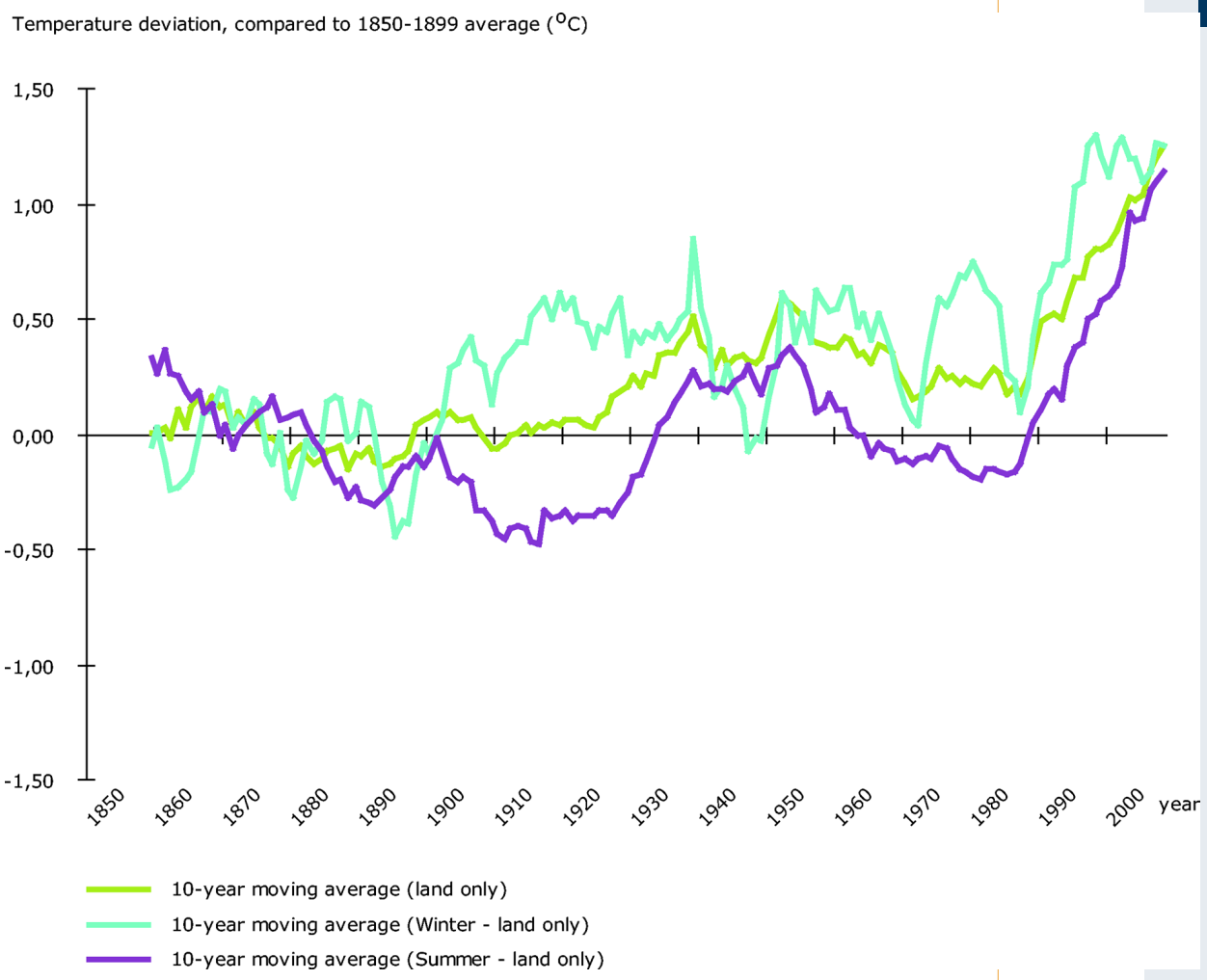


Figure 4: Figure 5: CSI012 European Land Temperature Deviations

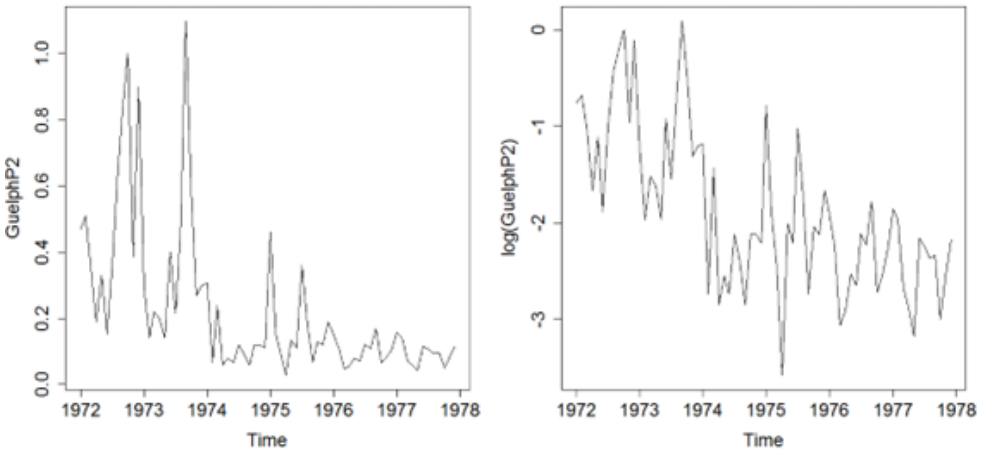
Discuss critically the statement and the above figure and how they might be related. Suggest how you might model such temperature data above to address the policy issues.

Solution

This example is again all about trends, expressed one presumes as a linear trend, and a slope in the quote. The interesting thing is that the slope is different in different time periods and over different lengths of time. Again, no uncertainty is quoted. The plot shows deviations using a 10 year moving average, so a smoothed plot, and we can also see differences between summer and winter, suggesting temperature change may indeed be occurring at different rates in different times of the year. How might we model absolute temperature and the deviations? We can certainly imagine therefore that temperature will have a strong seasonal cycle (so could use a harmonic regression), that a simple linear regression is not likely to be complex enough since the rate of change is changing, so could fit some sort of piecewise regression but what is the justification for the changepoints? Or we could fit some sort of smooth curve (not necessarily monotonic).

Task 6

The plot below shows the Monthly time series of phosphorous (P) concentrations in mg/l, Speed River, Guelph, Ontario between 1972.1–1977.1.



Comment subjectively on the patterns in the plots P in the River Guelph and how you might statistically model any trends/patterns over time observed.
The output below shows the summary information for the Mann Kendall score for all time points and for the data split by month (Seasonal). Use the information to carry out a Mann Kendall test and seasonal Mann Kendall test and compute the Mann Kendall correlation coefficient for each. Interpret your results.

```
All Time Points          Seasonal
D = 2546.482             D = 178.4556
S =  -1171               S = -99
Var(S) = 42292.33        Var(S) = 337
```

The output below shows the output for an estimate of Sen's slope for the River Guelph phosphorus data. Comment on how this slope estimate is computed and how it differs from a standard linear regression slope. Why might we want to use a Sen's slope rather than a standard regression slope in this case?

```
Sen's slope and intercept

slope:  -0.0248
95 percent confidence interval for slope
-0.0174 -0.0323

intercept: -0.956
nr. of observations: 72
```

Solution
The plots show a decreasing trend. The plot of the data on the original scale indicates non-constant variability but this seems to have been dealt with by the application of the log transformation. Subjectively there may be an indication of a change point in 1974 were there is a large jump in the level of P and so a changepoint analysis may be applied here.
The data are quite noisy but there is some indication of a seasonal pattern. This could be incorporated using a harmonic regression term within a linear regression. Alternatively, we could assess if there was a trend using a seasonal Mann Kendall test to check if there was any evidence of a monotonic change in P over the time period considered.

```
All Time Points          Seasonal
D = 2546.482             D = 178.4556
S =  -1171               S = -99
Var(S) = 42292.33        Var(S) = 337
```

All data:

- Tau $= S/D = -1171/2546.482 = -0.46$
- $Z = (-1171 + 1)/\sqrt{42292.33} = -5.7.$

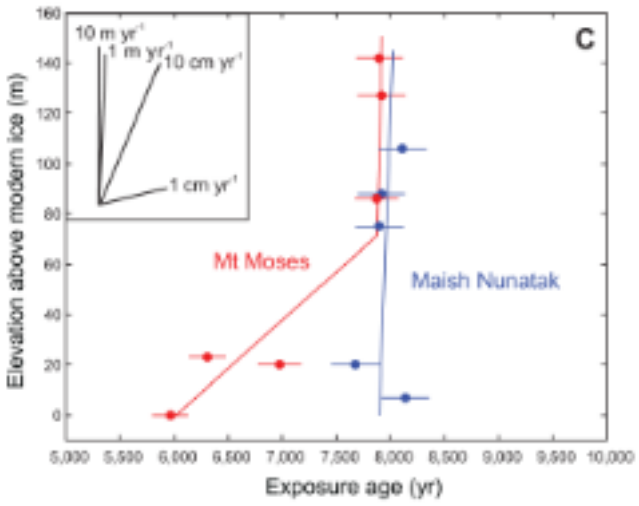
Seasonal:

- Tau $= S/D = -99/178.4556 = -0.55$
- $Z = (-99 + 1)/\sqrt{337} = -5.3.$

Both the Z score for all data together and the Z score after the seasonal adjustment are greater than the critical value of 1.96 (in absolute value terms). So there is evidence of a statistically significant monotonic trend in the data. The tau values of -0.46 for all data together, and -0.55 for the data taking into account the seasonal component both suggest there is a moderate negative monotonic trend. This supports the initial impression gained from inspection of the data.
Sen's slope is a non-parametric estimator of the regression slope. While a standard regression line fitted using OLS minimises the sum of squared residuals, Sen's slope is computed by calculating the median of the slopes which join each pair of points. It is thought to be more robust when there are doubts about the distribution of the data because it is non-parametric and therefore does not require the same distributional assumptions as OLS.

Task 7

The figure below shows a regression example from a recent paper in Science (Johnson et al, 2014) which investigated the rate of thinning of the Pine Island glacier. They fit a "2-segment, piecewise-linear age-elevation history to the Mt Moses data".



Explain what the phrase in bold above means, and write down the equation for a 2 stage regression, assuming a **known changepoint**. Explain how the parameters of the model can be estimated. Comment briefly on the figure above.

Solution
The model that the authors have fitted is one where there are two straight lines which are constrained to meet at a changepoint.
A change-point could be defined as a point that separates a series of observations into two groups, each following a different model, in this case a different slope and intercept.
Known changepoint:

$$f(x_i) = \begin{cases} \alpha_1 + \beta_1 x_i & x_i \leq \delta \\ \alpha_2 + \beta_2 x_i & \delta \leq x_i \end{cases}$$

constrained such that

$$\alpha_1 + \beta_1 \delta = \alpha_2 + \beta_2 \delta$$

Parameters can then be estimated by ordinary least squares. Plot shows the diverging line, no uncertainty estimates, nor justification of where the changepoint occurs.

2 Part B: Models for extremes

Task 8

If $X_1, \ldots, X_n$ is a sequence of independent standard exponential $Exp(1)$ variables,

(a) Show that $F(x) = 1 - e^{-x}$ for $x > 0$.
(b) Show that, for $a_n = 1$ and $b_n = \log(n)$, the limit distribution of M_n as $n \to \infty$ is the Gumbel distribution, corresponding to $\zeta = 0$ in the GEV family.

Solution

(a) If $X_1, \ldots, X_n \sim Exp(1)$,

- range of x is $X > 0$
- pdf of $Exp(\lambda)$ is $f(x) = \lambda e^{-\lambda x}$, so for $Exp(1)$, $f(x) = e^{-x}$

$$\begin{aligned} \therefore F(x) &= P(X \leq x) \\ &= \int_0^x f(t)dt \\ &= [-e^{-t}]_0^x \\ &= -e^{-x} - (-e^0) \\ &= 1 - e^{-x} \qquad (\text{for } x > 0) \end{aligned}$$

(b) Let $M_n = \max\{X_1, \ldots, X_n\}$, let $a_n = 1$ and let $b_n = \log(n)$. Then:

$$Pr\left\{\frac{M_n - b_n}{a_n} \leq z\right\} = F^n(a_n z + b_n) \to G(z) \text{ as } n \to \infty$$

$$
\begin{aligned}
F_n(a_n z + b_n) &= F^n(z + \log(n)) \\
&= [1 - e^{-(z+\log(n))}]^n \\
&\to \left[1 - \frac{e^{-z}}{n}\right]^n \quad \text{as } n \to \infty \\
&= -\exp(-\exp(-z))
\end{aligned}
$$

Note:

$\exp(x) = \lim_{n\to\infty} \left(1 + \frac{x}{n}\right.$

Task 9

If $X_1, \ldots, X_n$ is a sequence of independent standard exponential $Exp(\lambda)$ variables,

(a) Show that $F_n(x) = (1 - e^{-\lambda x})^n$ for $x > 0$, where $F_n(x)$ is the distribution of $X(n)$.

(b) Find the distribution function of $X_{(1)}$, and show that $X_{(1)}$ has an Exponential distribution with parameter $n\lambda$.

Solution

(a) $F(x) = 1 - e^{-\lambda x}$

$$
\begin{aligned}
F_n(x) &= [P(X \le x)]^n \\
&= \left[\int_0^x f(t)dt\right]^n \\
&= \left[\int_0^x \lambda e^{-\lambda t} dt\right]^n \\
&= \left[\left[e^{-\lambda t}\right]_0^x\right]^n \\
&= \left[1 - e^{-\lambda x}\right]^n
\end{aligned}
$$

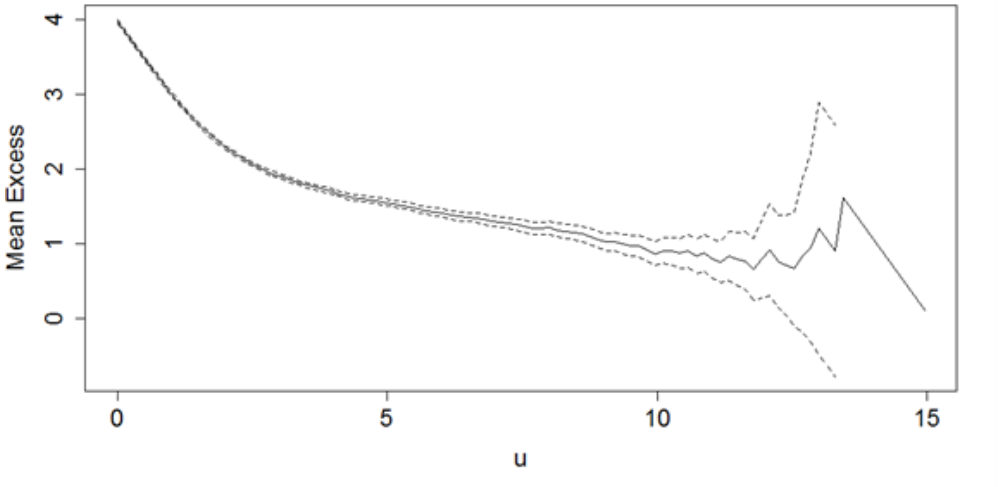
(b) Let $X_{(1)} = \min\{X_1, \ldots, X_n\}$. Then:

$$
\begin{aligned}
\Pr\left\{X_{(1)} \le t\right\} &= \Pr\{\min\{X_1, \ldots, X_n\} \le t\} \\
&= 1 - \Pr\{\min\{X_1, \ldots, X_n\} \ge t\} \\
&= 1 - \Pr\{X_i \ge t \text{ for all } i\} \\
&= 1 - (e^{-\lambda t})^n \\
&= 1 - e^{-\lambda n t}
\end{aligned}
$$

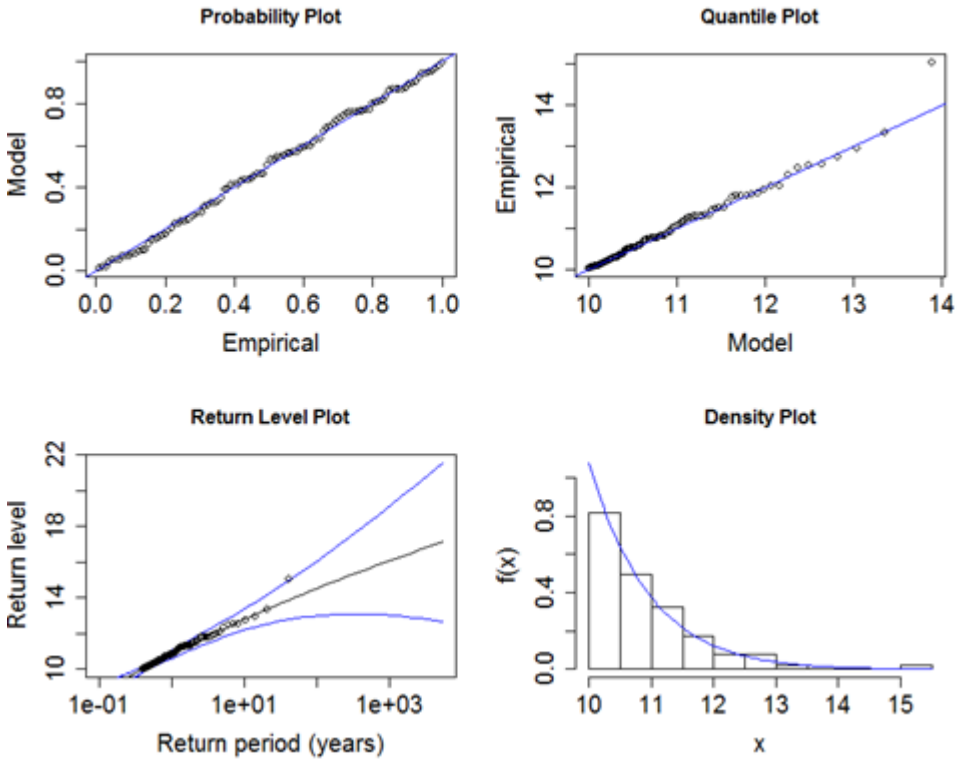
$$\therefore X_{(1)} \sim Exp(n\lambda)$$

Task 10

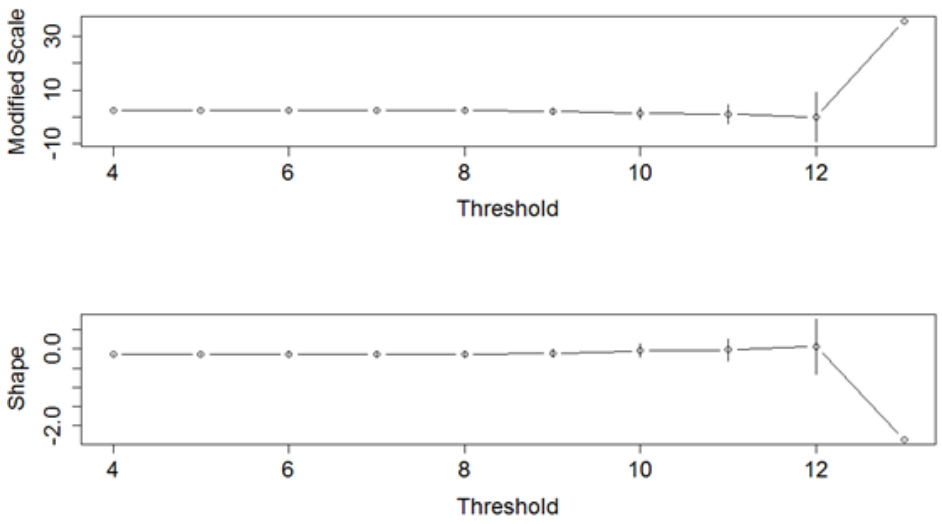
(a) Daily wind speed data are available for a location in the Netherlands. A POT modelling approach has been used to model the data. The figure below shows a mean residual life plot for the series. Comment on how you could use this plot to identify a suitable threshold for the POT model. What threshold would you select here?



(b) The figure below shows the diagnostic plots for a fitted POT model for the wind speed data. Comment on the model fit using these plots.



(c) The threshold selected was 10. Comment on this choice using the sensitivity analysis below.



(d) The R output for the model is shown below.

```
$threshold
[1] 10

$nexc
[1] 105

$conv
[1] 0

$nllh
[1] 92.17218

$mle
[1]  0.9278907 -0.0473220

$rate
[1] 0.00684485

$se
[1] 0.12364549 0.09084164
```

Estimate the 1 in 100 year wind speed.
Solution

(a) Looking for area where (after taking confidence bands into account) the mean excess is linearly related to the threshold. Not immediately obvious, somewhat subjective. I would say somewhere between 10 and 13. Start as low as possible so 10 would be a good threshold to start with.

(b) Quantile and probability plots both have good agreement between points (corresponding to the sample) and the theoretical line, therefore the fit seems reasonable. Histogram also seems to have good agreement between the sample and the solid line which represents the estimated density using the fitted model. Have to be careful when interpreting histograms as interpretation can

change with the number of bins used, particularly when the sample size is small. Return level plot, points lie within confidence interval in general. Another indication the model is suitable.

(c) The threshold value of 10 seems like a good threshold. There is very little variability in the sensitivity analysis fit, the parameter estimates seem constant across a range of thresholds, only at 12 and above do estimates seem to vary. Maybe could argue we could look at thresholds even lower than 10 —- need to balance what is truly extreme, while ensuring there is sufficient data available above the threshold to use to estimate our model.

Task 11

The generalized extreme value distribution (GEV) has three parameters and distribution function

$$G(z) = \exp\left\{-\left[1+\xi\frac{(z_i-\mu)}{\sigma}\right]^{\frac{-1}{\xi}}\right\}$$

GEV is often used when modelling block maxima. Identify the particular family and write down its distribution function when ξ is assumed equal to 0.
Show that when $G(z_p) = 1-p$, then

$$z_p = \begin{cases} \mu - \sigma\frac{[1-\{-\log(1-p)\}^{-\xi}]}{\xi} & \text{for } \xi \neq 0 \\ \mu - \sigma[\log\{-\log(1-p)\}] & \text{for } \xi = 0 \end{cases}$$

For the distribution G in the case $\xi = 0$, show that if z_p is plotted against $\log[-\log(1-p)]$, the plot should be linear.
Solution

$$G(z) = \exp\left\{-\left[1+\xi\frac{(z-\mu)}{\sigma}\right]^{\frac{-1}{\xi}}\right\}$$

Set $G(z_p) = 1-p$ and solve equation above for z_p:

$$\exp\left\{-\left[1+\xi\frac{(z_p-\mu)}{\sigma}\right]^{\frac{-1}{\xi}}\right\} = 1-p$$
$$\left\{-\left[1+\xi\frac{(z_p-\mu)}{\sigma}\right]^{\frac{-1}{\xi}}\right\} = \log(1-p)$$
$$-\left[1+\xi\frac{(z_p-\mu)}{\sigma}\right] = \log(1-p)^{-\xi}$$
$$\left[1+\xi\frac{(z_p-\mu)}{\sigma}\right] = -\log(1-p)^{-\xi}$$
$$\xi\frac{(z_p-\mu)}{\sigma} = \{-\log(1-p)^{-\xi}\}-1$$
$$\frac{(\mu-z_p)}{\sigma} = \frac{1}{\xi}[1-(-\log(1-p)^{-\xi})]$$
$$-z_p = -\mu+\frac{\sigma}{\xi}(1-(-\log(1-p)^{-\xi}))$$
$$z_p = \mu-\frac{\sigma}{\xi}(1-(-\log(1-p)^{-\xi}))$$

When $\xi = 0$, we have a Gumbel distribution:

$$\exp\left[-\exp\left(\frac{-(z_p-\mu)}{\sigma}\right)\right] = 1-p$$
$$\left[-\exp\left(\frac{-(z_p-\mu)}{\sigma}\right)\right] = \log(1-p)$$
$$\left(\frac{-(z_p-\mu)}{\sigma}\right) = \log(-\log(1-p))$$
$$-z_p+\mu = \sigma\log(-\log(1-p))$$
$$-z_p = -\mu+\sigma\log(-\log(1-p))$$
$$z_p = \mu-\sigma\log(-\log(1-p))$$

z_p is the return level associated with the return period $1/p$.
Let $s_p = \log(-\log(1-p))$. Then $z_p = \mu-\sigma\log(-\log(1-p)) = \mu-\sigma s_p$ when $\xi = 0$. Therefore, z_p is linearly related to $\log(-\log(1-p))$ in the case $\xi = 0$.