Environmental Statistics

Week 4: Time Series — Assessing Changes Over Time

Jafet Belmont and Craig Wilkie

Recap

- Last week, we discussed some different sampling methods and how to determine the sample size.
- We also looked at monitoring networks and how they were set up to collect data.
- Many monitoring networks measure the same variables over a long time period.
- This week we will therefore be looking at time series methods and their applications to environmental data.

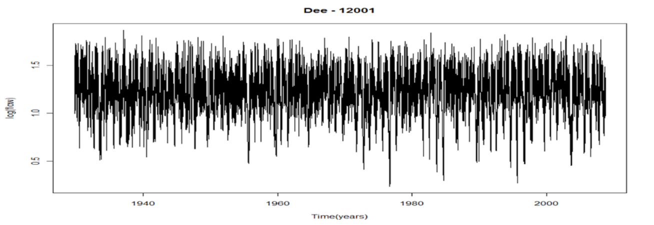
Time Series

What is a Time Series?

What is a Time Series?

- A **time series** is a sequence of measurements on the same object made over time.
- For example, we might measure the level of carbon dioxide (CO_2) in a town every day for a year.
- The purpose of making such measurements is to understand how our variable of interest has changed over time.
- For example, a government would be keen to know if air pollution levels are getting better or worse.

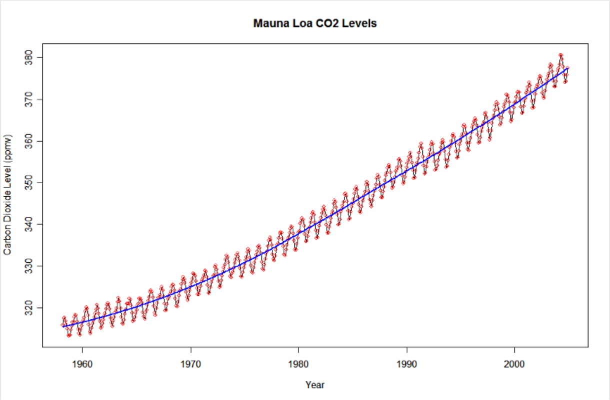
- We can write our set of time series data as $y_1, \ldots, y_T$, where y_i is the observation at time point i, and T is the total number of observations.
- Time series data are typically **not independent**. There will often be correlation between consecutive observations.
- This dependency structure must be taken into account when modelling.



Example: Mauna Loa

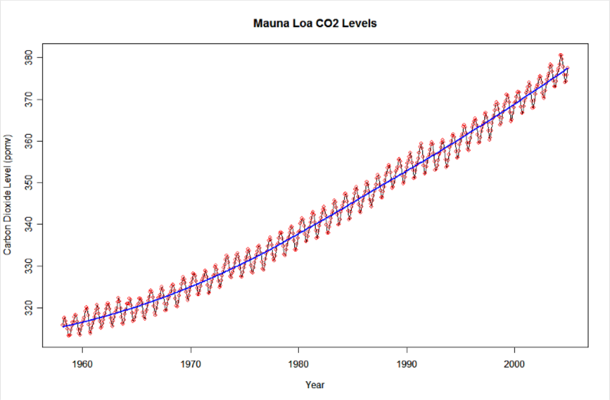
- Mauna Loa in Hawaii is one of the biggest and most active volcanoes in the world.
- CO_2 levels have been monitored since 1958.
- One of the first sites worldwide where increasing CO_2 levels were identified.

Example: Mauna Loa



University of Glasgow

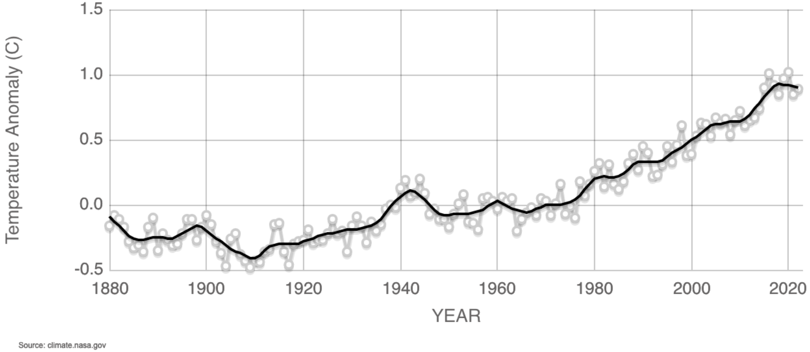
Example: Mauna Loa



- We can observe a clear trend, and also a seasonal pattern.
- It may be sensible to standardise the data and represent all observations in terms of '**anomalies**', i.e. their deviation from the starting point (1960 mean level).

Example: Global Temperature

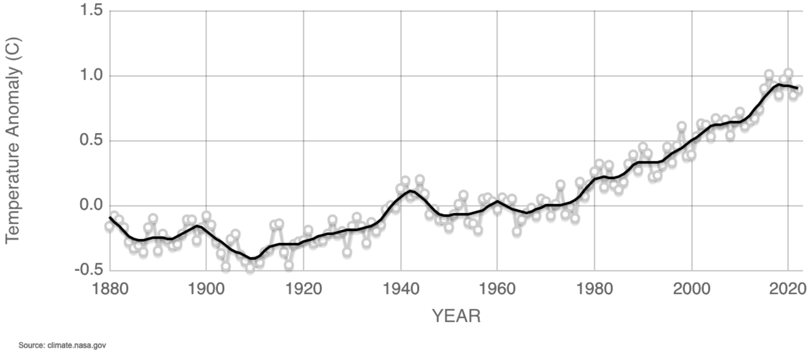
- The plot below shows the global temperature anomaly (the current value compared to the average from 1951–1980)
- How would you describe the change in temperature?



`https://climate.nasa.gov/vital-signs/global-temperature/`

Example: Global Temperature

- Often this is just simply ‘say what you see’.
- *“The overall temperature seems fairly stable until around 1980, but then rises substantially, reaching a peak at the present day.”*



https://climate.nasa.gov/vital-signs/global-temperature/

Assessing Change Over Time

Assessing Change

- We still have a number of questions to consider when we think about measuring or understanding changes.
- Is routine monitoring data useful/adequate/sufficient for environmental change detection?
- How much data do we need? How long should our time series be?
- Are the commonly used ecological tools for measuring environmental change statistically rigorous?
- How can we use statistical methods in a way that is easily understood by policy makers?

Ecological Trend

- The purpose of time series modelling is to identify any **trends** which exist in the dataset.
- But what exactly is a trend?

Ecological Trend

- The purpose of time series modelling is to identify any **trends** which exist in the dataset.
- But what exactly is a trend?
- It depends who you ask.

Ecological Trend

- The purpose of time series modelling is to identify any **trends** which exist in the dataset.
- But what exactly is a trend?
- It depends who you ask.
- The Joint Nature Conservation Council (JNCC) define it as *"a measurement of change derived from a comparison of the results of two or more statistics"*
- This is often considered to be the *ecological* definition of trend: a change (in terms of percentage or some index) between two timepoints.

Statistical Trend

- In statistics, the definition of a trend is often more wide-ranging:
  - A long-term change in the mean level (Chatfield, 1996)
  - Long-term movement (Kendall and Ord, 1990)
  - Long-term behaviour of the process (Chandler, 2002)
  - The non-random function $\mu(t) = E(Y(t))$ (Diggle, 1990)
- We may be interested in trends in mean, variance or extreme values.
- Trends are not limited to linear or monotonic patterns.

Simple Linear Trend

- We can represent a simple linear trend using the standard notation:

$$Y_t = \beta_0 + \beta_1 x_t + \epsilon_t.$$

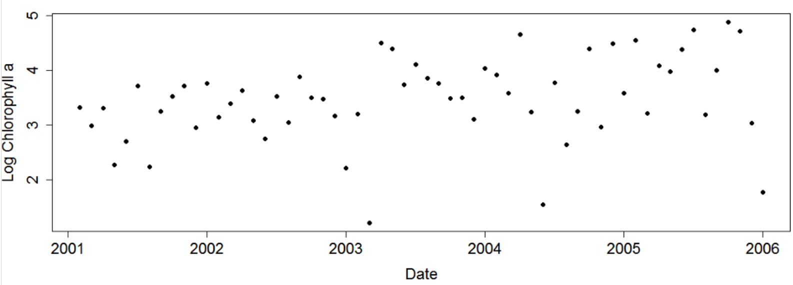
- Here, β_0 is an intercept and β_1 represents the slope (trend).
- This is just a standard linear model, with all the usual assumptions (normality, constant variance, independence etc.).
- This model therefore doesn't account for any seasonality or autocorrelation in our data.

Example: Chlorophyll

- We observe monthly chlorophyll levels in a lake between 2001 and 2006.

- We can fit a linear model of the form:

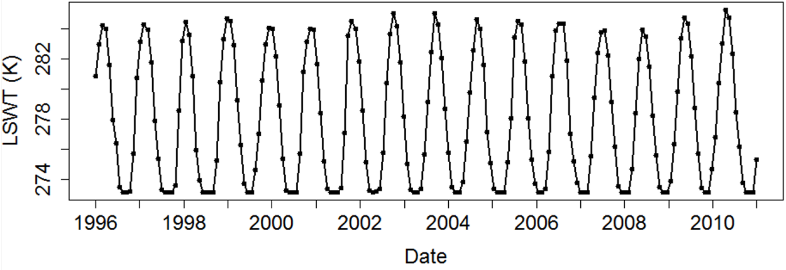
$$\text{LogChlorophyll} = \beta_0 + \beta_1 \text{ Date} + \text{error}.$$



Example: Lake Nam

- Lake Nam (Namtso) is a mountain lake in Tibet.

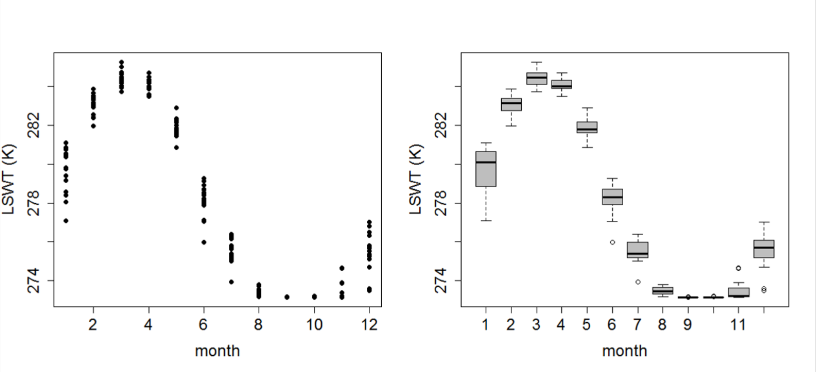
- The mean surface water temperature was measured monthly between 1996 and 2011.



Seasonal Patterns

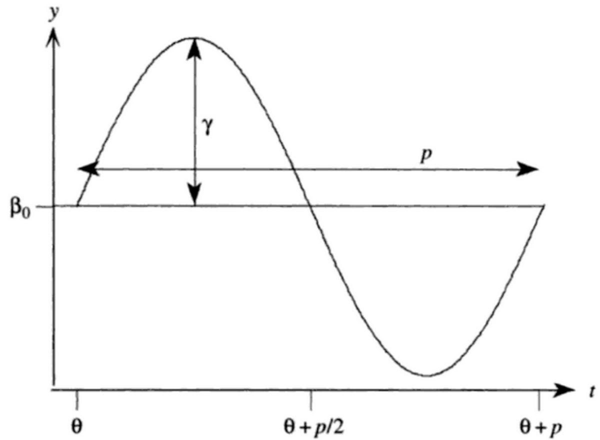
- Many environmental time series have some sort of **periodicity** (e.g. a monthly pattern in temperature).
- We can produce some form of seasonality plot to understand this better.
- The **period** is the time interval between consecutive peaks or troughs.
- A **seasonal component** of a dataset is a regular fluctuation with a period of one year or less.

Seasonal Patterns



- Plotting the data by month indicates a clear seasonal pattern.
- There is a peak in Month 3 and a trough in Months 9/10.

- The monthly pattern is very similar to a sine wave, and we can use this feature in our modelling.
- This is known as harmonic regression, and is suitable when we have a regular seasonal trend.



Harmonic Regression

- Harmonic regression is based on an equation of the form

$$Y_t = \beta_0 + \gamma \sin\left(\frac{2\pi[x_t - \theta]}{p}\right) + \epsilon_t.$$

- Here, γ is the amplitude of the wave, p is the period of the wave, and θ represents the 'position' on the wave (in radians).

Harmonic Regression

- Harmonic regression is based on an equation of the form

$$Y_t = \beta_0 + \gamma \sin\left(\frac{2\pi[x_t - \theta]}{p}\right) + \epsilon_t.$$

- Here, γ is the amplitude of the wave, p is the period of the wave, and θ represents the 'position' on the wave (in radians).

- However, it can often be more convenient to rewrite this in the form of a simple multiple regression model, taking advantage of the double angle formula.

Proof

- Given that $\sin(a-b) = \sin(a)\cos(b) - \cos(a)\sin(b)$, we can show that:

$$\gamma \sin\left(\frac{2\pi[x_t-\theta]}{p}\right) = \gamma \sin\left(\frac{2\pi x_t}{p} - \frac{2\pi\theta}{p}\right)$$
$$= \gamma\left[\sin\left(\frac{2\pi x_t}{p}\right)\cos\left(\frac{2\pi\theta}{p}\right) - \cos\left(\frac{2\pi x_t}{p}\right)\sin\left(\frac{2\pi\theta}{p}\right)\right]$$

- Since π, θ and p are known, we can create new regression parameters $\gamma_1 = \gamma\cos\left(\frac{2\pi\theta}{p}\right)$ and $\gamma_2 = \gamma\sin\left(\frac{2\pi\theta}{p}\right)$.

Harmonic Model

- The final harmonic regression model can thus be written:

$$Y_t = \beta_0 + \gamma_1 \sin\left(\frac{2\pi x_t}{p}\right) + \gamma_2 \cos\left(\frac{2\pi x_t}{p}\right) + \epsilon_t$$

- Our new parameters γ_1 and γ_2 control the seasonal trends, with p representing the period.

- β_0 is still the intercept term, which can also be interpreted as the overall mean.

Harmonic Model

- The final harmonic regression model can thus be written:

$$Y_t = \beta_0 + \gamma_1 \sin\left(\frac{2\pi x_t}{p}\right) + \gamma_2 \cos\left(\frac{2\pi x_t}{p}\right) + \epsilon_t$$

- Our new parameters γ_1 and γ_2 control the seasonal trends, with p representing the period.

- β_0 is still the intercept term, which can also be interpreted as the overall mean.

- Note that this is still a linear model, since it is linear in the coefficients.

Limitations of Harmonic Regression

- The standard harmonic regression assumes that we have the *same seasonal pattern* each year, but this may not always be appropriate.
- There are many more sophisticated models available if this assumption does not hold.
- Some are still based on sine and cosine waves, while others may use autocorrelation functions or a form of semiparametric smoothing.

Limitations of Harmonic Regression

- The seasonal variation can sometimes be so strong that it obscures the overall trend (or any other patterns).
- In most cases, we are not actually particularly interested in knowing about the seasonal trend, and we treat it as a nuisance factor to account for in our model.
- Our primary interest is usually in understanding the longer-term trends in our data.
- Therefore, we often try to remove or extract this seasonal pattern when analysing time series.

Time Series Model

- We can therefore think of our overall time series model in the following form:

$$X = \text{trend} + \text{seasonal component} + \text{error}.$$

- In terms of mathematical notation, we can write this as

$$X_t = m_t + s_t + \epsilon_t.$$

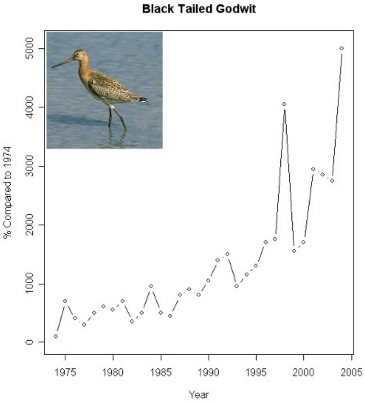
- Our error, ϵ_t, is assumed to be random and follow the Normal distribution: $\epsilon_t \sim \text{Normal}(0, \sigma^2)$.

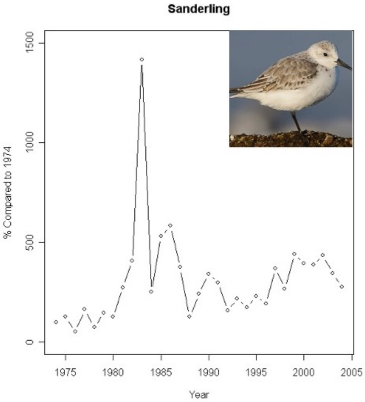
Understanding Trend

- We have now identified a method for isolating the trend in our model.
- However, we still have to work out the best way to explore and understand this trend.
- We want to know the size of the trend, but also have to assess whether it is linear, and also test for statistical significance.
- A variety of models and techniques exist for exploring our trend.

Non-Parametric Trend Estimation

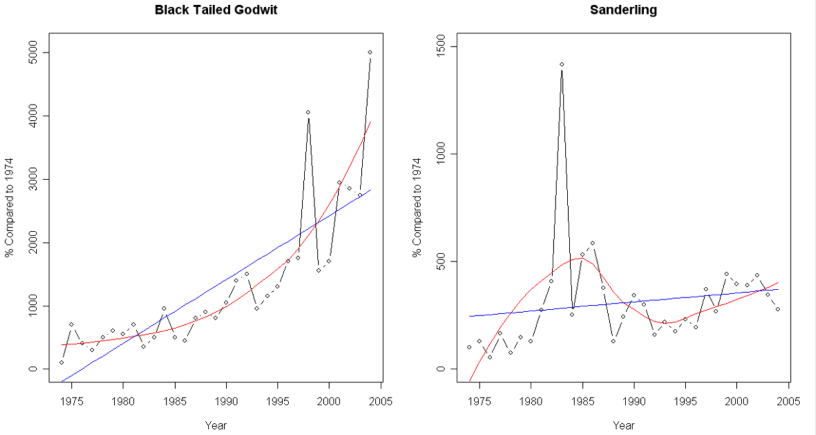
- We have collected annual data on the population of two birds between 1975 and 2005.
- What are the trends? Are they significantly different from zero?





Example: Bird Populations

- We have fitted two models to attempt to assess the trends for each bird.
- The blue line is a linear regression. The red line is a more flexible additive model.



Assessing significance

- Both models indicate the overall trend, but they do not test for significance.
- We therefore cannot be sure whether the changes are ‘genuine’ or are simply down to random variation.
- We can use non-parametric approaches to assess the trend in our data.
- Two such approaches are the Mann-Kendall test and the Seasonal Kendall test.

Mann-Kendall test

- The **Mann-Kendall test** is commonly used to detect trends in environmental, climate and hydrological data.
- It looks for a consistent increase or decrease in a trend over time (not necessarily linear).
- It is commonly used for short time series, where we may not have sufficient data for more sophisticated approaches.

Mann-Kendall statistic

Assume that we have an ordered dataset $z_1, \ldots, z_T$

1. Compute **all** possible differences $d = z_j - z_k$ where $j > k$
2. Create an indicator function $\text{sign}(z_j - z_k)$ such that:

$$\text{sign}(z_j - z_k) = \begin{cases} 1 & \text{if } (z_j - z_k) > 0 \\ 0 & \text{if } (z_j - z_k) = 0 \\ -1 & \text{if } (z_j - z_k) < 0 \end{cases}$$

3. The Mann-Kendall statistic, S, is given by:

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} \text{sign}(z_j - z_k)$$

Mann-Kendall hypothesis test

- Our test statistic, S, measures the size and direction of the trend:
  - A positive value of S suggests that the data are increasing over time (i.e. an upward trend).
  - A negative value of S suggests a downward trend.
  - $S = 0$ implies no trend.

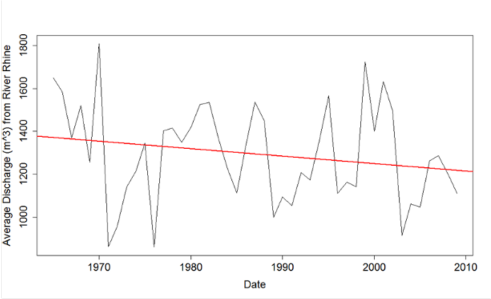
- We can carry out a hypothesis test to assess whether S is significantly different from zero:

$$H_0 : \text{our data are independent random realisations (no trend).}$$
$$H_1 : \text{there is a significant trend in our data.}$$

- We compare the test statistic to a standard normal distribution $Z_{(1-\alpha/2)}$.

Fitting in R

- We can use the `mk.test` function in R's `trend` package.
- Here we see a p-value of 0.16.
-



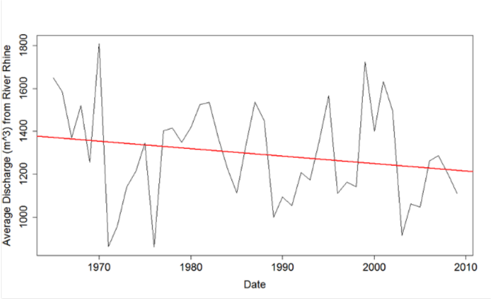

```
> mk.test(Q)
Mann-Kendall Test two-sided homogeneity test
Statistics for total series

H0: S = 0 (no trend)
HA: S != 0 (monotonic trend)

Statistics for total series
     S  varS    Z    tau  pvalue
1 -144 10450 -1.4 -0.145 0.16185
```

Fitting in R

- We can use the `mk.test` function in R's `trend` package.
- Here we see a p-value of 0.16.
- No evidence to reject H_0 (no trend present).




```
> mk.test(Q)
Mann-Kendall Test two-sided homogeneity test
Statistics for total series

H0: S = 0 (no trend)
HA: S != 0 (monotonic trend)

Statistics for total series
     S  varS    Z    tau  pvalue
1 -144 10450 -1.4 -0.145 0.16185
```

Kendall rank correlation coefficient

- We can also compute a rank correlation coefficient, τ, which measures the strength of our trend,

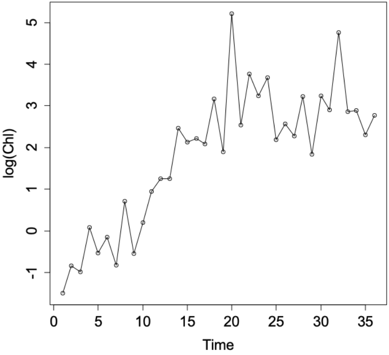
$$\tau = \frac{S}{D}.$$

- Here, $D = \frac{n(n-1)}{2}$, the number of pairwise comparisons used in the calculation of S.

- τ has a range $(-1, 1)$, similar to the standard correlation used in regression modelling.

Kendall rank correlation coefficient

- Chlorophyll levels in a lake have been measured over 36 years.
- Given that $S = 384$, compute τ to measure the strength of the trend.



$$D = \frac{n(n-1)}{2} = \frac{36 \times 35}{2} = 680$$

$$\tau = \frac{S}{D} = \frac{384}{680} = 0.61$$

Seasonal Kendall test

- The seasonal Kendall test accounts for seasonality by computing S for each of M seasons separately, then combining the results.
- For example, if we had monthly data, we might compute S separately for each month.
- Let S_j be the Mann-Kendall statistic for season j. Then, the overall statistic is given by:

$$S_k = \sum_{j=1}^{M} S_j$$

- Again, this can be compared to a standard normal distribution $Z_{(1-\alpha/2)}$.

Smoothing in Time Series

Smoothing in Time Series

- Environmental time series data are often complex and traditional parametric methods are difficult to implement.
- The relationship between our parameter of interest and time may not follow a linear pattern.
- We could simply keep adding polynomial functions, but this may become inefficient and lead to a model with too many parameters.
- It is often more elegant to consider an approach which uses **smoothing**.

Smoothing

- We can express the relationship between any response and explanatory variable as

$$y = f(x) + \varepsilon.$$

- Here, y is the response, x is our explanatory variable and $f()$ is a function that describes their relationship.
- Smoothing techniques are used to model $f()$ without specifying any specific statistical form of the underlying function.

Smoothing methods

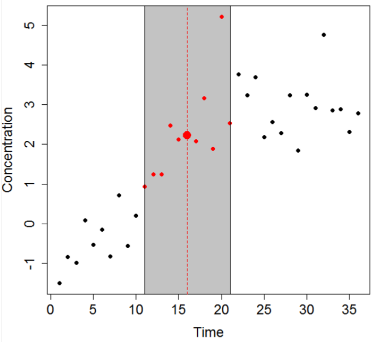
- There is a whole course on smoothing methods (Flexible Regression), and many of you will already have taken this.
- Therefore we will simply focus briefly on a couple of key methods which are used for environmental data.
- We will look at one method mainly used for descriptive purposes (LOWESS) and one which is used for estimation (penalised splines).

LOWESS

- LOWESS (LOcally WEighted Scatterplot Smoothing) is an approach which is often used to obtain a graphical illustration of our data.
- It involves carrying out a series of polynomial regressions on small regions of the data, and then combining them.
- The more datapoints we have in a region, the smoother our curve will be.
- This can be somewhat computationally intensive compared to simple moving average methods, but generally produces a smoother function.

LOWESS

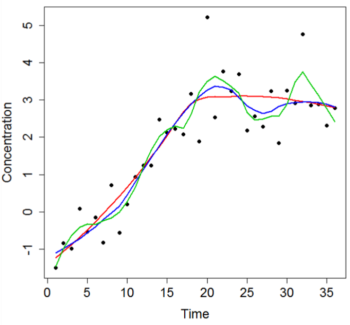
- Identify a target point, *x*.
- Construct a 'window' containing its *k* nearest neighbours.
- Fit a weighted polynomial to these *k* datapoints.



- We then choose a new target point and repeat until we have covered all timepoints.

Smoothing level

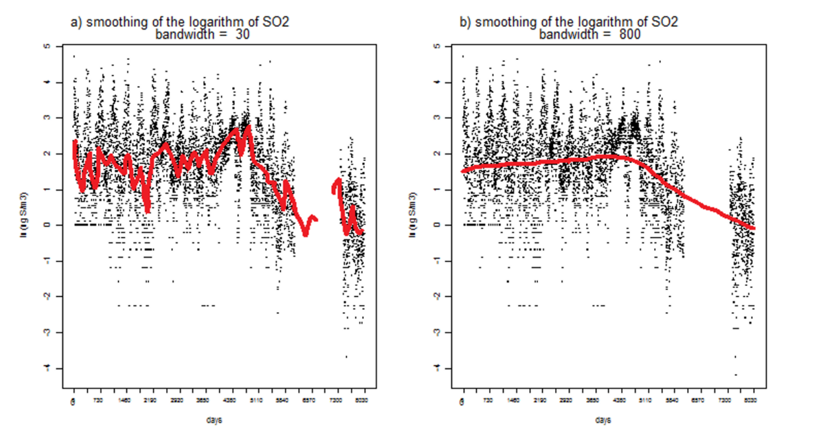
- We have to decide on the size of the window (i.e. the "bandwidth"). In R, the default is that each window contains two thirds of the data.
- We can fit these models in R using the `scatter.smooth` or `loess` functions.



- The different colours show different sizes of windows.
- The wider the window, the smoother the function (green narrowest, red widest).

Example: Air Quality

- SO_2 levels are measured daily over 30 years.
- The narrower bandwidth (left plot) leads to a gap where there are missing values. The wider bandwidth (right plot) leads to more smoothness (too smooth?).



LOWESS: Advantages and disadvantages

Advantages

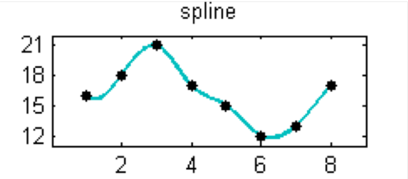
- Simple approach, with no need to specify the form of the relationship.
- Easy to fit in many statistical modelling packages.

Disadvantages

- Need to (manually) specify an appropriate bandwidth.
- Mainly suitable for exploratory analysis — no natural expression of uncertainties.
- Cannot be extended to model more complex relationships (like splitting into seasonal component and trend, or smooth interactions between variables).

Splines

- **Splines** are an alternative approach to constructing a smooth function.
- This approach uses piecewise polynomials to estimate the function $f(x)$.
- Spline functions are polynomial segments which are joined together smoothly at predefined subintervals.
- The points where the functions join together are known as **knots**.



Estimation with splines

- Our model takes the form:

$$Y_i = f(x_i) + \epsilon_i.$$

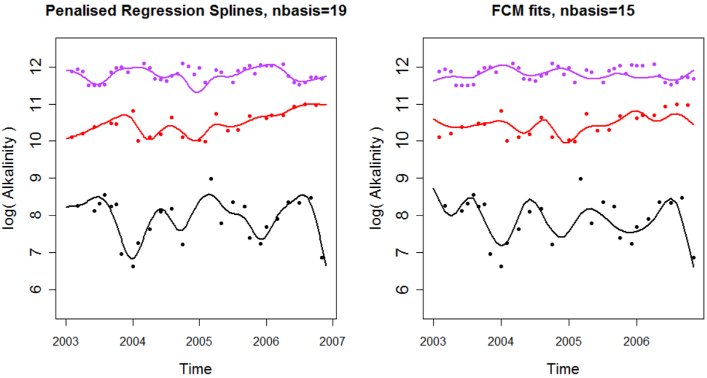
- We estimate the function $f()$ as

$$\hat{f}(x_i) = \sum_{k=0}^{p} \beta_k b_k(x_i)$$

- Here, $b_k()$ are a set of polynomial functions known as *basis functions* and β_k are their coefficients.
- We must decide in advance the value of p, which defines the number of basis functions used.

Estimation with splines

- Increasing the number of basis functions leads to a more 'wiggly' line.
- Too few basis functions might make the line too smooth, but too many might lead to overfitting.



Penalised Splines

- Choosing the correct number of basis functions can be difficult.
- Penalised splines (p-splines) avoids this issue. We can set a large number of basis functions, but then penalise the coefficients to encourage smoothness.
- This is a modified form of a standard linear regression, with a parameter λ that controls the smoothness of the estimator.

Additive Models

- Developing methods for estimating smooth functions is only one part of the process. We must also work out how to include these in our models.

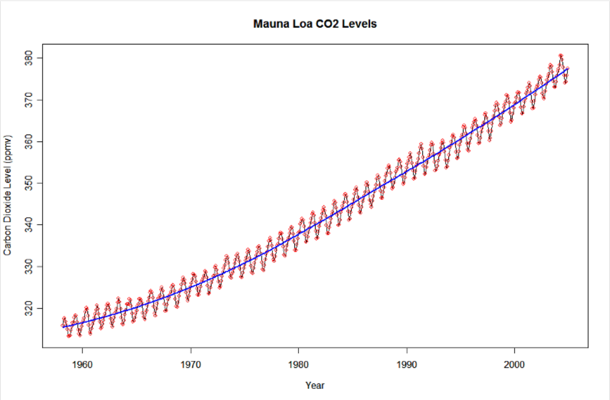
- Additive models are a general form of statistical model that allows us to incorporate smooth functions alongside linear terms.

$$y_i = \alpha + \sum_{j=1}^{k} g_j(x_{ij}) + \epsilon_{ij}$$

- Here $g_j()$ is a smooth function for the jth explanatory variable and α is the overall mean.

- Note that $g_j()$ could simply be a linear function for one or more variables.

Additive Models: Example



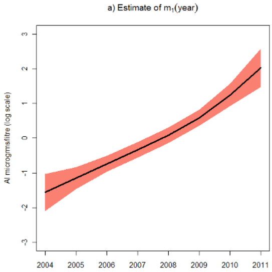
- Recall the Mauna Loa CO_2 example.
- Evidence of long-term trend plus seasonal pattern.
- We could model trend and seasonal pattern using splines.

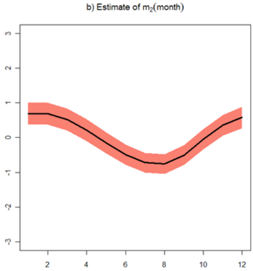
Additive Terms

- We could fit a model with smooth terms for both year (top plot) and month (bottom plot).
- We assume that the seasonal pattern does not change from year to year (i.e. no interaction).
- This can be written in the form

$$y = f_1(x_1) + f_2(x_2) + \epsilon$$

- Roughly linear increasing trend, but with a seasonal pattern featuring a peak in the winter.



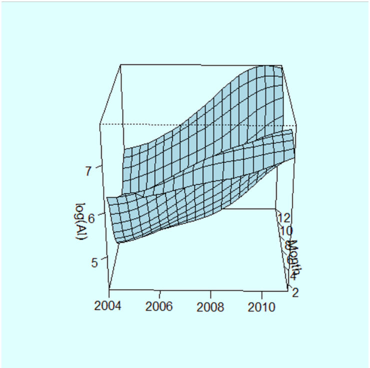


Bivariate Terms

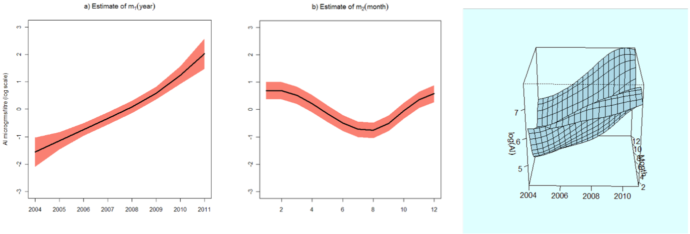
- Suppose we decided there was a month by year interaction.
- We would incorporate this via a bivariate term.
- This can be written in the form

$$y = f(x_1, x_2) + \epsilon$$

- Harder to interpret visually, but we can still see a similar pattern.



Additive versus Bivariate Smooths



- The "*Additive Terms*" model $y = f_1(x_1) + f_2(x_2) + \epsilon$ assumes that the long-term trend is identical for all months, and the seasonal pattern does not change over the years.
- The "*Bivariate Terms*" model $y = f(x_1, x_2) + \epsilon$ allows for different long-term patterns over the months, and for changing seasonality over the years.
  - More flexible but more computationally complex!

Additive Models: Advantages and Disadvantages

Advantages

- Highly flexible — allows fitting complex models.
- Nonparametric — no need to know the form of parametric relationships.
- No need to specify the number of splines correctly (just specify enough).
- Can include multiple variables.
- Commonly fitted in R.

Disadvantages

- Can be computationally complex (e.g. bivariate smooths).
- Nonparametric, so no coefficients to report — need to interpret plots of smooths.

Summary points

Summary points

- A **time series** is a sequence of measurements on the same object made over time.
- Time series data are typically **not independent**. There will often be correlation between consecutive observations.
- The purpose of time series modelling is to identify any **trends** which exist in the dataset.
- We can therefore think of our overall time series model in the following form:

$$X = \text{trend} + \text{seasonal component} + \text{error}.$$

- Our error, ϵ_t is assumed to be random, and follows the distribution $\epsilon_t \sim \text{Normal}(0, \sigma^2)$.

Summary points

- Many environmental time series have some sort of **periodicity** (e.g. a monthly pattern in temperature).
- The **period** is the time interval between consecutive peaks or troughs.
- A **seasonal component** of a dataset is a regular fluctuation with a period of one year or less.
- Harmonic regression is suitable when we have a regular trend, and can be written as

$$Y_t = \beta_0 + \gamma_1 \sin\left(\frac{2\pi x_t}{p}\right) + \gamma_2 \cos\left(\frac{2\pi x_t}{p}\right) + \epsilon_t$$

University of Glasgow

Summary points

- The existence of a trend can be assessed using the **Mann-Kendall test**.
- The **Kendall rank correlation coefficient** can tell us about the strength of the trend.
- The **Seasonal Kendall test** accounts for seasonality by computing the test statistic for each season separately and combining the results.

- **Smoothing methods** can be used to express relationships in terms of smooth functions:
  - **LOWESS** provides a graphical illustration of the data.
  - **Penalised splines** can be used to model nonlinear relationships, often as part of **additive models**.