STAT 443: Time Series and Forecasting

Chapter 1

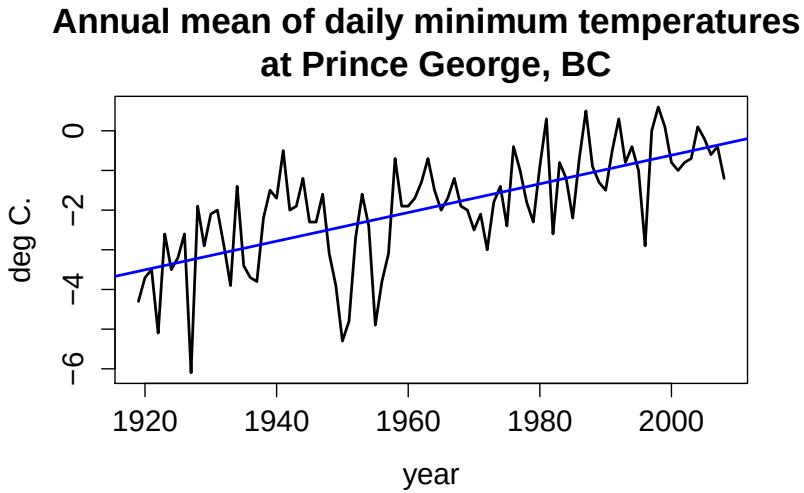
Exploratory techniques in time series analysis

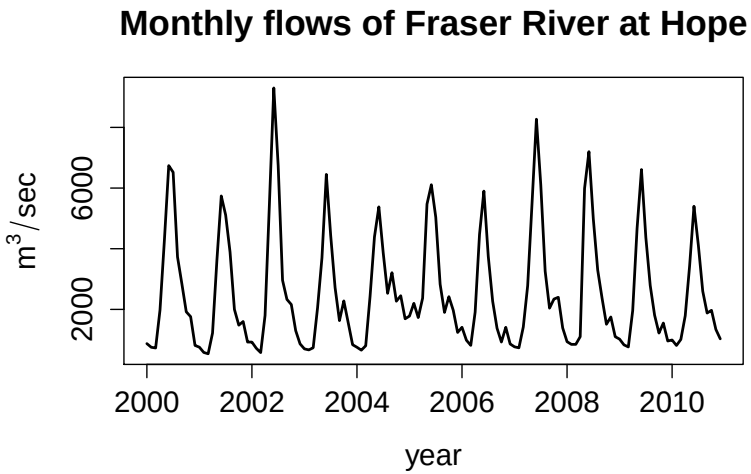
What is a time series?

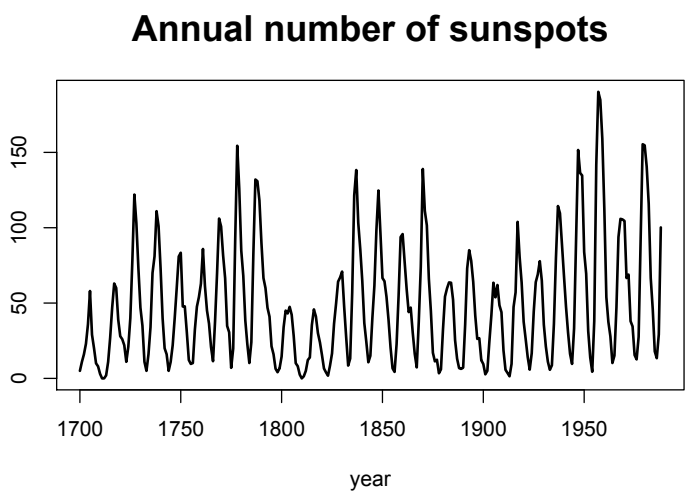
What is a time series?

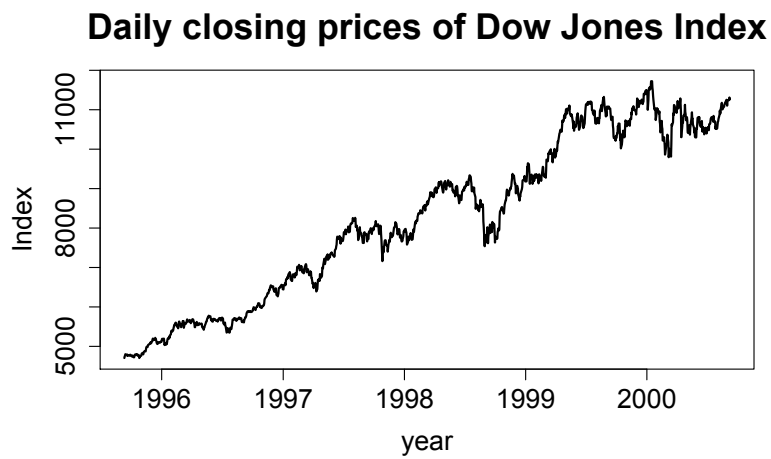
a collection of observations recorded sequentially in time

Examples of time series data









What are some applications where 'standard' time series data occur?

- Physical sciences (meteorology, hydrology, geophysics, etc.)
- Economics, finance and commerce (marketing)
- Engineering (e.g., electricity or water consumption)

- Time plot: it is customary to visualize time series by plotting observations against time and joining the dots

- Time series are inherently discrete, giving an approximation (via interpolation) to a continuous time process due to

  - data aggregation over a time interval (e.g., precipitation)
    or
  - data sampling/recording at a particular time (e.g., temperature)

- The choice of sampling frequency or time scale is important in view of the purpose of the analysis

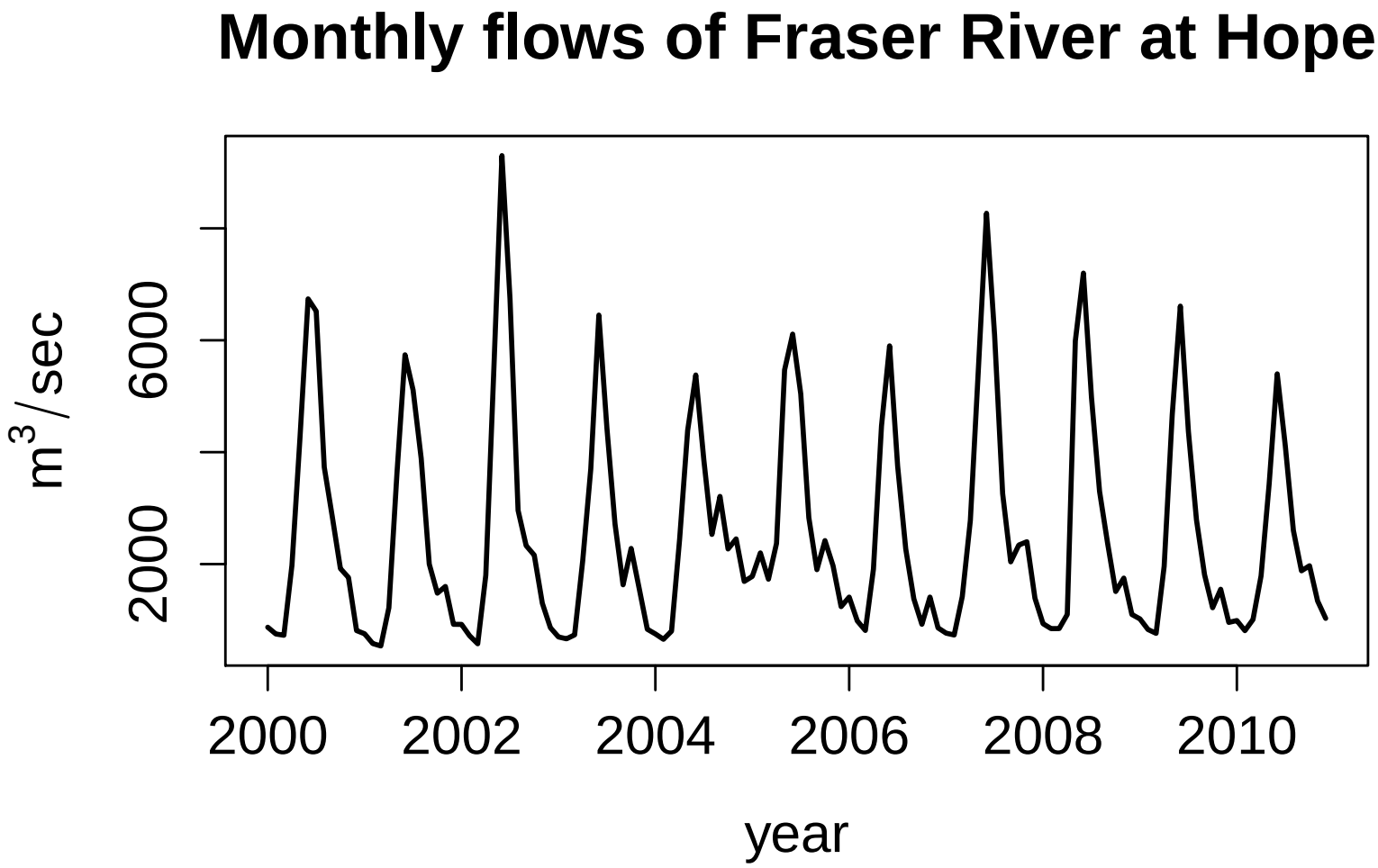
- In this course: observations are assumed to correspond to specific equally-spaced time intervals (e.g., daily, monthly, etc.)

Interlude on plotting time series in R

```
library(tseries)
library(zoo)

# EXAMPLE: Flows of Fraser River at Hope (station 08MF005)
# Data source: https://wateroffice.ec.gc.ca/index_e.html
dat <- read.csv("dataFraserRiver.csv")
```

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1912			485	1150	4990	6130	4780	3960	2160	1530	1060	761
1913	516	710	570	984	2690	8280	6480	5990	3930	1760	1010	736
1914	1050	787	906	2100	5730	7140	6530	3720	2140	1970	1760	1080
1915	886	749	806	2510	4140	4610	5020	3700	1690	1270	1190	585
1916	543	945	780	1280	3880	6580	6460	3930	2350	1620	1040	643
1917	589	543	482	767	4170	7880	6510	3720	2130	2100	1940	1590
1918	1660	1140	964	2020	5290	7970	6960	4320	2570	1650	1180	837
1919	668	584	497	1360	4050	6470	6930	4640	2420	1430	1320	1160
1920	1010	834	648	857	3980	7240	9600	5460	3130	3310	1860	1200
1921	935	899	824	1580	5330	9320	6610	4230	2710	2700	1940	1360
1922	908	677	590	848	4030	8130	5430	3450	2510	2300	1310	1170
1923	906	674	612	1410	4830	7830	5670	3490	2500	1290	939	841

```
flow <- dat[,-1]
flow <- c(t(flow))
flow <- na.omit(flow)
# removes first 2 missing values in 1912

flow.ts <- ts(flow, start=c(1912,3), frequency=12)
plot(flow.ts, xlab="year", ylab=expression(m^3/sec))
```


Monthly flows of Fraser River at Hope
m3/sec
2000
6000
2000
2002
2004
2006
2008
2010
year

Interlude on plotting time series in R (cont'd)

```
flow20xx <- window(flow.ts, start=c(2000,1), end=c(2010,12))
plot(flow20xx, xlab="year", ylab=expression(m^3/sec),
main="Monthly flows of Fraser River at Hope", lwd=2)

## with ggplot
autoplot(flow20xx, xlab="year", ylab="Flow (m^3/sec)")
```

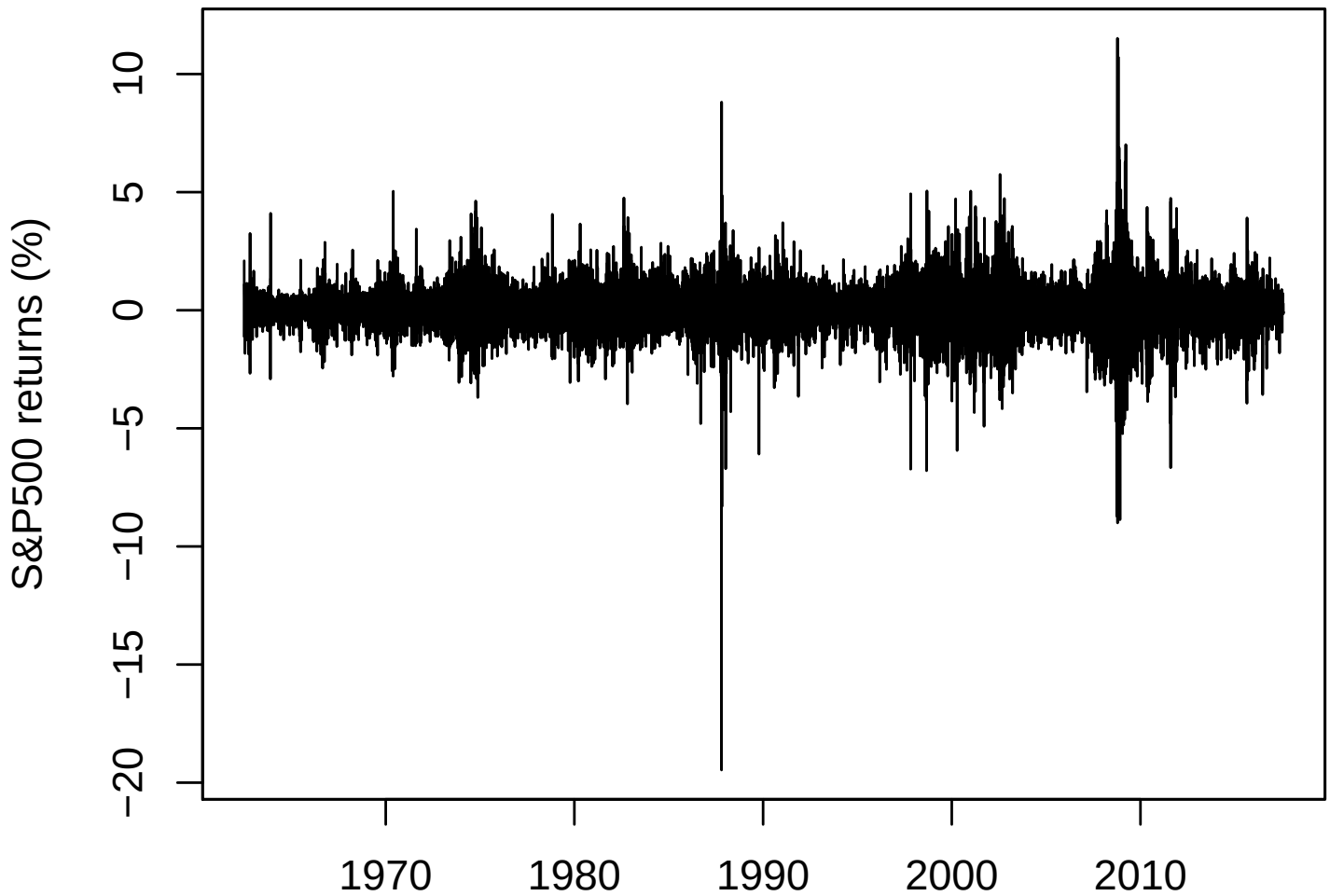
Other useful functions:

```
flow.annual <- aggregate(flow.ts, FUN = mean)
flow.max <- aggregate(flow.ts, FUN = max)

time(flow.ts)
```

Interlude on plotting time series in R (cont'd)

```
## Example: S&P500 returns time series

dat<-read.csv("dataFinancialReturns.csv")
# raw data file (1962/07/02 - 2017/07/31)

dat.indices<-unique(data.frame(dat$date,dat$Ret_M,dat$SP500))
colnames(dat.indices)<-c("date","retM","sp500")

rSP<- zoo(x=dat.indices$sp500,
order.by = strptime(dat.indices$date,"%Y%m%d"))

plot(100*rSP, xlab="", ylab="S&P500 returns (%)")
```

Interlude on plotting time series in R (cont'd)



A few useful R commands for time series plotting

- `plot.ts()`
- `ts()`
- `zoo()`
- `window()`
- `start()`
- `end()`
- `time()`
- `diff()`
- `lag()`
- `lag.plot()`
- `autoplot()`

Purposes of time series analysis

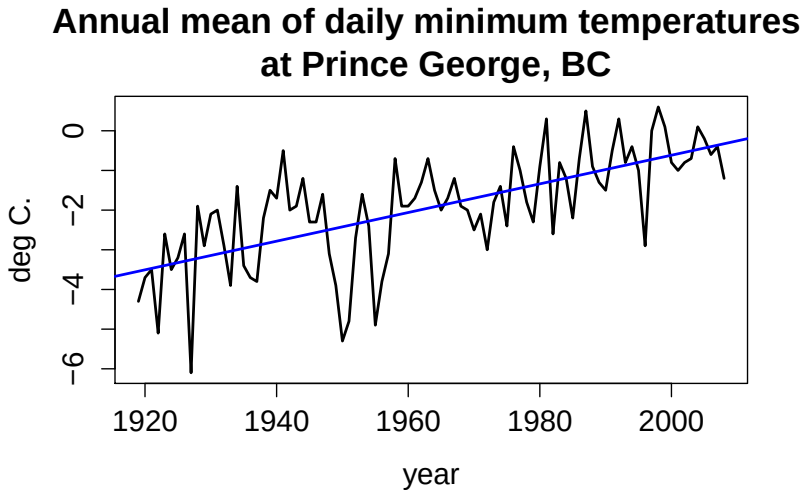
- Explanatory analysis and modelling: understand the past
- Forecasting: predict the future
- Basis for computer simulations

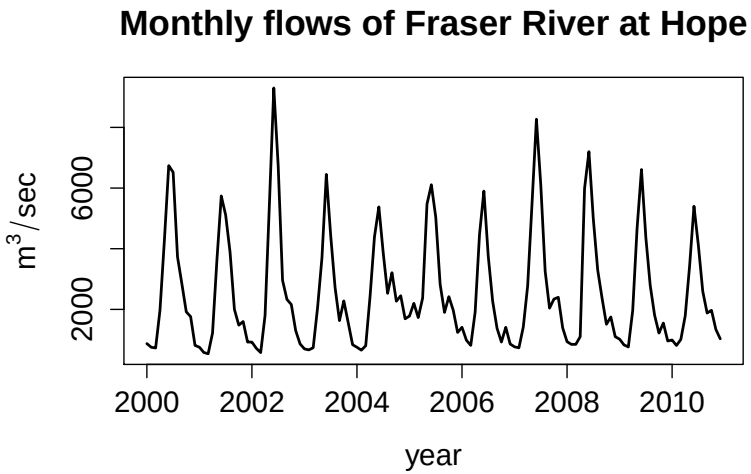
  e.g., to assess impact of policies or regulations

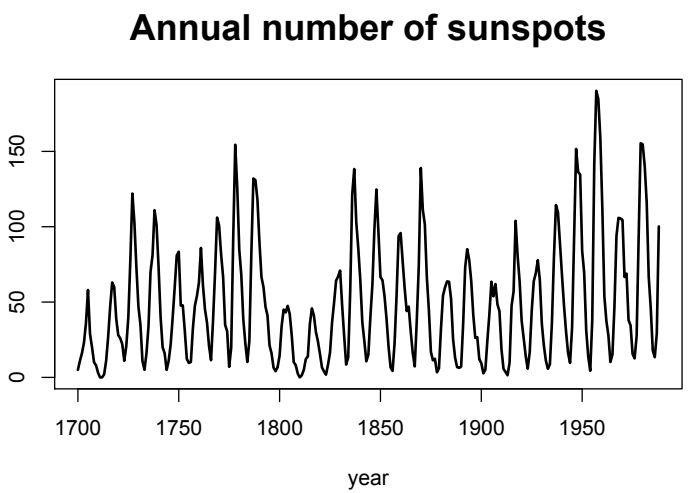
- Time series regression: assumptions of ordinary least squares regression framework are violated when the response and exploratory variables are time series
- Quality control and optimal management:

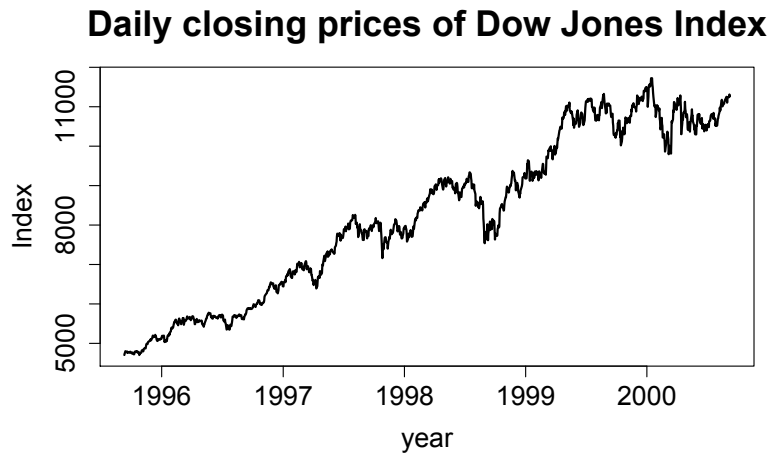
  e.g., quantitative risk management of investment portfolios by a bank; anomaly detection in datasets

Main features of time series









Main features of time series

- Serial or temporal dependence
  - tendency of observations close in time to be correlated
- Trend
  - a long term change in the mean of the series
- Seasonal effects
  - a regular variation with "season", where "season" might be month, day, year, etc.
  - such effects are presumed to be periodic, and in some sense predictable
- Cyclical variation

  oscillations of possibly unknown cause and variable period

Aim:

develop statistical models to explain these time series characteristics

Basic notation

- Unless otherwise stated, our time variable t is discrete

$$t = 0,\ \pm 1,\ \pm 2,\ \ldots$$

- Model: $\{X_t\}$
  - random, function of time t; $t = 0,\ \pm 1,\ \pm 2,\ \ldots$
  - discrete-time stochastic process or a time series model
  - can be thought of as the data generating process
- Data: $\{x_1, \ldots, x_n\}$ or $\{x_t\}_{t=1,\ldots,n}$ or $\{x_t\}$
  - historical time series
  - realization of a sequence of random variables

A little bit of terminology...

A purely random or white noise process is a sequence of independent and identically distributed (i.i.d. for short) random variables, which we denote $\{Z_t\}$.

- It is often further assumed that $Z_t \sim \mathcal{N}\left(0, \sigma^2\right)$
  - "$\sim$" reads "is distributed as"
  - $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2
- Many time series models involve a white noise process as a building block

Outline

A more detailed look at the main features of time series data

- Time series with a trend
- Time series with a trend and seasonal variation
- Assessing temporal dependence

Series with a trend

Confronted with a series with an obvious trend, we might want to

(i) measure the trend,

(ii) remove the trend, or

(iii) both (i) and (ii)

Series with a trend (cont'd)

One approach to this (for non-seasonal series) is curve fitting - i.e., fit a line or a curve through the points

In doing this we would usually estimate the unknown coefficients in the model we fit using the least squares estimation

Some examples of families of curves are:

- Linear: $m(t) = a + bt$
- Quadratic: $m(t) = a + bt + ct^2$
- Gompertz: $m(t) = a - br^t$, for $|r| < 1$
- Logistic:

$$m(t) = \frac{a}{(1 + be^{-ct})}$$

Series with a trend (cont'd)

Question: What are disadvantages of curve fitting?

- not dynamic: assumes that the same curve fits the data at all time points
- in cases where it works well, however, there is often little need for further analysis

Models for time series with a trend and seasonal variation

- Additive decomposition model

$$X_t = m_t + s_t + Z_t,$$

where, at time t, m_t: trend, s_t: seasonal effect, Z_t: white noise

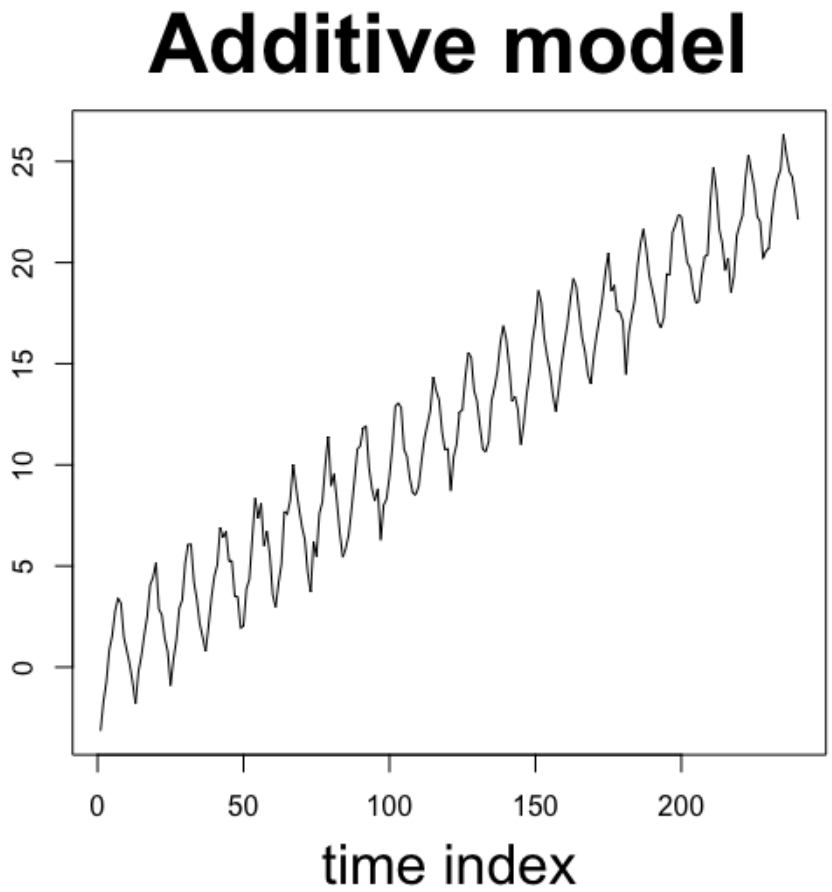
- Multiplicative seasonal effect model

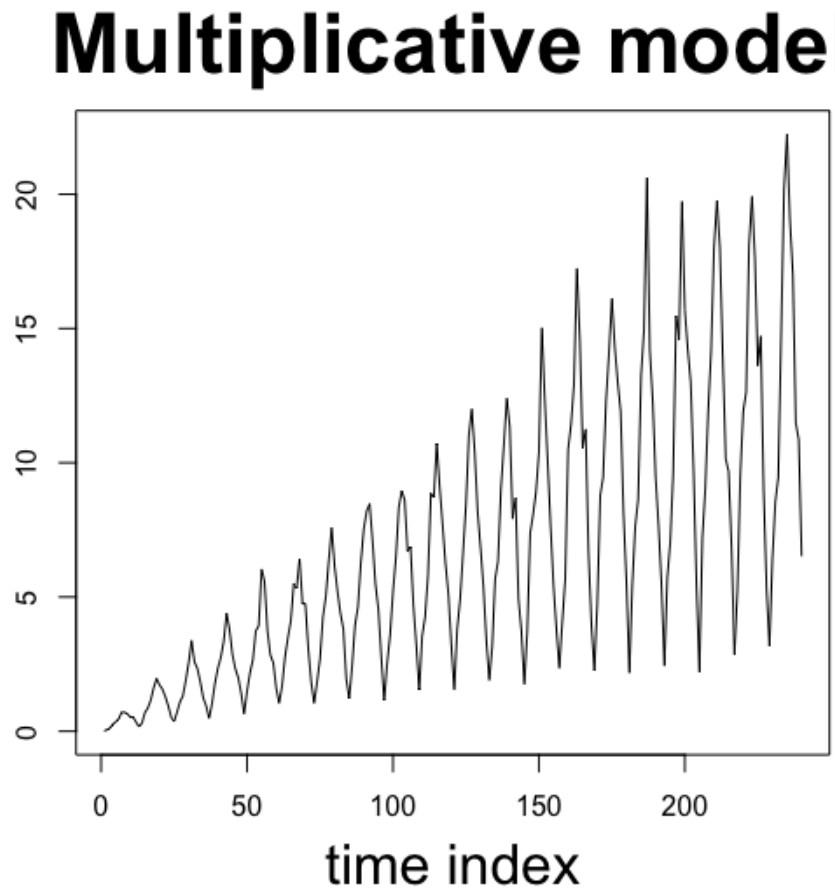
$$X_t = m_t \; s_t \; Z_t$$

(i.e., seasonal variation increases with the mean)

  - if X_t is positive, this can be re-formulated as an additive model for $\log X_t$

Models for time series with a trend and/or seasonal variation - Illustration





Common assumptions about the additive seasonal effect s_t:

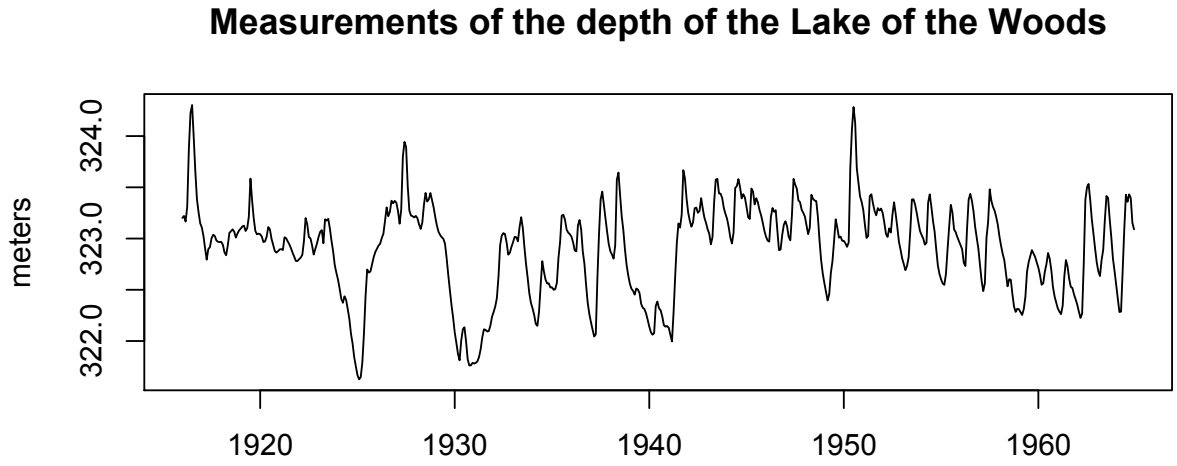
1. The function s_t has period p:

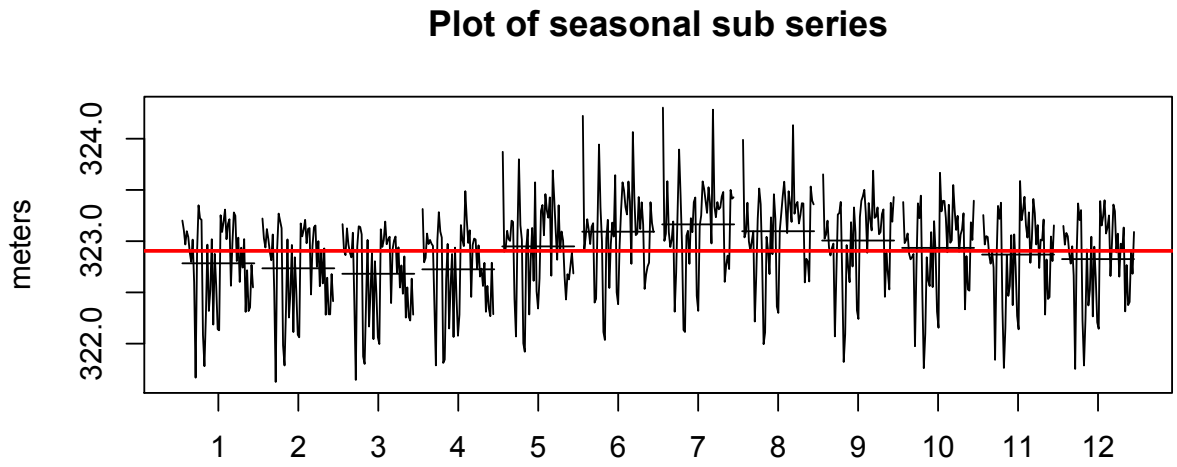
$$s_{t+p} = s_t \quad \text{for all t}$$

2. The sum of the seasonal effects over a complete cycle/period is zero:

$$\sum_{i=1}^{p} s_{t+i} = 0$$

A crude way to estimate a seasonal effect is to take the difference between the average in the period of interest (e.g., a given month or a quarter) and the overall average (in the additive case).





Lake of the Woods Example

- Data: Lake of the Woods monthly mean water levels (in meters) for the period Jan. 1916 to Dec. 1964
- The means across months are summarized below:

1	2	3	4	5	6	7	$\cdots$
322.78	322.74	322.68	322.73	322.95	323.09	323.16	$\cdots$

- The overall mean is 322.91
- Exercise: estimate seasonal effects s_t for $t = 1$ (Jan) and $t = 7$ (Jul) assuming the additive seasonal model

Questions:

Is the above method for estimating seasonal effects adequate in the presence of a trend? Why?

Estimation using smoothing techniques

- A smoothed series is obtained from the original series $\{x_t\}$ via

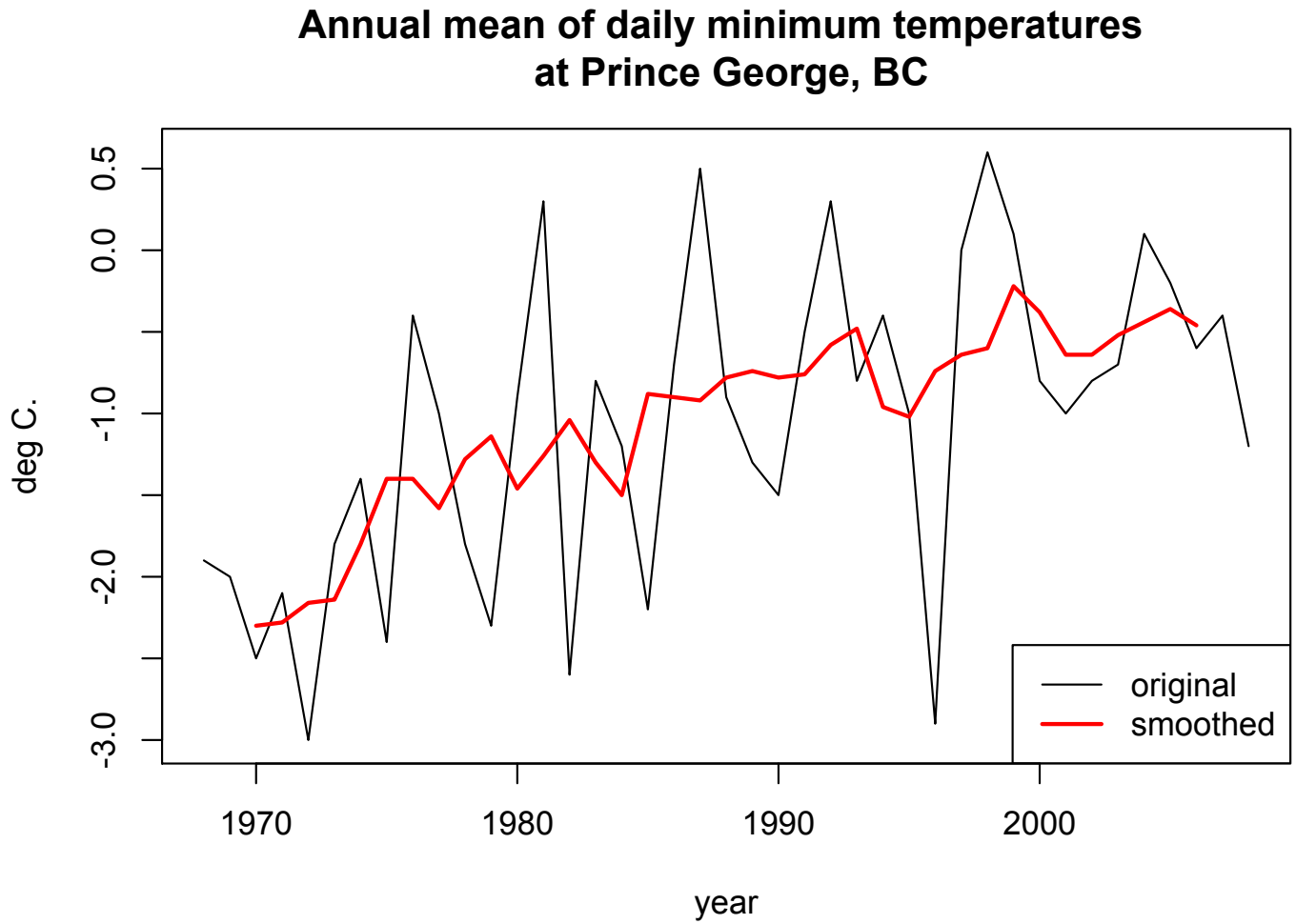
$$Sm_t := \sum_{r=-q}^{s} \omega_r \; x_{t+r},$$

where $\{\omega_r\}$ are weights (usually) satisfying $\sum_{r=-q}^{s} \omega_r = 1$

  - Symmetric weights: $s = q$, $\omega_r = \omega_{-r}$ for all r

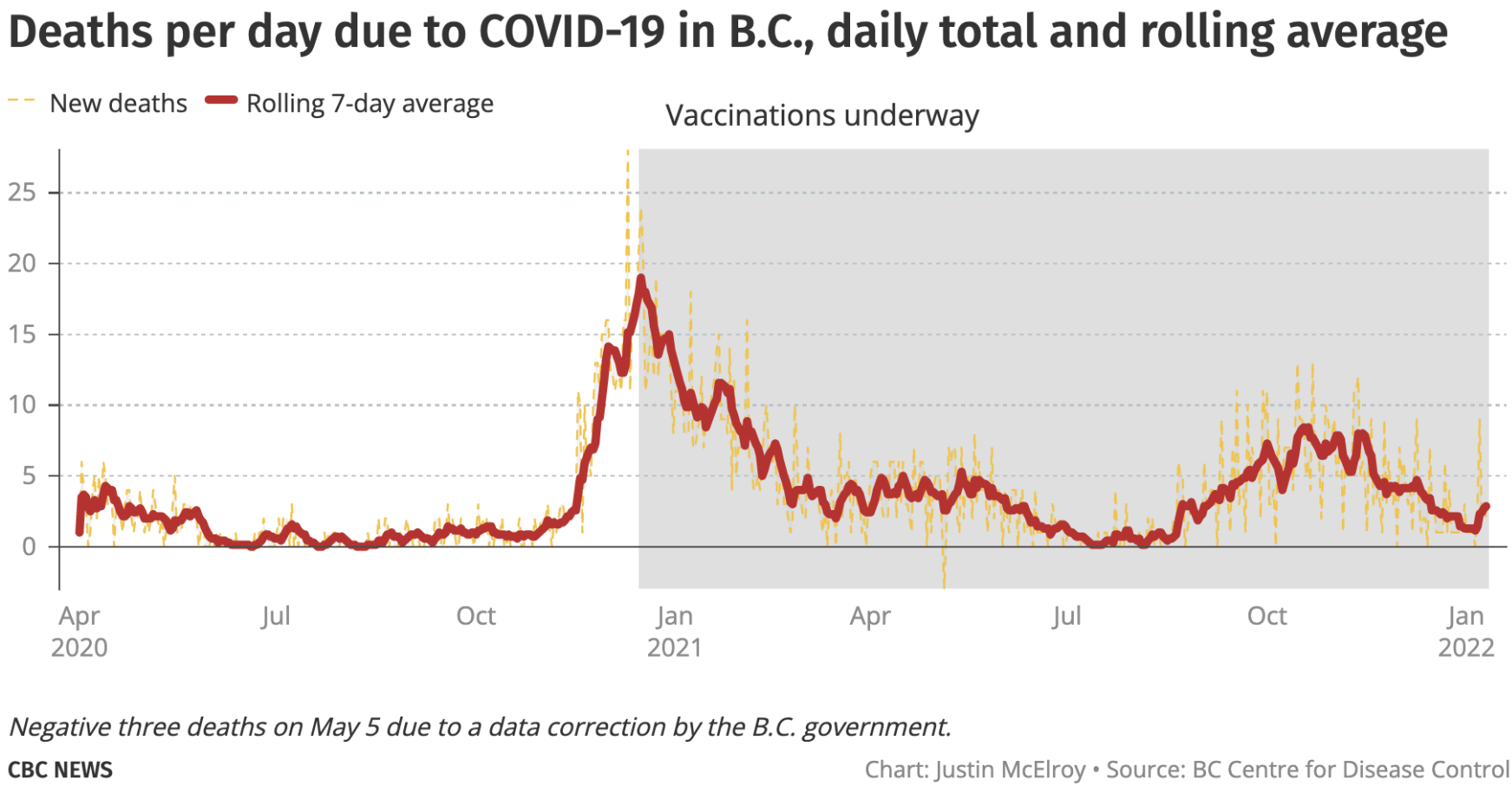
Smoothing can be used to

- identify and estimate the trend by reducing the local variation



Smoothing can be used to

- identify and estimate the trend by reducing the local variation



Smoothing can be used to

- identify and estimate the trend by reducing the local variation
- remove the seasonal effect, say of period p, by forming the moving average series

$$\frac{1}{p}\sum_{j=0}^{p-1} x_{t+j},\ \frac{1}{p}\sum_{j=1}^{p} x_{t+j},\ \frac{1}{p}\sum_{j=2}^{p+1} x_{t+j}, \ldots$$

Question: Why does this moving average series have no seasonal effect assuming the additive model?

Estimation using smoothing techniques (cont'd)

Note: when the period of the seasonal effect is even, the moving average series will not correspond to the time points of the original series, instead being positioned between two points of the original series

Possible solutions:

- Include an extra term and re-weight the end-points of the moving average window
  - e.g., quarterly data ($p = 4$)

$$Sm_t = \frac{0.5x_{t-2} + x_{t-1} + +x_t + x_{t+1} + 0.5x_{t+2}}{4}, \qquad t = 3, \ldots, n-2$$

- Centre the moving average series by forming a new series which averages consecutive values of the smoothed one
  - i.e., take a moving average of order 2 on the first moving average series

Estimation using smoothing techniques (cont'd)

- The (centred) moving average series can be taken as the trend estimate, denoted $\{\hat{m}_t\}$
- The additive seasonal effect at time t is then $\hat{s}_t = x_t - \hat{m}_t$
- Averaging these estimates for each period (e.g., month, quarter) gives (initial) estimates of the seasonal effects, denoted $\widehat{S}_j^{initial}$ $(j = 1, \ldots, p)$
- As $\widehat{S}_j^{initial}$ may not add up to exactly zero, adjust their values by subtracting $\frac{1}{p}\sum_{j=1}^{p} \widehat{S}_j^{initial}$.

  Let $\widehat{S}_j$ $(j = 1, \ldots, p)$ denoted the resulting seasonal index estimates.
- $\{x_t - \widehat{S}_j\}$ for index t corresponding to seasonal period j is then a seasonally adjusted series (under the additive model)

Remark: Under the multiplicative model, subtraction is simply replaced by division

Activity: Estimation of seasonal effects using smoothing

Data: quarterly energy consumption figures (in MWe) in the UK for the years 1975–1979

1. Comment on what you observe from the time plot.

2. Find the numbers indicated by "$\star$" in the above table.

3. Assuming an additive seasonal effect and making use of the filtered series, estimate the adjusted seasonal indices $\hat{S}_1$, $\hat{S}_2$, $\hat{S}_3$, $\hat{S}_4$.

4. Why would the method you applied in 3. be preferable here to the method first applied to the Lake of the Woods data that does not use smoothing?

5. When the filtered data are regressed against t, the fitted line is

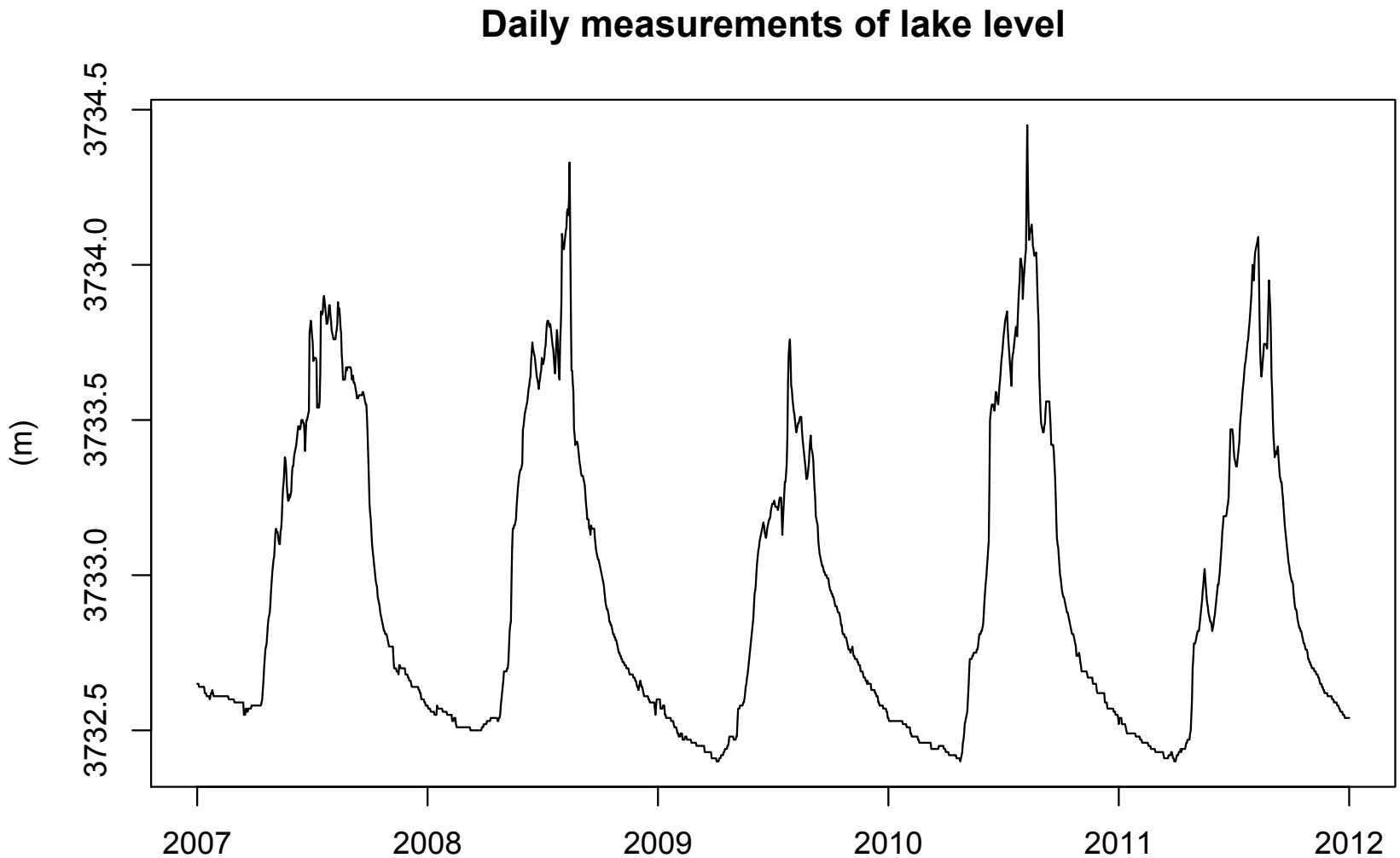
$$T_t = 776.18 + 6.98t, \qquad t = 1, 2, \ldots.$$

   Using this, forecast the energy consumption for the first two quarters of year 1980.

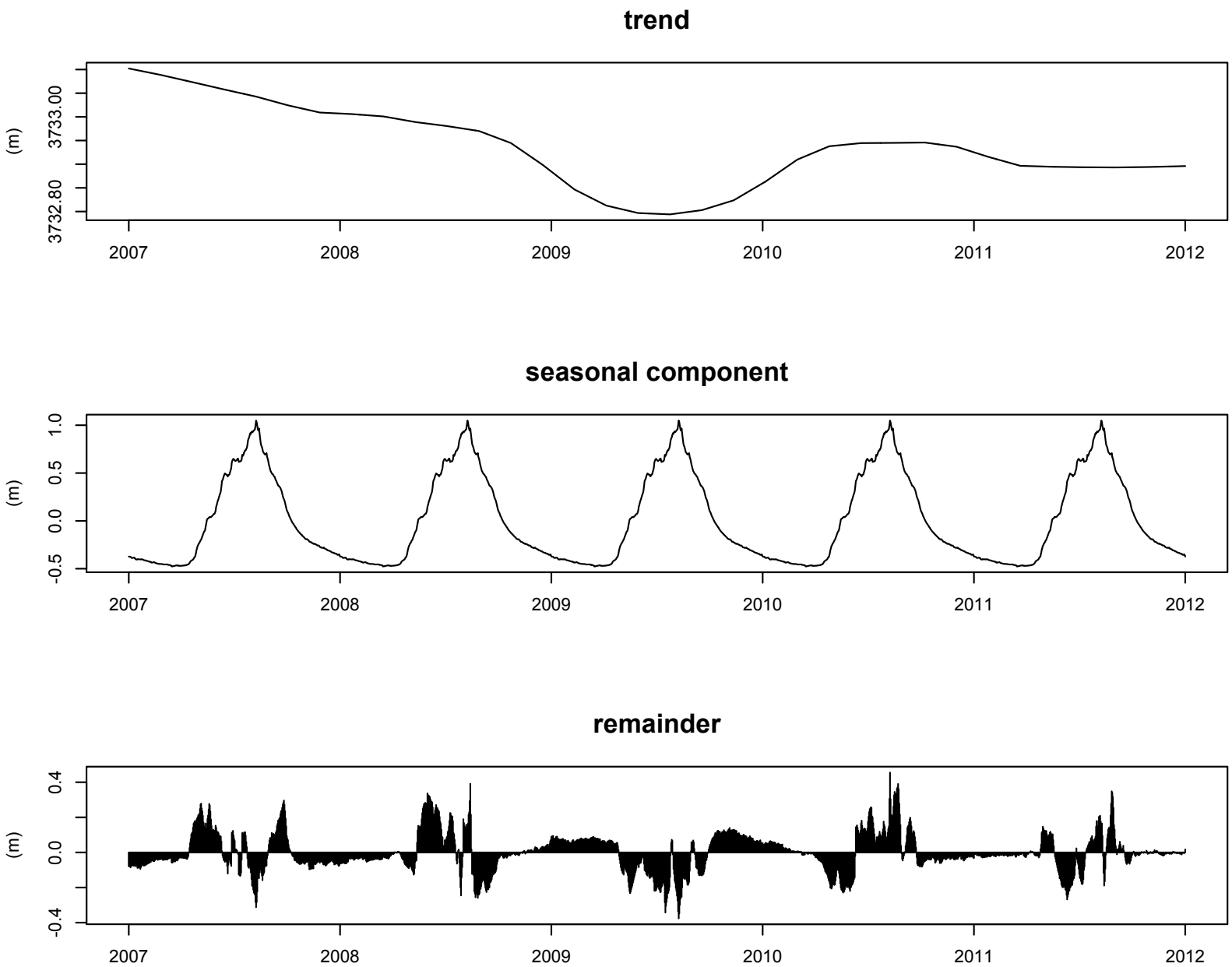
Estimation using smoothing techniques (cont'd)

- A second and a more sophisticated technique of decomposing a time series into the trend and seasonal component is via the locally weighted regression (known as the loess smoothing)
- Details are complicated here
- Reference: R. B. Cleveland, W. S. Cleveland, J.E. McRae, and I. Terpenning (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *J. Official Statistics* 6, pp. 3-73

Example


Daily measurements of lake level
(m)
3732.5
3733.0
3733.5
3734.0
3734.5
2007
2008
2009
2010
2011
2012

Example: Decomposition using loess smoothing



Assessing temporal dependence

- Temporal dependence is a key feature in time series data
- Our ultimate goal is to derive models with similar serial dependence properties as the ones observed in the data

What is a common measure of dependence between any two variables?

Interlude on the sample correlation

- Consider paired observations $(x_1, y_1), \ldots, (x_n, y_n)$
- The sample correlation coefficient is defined as

$$r := \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\,\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}},$$

  where $\overline{x}$ and $\overline{y}$ denote sample means of x's and y's, respectively

  - $-1 \leq r \leq 1$
  - r is a measure of degree of linear association between the two variables

Assessing temporal dependence (cont'd)

- For a time series $\{x_1, \ldots, x_n\}$, it is natural to look at the correlation between consecutive values:

$$(x_1, x_2),\ (x_2, x_3),\ \ldots, (x_{n-1}, x_n)$$

- The correlation coefficient between values of the series one lag apart, known as the autocorrelation coefficient at lag 1, is given by

$$\tilde{r}_1 := \frac{\sum_{t=1}^{n-1}(x_t - \overline{x}_{(-n)})(x_{t+1} - \overline{x}_{(-1)})}{\sqrt{\sum_{t=1}^{n-1}(x_t - \overline{x}_{(-n)})^2}\ \sqrt{\sum_{t=2}^{n}(x_t - \overline{x}_{(-1)})^2}},$$

where

$$\overline{x}_{(-n)} = \frac{1}{n-1}\sum_{t=1}^{n-1} x_t \quad \text{and} \quad \overline{x}_{(-1)} = \frac{1}{n-1}\sum_{t=2}^{n} x_t$$

Sample autocorrelation function

- For n large, $\overline{x}_{(-n)} \approx \overline{x}_{(-1)} \approx \overline{x} := \sum_{t=1}^{n} x_t/n$ and $n-1 \approx n$, which can be used to simplify the above formula for $\tilde{r}_1$:

$$r_1 := \frac{\sum_{t=1}^{n-1}(x_t - \overline{x})(x_{t+1} - \overline{x})}{\sum_{t=1}^{n}(x_t - \overline{x})^2} \quad (\approx \tilde{r}_1)$$

- r_1 will be used as the sample autocorrelation coefficient at lag 1

Sample autocorrelation function

- For n large, $\overline{x}_{(-n)} \approx \overline{x}_{(-1)} \approx \overline{x} := \sum_{t=1}^{n} x_t/n$ and $n-1 \approx n$, which can be used to simplify the above formula for $\tilde{r}_1$:

$$r_1 := \frac{\sum_{t=1}^{n-1}(x_t - \overline{x})(x_{t+1} - \overline{x})}{\sum_{t=1}^{n}(x_t - \overline{x})^2} \quad (\approx \tilde{r}_1)$$

- r_1 will be used as the sample autocorrelation coefficient at lag 1
- Similarly, the sample autocorrelation coefficient at lag h is defined as

$$r_h := \frac{\sum_{t=1}^{n-h}(x_t - \overline{x})(x_{t+h} - \overline{x})}{\sum_{t=1}^{n}(x_t - \overline{x})^2} \qquad (h = 1, 2, \ldots, n-1)$$

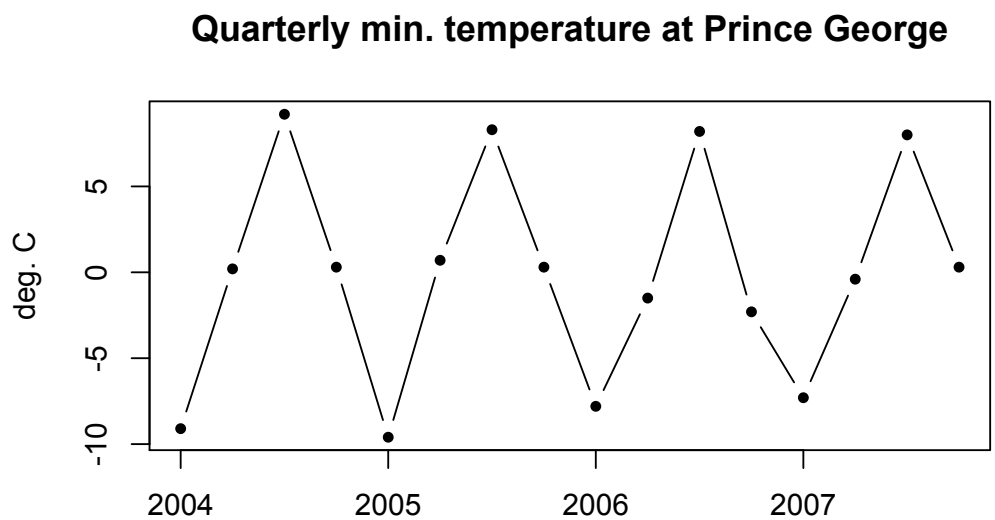
- r_h's viewed as a function of the lag h constitute the autocorrelation function (**acf**)

Activity: The Sample Autocorrelation

- Data: quarterly means of the daily minimum temperature at Prince George, BC during 2004–2007

Year	Quarter 1	2	3	4
2004	-9.1	0.2	9.2	0.3
2005	-9.6	0.7	8.3	0.3
2006	-7.8	-1.5	8.2	-2.3
2007	-7.3	-0.4	8.0	0.3
	-33.8	-1	33.7	-1.4

- Sample mean $\overline{x} = -0.2$
- Standard deviation $s = 6.2$



Activity: The Sample Autocorrelation (cont'd)

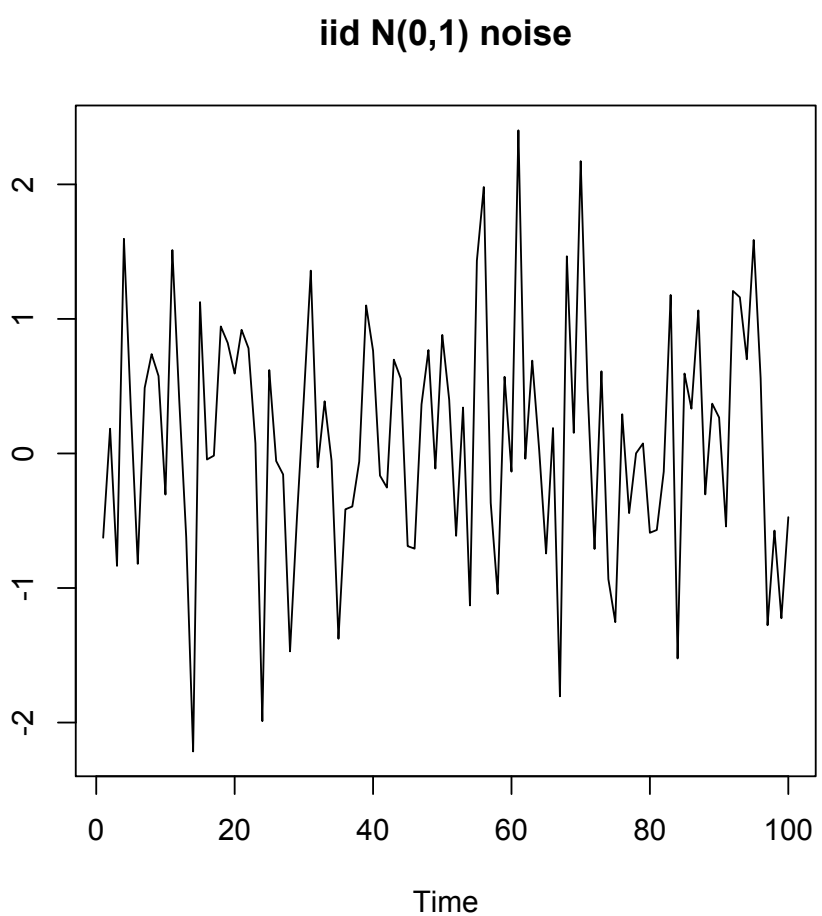
The correlogram

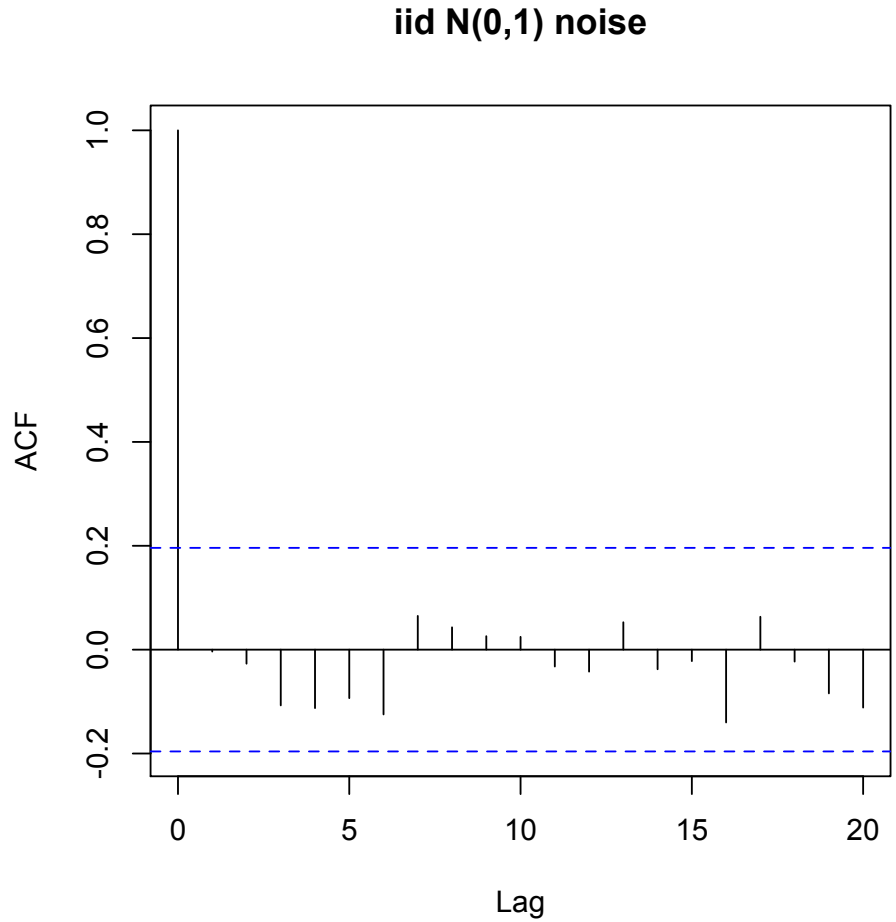
- The correlogram is the plot of the acf r_h against the lag h
- It is often a useful graphical tool to examine serial dependence in time series
- For a sequence of n independent observations with the same distribution, i.e. a completely random series,

$$r_h \sim \mathcal{N}(-1/n, 1/n), \qquad h \neq 0$$

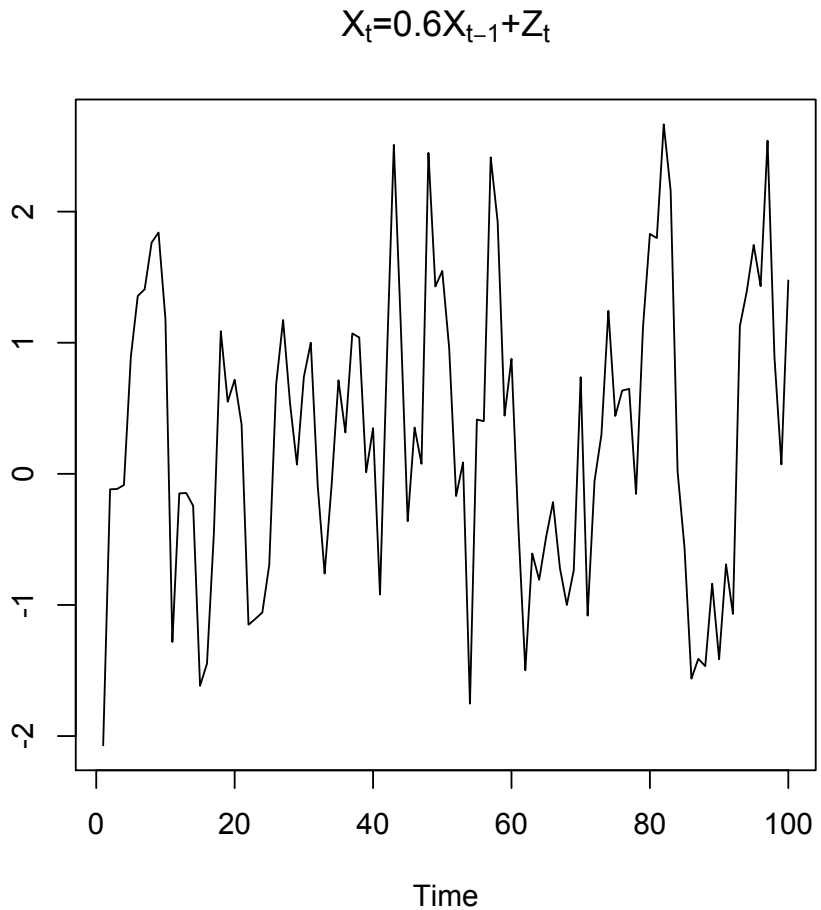
  approximately for n sufficiently large
- This can be used to construct confidence bounds to detect significant autocorrelations
- Note: for a completely random series, roughly one in twenty values of r_h are expected to lie outside the interval $(-1/n - 2/\sqrt{n}, -1/n + 2/\sqrt{n})$

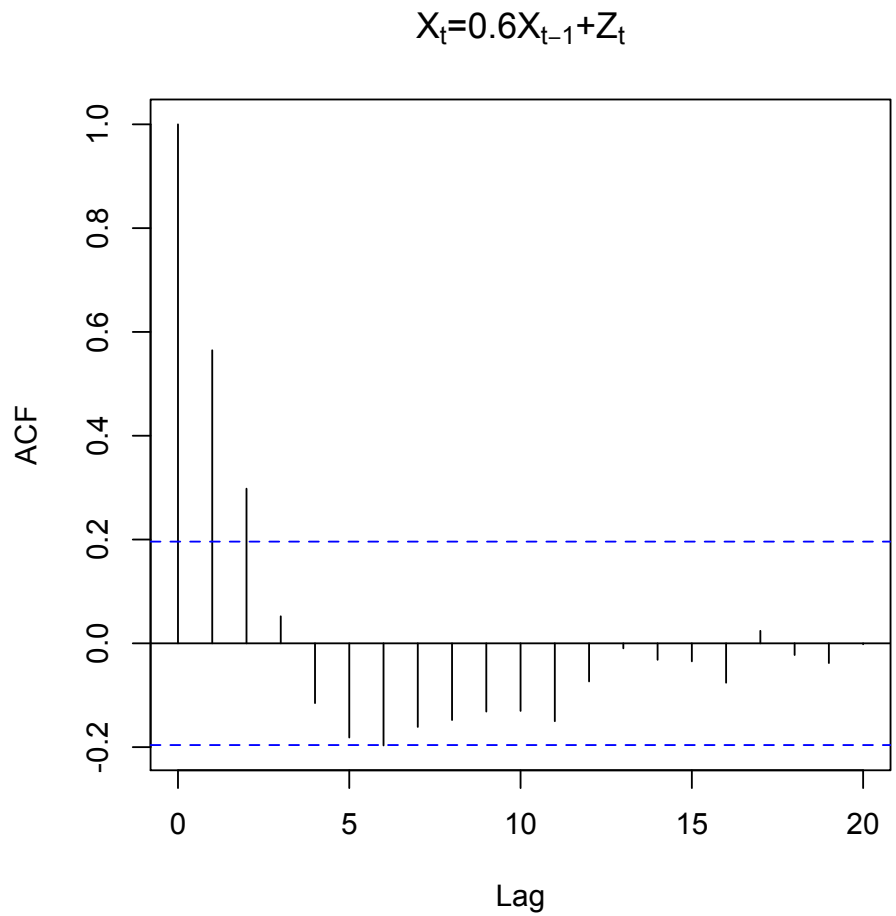
The correlogram: Examples



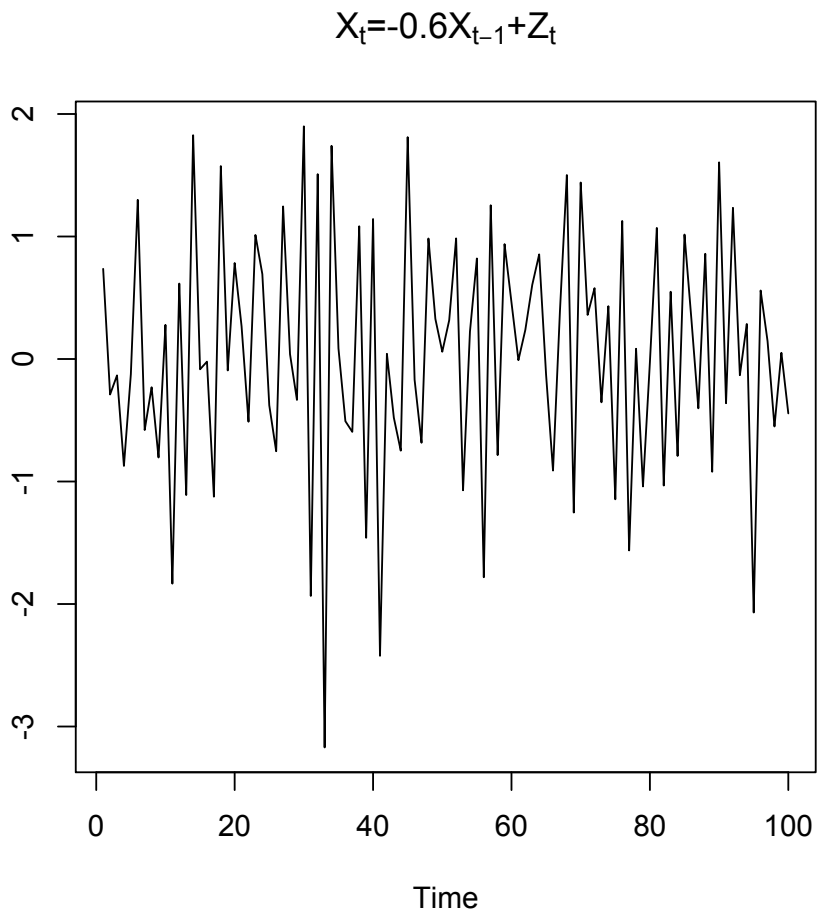


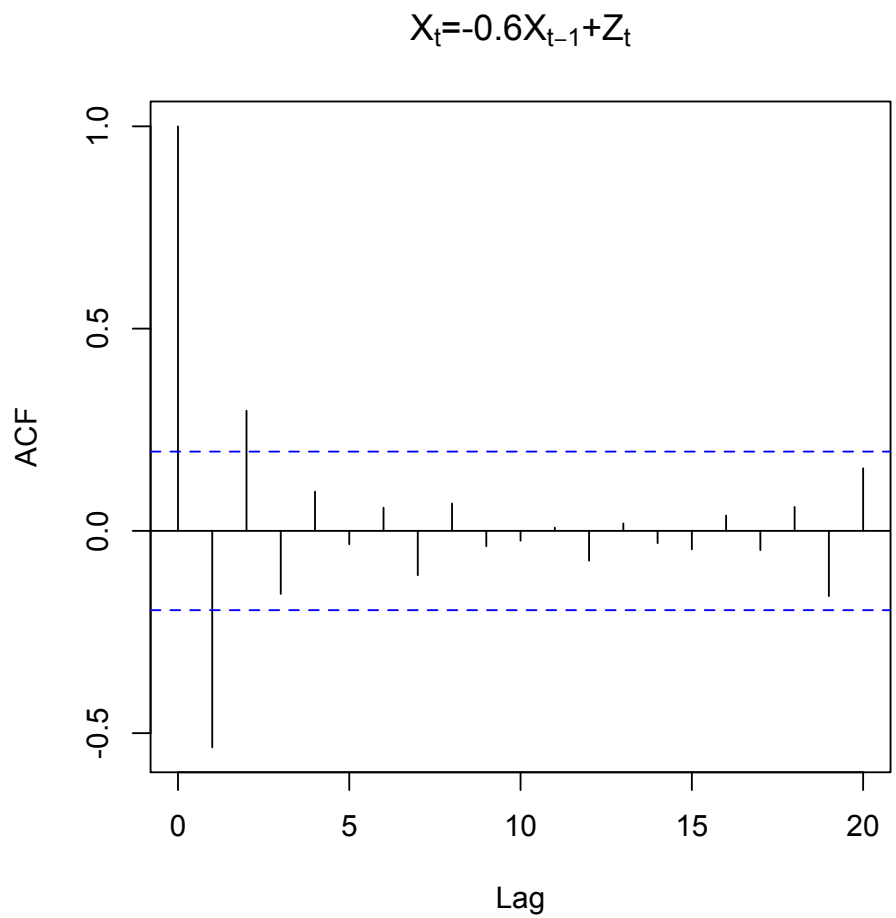
The correlogram: Examples



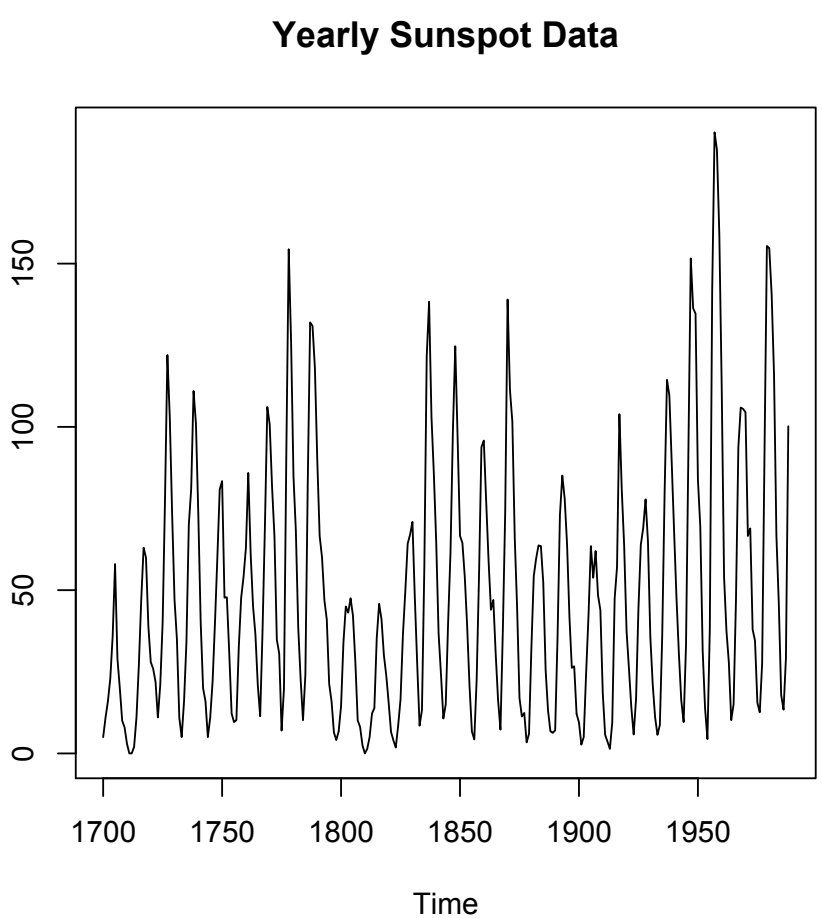


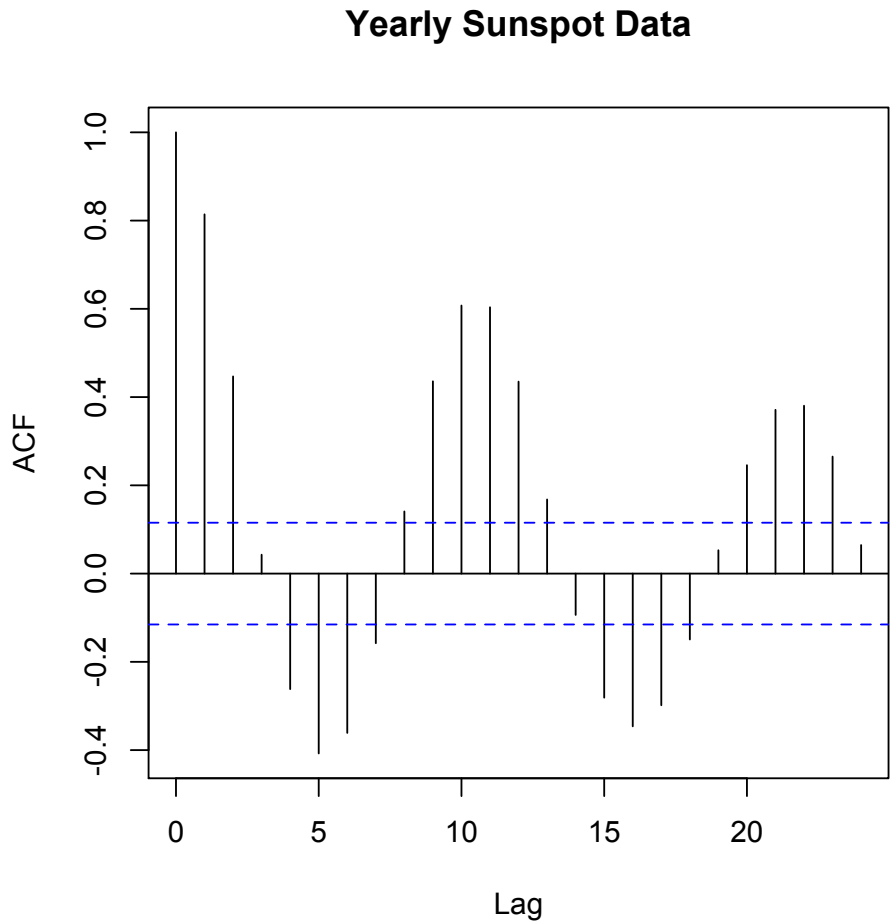
The correlogram: Examples



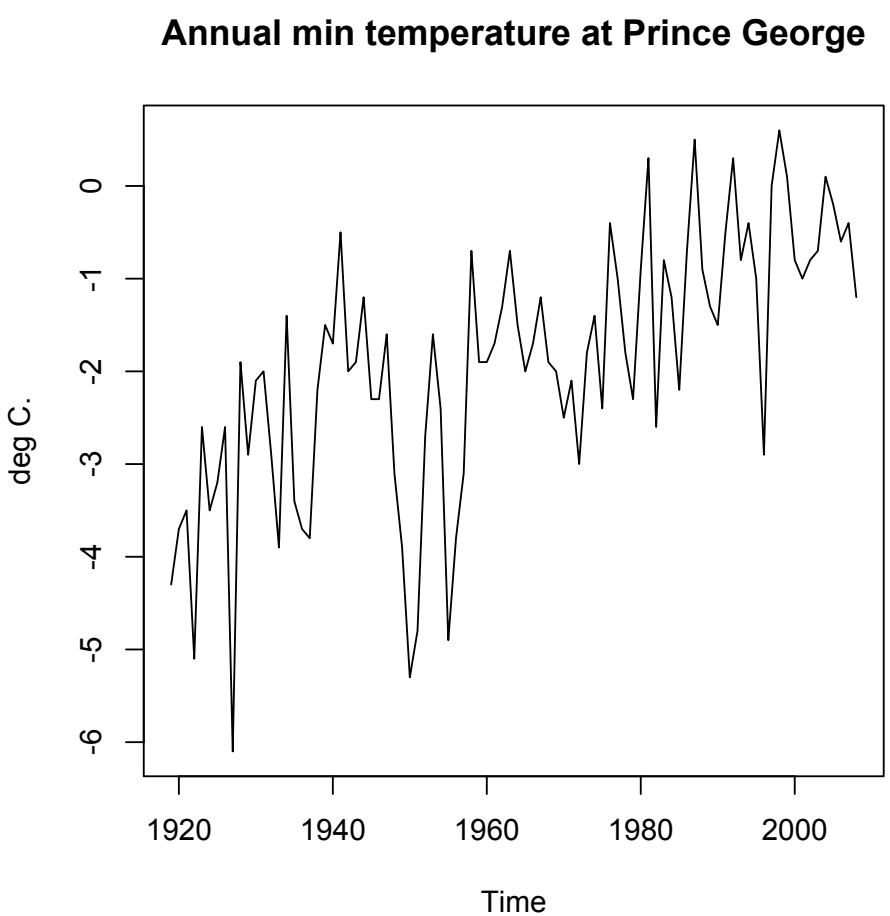


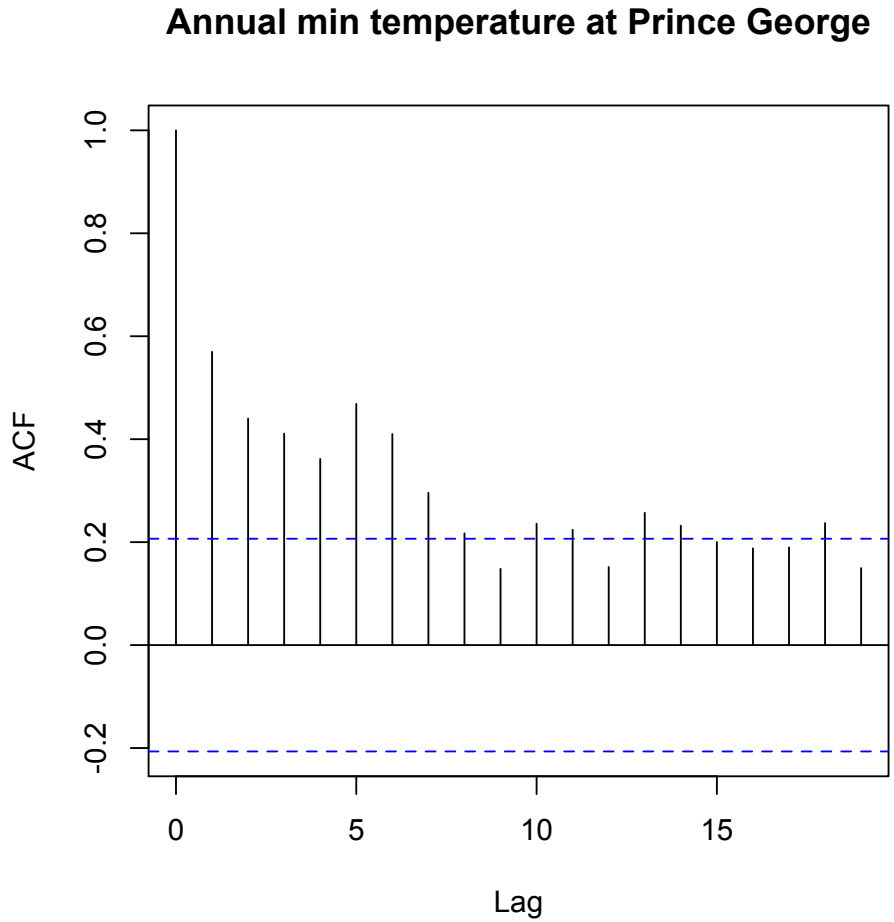
The correlogram: Examples





The correlogram: Examples





The correlogram - special cases in summary

1. IID noise: the sample acf tends to be within $\pm 2/\sqrt{n}$ bounds

2. For many stationary time series, the sample acf will decay with the lag (observations farther apart in time are less correlated)

3. If the series alternates (i.e. consecutive values tend to be on the opposite sides of the mean), so does its acf. The value of r_1 would be negative, while r_2 is likely to be positive

4. If the series has seasonal or cyclical fluctuations, its acf will oscillate at the same frequency. E.g., for quarterly data, r_4 could be expected to be positive and relatively large (cf. Activity)

5. If the series has a trend, the acf will have a very slow decay due to high correlation of the consecutive values (which tend to lie on the same side of the mean)

The correlogram - final remarks

- The correlogram is only informative for stationary time series
- Beware of the effect of outliers on the acf which can significantly affect its values
- In general, experience is required to glean much from an acf plot
- We will return to the discussion of the correlogram as a model selection tool

Summary

- Time series with a trend
  - trend estimation via curve fitting
  - trend estimation via smoothing
- Time series with seasonal variation but no trend
  - a simple method to estimate additive seasonal effects
- Time series with a trend and seasonal variation
  - Models (additive, multiplicative)
  - Estimation using smoothing techniques
    - via forming a moving average series (R function: `decompose`)
    - via Loess smoothing (R function: `stl`)
- Measuring temporal dependence
  - Sample autocorrelation function (acf)
  - Correlogram