NHS
Business Services Authority

# Hackathon November 2021
# 1. Day

# Time series

Data points are often not independent of each other (autocorrelation with previous ‘lags’)

=> Forecasts

=> Spurious correlations

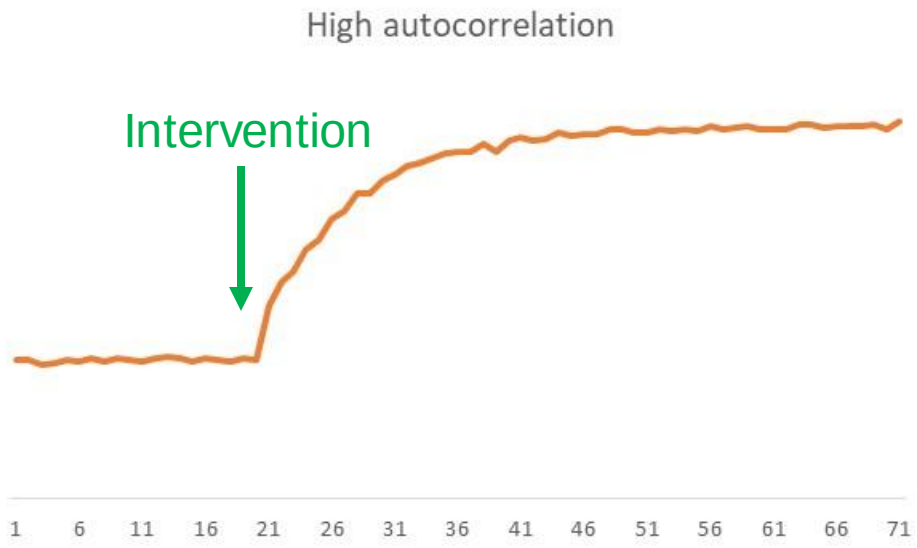
# Purposes of time series analysis

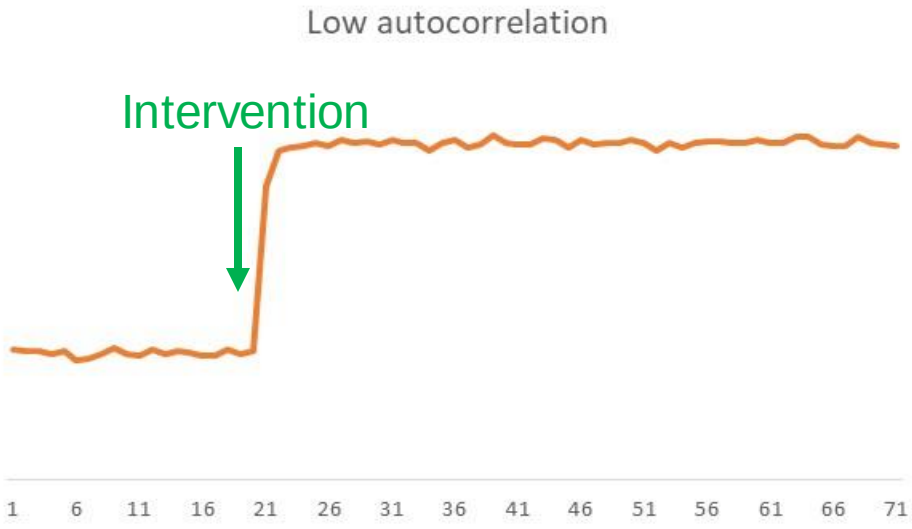
1. Exploring data and understanding underlying processes

    a. Is there autocorrelation?

# Autocorrelation

Intervention with delayed reaction



# Purposes of time series analysis

1. Exploring data and understanding underlying processes

    a. Is there autocorrelation?

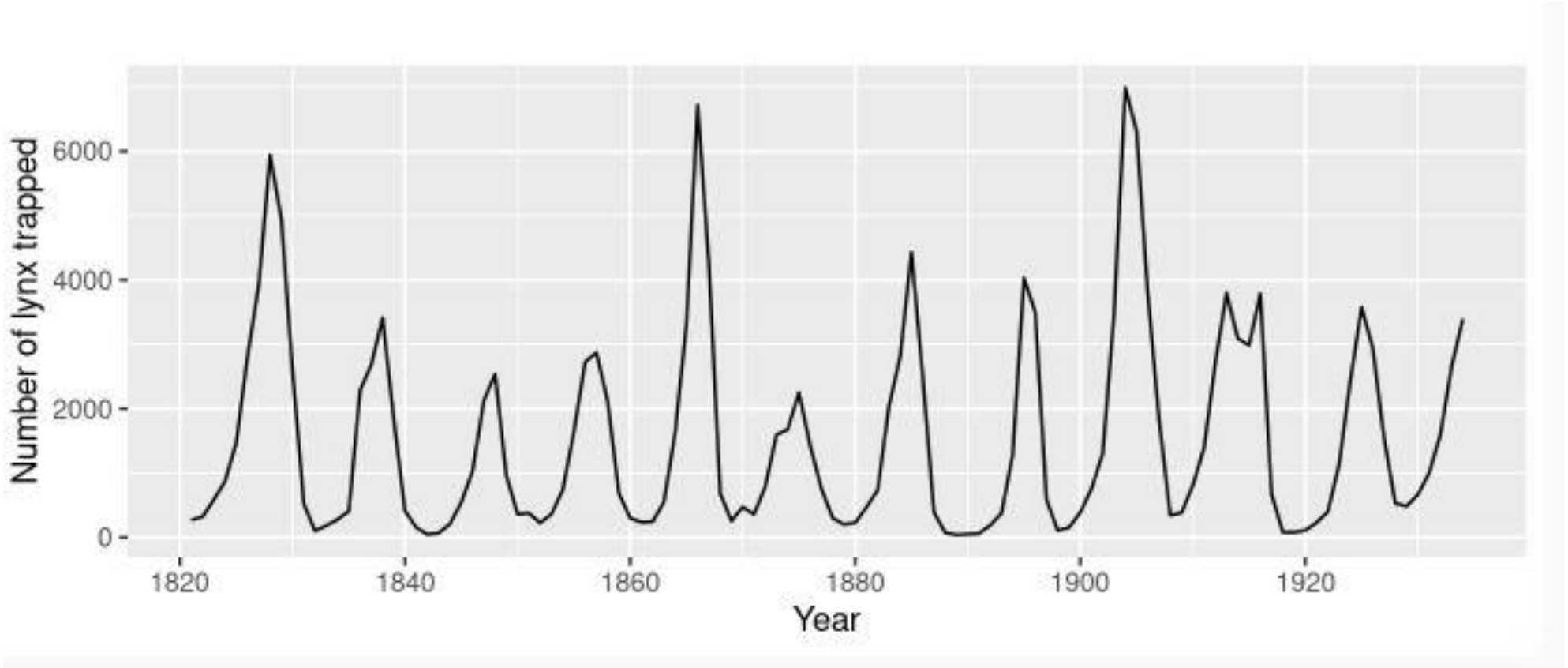
    b. Is there a trend?

    c. Is there seasonality?

    d. Are there cycles?

    Cycles differ from seasonality in so far as they can differ in length and variability.

# Cycles


Number of lynx trapped
6000
4000
2000
0
1820
1840
1860
1880
1900
1920
Year

# Purposes of time series analysis

2. Forecasting

   a. Estimation of confidence intervals

   b. Are reliable forecasts possible?

# Purposes of time series analysis

3. What correlates might influence the data?

   a. Spurious correlations

   b. Which time lag?

# Time series analysis

1. Day

Explore single time series including forecasting

2. Day

Multiple time series, correlations

# Time series analysis

Two main methods:

a. Exponential smoothing

(additive exponential smoothing is a special case of ARIMA, but it is more restricted; multiplicative exponential smoothing is different from ARIMA)

b. ARIMA models

c. (Machine learning)

# Tips

1. Always plot time series data first
2. How to deal with missing values (and outliers); e.g. LOCF, NOCB, na_kalman() in R package imputeTS, tsoutliers
3. Variance stabilizing data transformations: e.g., Box-Cox transformation, arcsin(2p-1) (proportions), log(x) (counts)

   BoxCox() in R package forecast; retransformation: median or mean forecast (!); InvBoxCox() biasadj=FALSE/TRUE
4. Training and test data
5. Automatic fitting of a time series model: ses() (exponential smoothing), decompose() (seasonal exponential smoothing), auto.arima() in package forecast
6. Customised fitting of ARIMA models
7. Stationary time series, tests for stationarity; R: adf.test(), kpss.test()
8. Differentiating of time series => ARMA models; re-differentiate forecasts
9. ACF and PACF
10. Measures of fit

    95% confidence interval of forecasts

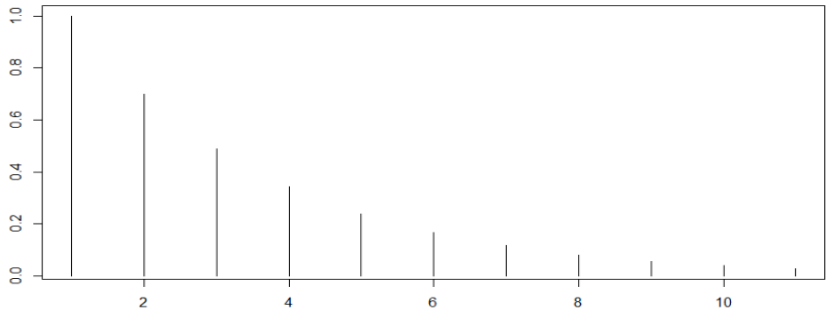
    Mincer-Zarnowitz Test / Wald test (only use when comparing fitted values to test data; not suitable for comparing fitted values to training data)

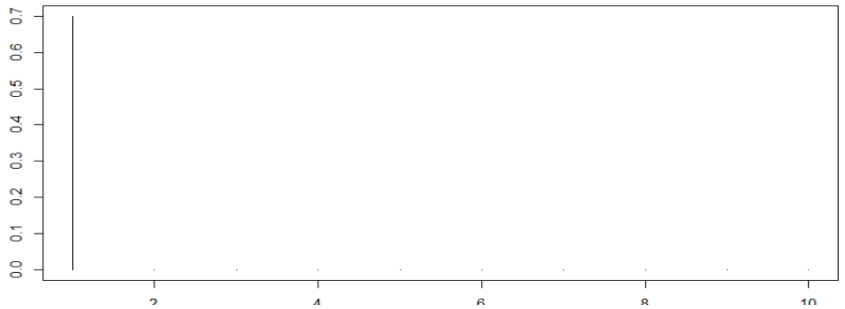
# AR() process

ACF ‘tapers off’

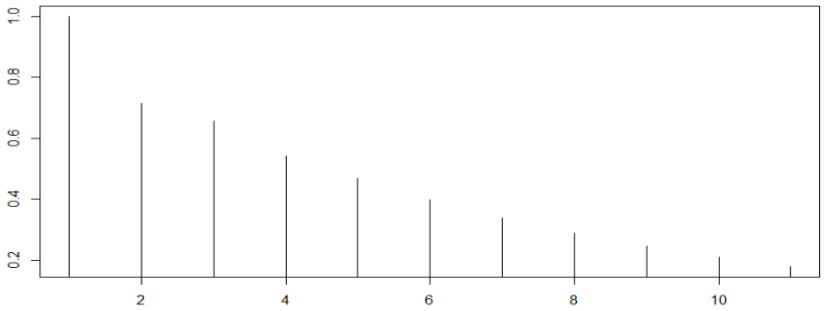
PACF indicates relevant lags in AR() process

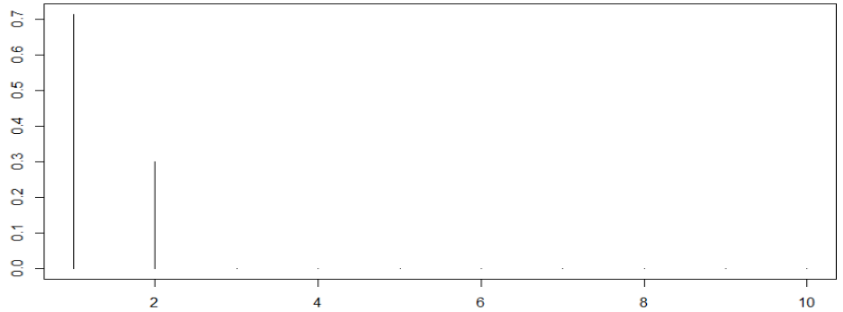
## AR(1)



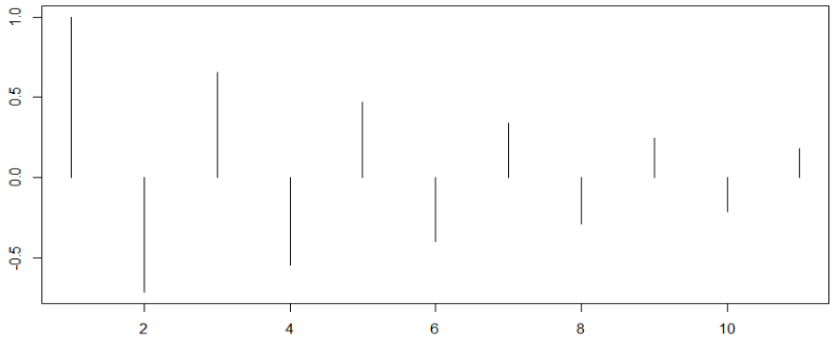

## AR(2)

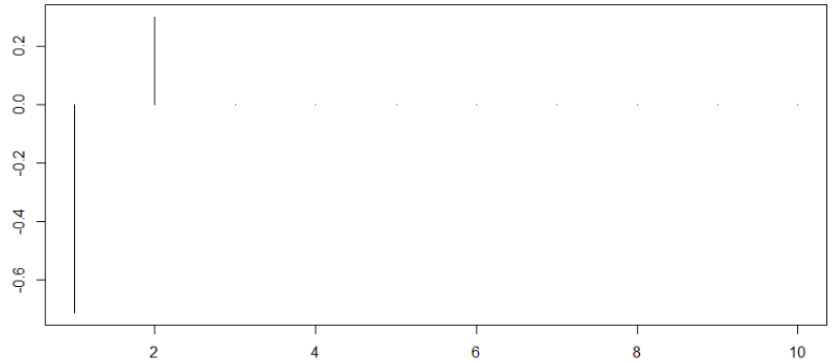



## AR(2)

with first coefficient negative



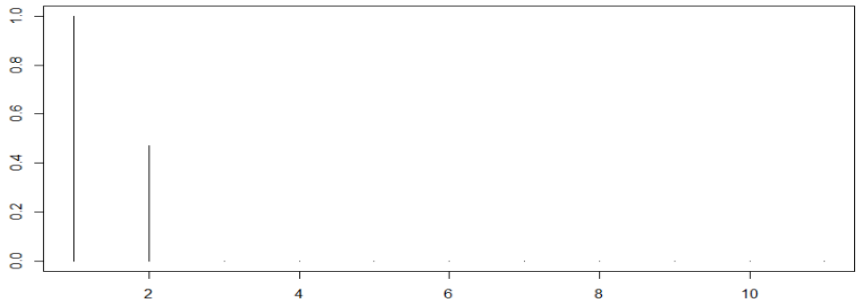



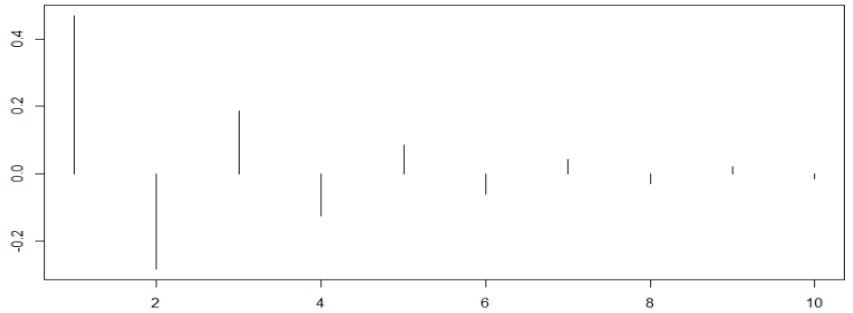
# MA() process

ACF indicates relevant lags in MA() process

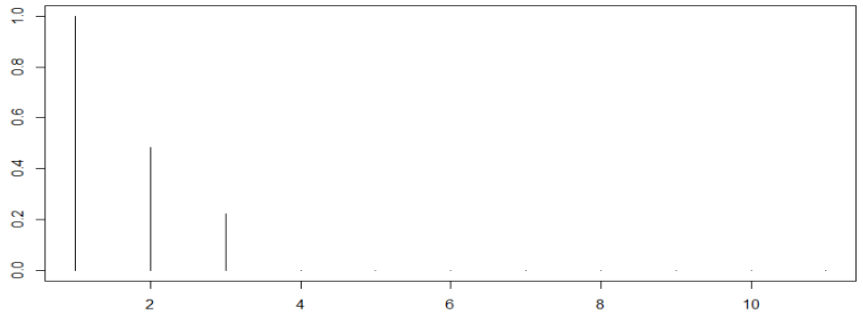
PACF 'tapers off'

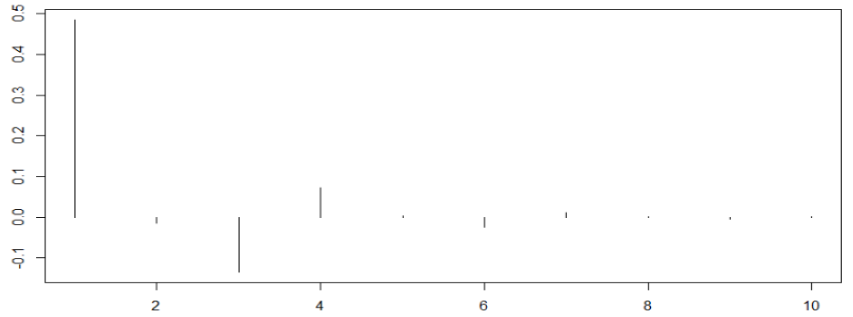
## MA(1)



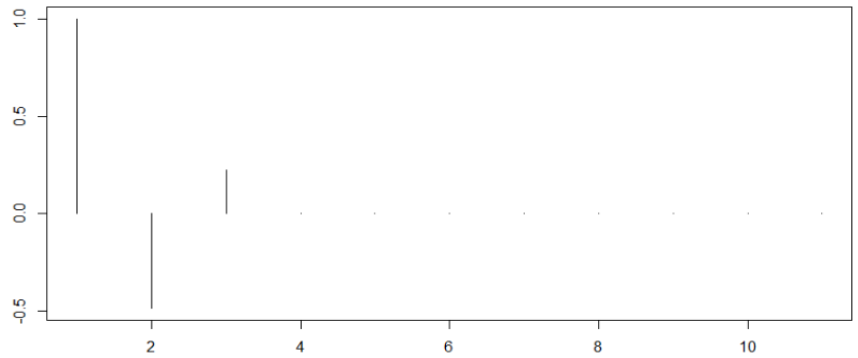

## MA(2)

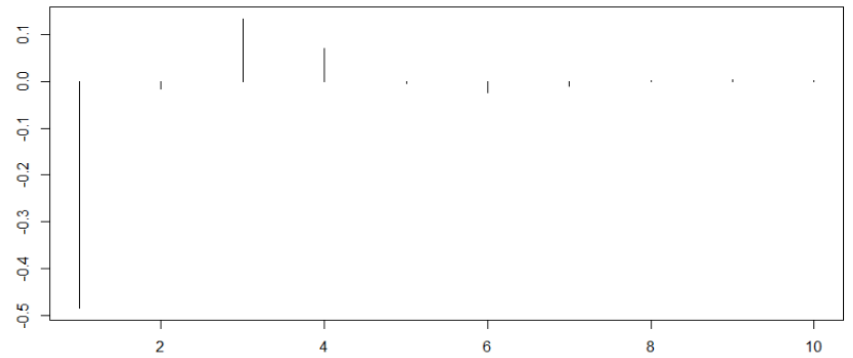



## MA(2)

with first coefficient negative



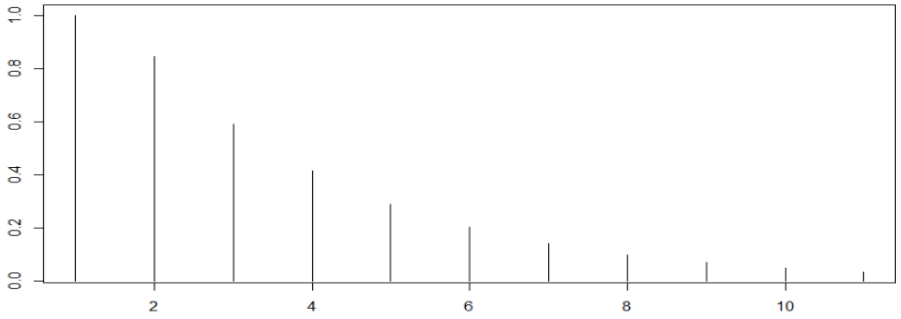



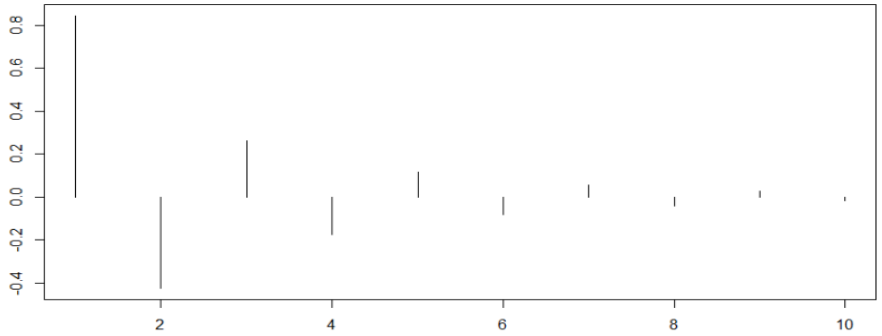
# ARMA() process

ACF "tapers off"

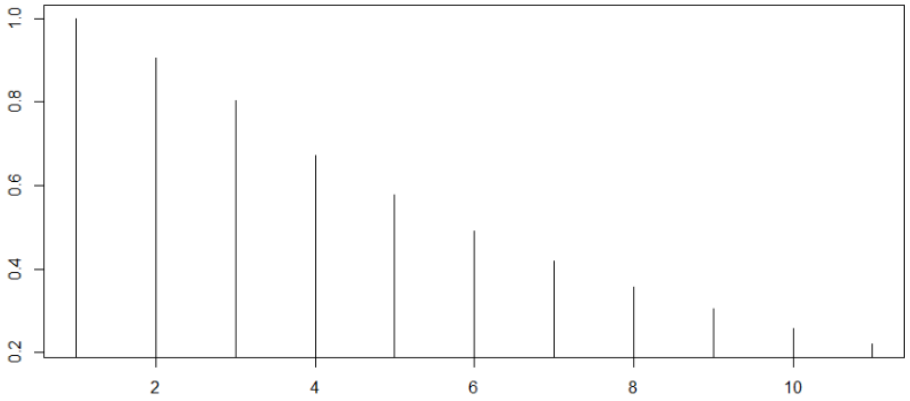
PACF 'tapers off'

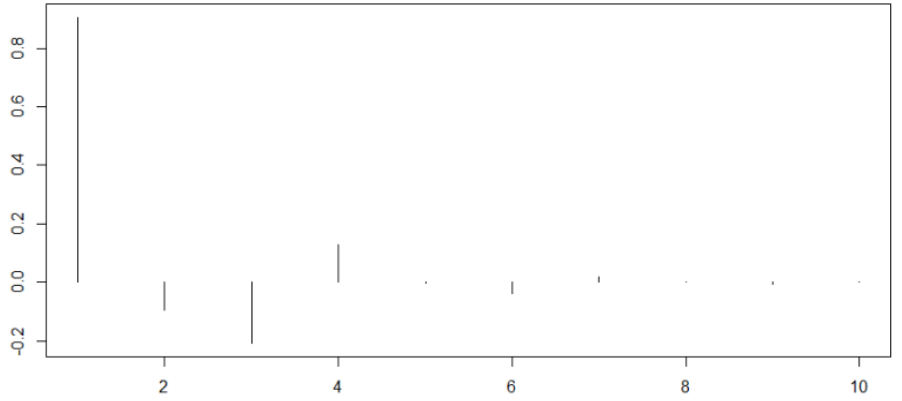
## ARMA(1,1)




## ARMA(2,2)



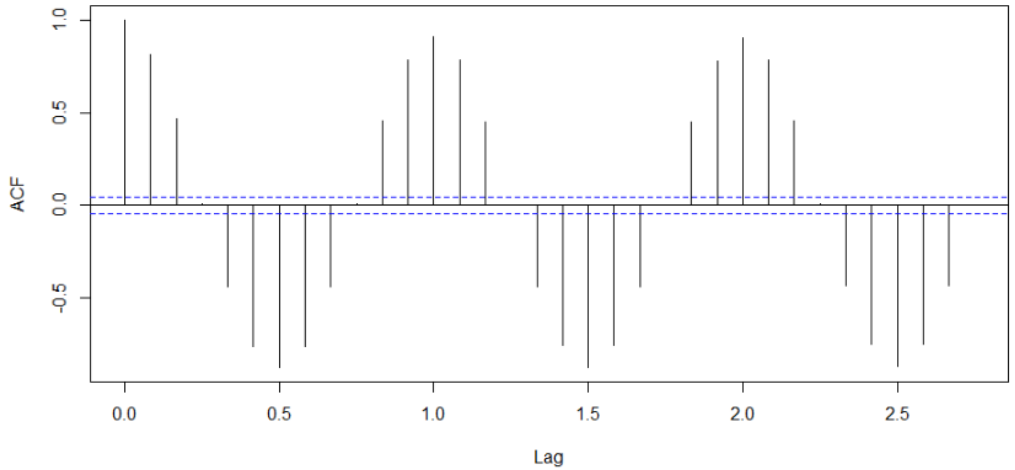



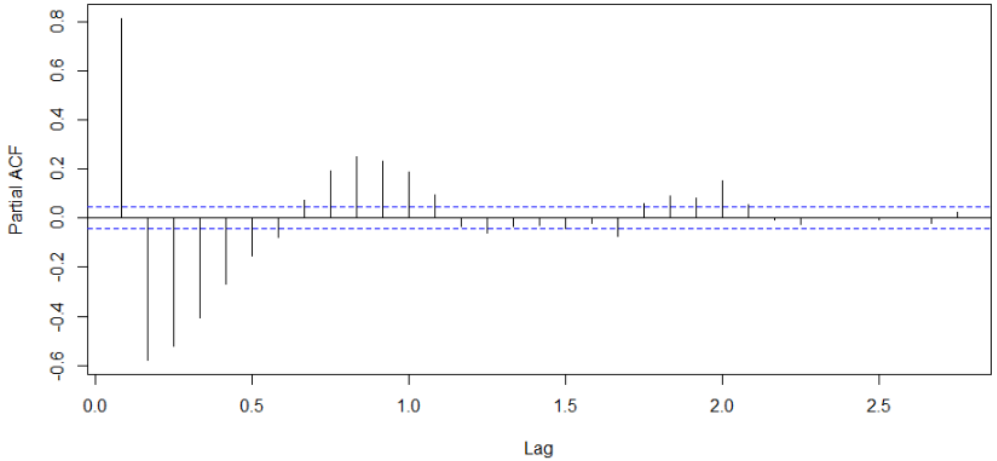
# ARIMA() seasonal process

ACF

PACF

**ARIMA(2,0,0)(2,1,0)[12]**



# Data sets

**COVID data**

owid-covid-data_28102021.xlsx

You can download a newer version:

covid-19-data/public/data at master · owid/covid-19-data · GitHub

**Mortality data**

AnnualMortalityEnglandWales19902020.xlsx

MortalityDataEngandWales_2011_2021.xlsx (weekly)

**Weather data**

WeatherDataFormatted.xlsx

**Other data**

EEGEyeStateData.xlsx

M3C.xlsx

lynx.txt

R library("datasets"), data("NelPlo")

# Some suggestions for first day

COVID data:

- Are reliable forecasts possible?
  => Measures of fit

Mortality data:

- Use time series analysis to estimate excess mortality

Weather data:

- Estimate climate changes

# Groups

1. Matt, Kirsty, Adnan, Kayoung
2. Grace, Mihail, Adam, (Larissa)
3. Lynn, Nadine, Steven, Graeme

Springer: Larissa

# Websites on Time Series Analysis

**General**

Forecasting: Principles and Practice (2nd ed) (otexts.com)

Time Series Analysis in R Part 1: The Time Series Object | DataScience+ (datascienceplus.com)

Time Series Analysis in R Part 2: Time Series Transformations | DataScience+ (datascienceplus.com)

📣 Announcing PyCaret's New Time Series Module | by Moez Ali | Nov, 2021 | Towards Data Science

R: package fable

**tsoutliers**

forecasting - Detecting Outliers in Time Series (LS/AO/TC) using tsoutliers package in R. How to represent outliers in equation format? - Cross Validated (stackexchange.com)

# Schedule

**Day 1 (Wednesday)**

| | |
|---|---|
| 9:00 – 9:45 | Short Introduction by Larissa |
| 10:00 – 10:30 | Intruduction to Git for version control meeting |
| 10:30 – 11:45 | Working in teams |
| 11:45 – 12:00 | Short get-together |
| 12:00 – 13:00 | Lunch break |
| 13:00 – 16:00 | Working in teams |
| 16:00 – 17:00 | Get together and reporting back |

**Day 2 (Thursday)**

| | |
|---|---|
| 9:00 – 9:30 | Short Introduction by Larissa |
| 9:30 – 11:30 | Working in teams |
| 11:30 – 12:00 | Short get-together |
| 12:00 – 13:00 | Lunch break |
| 13:00 – 16:00 | Working in teams |
| 15:45 – 17:00 | Final get together |