DATA ANALYTICS

Unit 3: Concept of stationarity, DF and ADF test, transformations, ARIMA

Jyothi R.

Department of Computer Science and Engineering

Auto Regressive Integrated Moving Average (ARIMA) Process

ARIMA model was proposed by Box-Jenkins (1970)

and so known as Box-Jenkins Methodology

It has three components and is represented as ARIMA(p,d,q):

1. Auto-regressive with lag p
2. Integration component (d)
3. Moving average (q)

Integration component's objective: To convert a nonstationary signal to stationary

Nonstationarity could arise from deterministic or stochastic trend

Introduction to ARIMA

Introduction to ARIMA

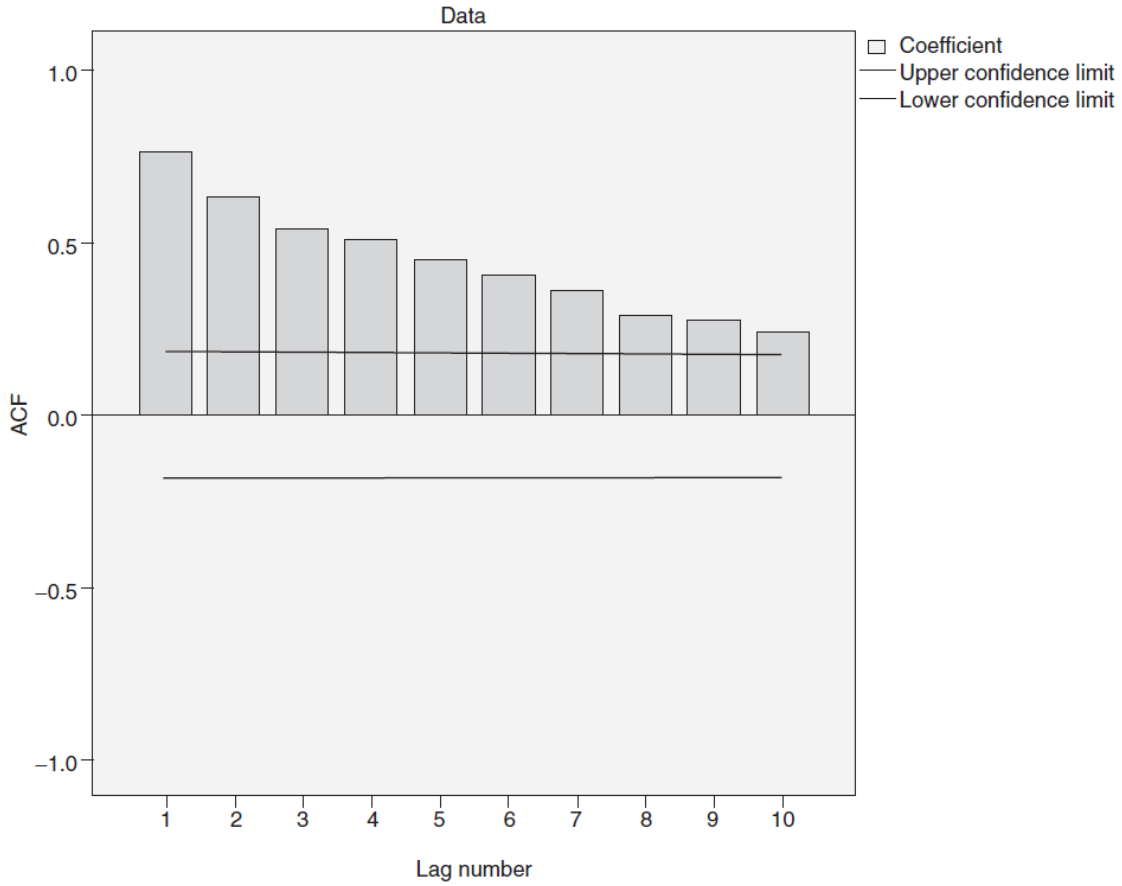
Terminologies

1.Differencing

2. Random walk model

Identifying nonstationarity - ACF

- ACF will not cut off to zero quickly; may show a very slow decline



Quantitative Test - Dickey Fuller (DF) Test

Consider AR(1) process defined below:

$$Y_{t+1} = \beta Y_t + \varepsilon_{t+1}$$

In Section 13.11, we proved that the AR(1) process can become very large when $\beta > 1$ and is non-stationary when $|\beta| = 1$. Dickey–Fuller test (Dickey and Fuller, 1979) is a hypothesis test in which the null hypothesis and alternative hypothesis are given by

H_0: $\beta = 1$ (the time series is non-stationary)

H_1: $\beta < 1$ (the time series is stationary)

The AR(1) can be written as

$$Y_{t+1} - Y_t = \Delta Y_t = (\beta - 1)Y_t + \varepsilon_{t+1} = \psi Y_t + \varepsilon_{t+1} \qquad (13.46)$$

Dickey Fuller Test

- $\psi = 0$ is same as $\beta = 1$. So, the Dickey–Fuller test can be written in terms of ψ;

H_0: ψ 0 (the time series is non-stationary)
H_A: ψ 0 (the time series is stationary)

- The test statistic is given by
- DF Test Statistic = $\dfrac{\psi}{S_e(\psi)}$

- where S_e is the standard error. Note that DF test statistic is not *t*-statistic since the null hypothesis is on non-stationary process.
- Critical values are derived based on simulation

Augmented Dickey–Fuller Test

- Dickey–Fuller test is valid only when the residual ε_{t+1} follows a white noise.
- When ε_{t+1} is not white noise, the actual series may not be AR(1); it may have more significant lags.
- To address this issue, we augment *p*-lags of the dependent variable *Y*.
- The model can be rewritten as:

$$\Delta Y_t = \psi Y_t + \sum_{i=0}^{p} \alpha_i \Delta Y_{t-i} + \varepsilon_{t+1}$$

- The above equation can be
- Again the null and alternative hypotheses are
- H_0: = 0 (the time series is non-stationary)
- H_0: ψ< 0 (the time series is stationary)

ψ

Stationarity and differencing

Transforming Non-Stationary Process to Stationary Process Using Differencing

- The first step in ARIMA is to identify the order of differencing (*d*) required to convert a non-stationary process into a stationary process.
- Many time-series data will be non-stationary due to factors such as trend and seasonality.
- If the non-stationary behaviour is due to trend, then it can be converted into a stationary process by de-trending the data.
- De-trending is usually achieved by fitting a trend line and subtracting it from the time series; this is known as **trend stationarity**.
- When the reason is not due to trend stationarity, then differencing the original time series may be useful for converting the non-stationary process into a stationary process (called **difference stationarity**).

Transformation

Transforming Non-Stationary Process to Stationary Process Using Differencing

- The first difference (d = 1) is the difference between consecutive values of the time series (Y_t and Y_{t-1})

- That is, the first difference ΔY_t is given by

$$\Delta y_t = Y_t - Y_{t-1}$$

- The second difference (d = 2) is the difference of the first differences and is given by

$$\nabla^2 Y_t = \nabla(\nabla Y_t) = Y_t - 2\,Y_{t-1} + Y_{t-2}$$

- In most cases, $d \leq 2$ will be sufficient to convert a non-stationary process to a stationary process.

Stationarity and differencing

ILLUSTRATION OF DIFFERENCING AFTER INFLATION ADJUSTMENT

Consumer Price Index, 1990=1.0

Auto sales ($B)

Deflated auto sales:
16.13 = 4.79 / 0.297

First difference of deflated auto sales:
0.51= 16.64 - 16.13, etc.

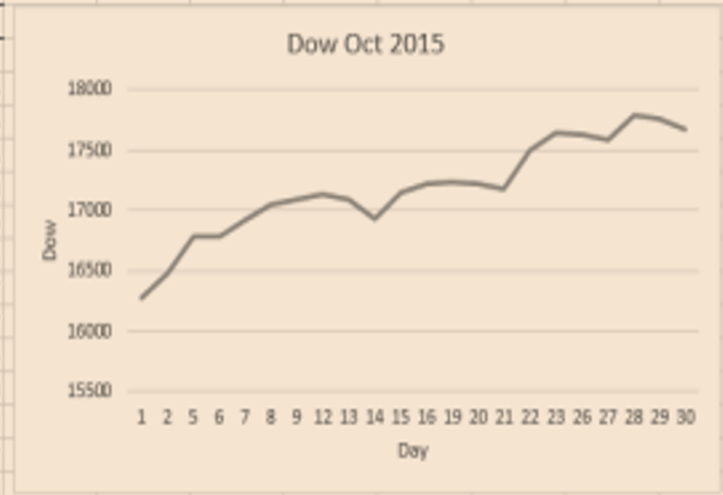
DATE	AUTOSALE	CPI	AUTOSALE/CPI	DIFF(AUTOSALE/CPI)
Jan-70	4.79	0.297	16.13	
Feb-70	4.96	0.298	16.64	0.51
Mar-70	5.64	0.300	18.80	2.16
Apr-70	5.98	0.302	19.80	1.00
May-70	6.08	0.303	20.07	0.27
Jun-70	6.55	0.305	21.48	1.41
Jul-70	6.11	0.306	19.97	-1.51

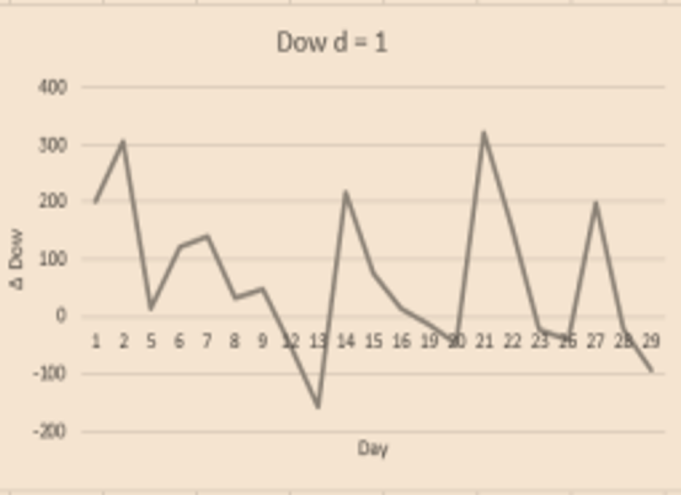
https://people.duke.edu/~rnau/411diff.htm

Stationarity and differencing

Figure 1, clearly shows an increasing trend. Differencing is a way to eliminate such trends.

Date	Dow	d = 1
1	16272.01	
2	16472.37	200.36
5	16776.43	304.06
6	16790.19	13.76
7	16912.29	122.10
8	17050.75	138.46
9	17084.49	33.74
12	17131.86	47.37
13	17081.89	-49.97
14	16924.75	-157.14
15	17141.75	217.00
16	17215.97	74.22
19	17230.54	14.57
20	17217.11	-13.43
21	17168.61	-48.50
22	17489.16	320.55
23	17646.70	157.54
26	17623.05	-23.65
27	17581.43	-41.62
28	17779.52	198.09
29	17755.80	-23.72
30	17663.54	-92.26





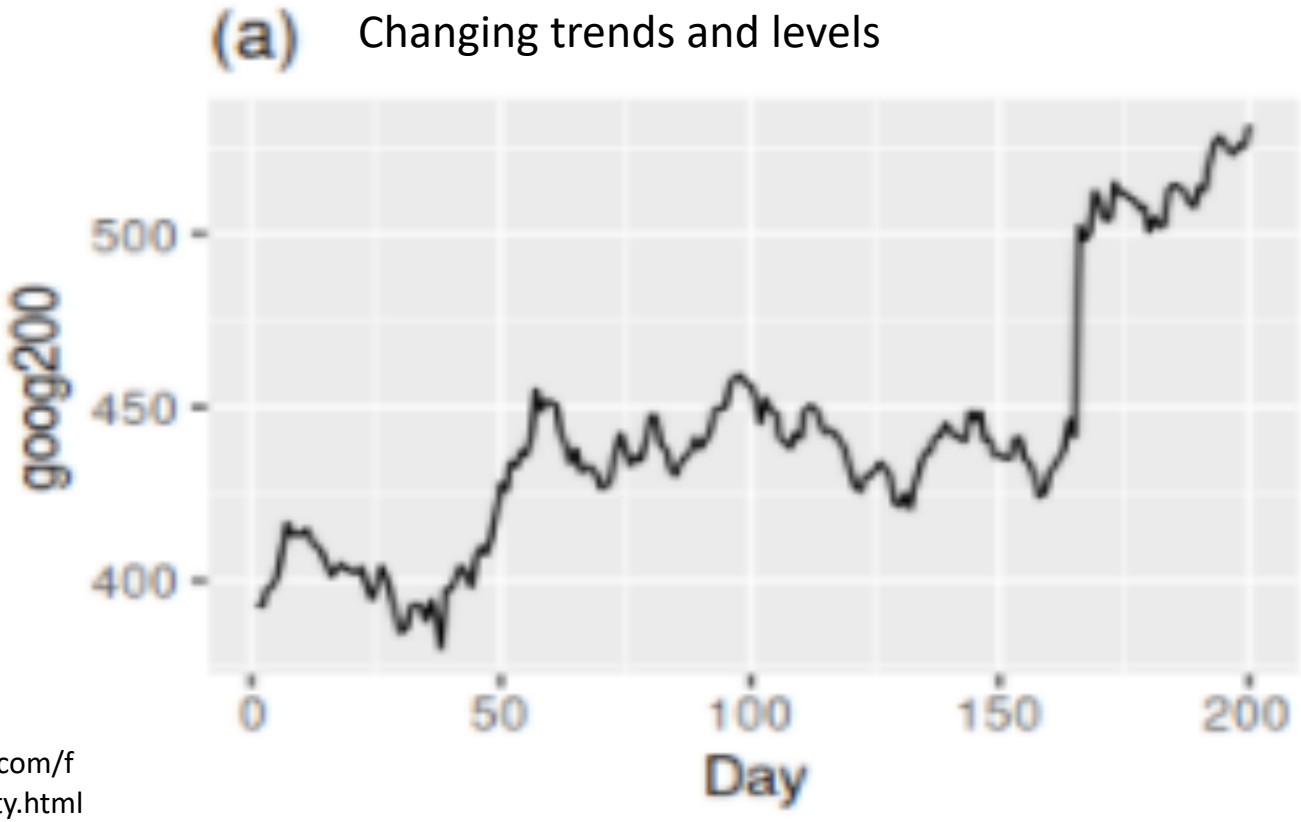
https://www.real-statistics.com/time-series-analysis/arima-processes/arima-differencing/

Stationarity and differencing

- A stationary time series is one whose properties do not depend on the time at which the series is observed
- Time series with trends, or with seasonality, are not stationary — the trend and seasonality will affect the value of the time series at different times.
- On the other hand, a white noise series is stationary — it does not matter when you observe it, it should look much the same at any point in time.
- Some cases can be confusing — a time series with cyclic behaviour but with no trend or seasonality is stationary.
- This is because the cycles are not of a fixed length, so before we observe the series we cannot be sure where the peaks and troughs of the cycles will be.
- In general, a stationary time series will have no predictable patterns in the long-term.

Which of these series are stationary?

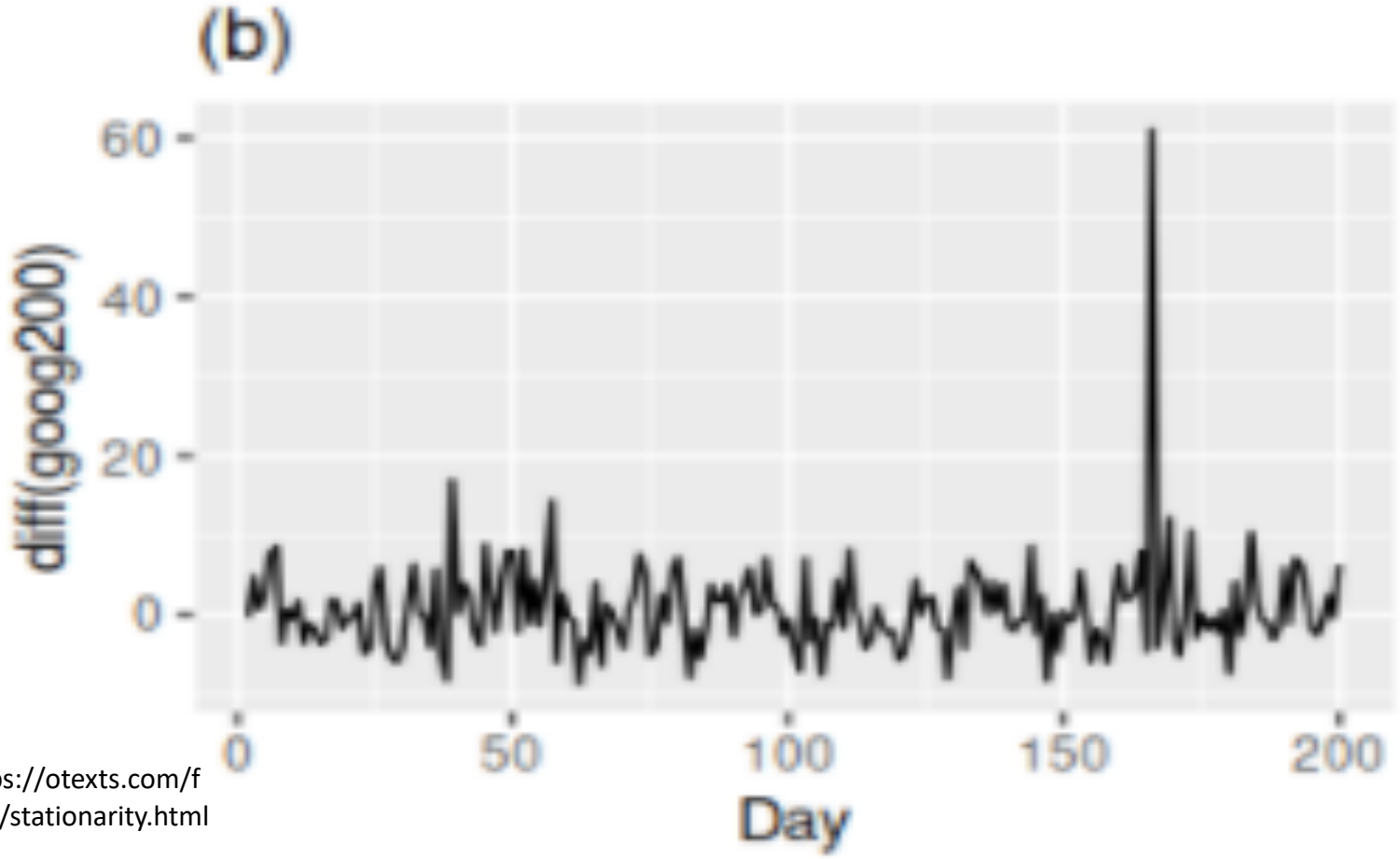
- Time plots will show the series to be roughly horizontal although some cyclic behaviour is possible, with constant variance.



https://otexts.com/fpp2/stationarity.html

Figure1. (a) Google stock price for 200 consecutive days

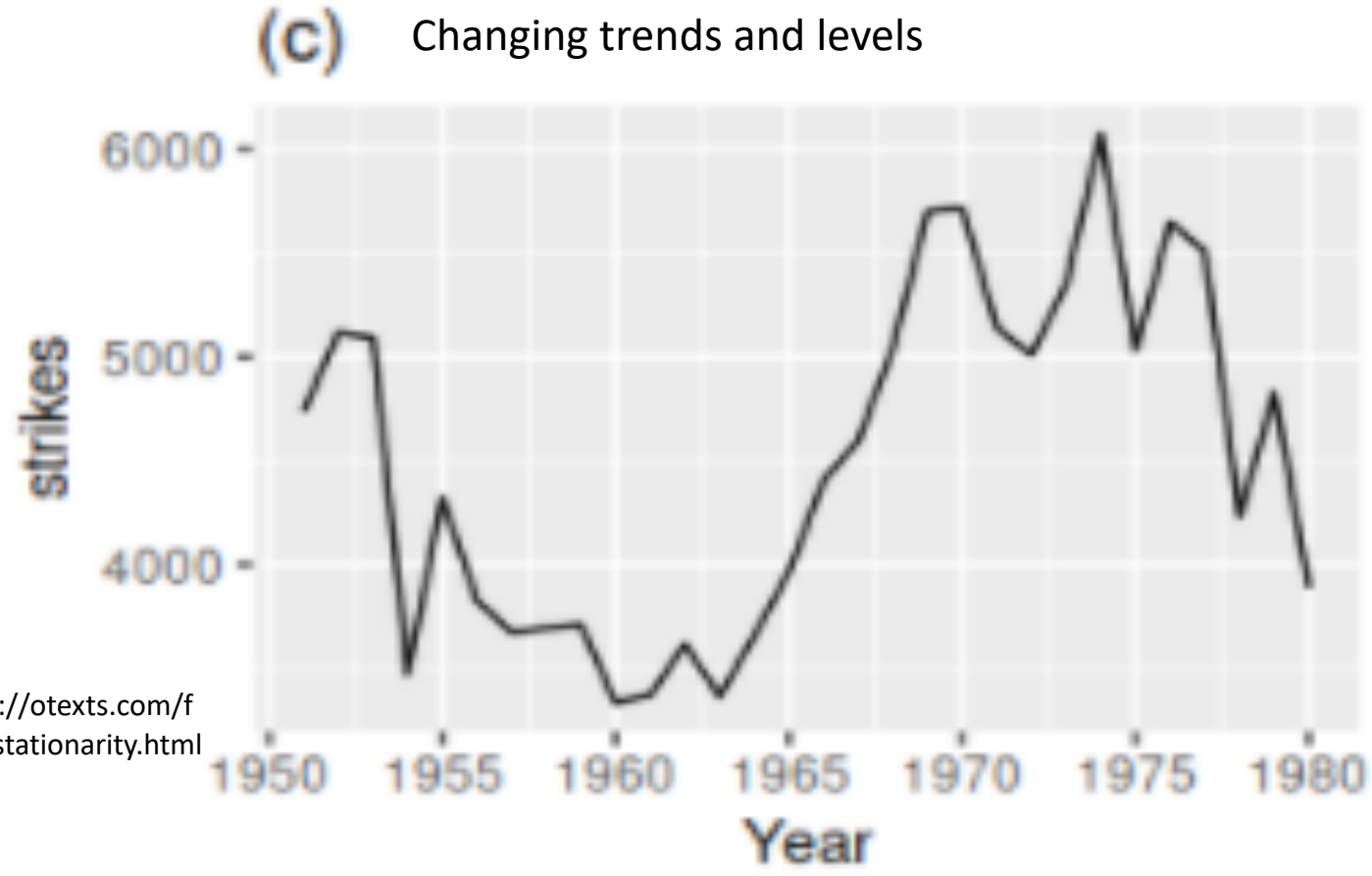
Which of these series are stationary?



https://otexts.com/fpp2/stationarity.html

Figure 2: (b) Daily change in the Google stock price for 200 consecutive days;

Which of these series are stationary?

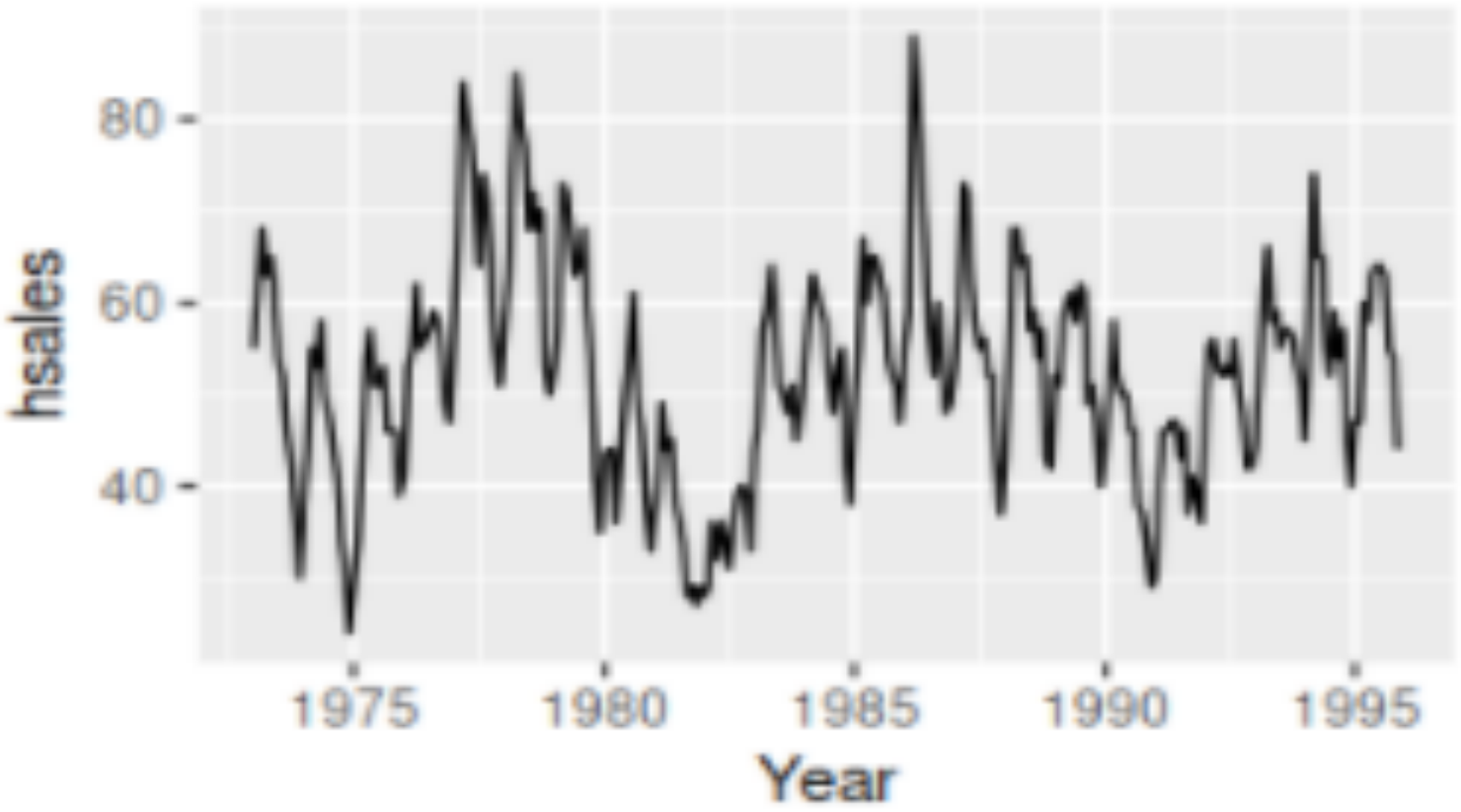


https://otexts.com/fpp2/stationarity.html

Figure 3 :(c) Annual number of strikes in the US

Which of these series are stationary?

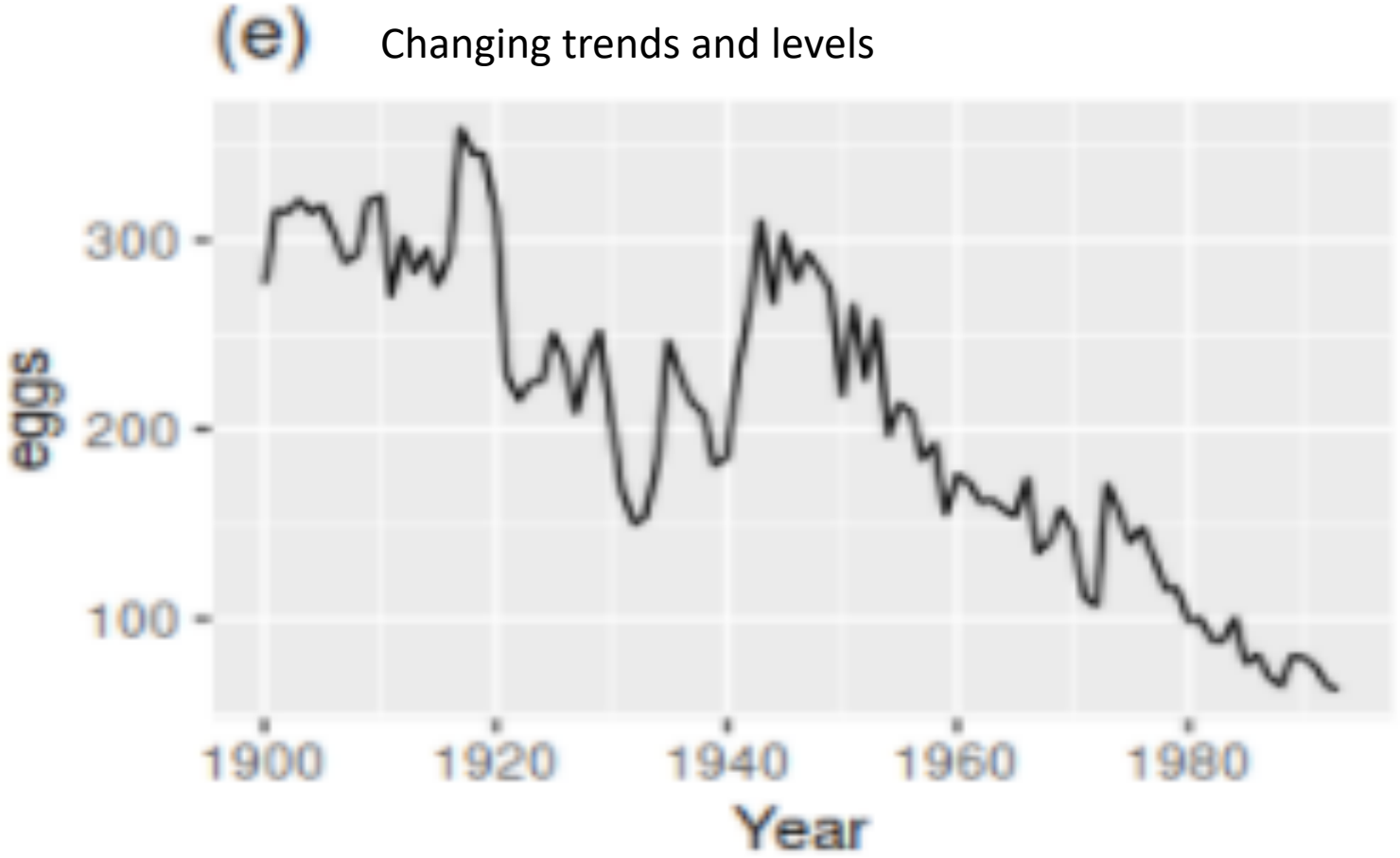
(d) Obvious seasonality



https://otexts.com/fpp2/stationarity.html

Figure 4: (d) Monthly sales of new one-family houses sold in the US

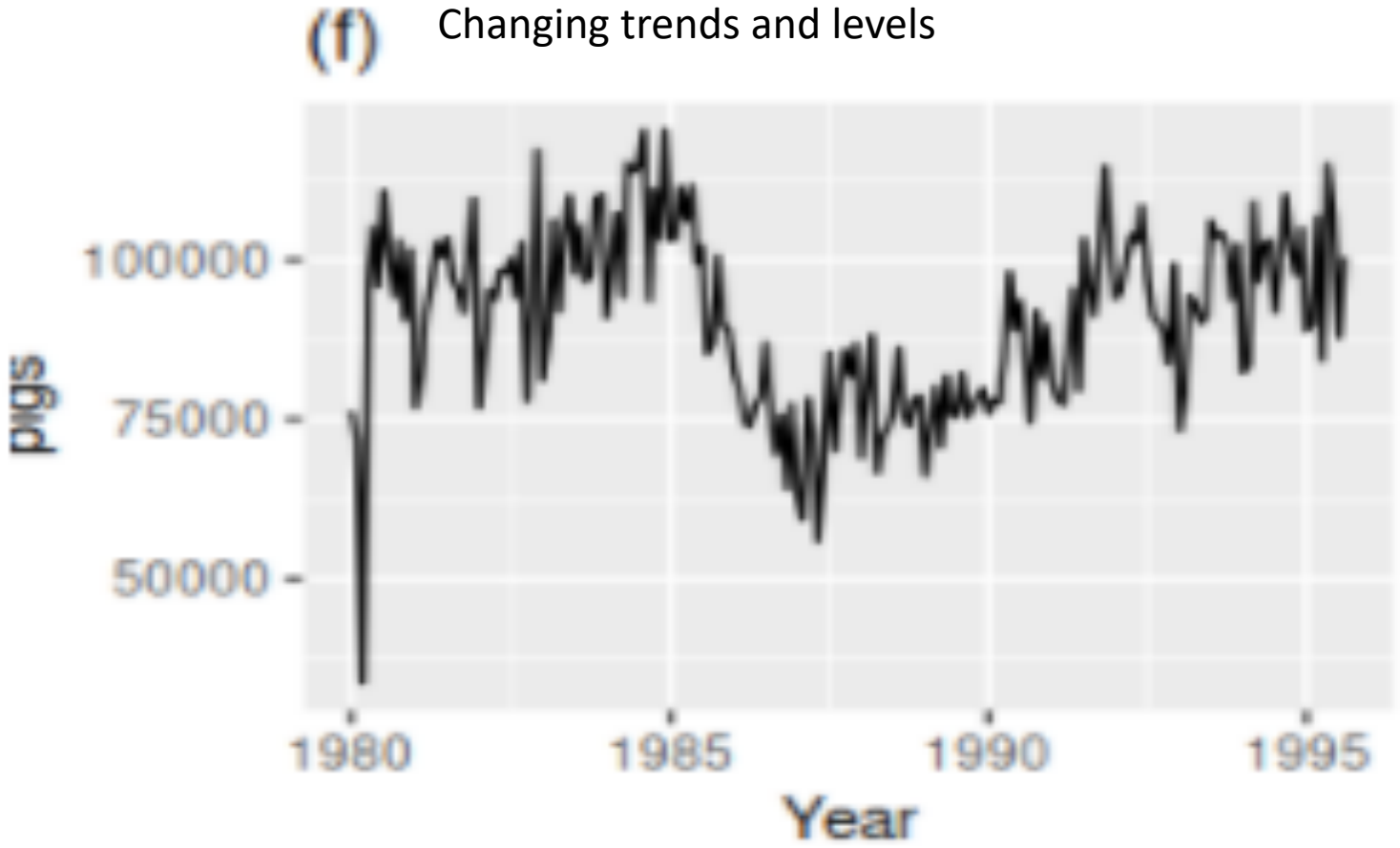
Which of these series are stationary?



https://otexts.com/fpp2/stationarity.html

Figure 5: (e) Annual price of a dozen eggs in the US (constant dollars);

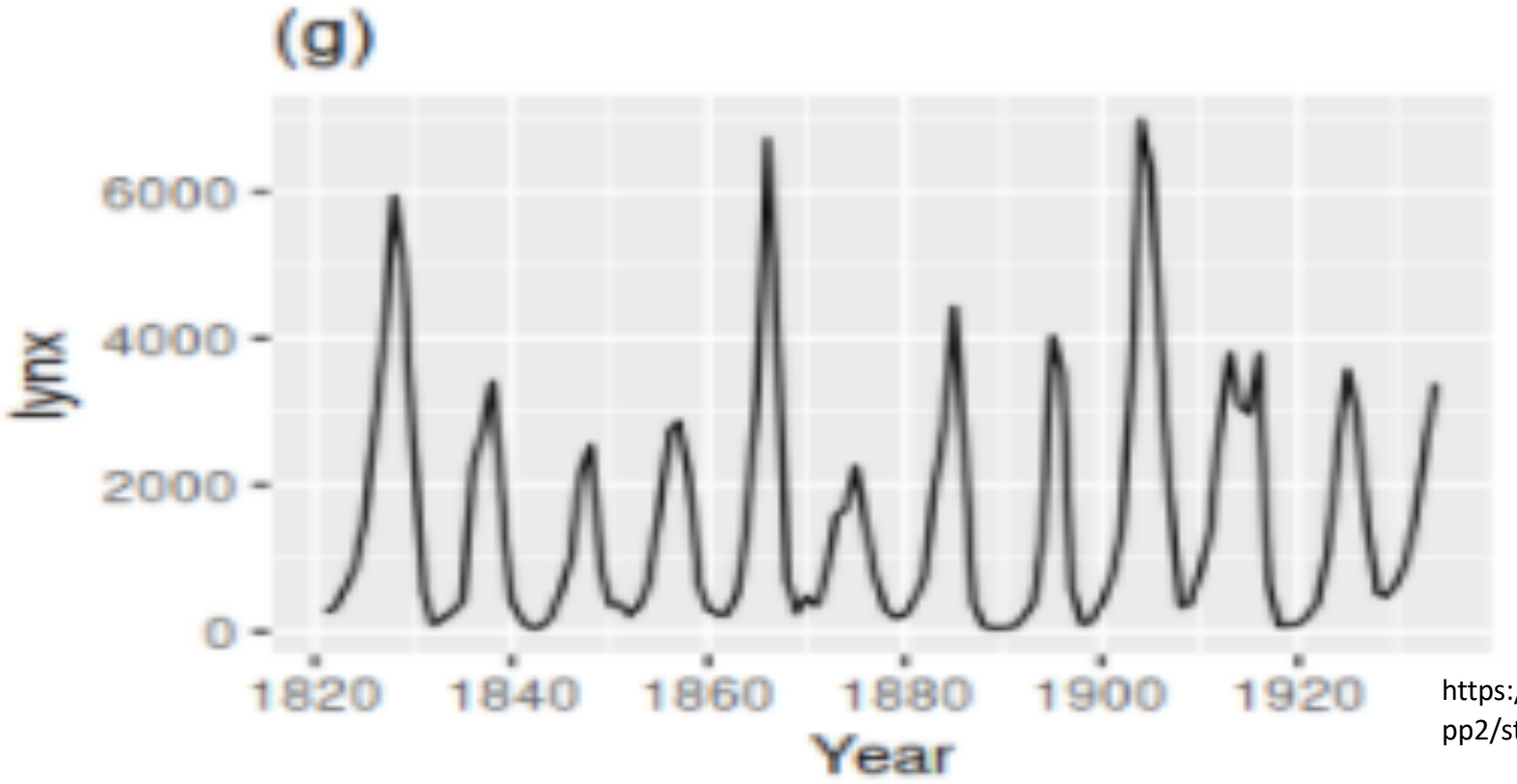
Which of these series are stationary?



https://otexts.com/fpp2/stationarity.html

Figure 6: (f) Monthly total of pigs slaughtered in Victoria, Australia;

Which of these series are stationary?

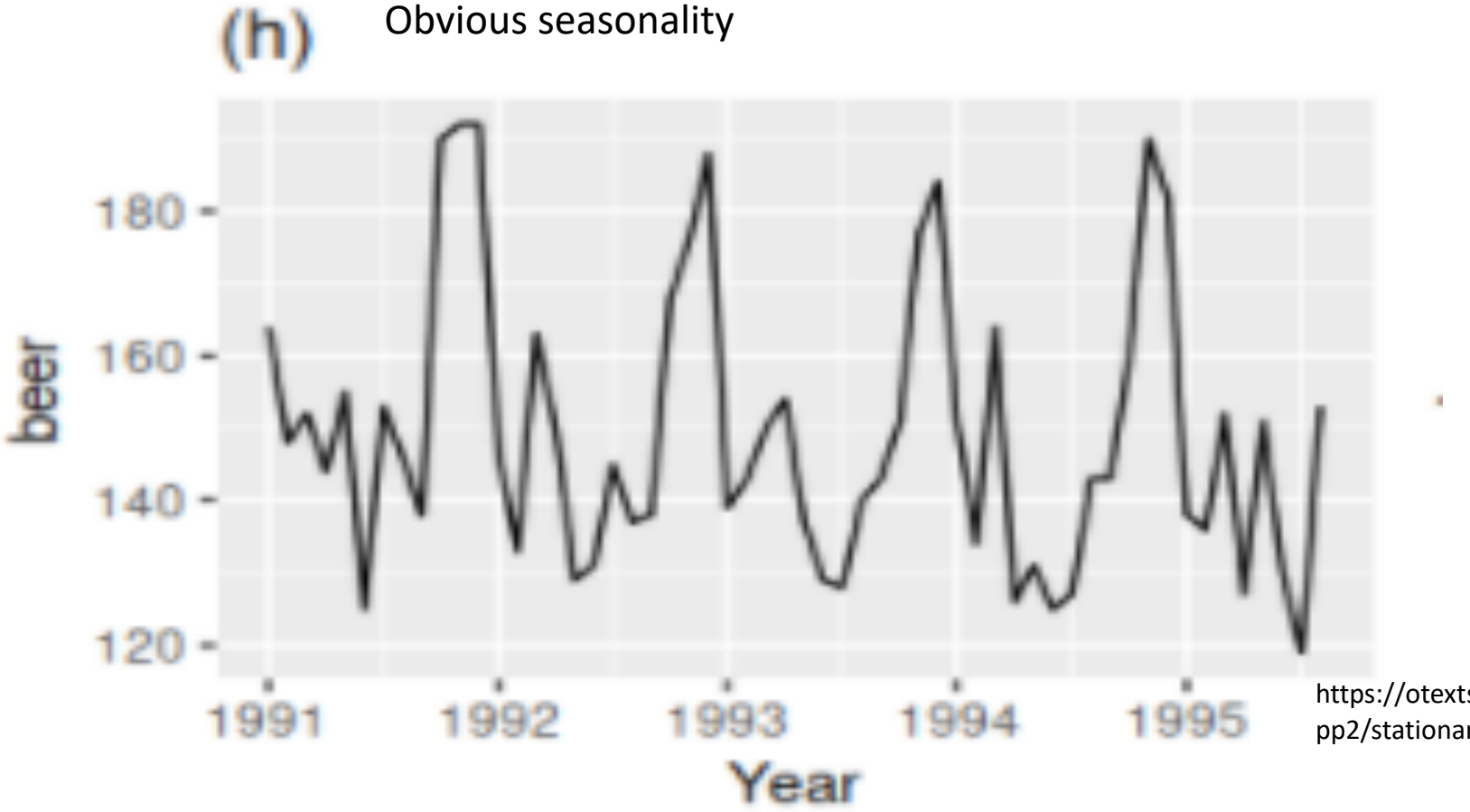


https://otexts.com/fpp2/stationarity.html

Figure 7: (g) Annual total of lynx trapped in the McKenzie River district of north-west Canada;

Which of these series are stationary?

Obvious seasonality

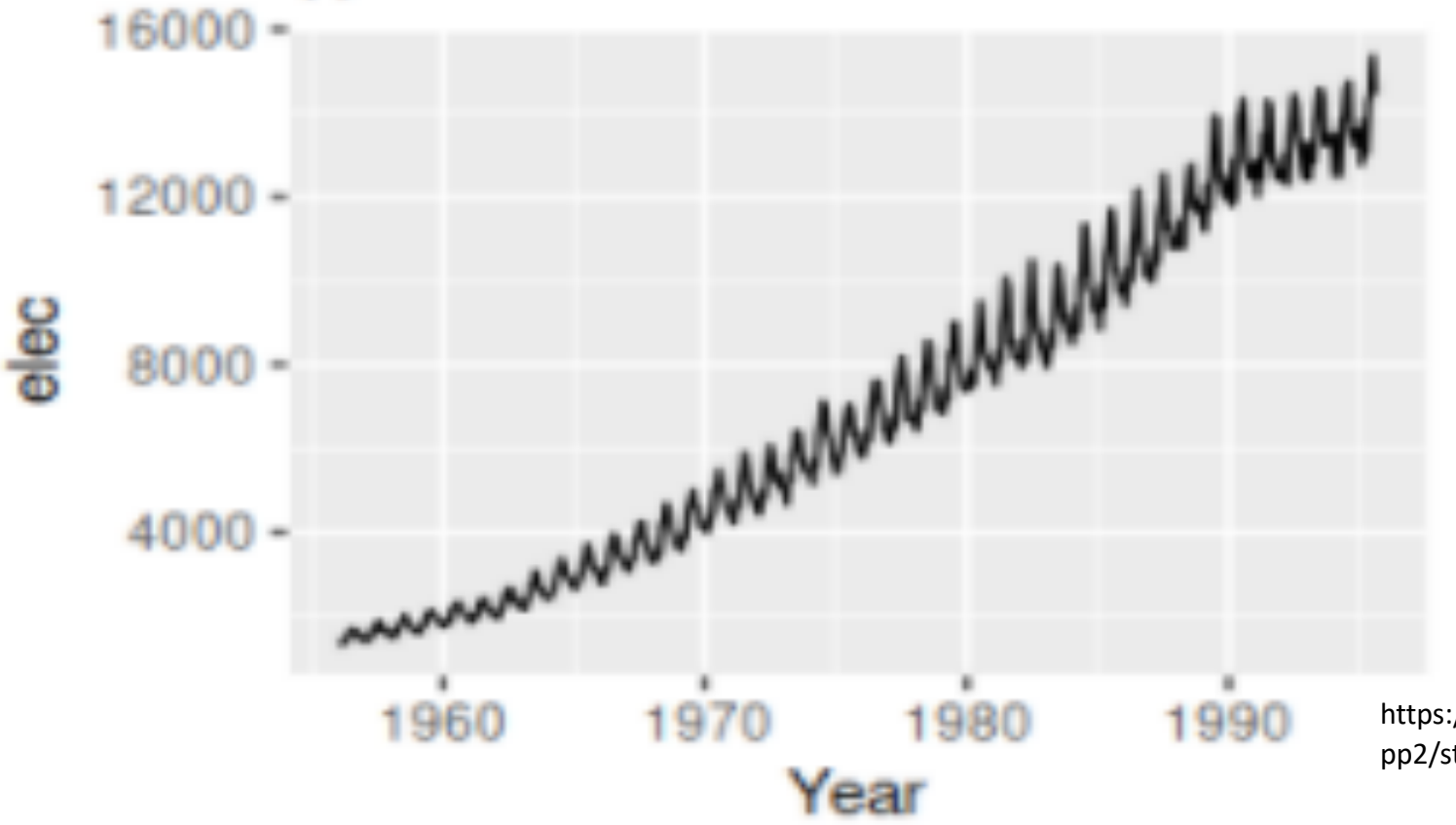


https://otexts.com/fpp2/stationarity.html

Figure 8: (h) Monthly Australian beer production;

Which of these series are stationary?

(i) Obvious seasonality and trend



https://otexts.com/fpp2/stationarity.html

Figure 9:(i) Monthly Australian electricity production.

Which of these series are stationary?

Consider the nine series plotted in Figure 1 to 9:

- Which of these do you think are stationary?

- Obvious seasonality rules out series (d), (h) and (i).
- Trends and changing levels rules out series (a), (c), (e), (f) and (i).
- Increasing variance also rules out (i).
- That leaves only (b) and (g) as stationary series.

- At first glance, the strong cycles in series (g) might appear to make it non-stationary.
- But these cycles are aperiodic — they are caused when the lynx population becomes too large for the available feed, so that they stop breeding and the population falls to low numbers, then the regeneration of their food sources allows the population to grow again, and so on.
- In the long-term, the timing of these cycles is not predictable. Hence the series is stationary.

Differencing

- In Figure 1 to 9: The Google stock price was non-stationary in panel (a)
- But the daily changes were stationary in panel (b). This shows one way to make a non-stationary time series stationary — compute the differences between consecutive observations. This is known as **differencing**.
- Transformations such as logarithms can help to stabilise the variance of a time series.
- Differencing can help stabilise the mean of a time series by removing changes in the level of a time series, and therefore eliminating or reducing trend and seasonality.
- By looking at the time plot of the data, the ACF plot is also useful for identifying non-stationary time series.
- For a stationary time series, the ACF will drop to zero relatively quickly, while the ACF of non-stationary data decreases slowly.
- Also, for non-stationary data, the value of r_1 is often large and positive.

Differencing

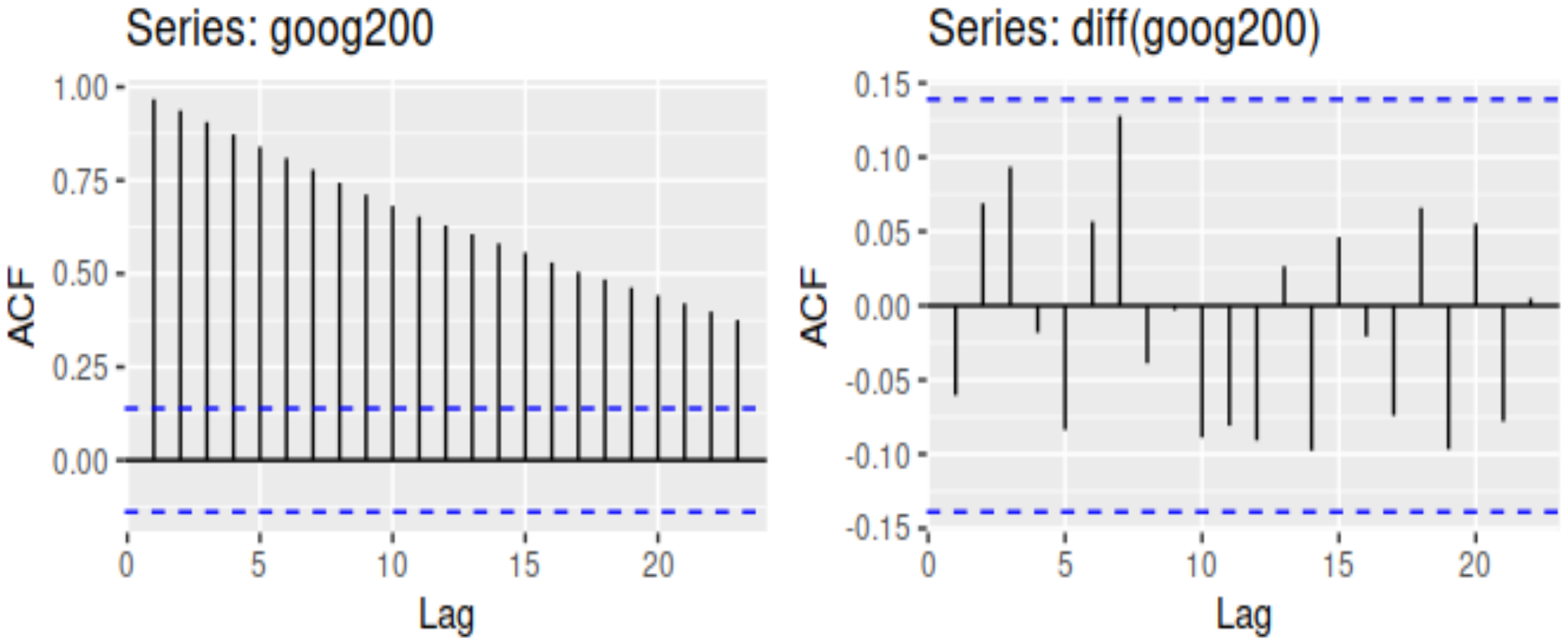


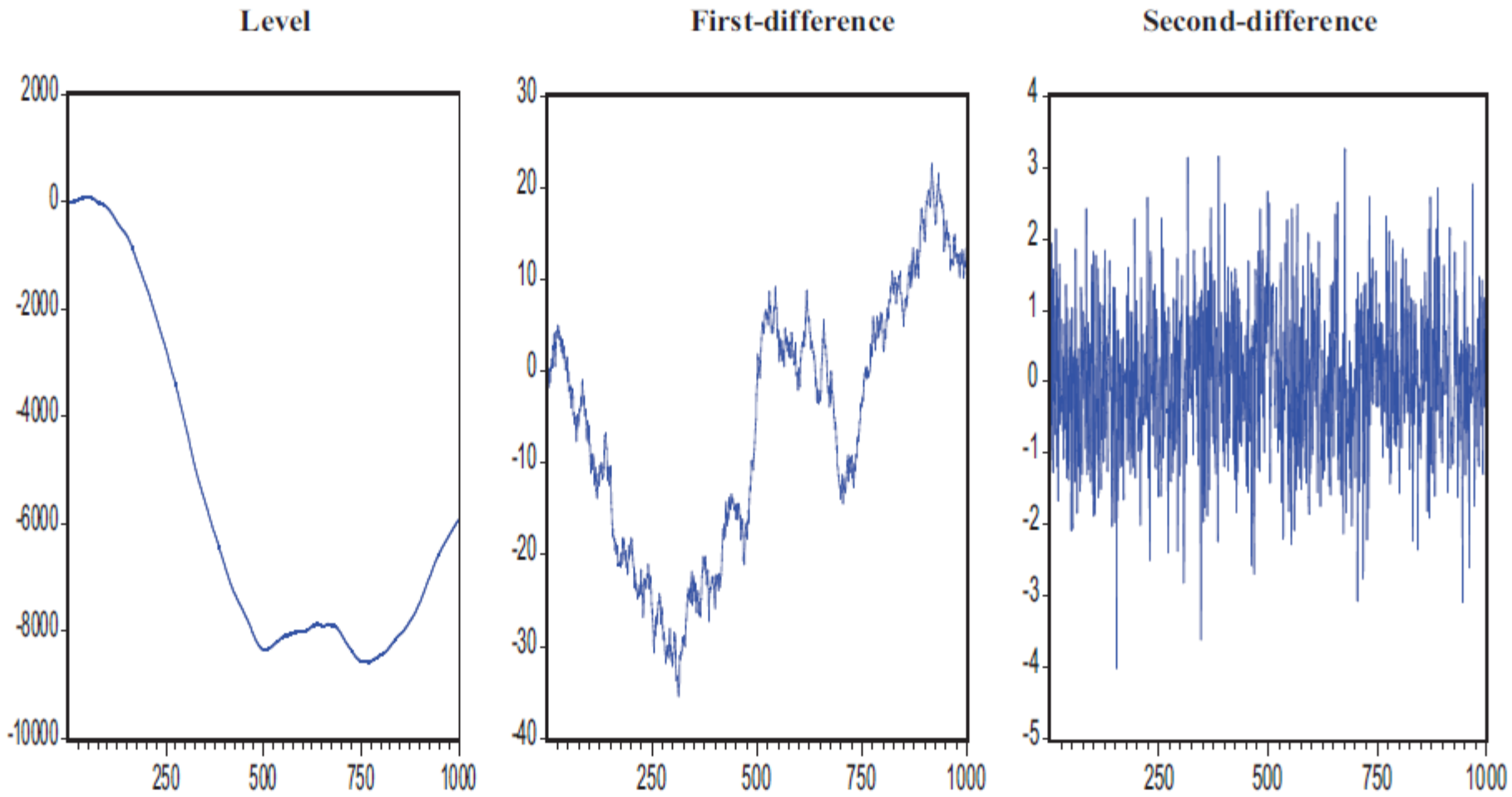
Figure 10:The ACF of the Google stock price (left) and of the daily changes in Google stock price (right).

Differencing

Figure 8.2:The ACF of the Google stock price (left) and of the daily changes in Google stock price (right).

- The ACF of the differenced Google stock price looks just like that of a white noise series.
- There are no autocorrelations lying outside the 95% limits
- This suggests that the *daily change* in the Google stock price is essentially a random amount which is uncorrelated with that of previous days.

Differencing- Example plot



Random Walk Model

- The differenced series is the *change* between consecutive observations in the original series, and can be written as $y'_t = y_t - y_{t-1}$.

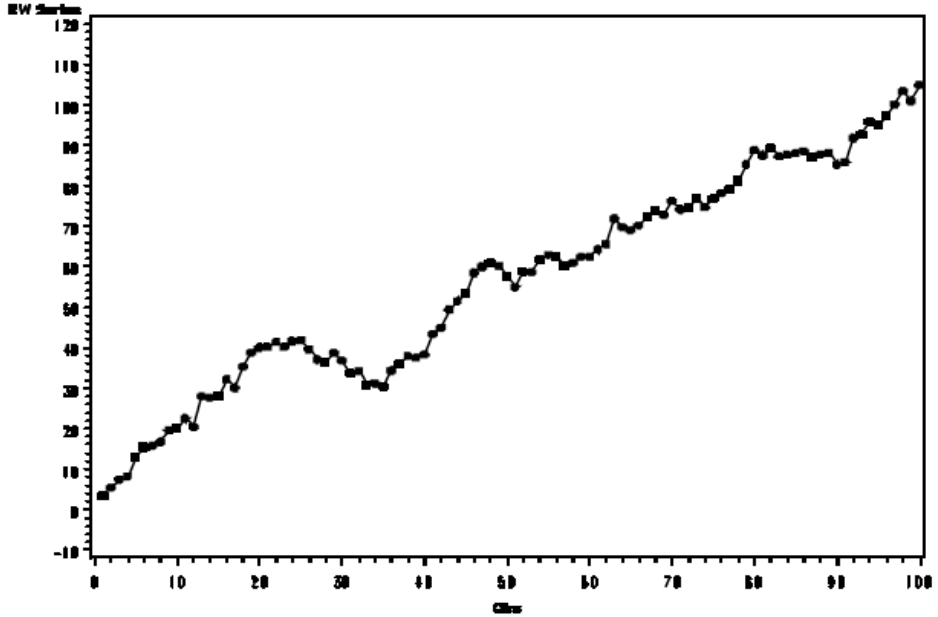
- The differenced series will have only T−1 values, since it is not possible to calculate a difference y'_1 for the first observation.
- When the differenced series is white noise, the model for the original series can be written as

  $y_t - y_{t-1} = \varepsilon_t$, where ε_t denotes white noise.

- Rearranging this leads to the "random walk" model $y_t = y_{t-1} + \varepsilon_t$

Random Walk Model

- Random walk models are widely used for non-stationary data, particularly financial and economic data.



Random walks typically have:

- long periods of apparent trends up or down
- sudden and unpredictable changes in direction.

- The forecasts from a random walk model are equal to the last observation, as future movements are unpredictable, and are equally likely to be up or down.
- Thus, the random walk model underpins naïve forecasts,

DATA ANALYTICS

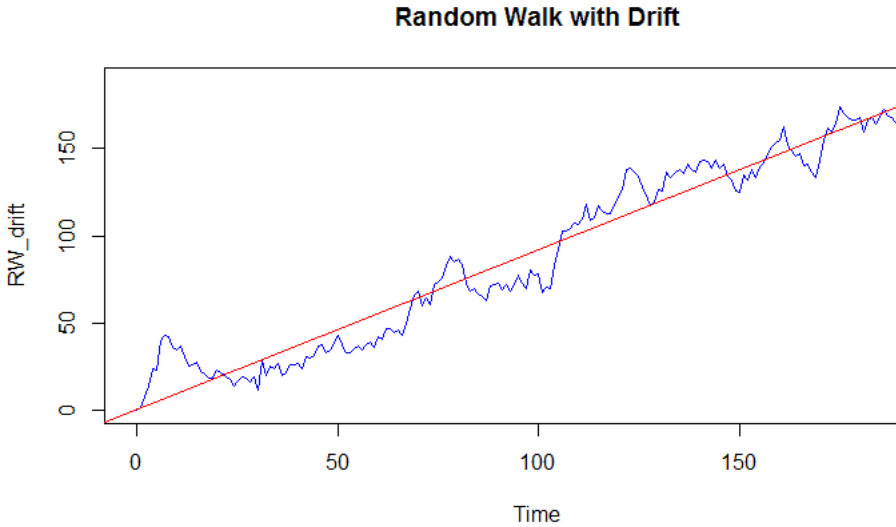
Random Walk Model

- A closely related model allows the differences to have a non-zero mean.

Then

$$y_t - y_{t-1} = c + \varepsilon_t \quad \text{or} \quad y_t = c + y_{t-1} + \varepsilon_t.$$

- The value of c is the average of the changes between consecutive observations.
- If c is positive, then the average change is an increase in the value of y_t.
- Thus, y_t will tend to drift upwards.
- However, if c is negative, y_t will tend to drift downwards.
- This is the model behind the drift method



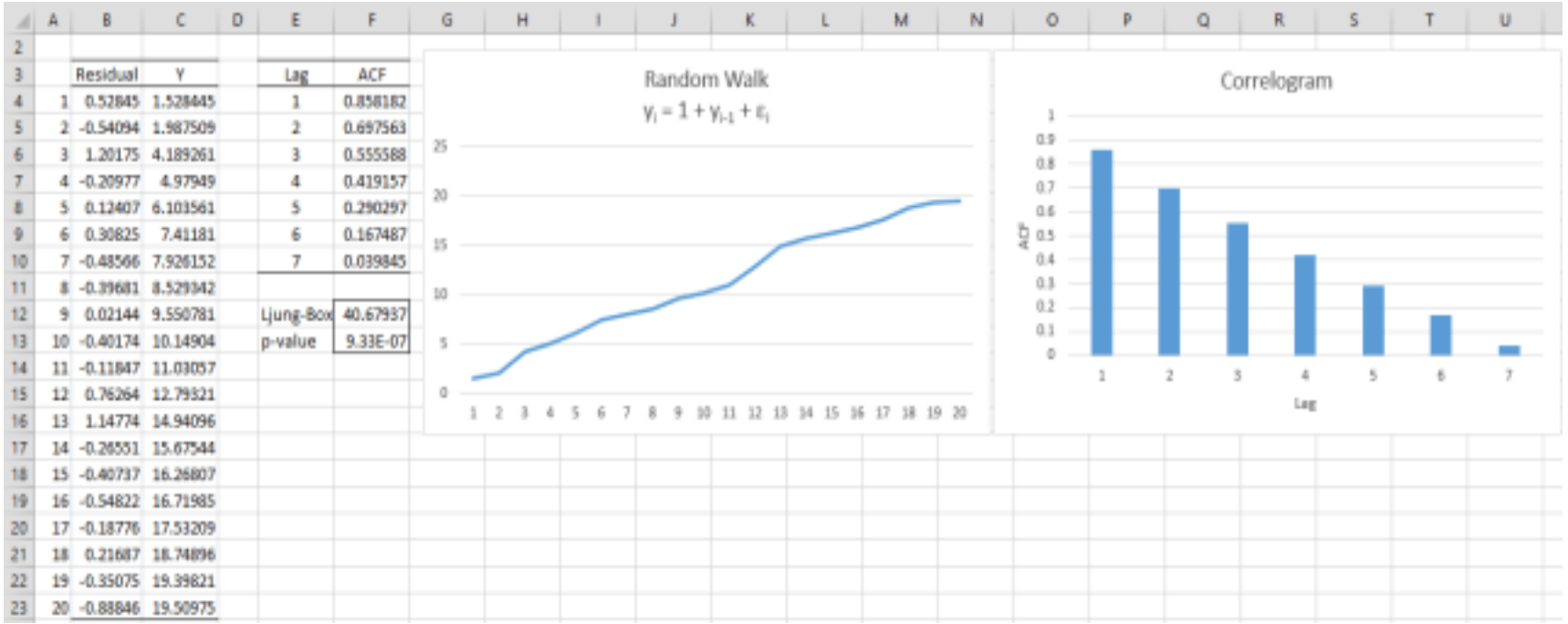
DATA ANALYTICS

Random Walk Model

Example 1: Graph the random walk with drift $y_i = y_{i-1} + \varepsilon_i$ where the $\varepsilon_i \sim N(0,.5)$.

The graph is shown in Figure 1. All the cells in column B contain the formula =NORM.INV(RAND(),0,.5), cell C4 contains the formula =1+B4 and cell C5 contains the formula =1+B5+C4.

As we can see, the graph shows a clear upward trend and the ACF shows a slow descent.



	Residual	Y	Lag	ACF
1	0.52845	1.528445	1	0.858182
2	-0.54094	1.987509	2	0.697563
3	1.20175	4.189261	3	0.555588
4	-0.20977	4.97949	4	0.419157
5	0.12407	6.103561	5	0.290297
6	0.30825	7.41181	6	0.167487
7	-0.48566	7.926152	7	0.039845
8	-0.39681	8.529342		
9	0.02144	9.550781	Ljung-Box	40.67937
10	-0.40174	10.14904	p-value	9.33E-07
11	-0.11847	11.03057		
12	0.76264	12.79321		
13	1.14774	14.94096		
14	-0.26551	15.67544		
15	-0.40737	16.26807		
16	-0.54822	16.71985		
17	-0.18776	17.53209		
18	0.21687	18.74896		
19	-0.35075	19.39821		
20	-0.88846	19.50975		

First differences are taken between the y values as shown in Figure 2. E.g. cell C5 contains the formula = B5-B4 (where column B replicates the values in column C from Figure 1). We see from the chart that the trend has been eliminated. We also see from the Ljung-Box test (cell F13) that the ACF values for the first 7 lags are statistically equal to zero, consistent with a purely random process.

https://www.real-statistics.com/time-series-analysis/stochastic-processes/random-walk/

What ARIMA stands for

- A series which needs to be differenced to be made stationary is an "integrated" (**I**) series
- Lags of the stationarized series are called "auto- regressive" (**AR**) terms
- Lags of the forecast errors are called "moving average" (**MA**) terms
- We've already studied these time series tools separately: differencing, moving averages, lagged values of the dependent variable in regression

ARIMA models put it all together

- Generalized random walk models: fine-tuned to eliminate all residual autocorrelation

- Generalized exponential smoothing models: that can incorporate long-term trends and seasonality

- Stationarized regression models: that use lags of the dependent variables and/or lags of the forecast errors as regressors.

- A general class of forecasting models for time series that can be stationarized by transformations such as differencing, logging, and or deflating.

ARIMA(*p, d, q*) Model Building

- The first step in ARIMA(*p, d, q*) is the model identification, that is, identifying the values of *p, d,* and *q*.
- Box and Jenkins (1970) proposed the following procedure to build the ARIMA(*p, d, q*) model.
- The main objective of model identification stage is to identify the right values of
- *p* (auto-regressive lags),
- *d* (order of differencing), and
- *q* (moving average lags).

ARIMA(*p, d, q*) Model Building

- The following flow chart can be used during the model identification stage
  1. The first step is to plot the ACF and PACF to identify whether the time series is stationary or not.
  2. If the time series is stationary then $d = 0$ and the model is ARIMA(p, 0, q) or ARMA(p, q) model.
  3. If the time series is non-stationary then it has to be converted into a stationary process by identifying the order of differencing.
  4. Once the value of d is known that will make the process stationary, then p and q are identified for the stationary process.

ARIMA(*p, d, q*) Model Building

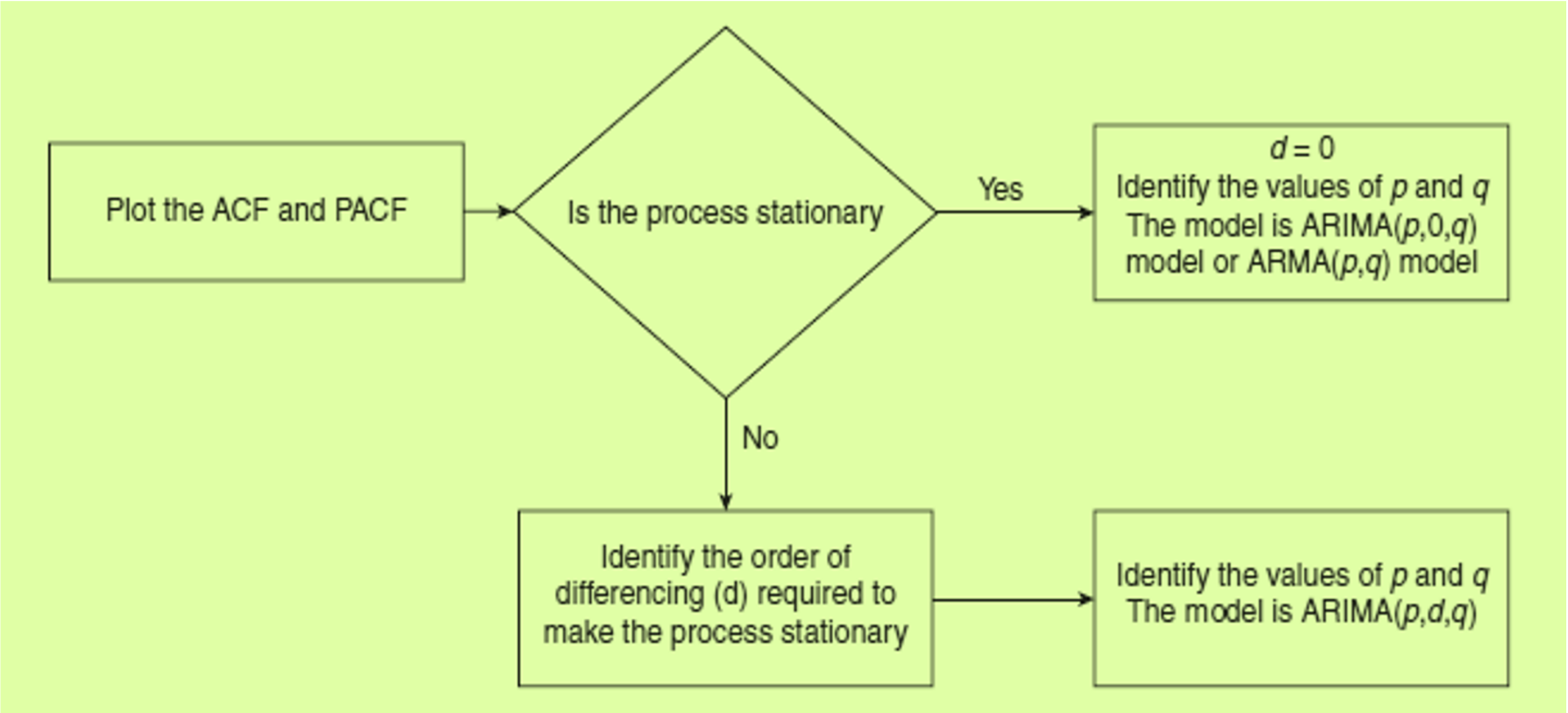


FIGURE 13.14 Model identification in ARIMA model.

Parameter Estimation and Model Selection

1. Once the model is identified (values of *p, d,* and *q*),
2. the next step in ARIMA model building is the parameter estimation.
3. That is, the estimation of coefficients in AR and MA components which are achieved using ordinary least squares.
4. The model selection may be carried using several criteria such as RMSE, MAPE, Akaike Information Criteria (AIC), or Bayesian Information Criteria (BIC).
5. AIC and BIC are measures of distance from the actual values to the forecasted values.

Parameter Estimation and Model Selection

- AIC is given by AIC = $-2LL + 2K$

  where LL is the log likelihood function and K is the number of parameters estimated (in this case $p + q$).

- BIC is given by BIC = $-2LL + K \ln(n)$

  In BIC equation, n is the number of observations in the sample. BIC assigns higher penalty compared to AIC for every additional variable added to the model.

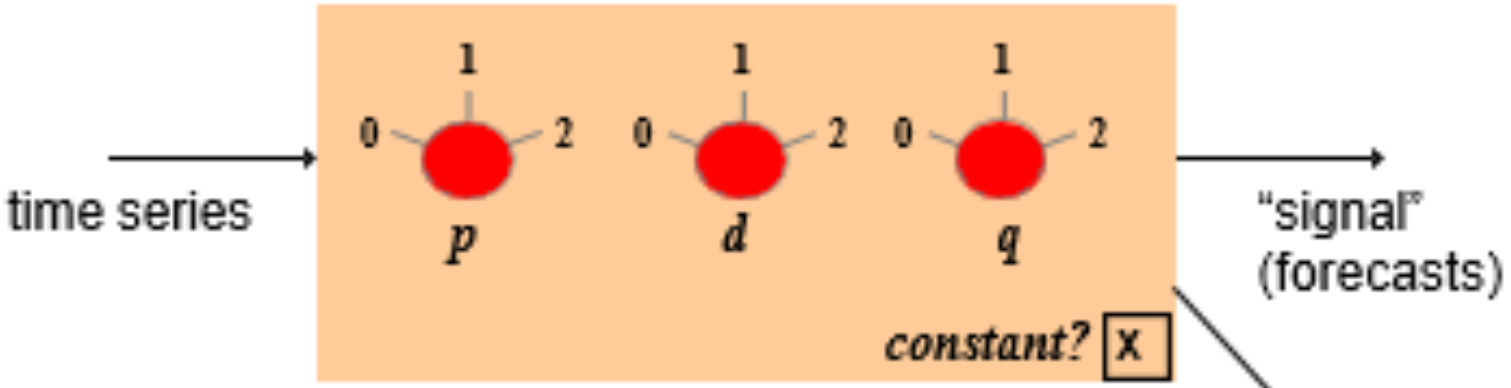
Lower values of AIC and BIC are preferred.

Model Validation

- ARIMA model is a regression model and thus has to satisfy all the assumptions of regression.

- The residual should be white noise. We can also perform a goodness of fit test using Ljung–Box test (coming up tomorrow!) before accepting the model.

The ARIMA "filtering box" – Extensions(to build projects in Time Series)



Objective: adjust the knobs until the residuals are "white noise" (uncorrelated)

ARIMA models we have already met

1. ARIMA(0,0,0)+c = mean (constant) model
2. ARIMA(0,1,0) = RW model
3. ARIMA(0,1,0)+c = RW with drift model
4. ARIMA(1,0,0)+c = regress Y on Y_LAG1
5. ARIMA(1,1,0)+c = regr. Y_DIFF1 on Y_DIFF1_LAG1
6. ARIMA(2,1,0)+c = ” ” plus Y_DIFF_LAG2 as well

ARIMA forecasting equation

- Let Y denote the original series
- Let y denote the differenced (stationarized) series

No difference (d=0): $y_t = Y_t$

First difference (d=1): $y_t = Y_t - Y_{t-1}$

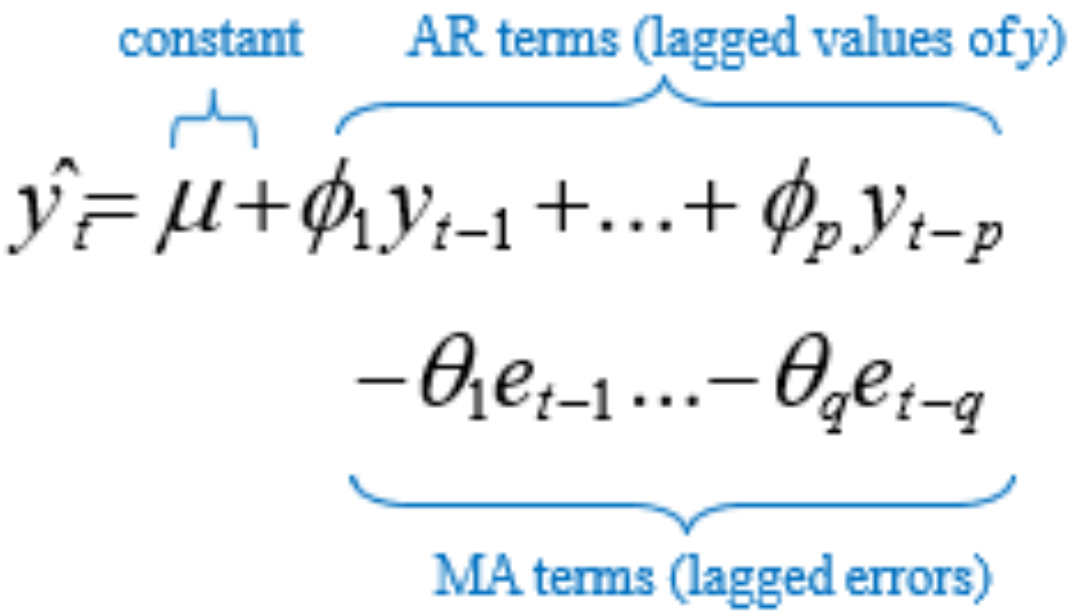
Second difference (d=2): $y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2})$

$= Y_t - 2Y_{t-1} + Y_{t-2}$

Note that the second difference is not just the change relative to two periods ago, i.e., it is not $Yt - Yt\text{-}2$. Rather, it is the change-in-the-change,

Forecasting equation for *y*

Not as bad as it looks! Usually $p+q <= 2$ and either $p=0$ or $q=0$ (pure AR or pure MA model)



By convention, the AR terms are + and the MA terms are -

Undifferencing the forecast

- The differencing (if any) must be reversed to obtain a forecast for the original series:

$$\text{If } d = 0: \quad \hat{Y}_t = \hat{y}_t$$

$$\text{If } d = 1: \quad \hat{Y}_t = \hat{y}_t + Y_{t-1}$$

$$\text{If } d = 2: \quad \hat{Y}_t = \hat{y}_t + 2Y_{t-1} - Y_{t-2}$$

- Fortunately, your software will do all of this automatically!

Do you need both AR and MA terms?

- In general, you don't: usually it suffices to use only one type or the other.
- Some series are better fitted by AR terms, others are better fitted by MA terms (at a given level of differencing).

- Rough rules of thumb:
  - If the stationarized series has positive autocorrelation at lag 1, AR terms often work best.
  - If it has negative autocorrelation at lag 1, MA terms often work best.
  - An MA(1) term often works well to fine-tune the effect of a nonseasonal difference, while an AR(1) term often works well to compensate for the lack of a nonseasonal difference, so the choice between them may depend on whether a difference has been used.

Interpretation of AR terms

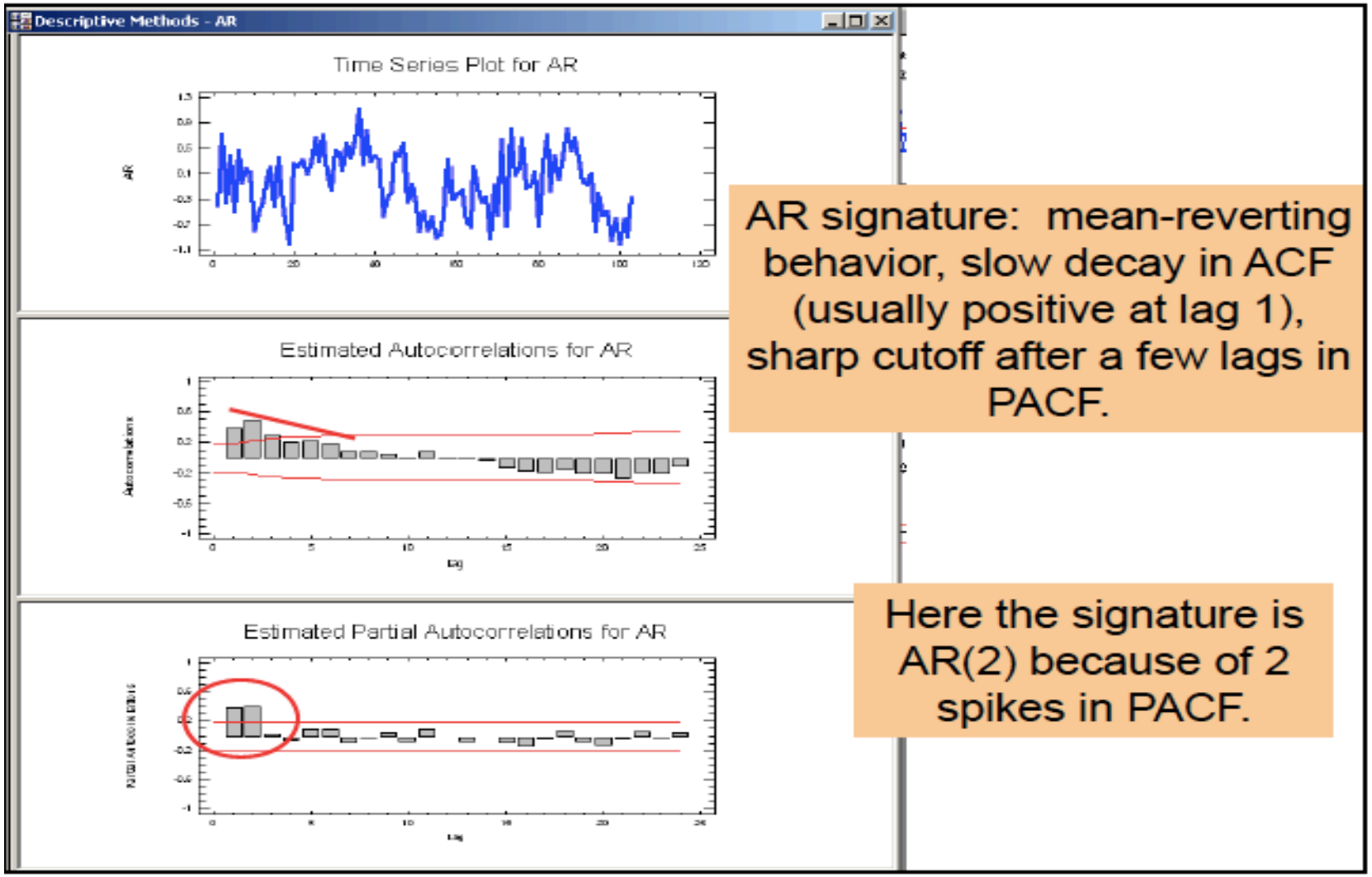
- A series displays autoregressive (AR) behavior if it apparently feels a “restoring force” that tends to pull it back toward its mean.
  - In an AR(1) model, the AR(1) coefficient determines how fast the series tends to return to its mean. If the coefficient is near zero, the series returns to its mean quickly; if the coefficient is near 1, the series returns to its mean slowly.
  - In a model with 2 or more AR coefficients, the sum of the coefficients determines the speed of mean reversion, and the series may also show an oscillatory pattern.

AR and MA "signatures"

- ACF that dies out gradually and PACF that cuts off sharply after a few lags => **AR signature**
  - An AR series is usually positively autocorrelated at lag 1 (or even borderline nonstationary)

- ACF that cuts off sharply after a few lags and PACF that dies out more gradually => **MA signature**
  - An MA series is usually negatively autcorrelated at lag 1 (or even mildly overdifferenced)

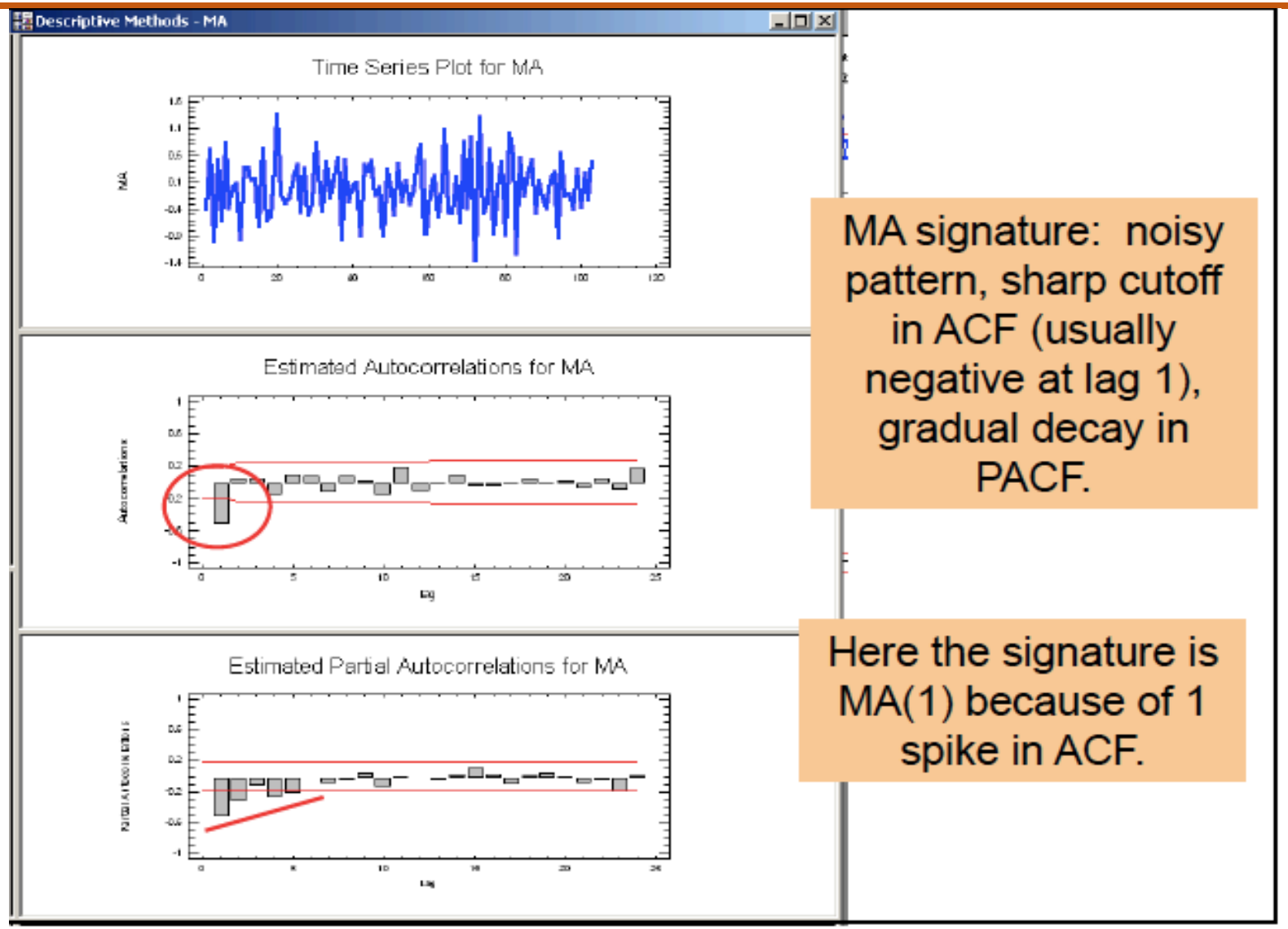
AR and MA "signatures"



AR signature: mean-reverting behavior, slow decay in ACF (usually positive at lag 1), sharp cutoff after a few lags in PACF.

Here the signature is AR(2) because of 2 spikes in PACF.

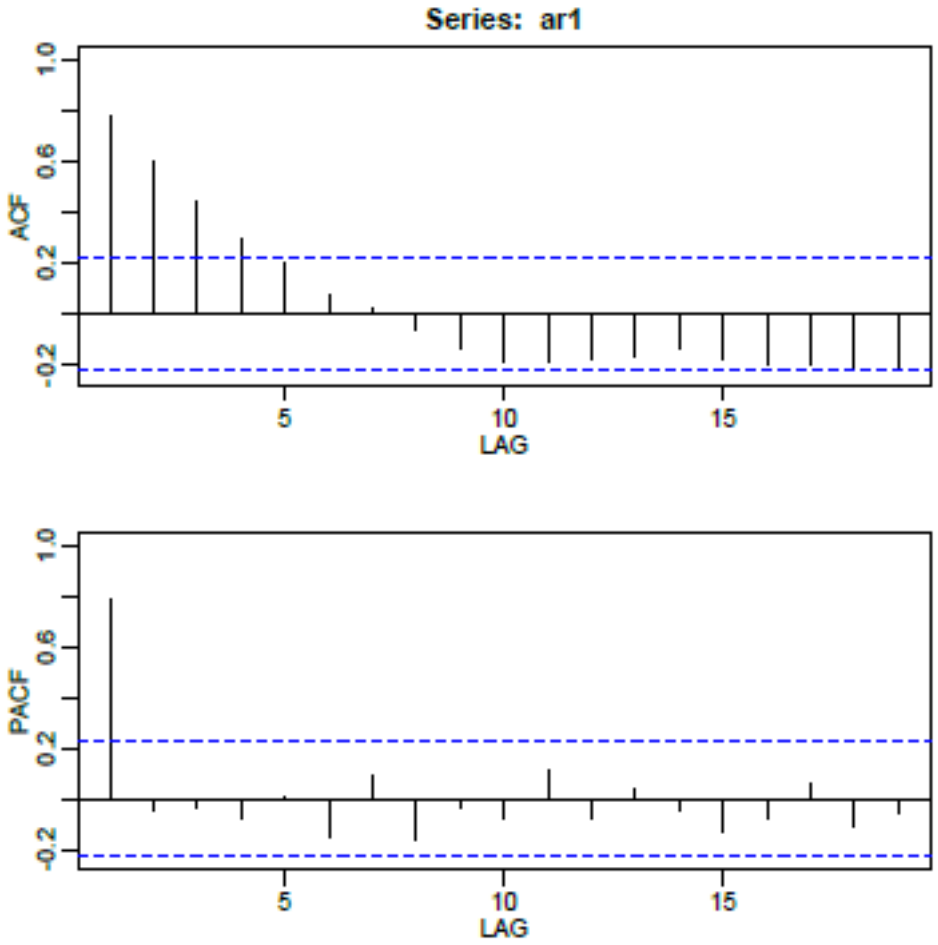
AR and MA "signatures"



MA signature: noisy pattern, sharp cutoff in ACF (usually negative at lag 1), gradual decay in PACF.

Here the signature is MA(1) because of 1 spike in ACF.

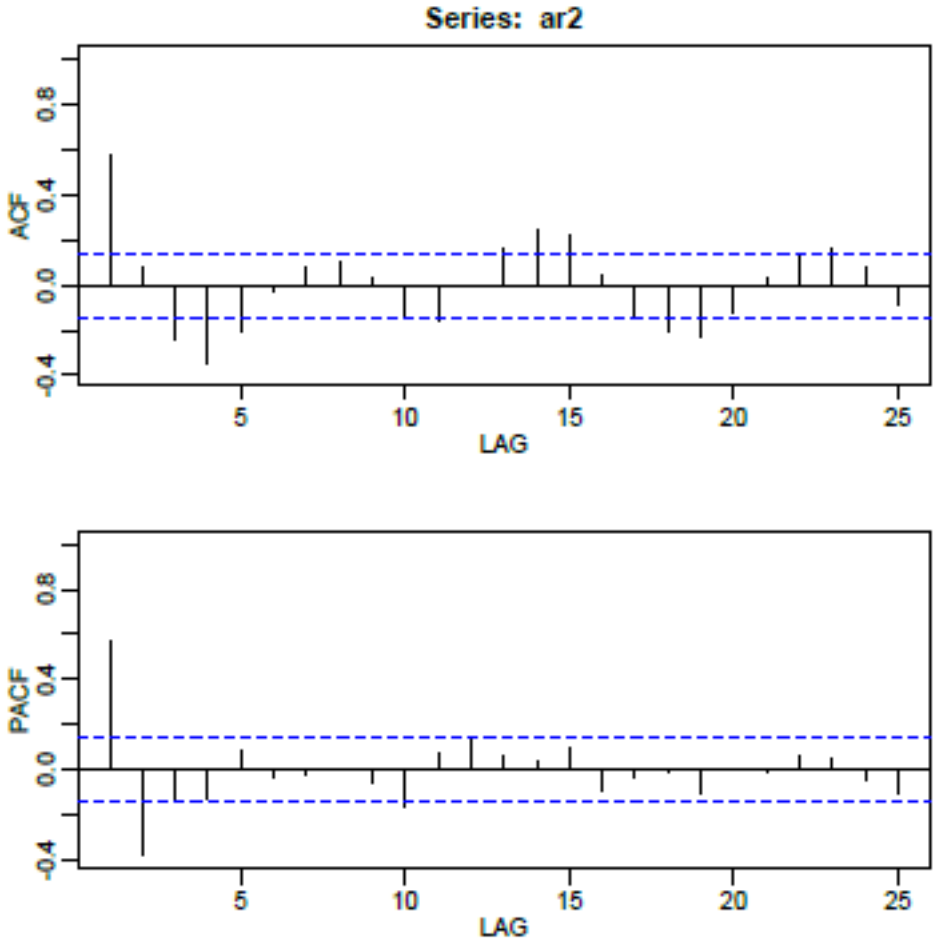
Identifying ARIMA models



The ACF and PACF should be considered together. It can sometimes be tricky going, but a few combined patterns do stand out.

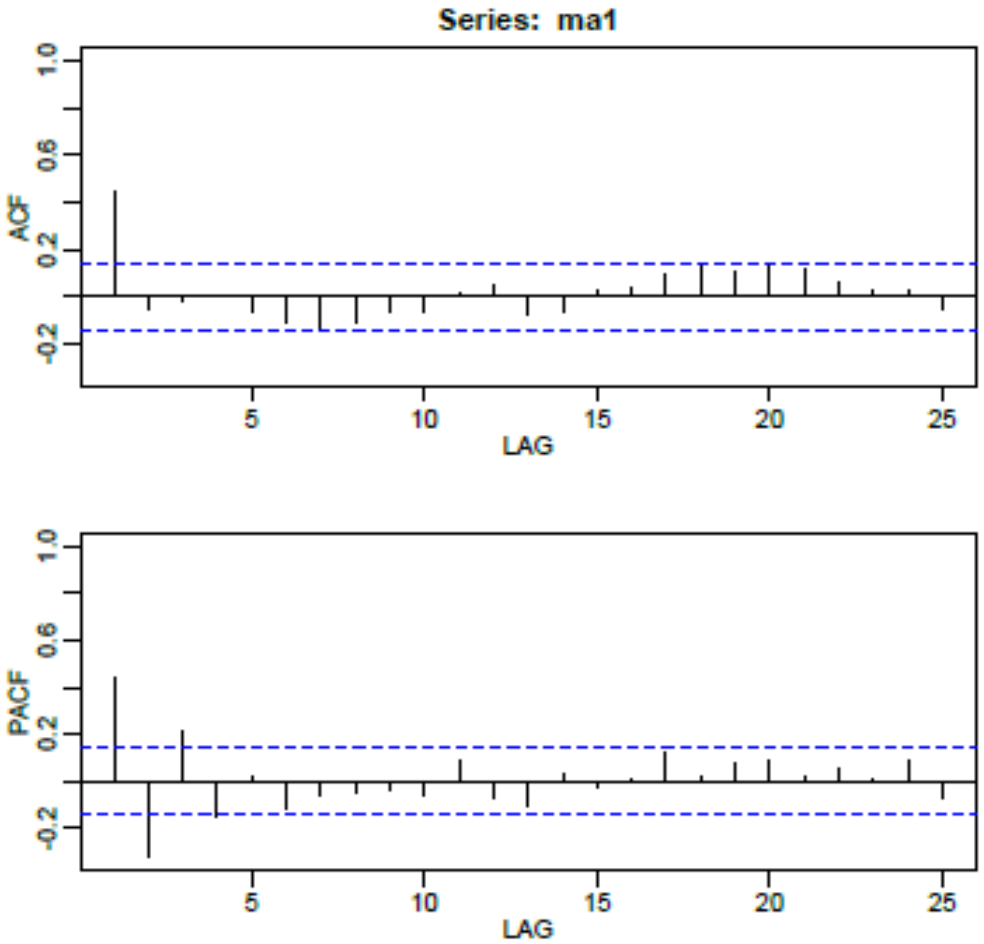
AR models have theoretical PACFs with non-zero values at the AR terms in the model and zero values elsewhere. The ACF will taper to zero in some fashion.

Identifying ARIMA models



An AR(2) has a sinusoidal ACF that converges to 0.

Identifying ARIMA models



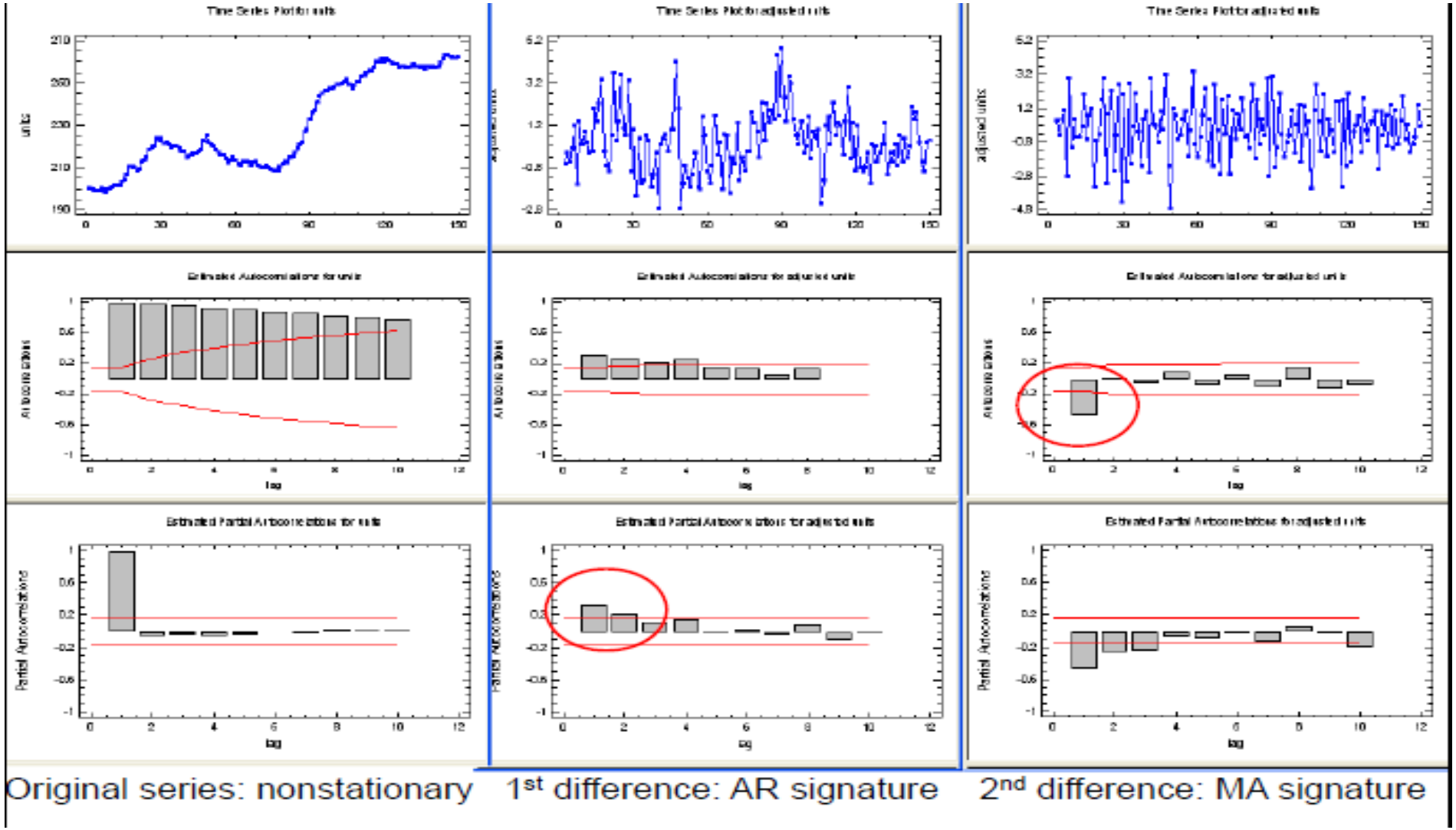
MA models have theoretical ACFs with non-zero values at the MA terms in the model and zero values elsewhere.

AR or MA? It depends!

- Whether a series displays AR or MA behavior often depends on the extent to which it has been differenced.
- An “underdifferenced” series has an AR signature (positive autocorrelation)
- After one or more orders of differencing, the autocorrelation will become more negative and an MA signature will emerge
- Don’t go too far: if series already has zero or negative autocorrelation at lag 1, don’t difference again

DATA ANALYTICS

An example



Model-fitting steps

1. Determine the order of differencing

2. Determine the numbers of AR & MA terms

3. Fit the model—check to see if residuals are "white noise," highest-order coefficients are significant (w/ no "unit "roots"), and forecasts look reasonable.
   If not, return to step 1 or 2.

References

Text Book:

"Business Analytics, The Science of Data-Driven Making", U. Dinesh Kumar, Wiley 2017 (Ch. 13.14 -13.14.4)

Image Courtesy

https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Time+Series+Analysis:+The+Basics

https://otexts.com/fpp2/stationarity.html

THANK YOU

Jyothi R

Assistant Professor, Department of Computer Science

jyothir@pes.edu