DATA ANALYTICS

Unit 3: Introduction to Time series data

Jyothi R.
Department of Computer Science and Engineering

INTRODUCTION TO FORECASTING

1. Moving Average
2. Single Exponential Smoothing (ES)
3. Double Exponential Smoothing – Holt's Method
4. Triple Exponential Smoothing (Holt-Winter Model)

Introduction to Forecasting

- These models are applicable to time series data with seasonal, trend, or both seasonal and trend component and stationary data

- Forecasting methods discussed in this chapter can be classified as:
  - Averaging methods
  - Equally weighted observations
  - Exponential Smoothing methods
  - Unequal set of weights to past data, where the weights decay exponentially from the most recent to the most distant data points

- All methods in this group require that certain parameters to be defined
  - These parameters (with values between 0 and 1) will determine the unequal weights to be applied to past data

Moving Average

- If a time series is generated by a constant process subject to random error, then mean is a useful statistic and can be used as a forecast for the next period
- Averaging methods are suitable for stationary time series data where the series is in equilibrium around a constant value (the underlying mean) with a constant variance over time

Mean: Uses the average of all the historical data as the forecast $F_{t+1} = \frac{1}{t}\sum_{i=1}^{t} y_i$

- When new data becomes available , the forecast for time t+2 is the new mean including the previously observed data plus this new observation

$$F_{t+2} = \frac{1}{t+1}\sum_{i=1}^{t+1} y_i$$

- This method is appropriate when there is no noticeable trend or seasonality

Averaging Methods

- The moving average for time period t is the mean of the “k” most recent observations
- The constant number k is specified at the outset
- The smaller the number k, the more weight is given to recent data points
- The greater the number k, the less weight is given to more recent data points

$$F_{t+1} = \hat{y}_{t+1} = \frac{(y_t + y_{t-1} + y_{t-2} + \cdots + y_{t-k+1})}{K}$$

$$F_{t+1} = \frac{1}{k} \sum_{i=t-k+1}^{t} y_i$$

- A large k is desirable when there are wide, infrequent fluctuations in the series

- A small k is most desirable when there are sudden shifts in the level of series

Moving Averages

- For quarterly data, a four-quarter moving average, MA(4), eliminates or averages out seasonal effects

- For monthly data, a 12-month moving average, MA(12), eliminate or averages out seasonal effect

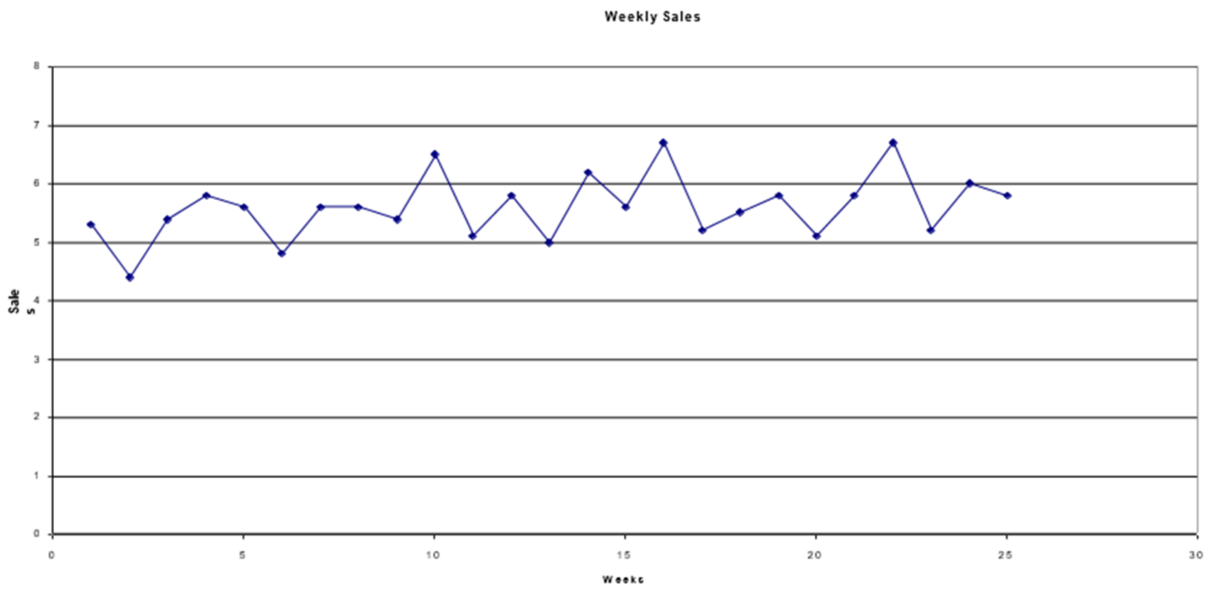
- Equal weights are assigned to each observation used in the average

- Each new data point is included in the average as it becomes available, and the oldest data point is discarded

- The moving average model does not handle trend or seasonality very well although it can do better than the total mean

Example: Weekly Department Store Sales

- The weekly sales figures (in millions of dollars) presented in the following table are used by a major department store to determine the need for temporary sales personnel



Week (t)	Sales (y)
1	5.3
2	4.4
3	5.4
4	5.8
5	5.6
6	4.8
7	5.6
8	5.6
9	5.4
10	6.5
11	5.1
12	5.8
13	5
14	6.2
15	5.6
16	6.7
17	5.2
18	5.5
19	5.8
20	5.1
21	5.8
22	6.7
23	5.2
24	6
25	5.8

Example: Weekly Department Store Sales

- Use a three-week moving average (k=3) for the department store sales to forecast for the week 24 and 26.

Week	Sales
21	5.8
22	6.7
23	5.2
24	6
25	5.8

$$\hat{y}_{24} = \frac{(y_{23}+y_{22}+y_{21})}{3} = \frac{5.2+6.7+5.8}{3} = 5.9$$

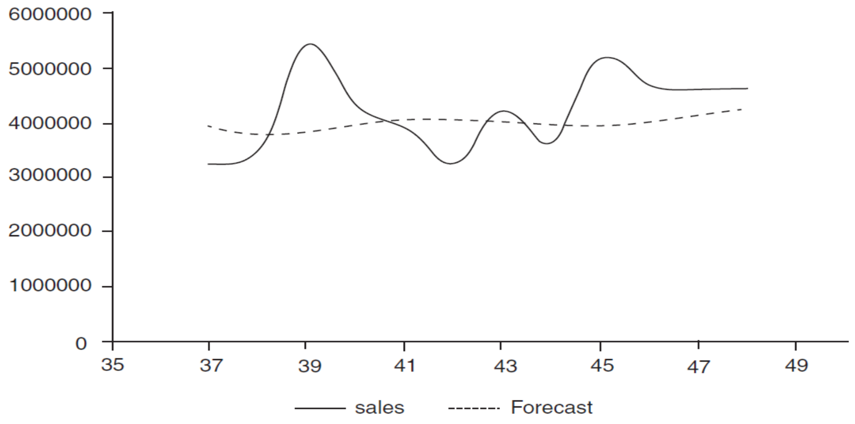


FIGURE 13.2 Plot of actual sales forecasted sales using moving average.

- The forecast error is
  - $e_{24} = y_{24} - \hat{y}_{24} = 6 - 5.9 = 0.1$

Weighted Average

$$F_{t+1} = \sum_{k=t+1-N}^{t} W_k \times Y_k$$

$$\sum_{k=t+1-N}^{t} W_k = 1$$

Week	Sales
21	5.8
22	6.7
23	5.2
24	6
25	5.8

Exponential smoothing methods

- The simplest exponential smoothing method is the single smoothing (SES) method where only one parameter needs to be estimated

- Holt’s method makes use of two different parameters and allows forecasting for series with trend

- Holt-Winters’ method involves three smoothing parameters to smooth the data, the trend, and the seasonal index

Simple Exponential Smoothing

- Formally, the exponential smoothing equation is

$$F_{t+1} = \alpha\, y_t + (1-\alpha) F_t$$

  - F_{t+1} = forecast for the next period.
  - α = smoothing constant.
  - y_t = observed value of series in period t.
  - F_t = old forecast for period t.
- The forecast F_{t+1} is based on weighting the most recent observation y_t with a weight α and weighting the most recent forecast F_t with a weight of $1-\alpha$

Why 'exponential'?

$$\begin{aligned} F_{t+1} &= \alpha\, y_t + (1-\alpha) F_t \\ &= \alpha\, y_t + (1-\alpha)[\alpha\, y_{t-1} + (1-\alpha) F_{t-1}] \\ &= \alpha\, y_t + \alpha(1-\alpha) y_{t-1} + (1-\alpha)^2 F_{t-1} \end{aligned}$$

$$F_{t+1} = \alpha\, y_t + \alpha(1-\alpha) y_{t-1} + \alpha(1-\alpha)^2 y_{t-2} + \alpha(1-\alpha)^3 y_{t-3} + \cdots + \alpha(1-\alpha)^{t-1} y_1$$

Influence of the exponential factor

Alpha in (0,1) and not equal to either 0 or 1

When is alpha small and when large?

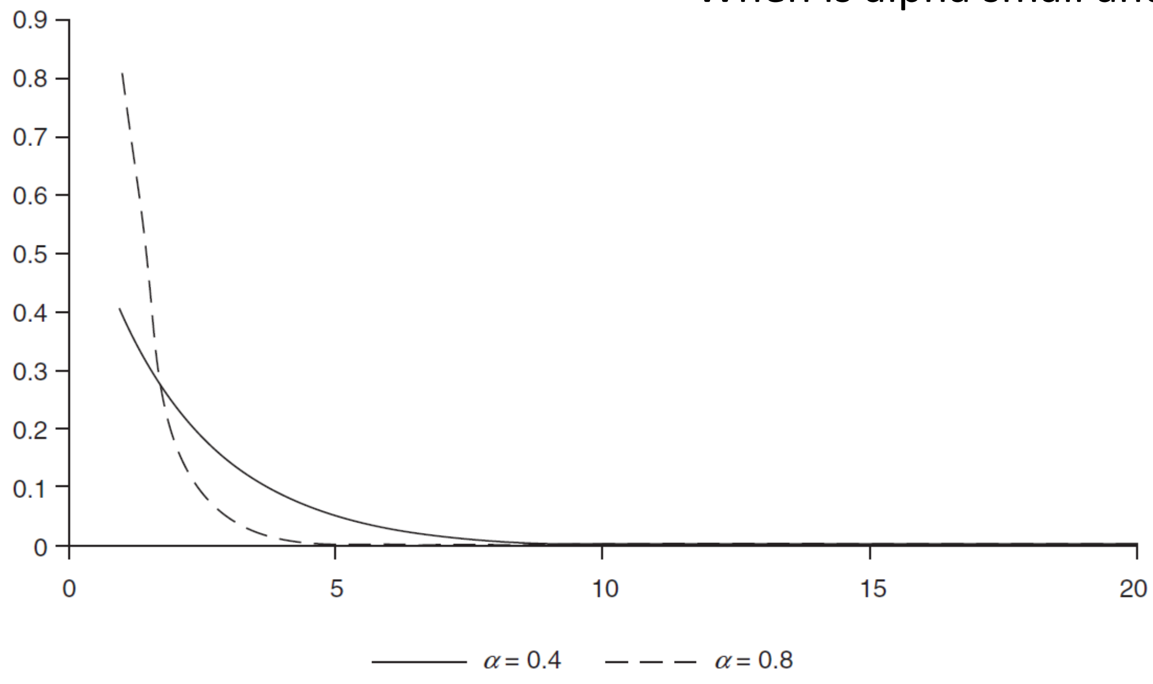


FIGURE 13.3 Exponential decay of weights to older observations.

Some pros and cons of Single Exponential Smoothing

Advantages:
1. It uses all the historic data unlike the moving average where only the past few observations are considered to predict the future value.

2. It assigns progressively decreasing weights to older data.

Some disadvantages of smoothing methods are:
1. Increasing *n* makes forecast less sensitive to changes in data.

2. It always lags behind trend as it is based on past observations. The longer the time period *n*, the greater the lag as it is slow to recognize the shifts in the level of the data points.

3. Forecast bias and systematic errors occur when the observations exhibit strong trend or seasonal patterns.

Holt's two parameter exponential smoothing

Holt's two parameter exponential smoothing method is an extension of simple exponential smoothing It adds a growth factor (or trend factor) to the smoothing equation as a way of adjusting for the trend.

Level (or Intercept) equation (L_t):

$$L_t = \alpha \times Y_t + (1 - \alpha) \times F_t \tag{13.12}$$

The trend equation is given by (T_t)

$$T_t = \beta \times (L_t - L_{t-1}) + (1 - \beta) \times T_{t-1} \tag{13.13}$$

α and β are the smoothing constants for level and trend, respectively, and $0 < \alpha < 1$ and $0 < \beta < 1$.

The forecast at time $t + 1$ is given by

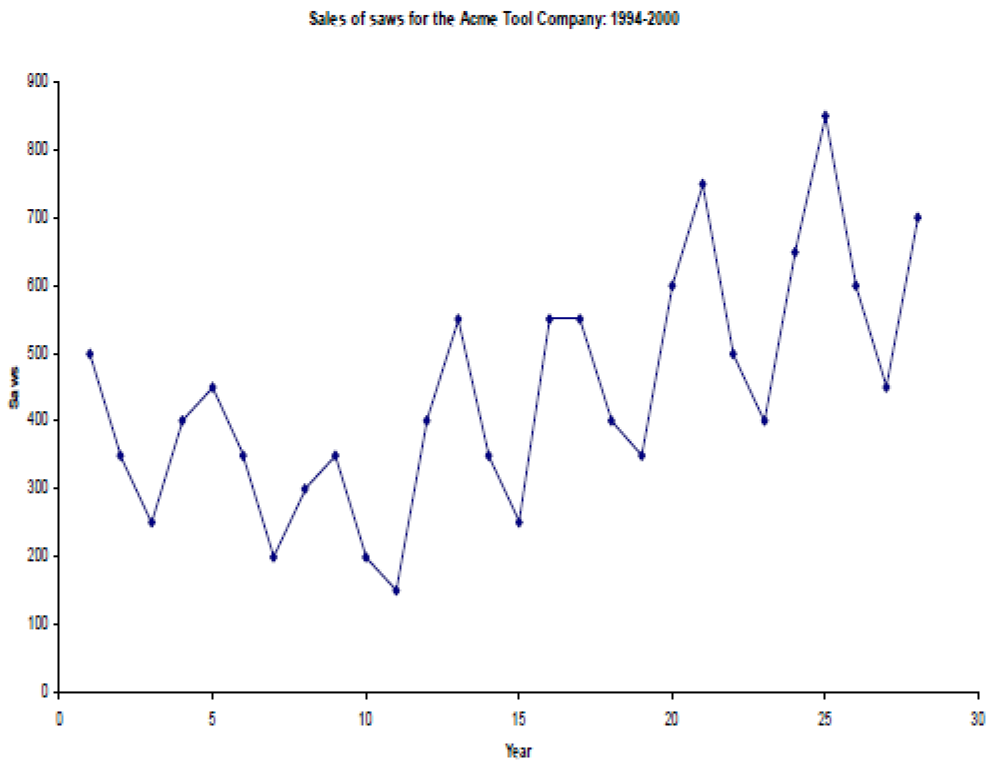
$$F_{t+1} = L_t + T_t \tag{13.14}$$

$$F_{t+n} = L_t + nT_t \tag{13.15}$$

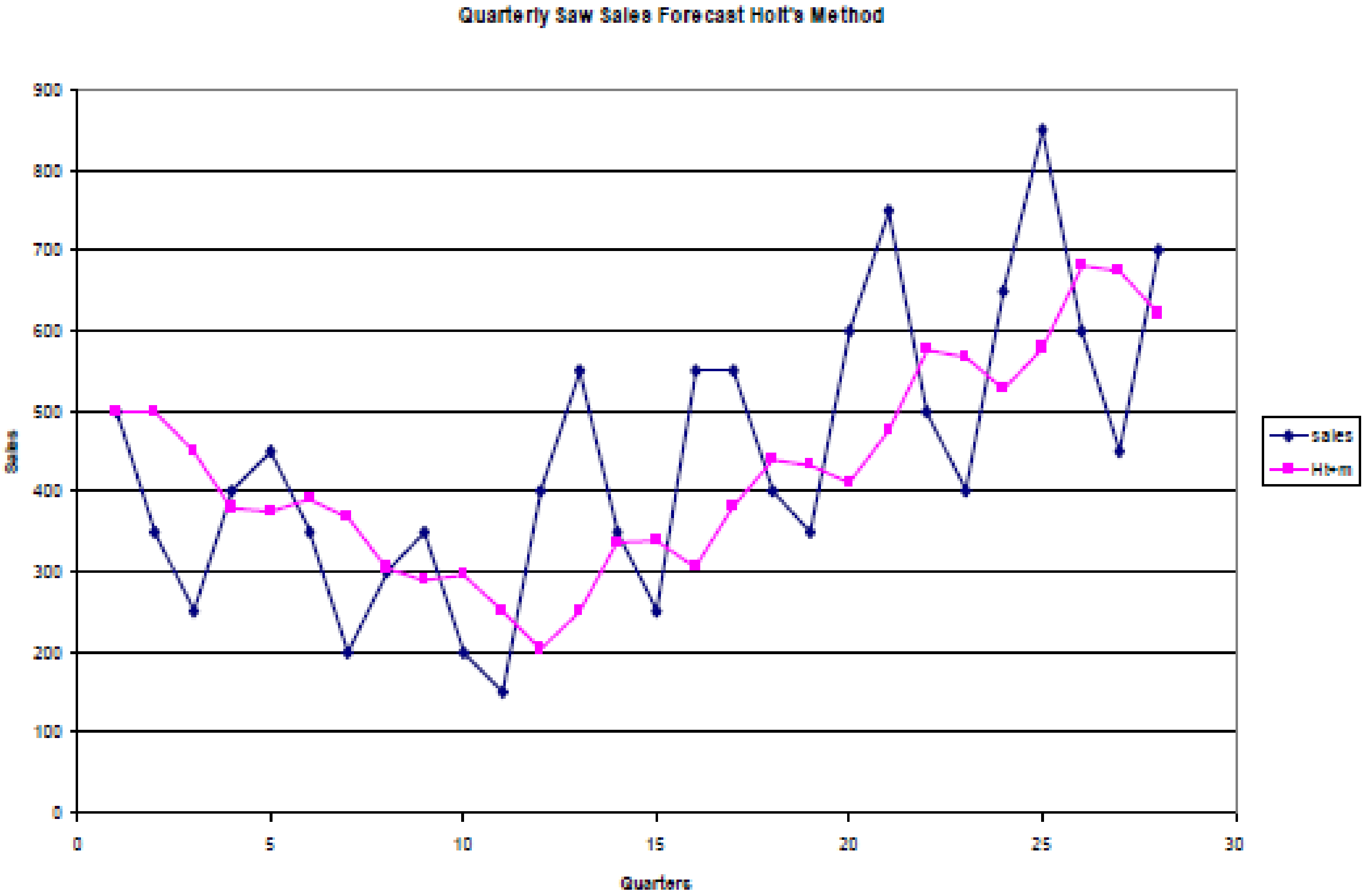
where L_t is the level which represents the smoothed value up to and including the last data, T_t is the slope of the line or the rate of increase or decrease at period t, n is the number of time periods into the future.

Initial value of L_t is usually taken same as Y_t (that is, $L_t = Y_t$). The starting value of T_t can be taken as $(Y_t - Y_{t-1})$ or the difference between two previous actual values of observations prior to the period for which forecasting is carried out. Another option for T_t is $(Y_t - Y_1)/(t - 1)$.

Holt’s exponential smoothing - example



Holt’s exponential smoothing - example



Alpha = 0.3
Beta = 0.1

Triple Exponential Smoothing (Holt Winter's Method)

Level (or Intercept) equation:

$$L_t = \alpha \frac{Y_t}{S_{t-c}} + (1-\alpha)\left[L_{t-1} + T_{t-1}\right]$$

Trend equation:

$$T_t = \beta \times (L_t - L_{t-1}) + (1 - \beta)\, T_{t-1}$$

Seasonal equation:

$$S_t = \gamma \frac{Y_t}{L_t} + (1-\gamma) S_{t-c}$$

The forecast F_{t+1} using triple exponential smoothing is given by

$$F_{t+1} = \left[L_t + T_t\right] \times S_{t+1-c}$$

Triple Exponential Smoothing (Holt Winter's Method)

Level (or Intercept) equation:

$$L_t = \alpha \frac{Y_t}{S_{t-c}} + (1-\alpha)\left[L_{t-1} + T_{t-1}\right]$$

Trend equation:

$$T_t = \beta \times (L_t - L_{t-1}) + (1 - \beta)\, T_{t-1}$$

Seasonal equation:

$$S_t = \gamma \frac{Y_t}{L_t} + (1-\gamma) S_{t-c}$$

The forecast F_{t+1} using triple exponential smoothing is given by

Note: this is a multiplicative model

$$F_{t+1} = \left[L_t + T_t\right] \times S_{t+1-c}$$

Initializations for Holt Winter's Method

$$L_t = Y_t$$

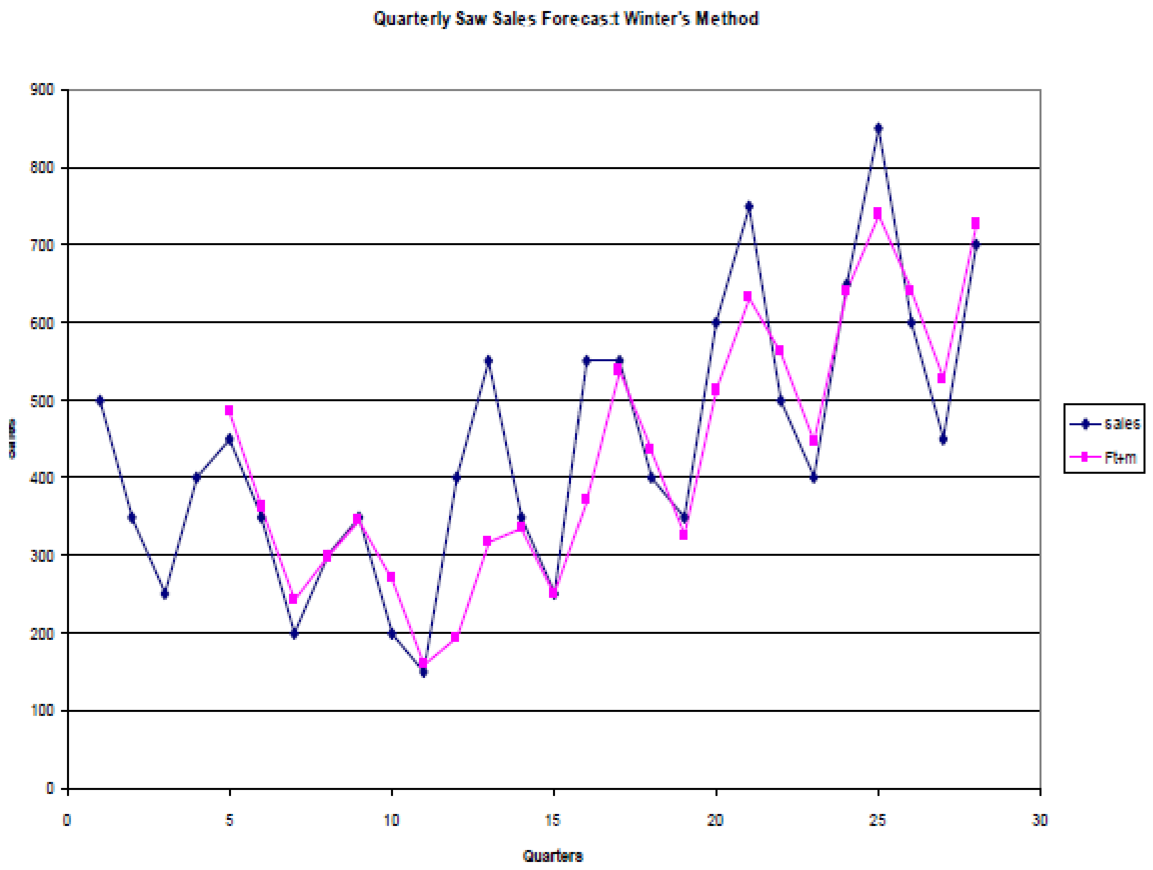
$$L_t = \frac{1}{c}\left(Y_1 + Y_2 + \cdots + Y_c\right)$$

$$T_t = \frac{1}{c}\left[\frac{Y_t - Y_{t-c}}{12} + \frac{Y_{t-1} - Y_{t-c-1}}{12} + \frac{Y_{t-2} - Y_{t-c-2}}{12} + \ldots + \frac{Y_{t-c+1} - Y_{t-2c+1}}{12}\right]$$

DATA ANALYTICS

Example of Holt Winter's Method

$\alpha = 0.4, \ \beta = 0.1, \ \gamma = 0.3$
and RMSE = 83.36



Class Project

- Choice of problem
- Choice of data
- Literature review
- Outcome

References

Text Book:

- "Business Analytics, The Science of Data-Driven Making", U. Dinesh Kumar, Wiley 2017 Chapter 13.4-13.7

Additional reference (for the interested reader)

- "Introduction to Time Series and Forecasting", Second Edition

Peter J. Brockwell, Richard A. Davis Springer 2002

Image Courtesy

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

THANK YOU

Jyothi R.
Assistant Professor,Department of Computer Science
jyothir@pes.edu

DATA ANALYTICS

Unit 3: Forecasting with Exponential, Croston's and Regression

Jyothi R.
Department of Computer Science and Engineering

TRIPLE EXPONENTIAL SMOOTHING (HOLT-WINTER MODEL)

- Moving averaging and single and double exponential smoothing techniques can handle data as long as the data do not have any seasonal component associated with it.

- When there is seasonality in the time-series data, techniques such as moving average, exponential smoothing, and double exponential smoothing are no longer appropriate.

- In most cases, the fitted error values (actual demand minus forecast) associated with simple exponential smoothing and Holt's method will indicate systematic error patterns that reflect the existence of seasonality.

- For example, presence of seasonality may result in all positive errors, except for negative values that occur at fixed intervals.

- Such pattern in error would imply existence of seasonality.

- Such time series data require the use of a seasonal method to eliminate the systematic patterns in error.

TRIPLE EXPONENTIAL SMOOTHING (HOLT-WINTER MODEL)

Multiplicative Model

Level (or Intercept) equation:

$$L_t = \alpha \frac{Y_t}{S_{t-c}} + (1-\alpha)\left[L_{t-1} + T_{t-1}\right]$$

Trend equation:

$$T_t = \beta \times (L_t - L_{t-1}) + (1 - \beta)\, T_{t-1}$$

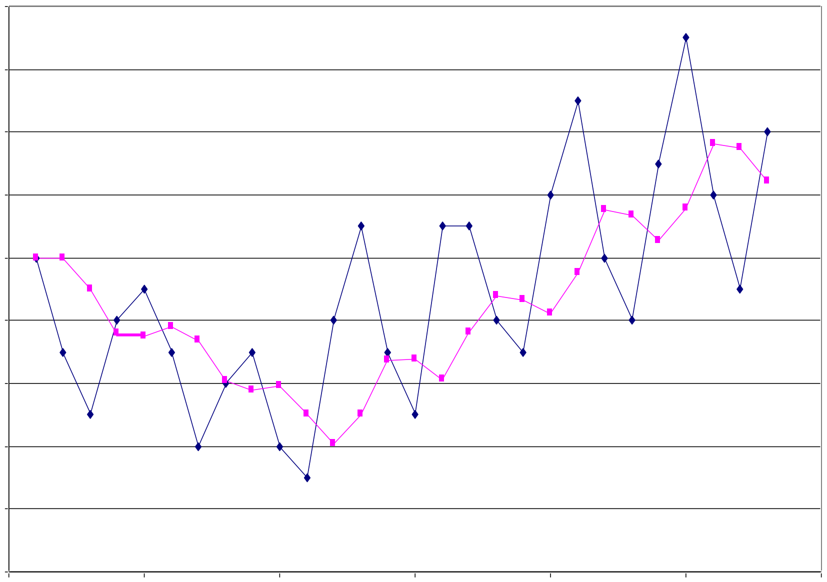
Seasonal equation:

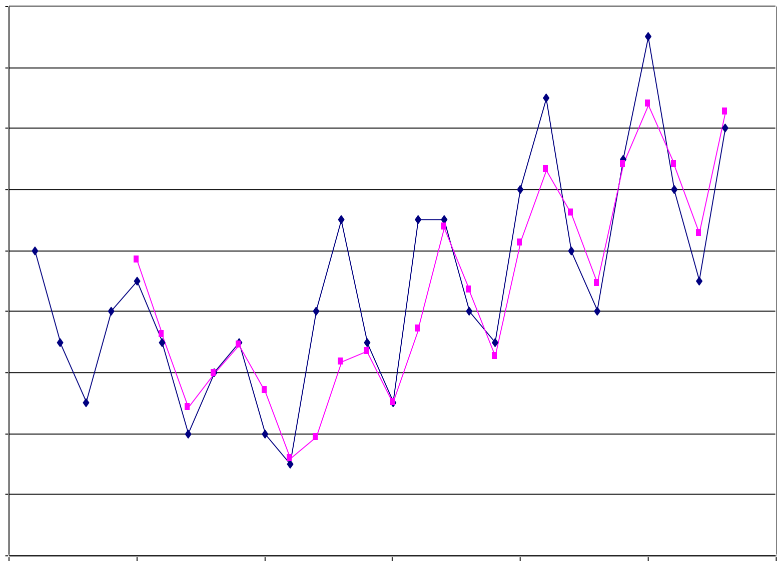
$$S_t = \gamma \frac{Y_t}{L_t} + (1-\gamma) S_{t-c}$$

The forecast F_{t+1} using triple exponential smoothing is given by

$$F_{t+1} = \left[L_t + T_t\right] \times S_{t+1-c}$$

Exponential Models





- RMSE for this application is: $\alpha = .3$ and $\beta = .1$

  RMSE = 155.5

- The plot also showed the possibility of seasonal variation that needs to be investigated.

- RMSE for this application is:

  $\alpha = 0.4$, $\beta = 0.1$, $\gamma = 0.3$ and

  RMSE = 83.36

- Note the decrease in RMSE.

Holt-Winter's Model

Additive Model

- The seasonal component in Holt-Winters' method.
- The basic equations for Holt's Winters' additive method are:

$$L_t = \alpha(y_t - S_{t-s}) + (1-\alpha)(L_{t-1} + T_{t-1})$$

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$$

$$S_t = \gamma(y_t - L_t) + (1-\gamma)S_{t-s}$$

$$F_{t+m} = L + mT_{t-1} + S_{t+m-s}$$

- The initial values for L_s and T_s are identical to those for the multiplicative method.
- To initialize the seasonal indices we use

$$S_1 = y_1 - L_s, \quad S_2 = y_2 - L_s, \cdots, S_s = Y_s - L_s$$

Computing the Seasonality Index

- Several techniques exist to calculate the initial seasonality index (Winters, 1961; Makridakis *et al.*, 1998;Taylor 2011).

- The initial seasonality index can be calculated using a technique called method of simple.

- Several variations to the procedure discussed in next section exist, such as ratio-to-moving average.

Predicting Seasonality Index Using Method of Averages

- The following steps are used for predicting the seasonality index using method of averages:
- **STEP 1**
- Calculate the average of value of *Y* for each season that is, if the data is monthly data, then we need to calculate the average for each month based on the training data.
- Let these averages be $\bar{Y}_{1,}$ $\bar{Y}_{2,}$ $\bar{Y}_{3......}$ $\bar{Y}_c$

- **STEP 2**
- Calculate the average of the seasons' averages calculated in step 1 (say $\bar{Y}$).

- **STEP 3**
- The seasonality index for season *k* is given by the ratio $\bar{Y}_k / \bar{Y}$.
- to the procedure explained above is first divide the value of Y_t with its yearly average and calculate the seasonal average
- We will use first 3 years of data to calculate the seasonality index for various months.

Predicting Seasonality Index Using Method of Averages

- The seasonality index based on first 3 years of data using method of averages.

- Seasonality index can be interpreted as percentage change from the trend line.
- For example, the seasonality index for January is approximately 1.088 or 108.8% (textbook example).

- This implies that in January, the demand will be approximately 8.8% more from the trend line. The seasonality index for March is 0.8885 or 88.85% (textbook example).

Predicting Seasonality Index Using Method of Averages

- TABLE 1: Seasonality index using method of averages

Month	Sale Quantity (2012)	Sale Quantity (2013)	Sale Quantity (2014)	Monthly Average $\bar{Y}_k$	Seasonality Index $\bar{Y}_k / \bar{\bar{Y}}$
January	3002666	4447581	4634047	4028098.00	1.087932
February	4401553	3675305	3772879	3949912.33	1.066815
March	3205279	3477156	3187110	3289848.33	0.888541
April	4245349	3720794	3093683	3686608.67	0.9957
May	3001940	3834086	4557363	3797796.33	1.02573
June	4377766	3888913	3816956	4027878.33	1.087872
July	2798343	3871342	4410887	3693524.00	0.997568
August	4303668	3679862	3694713	3892747.67	1.051375
September	2958185	3358242	3822669	3379698.67	0.912808
October	3623386	3361486	3689286	3558053.33	0.960979
November	3279115	3670362	3728654	3559377.00	0.961337
December	2843766	3123966	4732677	3566803.00	0.963342
Average of monthly averages				3702528.22	

Predicting Seasonality Index Using Method of Averages

- This implies that the demand in March will be 11.15% less from the trend line.
- Note that, multiplicative model is used in this example.
- To start the triple exponential smoothing, we need to set the starting values of level and trend.

  $L_{36} = Y_{36}/S_{36} = 4732677/0.9633 = 4912983.494$

- The initial value of trend (T_{36}) can be calculated based on second and third year by using

$$T_{36} = \frac{1}{12}\left[\frac{Y_{36} - Y_{24}}{12} + \frac{Y_{35} - Y_{23}}{12} + \frac{Y_{34} - Y_{22}}{12} + \cdots + \frac{Y_{25} - Y_{13}}{12}\right]$$

$$T_{36} = \frac{1}{12}\left[\frac{4732677 - 3123966}{12} + \frac{3728654 - 3670362}{12} + \ldots + \frac{4634047 - 4447581}{12}\right] = 21054.35$$

Predicting Seasonality Index Using Method of Averages

- The forecast for period 37 using triple exponential smoothing is given by

$$F_{37} = [L_{36} + T_{36}] \times S_{37-12} = [L_{36} + T_{36}] \times S_{25}$$

- The seasonal index S_{25} (seasonality index for January) is 1.088.
- Substituting the values of L_{36}, T_{36} and S_{25}, we get

$$F_{37} = [4912983.494 + 21054.35] \times 1.088 = 5368233.2$$

Predicting Seasonality Index Using Method of Averages

- TABLE 3: Forecasting using triple exponential smoothing (values differ for different round off values of parameters)

Month t	Actual Demand	L_{t-1}	T_{t-1}	S_t	F_t	$(Y_t - F_t)^2$	$\|Y_t - F_t\|/Y_t$
37	3216482	4912983.49	21054.35	1.09	5367895.97	4.62858E+12	0.668872
38	3453239	4301229.28	−295349.93	1.07	4273531.48	6.7288E+11	0.237543
39	5431651	3759825.78	−418376.71	0.89	2969014.38	6.06458E+12	0.453386
40	4241851	4228345.39	25071.45	1.00	4235127.90	45200134.5	0.001585
41	3909887	4255577.53	26151.79	1.03	4391900.21	2.32337E+11	0.123281
42	3216437	4131354.31	−49035.71	1.09	4441041.44	1.49966E+12	0.380733
43	4222004	3722098.55	−229145.74	1.00	3484457.63	5.43975E+11	0.174691
44	3621034	3729543.06	−110850.61	1.05	3804603.81	33697874118	0.050695
45	5162201	3562820.55	−138786.56	0.91	3125486.18	4.14821E+12	0.394544
46	4627176	4138038.05	218215.47	0.96	4186269.09	1.94399E+11	0.095286
47	4623945	4503072.73	291625.07	0.96	4609319.06	213918263.4	0.003163
48	4599368	4799566.34	294059.34	0.96	4906905.01	94579010434	0.066865

Predicting Seasonality Index Using Method of Averages

- The forecast for the period 37 to 48 for the data in Table 2 is given in Table 3 .
- Note that the values such as seasonality index are rounded to two decimals, the forecast values will be different if the actual seasonality index values are used.
- The RMSE and MAPE using triple exponential smoothing are 1228588.29 and 0.2208 (22.08%), respectively.
- The values of $a = 0.32$, $b = 0.5$, and $\gamma = 1$ are used for calculating the level, trend, and seasonal components.

- It is important to note that the exponential smoothing techniques are very sensitive to initial values of level, trend, and seasonal index

CROSTON'S FORECASTING METHOD FOR INTERMITTENT DEMAND

- Products such as spare parts may have intermittent demands.
- Exponential smoothing models discussed so far in the chapter will produce biased estimate when used for intermittent demand.
- Croston (1972) developed a model that uses two separate exponential smoothing equations for predicting mean time between demands and the magnitude of demand whenever the demand occurs.

- That is, Croston's method has two components:
  - (a) Predicting time between demand and
  - (b) magnitude of the demand.

- The primary objective of Croston's method is to forecast mean demand per period.

CROSTON'S FORECASTING METHOD FOR INTERMITTENT DEMAND

- Let
- Y_t = Demand at time t (Y_t may take value 0)
- F_t = Forecasted demand
- TD_t = Time between the latest and the previous non-zero demand in period t
- FTD_t = Forecasted time between demand at period t

The following steps are used for forecasting demand:

- If $Y_t = 0$ then $F_{t+1} = F_t$ and $FTD_{t+1} = FTD_t$ **Eqn 8**

Eqn 9 If $Y_t \neq 0$ the $F_{t+1} = \alpha \times Y_t + (1-\alpha)F_t$ and $FTD_{t+1} = \beta \times TD_t + (1-\beta) \times FTD_t$

CROSTON'S FORECASTING METHOD FOR INTERMITTENT DEMAND

- *α and β* are smoothing constants for forecasted demand and forecasted time between demands, respectively
- Once the forecasted demand and time between demands are known, then the mean demand per period D_{t+1}, is given by

Eqn 10: $$D_{t+1} = \frac{F_{t+1}}{FDT_{t+1}}$$

CROSTON’S FORECASTING METHOD FOR INTERMITTENT DEMAND

- Quarterly demand for spare parts of avionics system of an aircraft
- Use the demand during the quarters 1 to 4 as training data to forecast demand for periods 5 to 16 using Croston’s method.

CROSTON'S FORECASTING METHOD FOR INTERMITTENT DEMAND

- Quarterly demand for spare parts of avionics system of an aircraft is shown in Table 4:
- Use the demand during the quarters 1 to 4 as training data to forecast demand for periods 5 to 16 using Croston's method.

Quarter	1	2	3	4	5	6	7	8
Demand	20	12	0	18	16	0	20	22
Quarter	9	10	11	12	13	14	15	16
Demand	0	28	0	0	30	26	0	34

TABLE 13.8 Quarterly demand for avionic system spares

CROSTON'S FORECASTING METHOD FOR INTERMITTENT DEMAND

- Procedure used for starting values of F_t and FTD_t is shown in the table here:
- TD_4 = 2 since the elapsed time from the previous demand and current demand period is 2 (4 – 2).
- The forecasted time between demand is the average TD_t values up to t = 4.
- So, FTD_4 = (1+2)/2 = 1.5.
- The forecasted demand F_4 for t = 4 is (20 + 12 + 18)/3 = 16.67.
- Note that the total value is divided by 3 (not 4) since only 3 quarters had non-zero demand.
- So, the starting values for Croston's method are.

Quarter	Demand	TD_t	FTD_t	F_t
1	20			
2	12	1		
3	0			
4	18	2	1.5	16.67

$TD_4 = 2$, $FTD_4 = 1.5$, and $F_4 = 16.67$

Let $\alpha = \beta = 0.2$. Then

$$F_5 = 0.2 \times 18 + (1 - 0.2) * 16.67 = 16.936$$

$$FTD_5 = 0.2 \times 2 + (1 - 0.2) \times 1.5 = 1.6$$

CROSTON'S FORECASTING METHOD FOR INTERMITTENT DEMAND

- Forecasted demand for periods 5 to 16 using Croston's method.

Quarter	Demand	TD_t	FTD_t	F_t	$D_t = (F_t/FTD_t)$
1	20				
2	12	1			
3	0				
4	18	2	1.5000	16.67	11.11333
5	16	1	1.6000	16.936	10.585
6	0		1.4800	16.7488	11.31676
7	20	2	1.4800	16.7488	11.31676
8	22	1	1.5840	17.39904	10.98424
9	0		1.4672	18.31923	12.48585
10	28	2	1.4672	18.31923	12.48585
11	0		1.5738	20.25539	12.8707
12	0		1.5738	20.25539	12.8707
13	30	3	1.5738	20.25539	12.8707
14	26	1	1.8590	22.20431	11.94417
15	0		1.6872	22.96345	13.61034
16	34	2	1.6872	22.96345	13.61034

CROSTON'S FORECASTING METHOD FOR INTERMITTENT DEMAND

- **Example: lubricant sales**
- Several years ago, an oil company requested forecasts of monthly lubricant sales
- One of the time series is shown in the table below.
- The data contain small counts, with many months registering no sales at all, and only small numbers of items sold in other months.

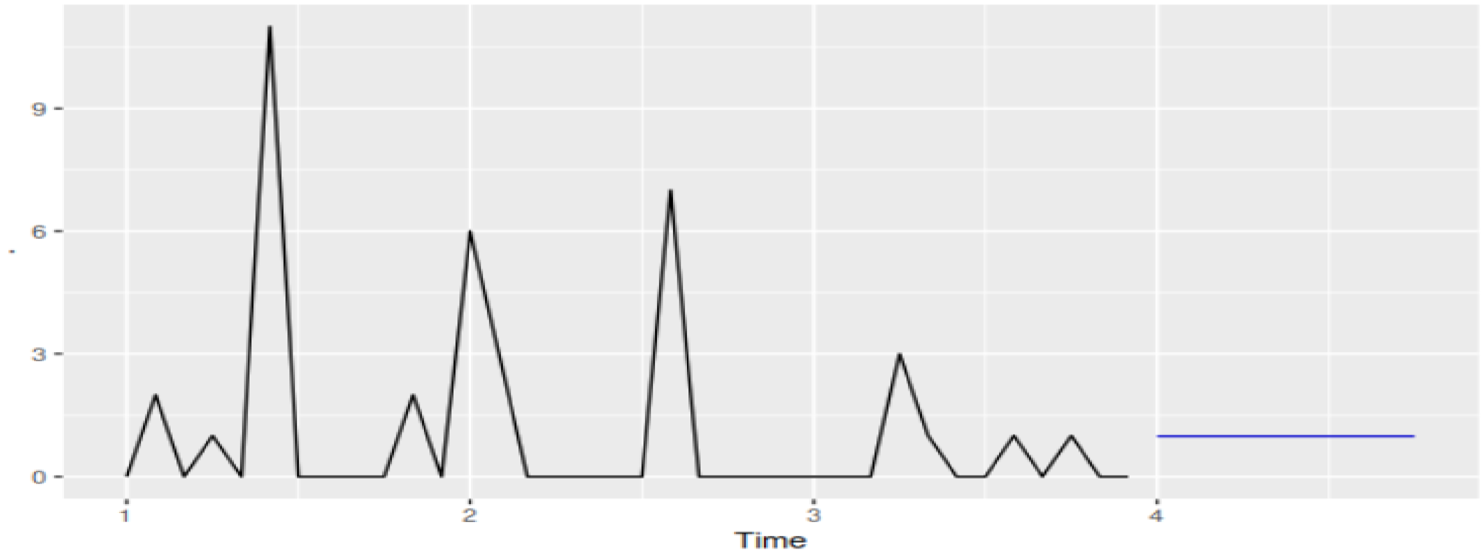
Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	0	2	0	1	0	11	0	0	0	0	2	0
2	6	3	0	0	0	0	0	7	0	0	0	0
3	0	0	0	3	1	0	0	1	0	1	0	0

- There are 11 non-zero demand values in the series, denoted by q.
- The corresponding arrival series a is also shown in the following table.

i	1	2	3	4	5	6	7	8	9	10	11
q	2	1	11	2	6	3	7	3	1	1	1
a	2	2	2	5	2	1	6	8	1	3	2

- Applying Croston's method gives the demand forecast 2.750 and the arrival forecast 2.793.
- So the forecast of the original series is $\hat{y}_{T+h|T} = 2.750/2.793 = 0.985$

CROSTON'S FORECASTING METHOD FOR INTERMITTENT DEMAND



- An implementation of Croston's method with more facilities (including parameter estimation) is available in the **tsinterminant package** for R.
- Forecasting models that deal more directly with the count nature of the data are described in Christou & Fokianos (2015).

Forecasting Stories 1: The Power of a Seasonality Index

- Read the second entry in a series on time series analysis and seasonality, and see how, through 2 simple use cases, the power of a seasonality index is uncovered.
- The strange fact was we had **performed poorly on all weeks**. Following are the weekly attainment figures: Week 1: 75% Week 2: 77% Week 3: 79% Week 4: 81%. What was amiss? What is the real-life forecasting story?

- **CASE 1 : Forecast over-indexed in April**

April Forecast and Actuals Comparison				
	FY19 Forecast	FY19 Actuals	FY18 Actuals	FY17 Actuals
April Average(Mn)	16.5	14.3	15.5	15.6
Year Average(Mn)	18.6	18.4	19.1	19.5
April Seasonality	89%	78%	81%	80%

Forecasting Stories 1: The Power of a Seasonality Index

- Here is what happened: As we can see from the image,
- **April forecast** seasonality was **over indexed by 11%**, i.e. at 89% of yearly average while actuals were trending towards 78%.
- What does a seasonality index mean?

Seasonality Index Calculation

	FY19 Forecast	FY19 Actuals	FY18 Actuals	FY17 Actuals
Week 1(Mn)	18.2	17.1	18.5	17.9
Week 2(Mn)	18.4	17.0	17.6	18.7
Week 3(Mn)	19.6	16.5	17.5	18.6
Year Average(Mn)	18.6	18.4	19.1	19.5

Week 1	98%	93%	97%	92%
Week 2	99%	=F73/F$75	92%	96%
Week 3	105%	90%	91%	95%

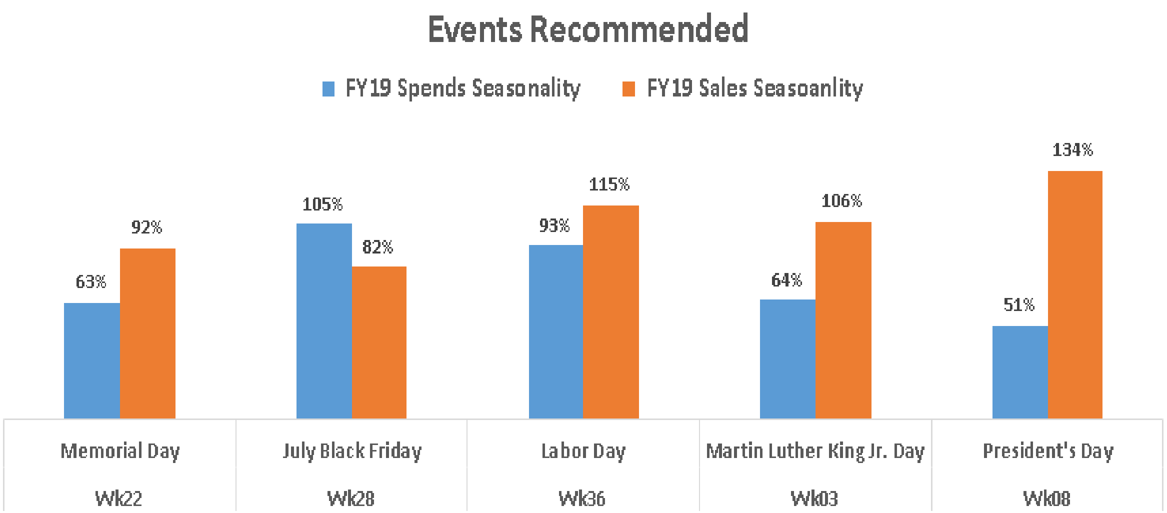
Forecasting Stories 1: The Power of a Seasonality Index

- It simply refers to all variables **normalized to a range close to 1**, so that all variables are comparable. As explained in the image, we divide each number by their yearly average to calculate the index. This way, the average of all values in the entire variable column is always 1.
- Hence interpreting the April seasonality, April being holiday is low performing month for this product.
- The forecast does partially take this into account, with 89% target compared to average of the year.
- However, actual performance across years can be seen ~80%.
- Hence the targets or f**orecast need to be even lower to be realistic**.

- DATA ANALYTICS

Forecasting Stories 2: The Power of a Seasonality Index

- Have a look at the spends and sales seasonality indices:



- What conclusions can you draw?
  - The first observation was that we are spending a much higher proportion of marketing budget on July Black Friday, whereas President's day week results in higher sales.
  - If we **reallocate the spends** from July Black Friday day to President's week, we would end up with a higher ROI without adding a single penny to the budget.

• Forecasting Stories 2: The Power of a Seasonality Index

Forecast over-indexed in April

- For President's day, the sales seasonality index is 134% while your marketing spends index is 51%!
- And hence we should **reallocate the spends** from July Black Friday day to President's week.

Comparing 2 seasonality indexes can give some power-packed insights

- Numbers after all, are good or bad only when they are compared against another.
- The rest of the story is based on events, specially holidays in the US Calendar.
- Not all of the following are holidays, and different events impact sales in different ways.
- Also, some events are more important than others.

Regression for Forecasting

- Parker and Segura (1971) claimed regression can predict more accurately than exponential smoothing
- Regression is particularly useful when there is one or more explanatory variable in addition to the dependent variable Y_t

The forecast value at time *t* can be written as

$$F_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_n X_{nt} + \varepsilon_t$$

Here F_t is the forecasted value of Y_t, and X_{1t}, X_{2t}, etc. are the predictor variables measured at time t.

References

Text Book:

“Business Analytics, The Science of Data-Driven Making”, U. Dinesh Kumar, Wiley 2017 Chapter-13.7-13.9

Image and Case Study

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

DATA ANALYTICS

Unit 3: Forecasting with Regression, Stationary Signals and ARMA

Jyothi R., Gowri Srinivasa
Department of Computer Science and Engineering

Regression for Forecasting

- Parker and Segura (1971) claimed regression can predict more accurately than exponential smoothing
- Regression is particularly useful when there is one or more explanatory variable in addition to the dependent variable Y_t

The forecast value at time *t* can be written as

$$F_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_n X_{nt} + \varepsilon_t$$

Here F_t is the forecasted value of Y_t, and X_{1t}, X_{2t}, etc. are the predictor variables measured at time *t*.

Forecasting with Regression – An Example

$$F_t = \beta_0 + \beta_1 \text{ promotion expenses at time } t + \beta_2 \text{ competition promotion at time } t$$

Model	*R*	*R*-Square	Adjusted *R*-Square	Std. Error of the Estimate	Durbin–Watson
1	0.928	0.862	0.853	207017.359	1.608

Note

- We need a high R2 value for forecasting applications
- Durbin-Watson Statistic D = 1.608

  Recall: D=2 $\Rightarrow$ autocorrelation; 1.608 $\Rightarrow$ no autocorrelation among the errors
- The presence of autocorrelation may lead to the inclusion of nonsignificant variables in the equation (since the standard error of the regression coefficient is underestimated when autocorrelation errors are present)

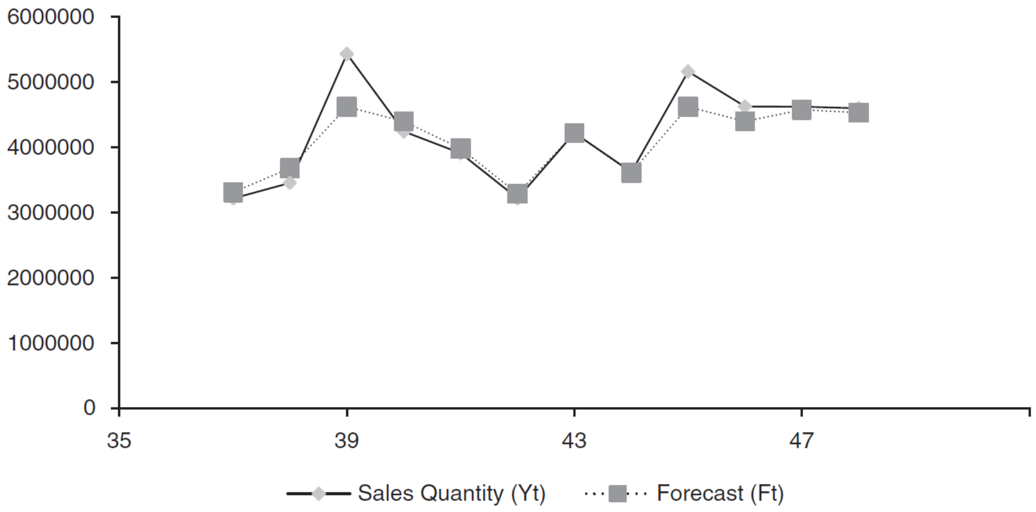
Model		Unstandardized Coefficients		Standardized Coefficients	*t*	Sig.
		B	Std. Error	Beta		
1	(Constant)	808471.843	278944.970		2.898	0.007
	Promotion Expenses	22432.941	1953.674	0.825	11.482	0.000
	Competition Promotion	−212646.036	77012.289	−0.198	−2.761	0.009

Forecasting with Regression – An Example

$$F_t = 808471.843 + 22432.941 X_{1t} - 212646.036 X_{2t}$$

$$X_{1t} = \text{Promotion expenses at time } t$$

$$X_{2t} = \begin{cases} 1 & \text{Competition is on promotion} \\ 0 & \text{Otherwise} \end{cases}$$



- Sales increases when promotions expenses increase and the sales decrease when the competition is on the promotion.

Method	MAPE	RMSE
Moving Average	734725.84	14.03%
Exponential Smoothing	742339.22	13.94%
Regression	302969	**4.19%**

Forecasting with Regression – Seasonality

STEP 1

Estimate the seasonality index (using techniques such as method of averages or ratio to moving average).

STEP 2

De-seasonalize the data using either additive or multiplicative model. For example, in multiplicative model, the de-seasonalized data $Y_{d,t} = Y_t / S_t$, where $Y_{d,t}$ is the de-seasonalized data and S_t is the seasonality index for period t.

STEP 3

Develop a forecasting model on the de-seasonalized data ($F_{d,t}$).

STEP 4

The forecast for period $t + 1$ is $F_{t+1} = F_{d,\,t+1} \times S_{t+1}$.

Autoregressive Models

Auto-regression simply means regression of a variable on itself measured at different time periods. One of the fundamental assumptions of AR model is that the time series is assumed to be a stationary process.

If a time-series data, Y_t, is stationary, then it satisfies the following conditions:
1. The mean values of Y_t at different values of t are constant.
2. The variances of Y_t at different time periods are constant (Homoscedasticity).
3. The covariances of Y_t and Y_{t-k} for different lags depend only on k and not on time t

When the time series data is not stationary (that is, any one of the above conditions are not satisfied), then we have to convert the non-stationary times-series data to stationary data before applying AR models

Another important concept associated with forecasting based on regression-based models is the white noise of residuals. White noise is a process of residuals that are uncorrelated and follow normal distribution with mean 0 and constant standard deviation. In AR models, one of the important assumptions that we make is that the errors follow a white noise.

Stationarity

- A strictly stationary process is one where the distribution of its values remains the same as time proceeds, implying that the probability lies in a particular interval is the same now as at any point in the past or the future.

- However we tend to use the criteria relating to a 'weakly stationary process' to determine if a series is stationary or not.

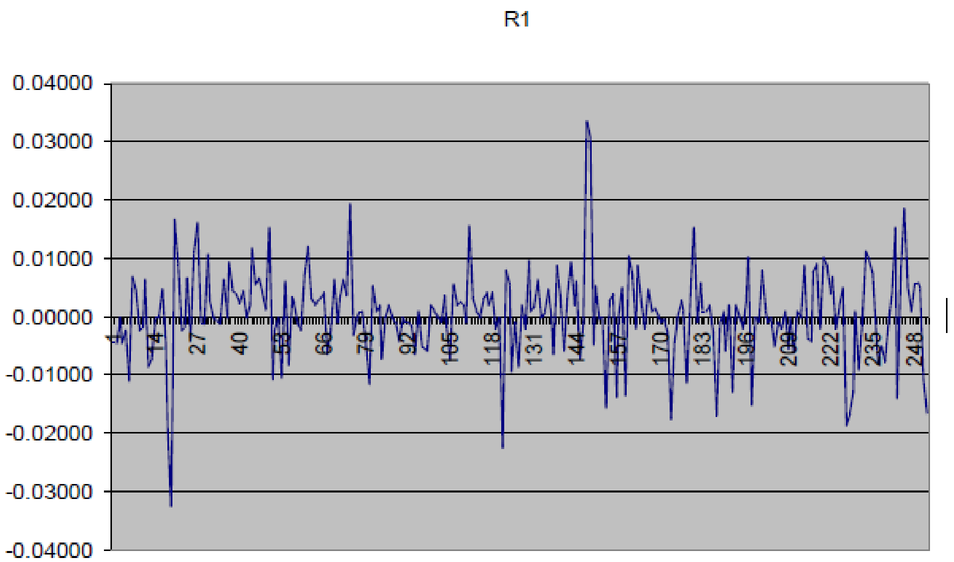
Weakly Stationary Series

- A stationary process or series has the following properties:
  - constant mean
  - constant variance
  - constant auto covariance structure

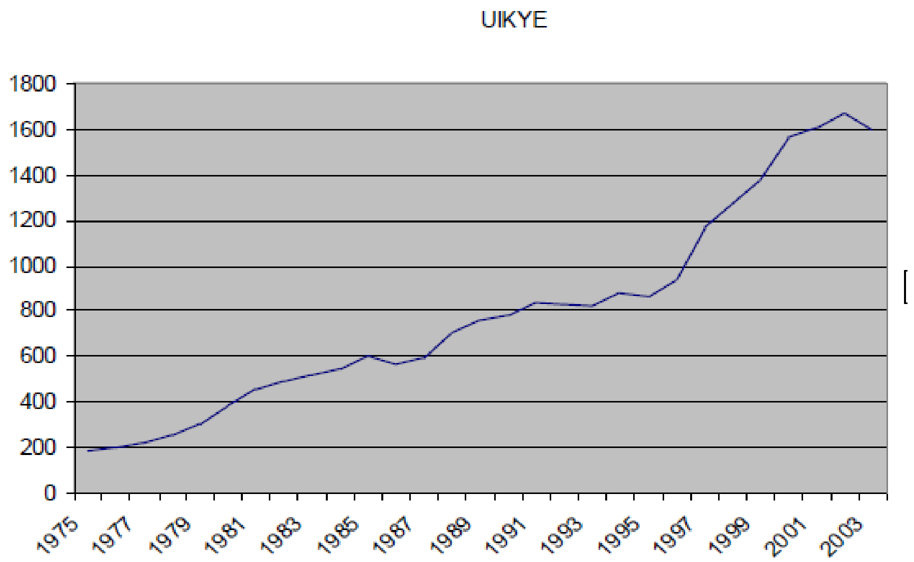
- $E(y_t) = \mu$
- $E(y_t - \mu)^2 = \sigma^2$
- $E(y_{t1} - \mu)(y_{t2} - \mu) = \gamma_{t2-t1}, \forall t_1, t_2$

- The latter refers to the covariance between y(t-1) and y(t-2) being the same as y(t-5) and y(t-6).

Stationary and NonStationary Series



Stationary Series



Non-stationary Series

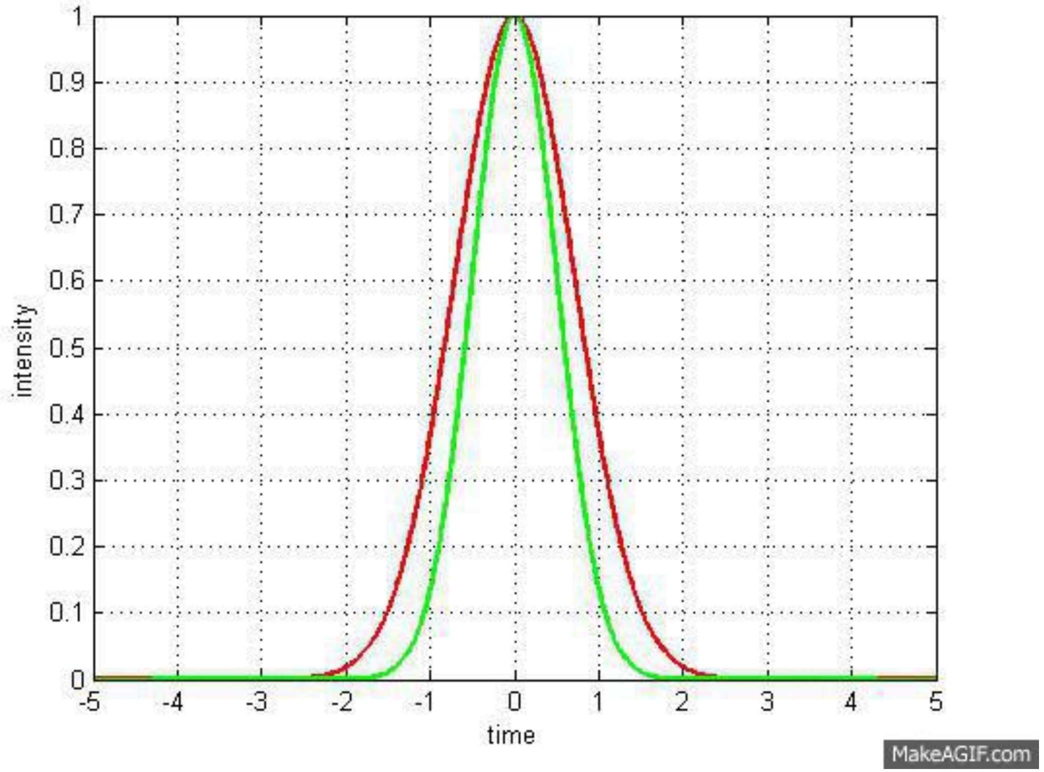
Implications of Nonstationary Data

- If the variables in an OLS regression are not stationary, they tend to produce regressions with high R-squared statistics and low Durbin-Watson statistics, indicating high levels of autocorrelation.
- This is caused by the drift in the variables often being related, but not directly accounted for in the regression, hence the omitted variable effect.
- It is important to determine if our data is stationary before the regression.
- This can be done in a number of ways:
  - plotting the data
  - assessing the autocorrelation function
  - Using a specific test on the significance of the autocorrelation coefficients.
  - Specific tests such as DF, ADF, etc. (to be covered later)

Autocorrelation Function (ACF) at lag *k*

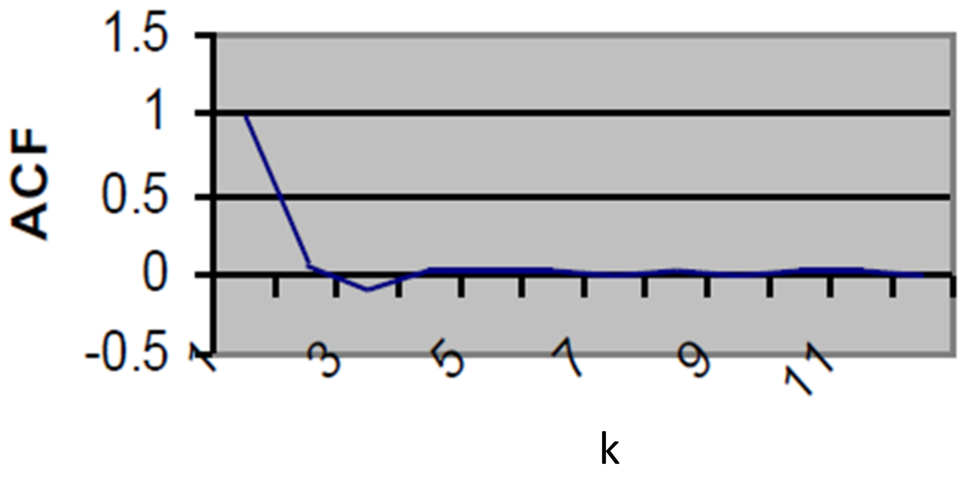
$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\text{covariance at lag k}}{\text{variance}}$$

$$\rho_k = \frac{\sum_{t=k+1}^{n}\left(Y_{t-k} - \bar{Y}\right)\left(Y_t - \bar{Y}\right)}{\sum_{t=1}^{n}\left(Y_t - \bar{Y}\right)^2}$$



Correlogram

- The sample Correlogram is the plot of the ACF against k.
- As the ACF lies between -1 and +1, the Correlogram also lies between these values.



- It can be used to determine stationarity, if the ACF falls immediately from 1 to 0, then equals about 0 thereafter, the series is stationary.
- If the ACF declines gradually from 1 to 0 over a prolonged period of time, then it is not stationary.

Statistical Significance of the ACF

- The Q statistic can be used to determine if the sample ACFs are jointly equal to zero.

$$Q = n \sum_{k=1}^{m} \hat{\rho}_k^2$$

- n -> sample size
- m -> lag length
- $\chi^2(m)$ ->degrees of freedom

- If jointly equal to zero we can conclude that the series is stationary.
- It follows the chi-squared distribution, where the null hypothesis is that the sample ACFs jointly equal zero.

Q-statistic Example

- The following information, from a specific variable can be used to determine if a time series is stationary or not.

$$\sum_{k=1}^{4} \hat{\rho}_k^2 = 0.32$$

$$n = 60$$

$$Q = 60 * 0.32 = 19.2$$

$$\chi^2(4) = 9.488$$

$$19.2 > 9.488 \rightarrow reject - H_0$$

- The series is not stationary as the ACFs are jointly significantly different to 0.

Partial ACF

- The Partial Autocorrelation Function (PACF) is similar to the ACF, however it measures correlation between observations that are *k* time periods apart, after controlling for correlations at intermediate lags.

- First order (i.e., k=1), AC and PAC are the same. For second order (k=2),

$$\frac{\text{Covariance}(y_t, y_{t-2} \mid y_{t-1})}{\sqrt{\text{Variance}(y_t \mid y_{t-1})\text{Variance}(y_{t-2} \mid y_{t-1})}}$$

- This can also be used to produce a partial Correlogram, which is used in Box-Jenkins methodology (covered later).

Autoregressive Process

Auto-regression simply means regression of a variable on itself measured at different time periods.
One of the fundamental assumptions of AR model is that the time series is assumed to be a stationary process.

If a time-series data, Y_t, is stationary, then it satisfies the following conditions:
1. The mean values of Y_t at different values of t are constant.
2. The variances of Y_t at different time periods are constant (Homoscedasticity).
3. The covariances of Y_t and Y_{t-k} for different lags depend only on k and not on time t.

When the time series data is not stationary (that is, any one of the above conditions are not satisfied), then we have to convert the non-stationary times-series data to stationary data before applying AR models.

Another important concept associated with forecasting based on regression-based models is the white noise of residuals. White noise is a process of residuals are uncorrelated and follow normal distribution with mean 0 and constant standard deviation. In AR models, one of the important assumptions that we make is that the errors follow a white noise.

Autoregressive Process

$$Y_{t+1} = \beta Y_t + \varepsilon_{t+1} \quad \text{which can be re-written as} \quad Y_{t+1} - \mu = \beta \times (Y_t - \mu) + \varepsilon_{t+1}$$

$$Y_{t+1} - \mu = \beta \times \left[\beta \times (Y_{t-1} - \mu) + \varepsilon_t \right] + \varepsilon_{t+1}$$

$$Y_{t+1} - \mu = \beta^t (Y_0 - \mu) + \beta^{t-1} \varepsilon_1 + \beta^{t-2} \varepsilon_2 + \cdots + \beta \varepsilon_t + \varepsilon_{t+1}$$

$$Y_{t+1} - \mu = \beta^t (Y_0 - \mu) + \sum_{k=1}^{t-1} \beta^{t-k} \times \varepsilon_k + \varepsilon_{t+1}$$

If $|\beta| > 1$, then $[\beta t\ (Y0 - \mu)]$ will result in infinitely large value of $Yt+1$ as the value of t increases and is not very useful for practical applications. The value of $|\beta| = 1$ would imply that the future value of Y depends on the entire past (and will lead to non-stationarity). For practical applications, the value of $|\beta|$ should be less than one.

The second part of the equation can also become infinitely large if the errors do not follow a white noise. When the errors are white noise then the expected value of sum($\beta_{t-k} e_k$) is zero.

Estimating β

$$\sum_{t=2}^{n} \varepsilon_t^2 = \sum_{t=2}^{n} [(Y_t - \mu) - \beta \times (Y_{t-1} - \mu)]^2 \tag{13.34}$$

Taking first-derivative of Eq. (13.34) with respect to β and equating that to zero, the estimate of β is given by

$$\hat{\beta} = \frac{\sum_{t=2}^{n} (Y_t - \mu)(Y_{t-1} - \mu)}{\sum_{t=2}^{n} (Y_{t-1} - \mu)^2} \tag{13.35}$$

AR Model Identification

Autocorrelation : $$\rho_k = \frac{\sum_{t=k+1}^{n}\left(Y_{t-k}-\bar{Y}\right)\left(Y_t-\bar{Y}\right)}{\sum_{t=1}^{n}\left(Y_t-\bar{Y}\right)^2}$$

Partial Autocorrelation: Correlation between Y_t and Y_{t-k} when the influence of all intermediate values is removed from both Y_t and Y_{t-k}

Plots of autocorrelation and partial autocorrelation for different values of k are called the ACF and PACF respectively

H_0: $\rho_k = 0$ and H_A: $\rho_k \neq 0$, where ρ_k is the auto-correlation of order k

H_0: $\rho_{pk} = 0$ and H_A: $\rho_{pk} \neq 0$, where ρ_{pk} is the partial auto-correlation of order k

The null hypothesis is rejected when $|\rho_k| > 1.96/\sqrt{n}$ and $|\rho_{pk}| > 1.96/\sqrt{n}$. To select the appropriate p in the auto-regressive model, the following thumb rule may be used. The number of lags is p when

1. The partial auto-correlation, $|\rho_{pk}| > 1.96/\sqrt{n}$ for first p values (first p lags) and cuts off to zero.
2. The auto-correlation function (ACF), ρ_k, decreases exponentially.

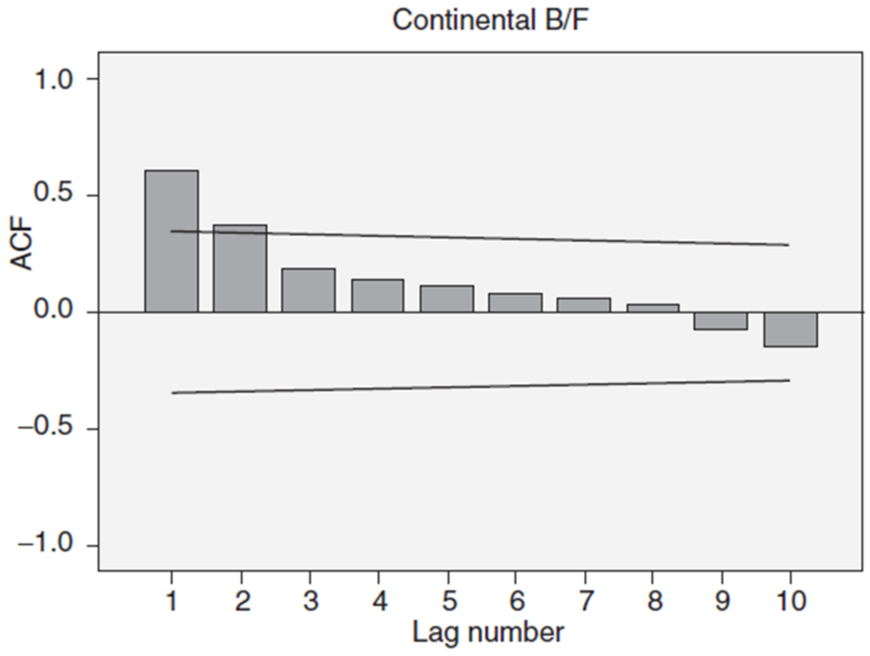
AR Model - Example

Build an auto-regressive model based on the first 30 days of data and forecast the demand for continental breakfast on days 31 to 37. Comment on the accuracy of the forecast.

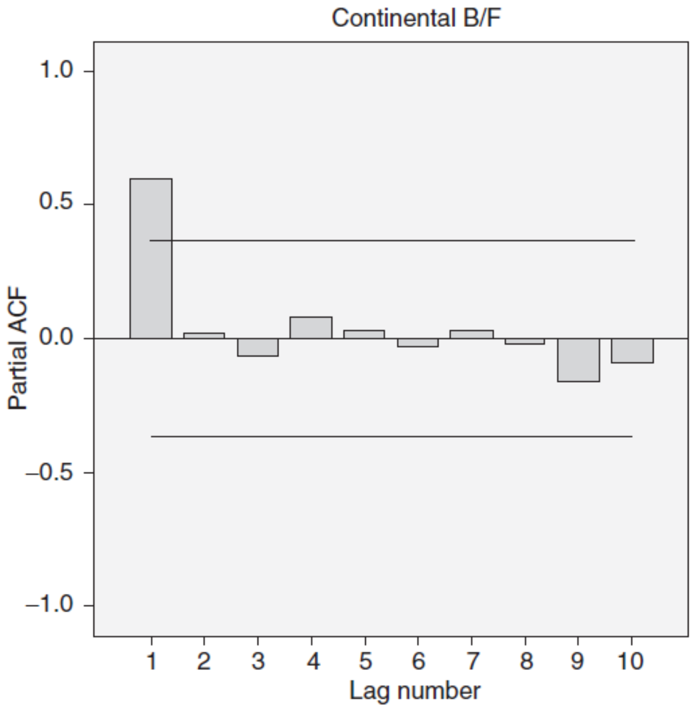
Day	Demand CB	Day	Demand CB
1	25	20	43
2	25	21	41
3	25	22	46
4	35	23	41
5	41	24	40
6	30	25	32
7	40	26	41
8	40	27	41
9	40	28	40
10	40	29	43
11	40	30	46
12	40	31	45
13	44	32	45
14	49	33	46
15	50	34	43
16	45	35	40
17	40	36	41
18	42	37	41
19	40		

Identifying *p*, the order of the AR Model

The first step in AR model building is the identification of the right value of p using ACF and PACF plots. ACF and PACF based on the first 30 observations are given in Figures 13.5 and 13.6, respectively. The horizontal lines in the plot represent the upper and lower critical values for ρ_k and ρ_{pk}. The correlation values (vertical bars) beyond the critical values will result in rejection of the null hypothesis.



ACF



PACF

Results for AR(1)

Model	Model Fit Statistics			
	R-Square	RMSE	MAPE	Normalized BIC
Continental B/F-Model_1	0.373	5.133	10.518	3.498

$$(F_{t+1} - 38.890) = 0.731(Y_t - 38.890)$$

Day	Y_t	F_t	$(Y_t - F_t)^2$	$\lvert Y_t - F_t \rvert / Y_t$
31	45	44.08741	0.832821	0.02028
32	45	43.35641	2.701388	0.036524
33	46	43.35641	6.988568	0.057469
34	43	44.08741	1.182461	0.025289
35	40	41.89441	3.588789	0.04736
36	41	39.70141	1.686336	0.031673
37	41	40.43241	0.322158	0.013844

MAPE 1.5721
RMSE 0.0332 (3.32%)

$$(F_{t+k} - 38.890) = 0.731(F_{t+k-1} - 38.890)$$

Day	Y_t	F_t	$(Y_t - F_t)^2$	$\lvert Y_t - F_t \rvert / Y_t$
31	45	44.0874	0.8328	0.0203
32	45	42.6893	5.3393	0.0513
33	46	41.6673	18.7723	0.0942
34	43	40.9202	4.3256	0.0484
35	40	40.3741	0.1399	0.0094
36	41	39.9749	1.0509	0.0250
37	41	39.6830	1.7344	0.0321

MAPE 2.1446
RMSE 0.04009 (4.009%)

Moving Average (MA) process

- In this simple model, the dependent variable is regressed against lagged values of the past terms or error terms. MA(1) is given by: $Y_{t+1} = \mu + \alpha_1 \varepsilon_t + \varepsilon_{t+1}$

$$Y_{t+1} = \alpha_1 \varepsilon_1 + \varepsilon_{t+1}$$

- MA(q) is given by:

$$Y_{t+1} = \mu + \alpha_1 \varepsilon_t + \alpha_2 \varepsilon_{t-1} + \cdots + \alpha_q \varepsilon_{t-q+1} + \varepsilon_{t+1}$$

- Order q of a MA process:

1. Auto-correlation value, $|\rho_p| > 1.96/\sqrt{n}$ for first q values (first q lags) and cuts off to zero.
2. The partial auto-correlation function, ρ_{pk}, decreases exponentially.

AR(p) and MA(q)

- Before conducting a regression, we need to consider whether the variables are stationary or not.
- The ACF and Correlogram is one way of determining if a series is stationary, as is the Q- statistic
- An AR(p) process involves the use of p lags of the dependent variable as explanatory variables
- A MA(q) process involves the use of q lags of the error term

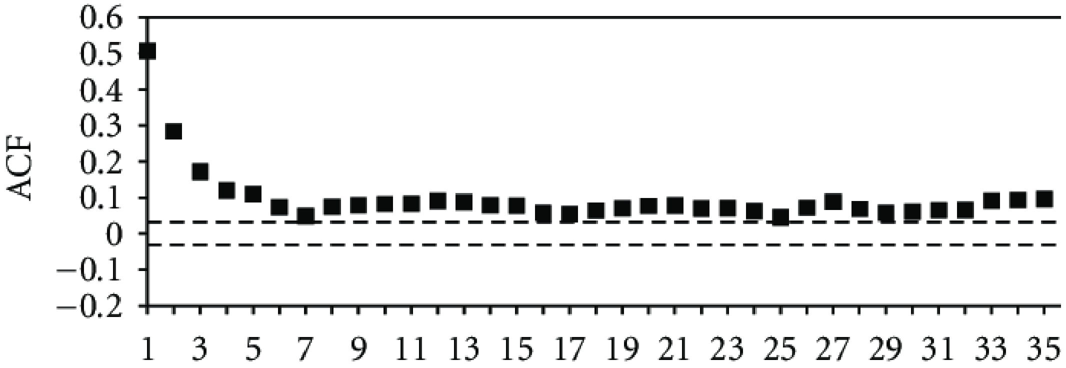
ARMA(p,q): Autoregressive, Moving Average

$$Y_{t+1} = \overbrace{\beta_1 Y_t + \beta_2 Y_{t-1} + \ldots + \beta_p Y_{t-p+1}}^{\text{Auto Regressive Part}} + \overbrace{\alpha_1 \varepsilon_t + \alpha_2 \varepsilon_{t-1} + \ldots + \alpha_q \varepsilon_{t-q+1}}^{\text{Moving Average Part}} + \varepsilon_{t+1}$$

1. Auto-correlation value, $|\rho_p| > 1.96/\sqrt{n}$ for first q values (first q lags) and cuts off to zero.
2. Partial auto-correlation function, $|\rho_{pk}| > 1.96/\sqrt{n}$ for first p values and cuts off to zero.

ARMA(p,q): Parameter selection

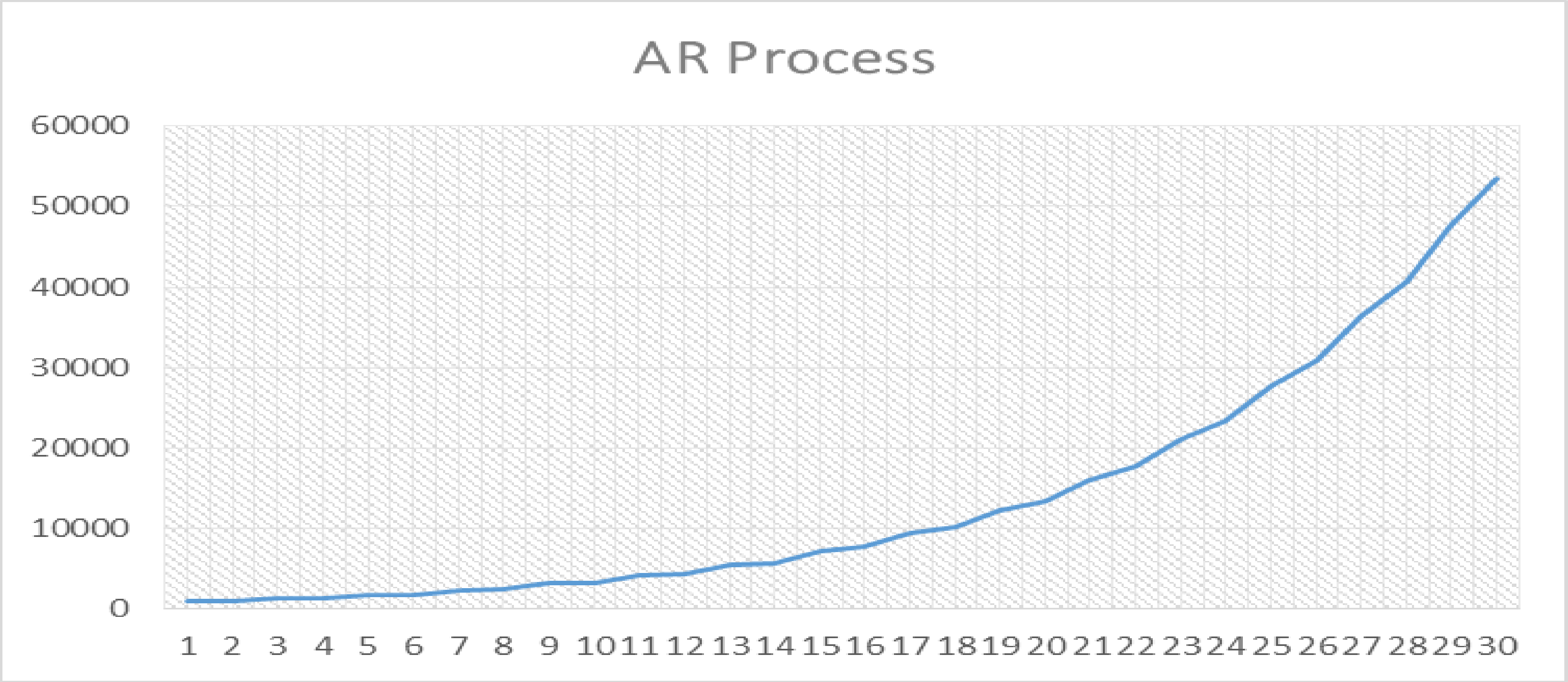
Model	ACF	PACF
AR (p)	Spikes decay towards zero. Coefficients may oscillate.	Spikes decay to zero after lag p
MA (q)	Spikes decay to zero after lag q	Spikes decay towards zero. Coefficients may oscillate.
ARMA (p, q)	Spikes decay (either direct or oscillatory) to zero beginning after lag q	Spikes decay (either direct or oscillatory) to zero beginning after lag p



AR & MAModels

- Autoregressive AR process:
  - Series current values depend on its own previous values
  - AR(p) - Current values depend on its own p-previous values
  - P is the order of AR process
- Moving average MA process:
  - The current deviation from mean depends on previous deviations
  - MA(q) - The current deviation from mean depends on q- previous deviations
  - q is the order of MA process
- Autoregressive Moving average ARMA process

AR Model

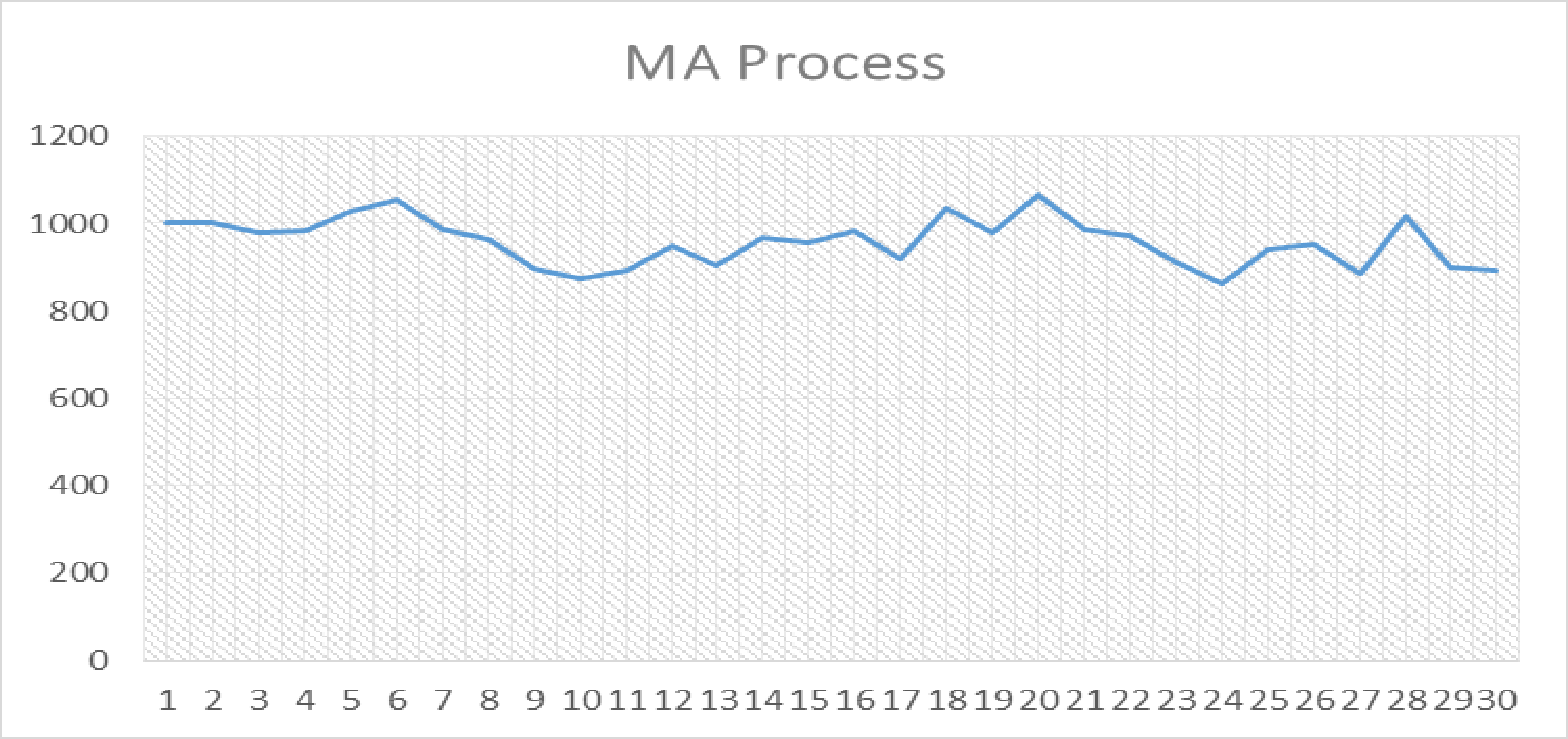


AR(1) $y_t = a1^* \, y_{t-1}$

AR(2) $y_t = a1^* \, y_{t-1} + a2^* \, y_{t-2}$

AR(3) $y_t = a1^* \, y_{t-1} + a2^* \, y_{t-2} + a3^* \, y_{t-2}$

MA Model



MA(1) $\varepsilon_t = b1*\varepsilon_{t-1}$

MA(2) $\varepsilon_t = b1*\varepsilon_{t-1} + b2*\varepsilon_{t-2}$

MA(3) $\varepsilon_t = b1*\varepsilon_{t-1} + b2*\varepsilon_{t-2} + b3*\varepsilon_{t-3}$

ARMA(p, q) – An example

Monthly demand for avionic system spares used in Vimana 007 aircraft is provided. Build an ARMA model based on the first 30 months of data and forecast the demand for spares for months 31 to 37. Comment on the accuracy of the forecast.

Month	Demand for Spares	Month	Demand for Spares
1	457	20	516
2	439	21	656
3	404	22	558
4	392	23	647
5	403	24	864
6	371	25	610
7	382	26	677
8	358	27	609
9	594	28	673
10	482	29	400
11	574	30	443
12	704	31	503
13	486	32	688
14	509	33	602
15	537	34	629
16	407	35	823
17	523	36	671
18	363	37	487
19	479		

Example: Step1 – Plot ACF, PACF

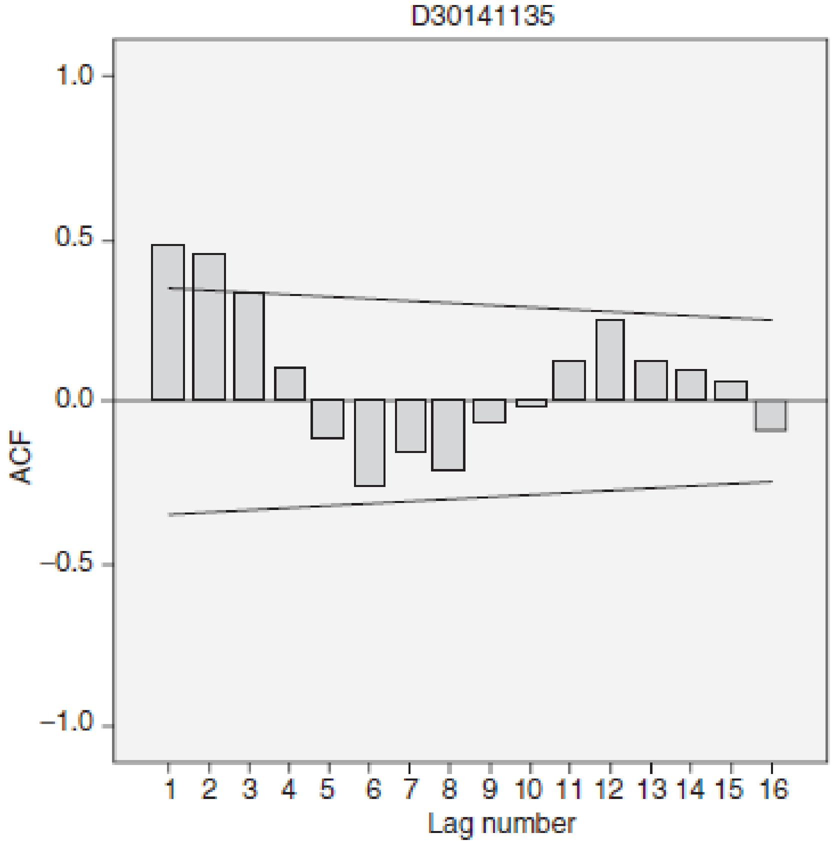


FIGURE 13.9 ACF plot for avionic system spares demand

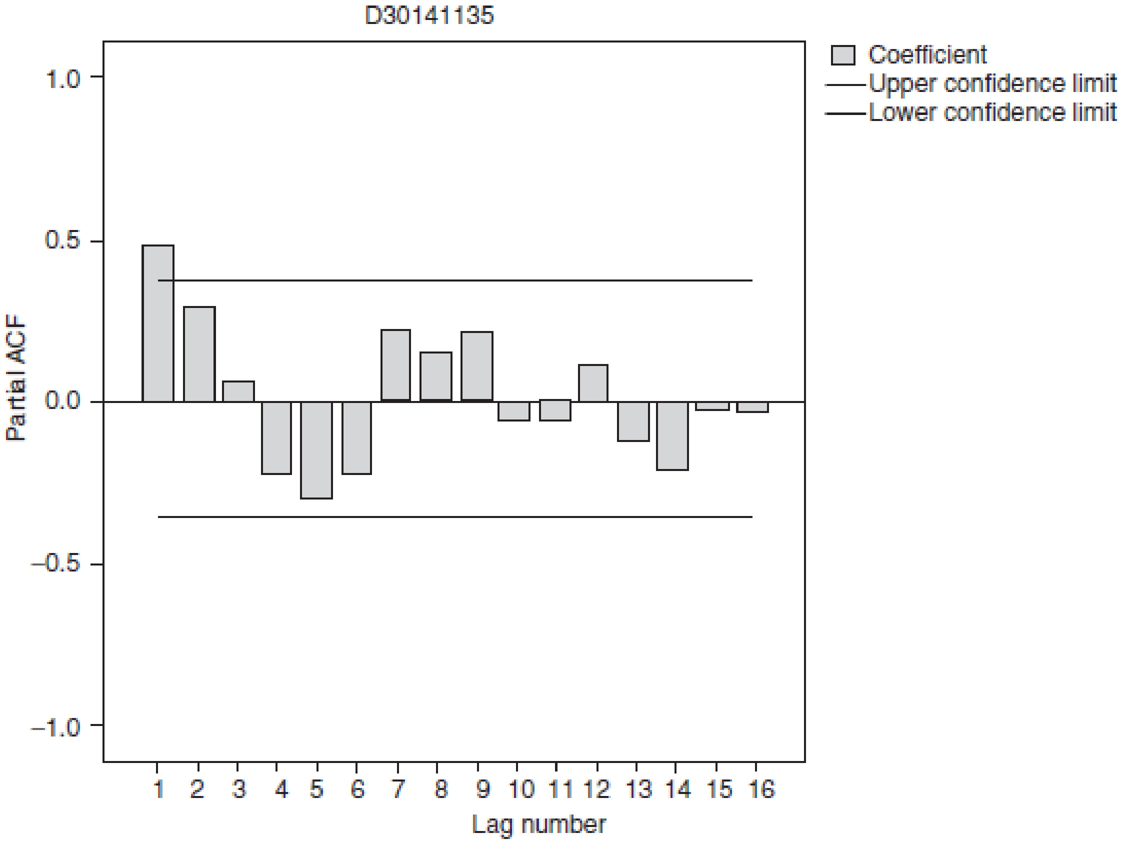


FIGURE 13.10 PACF plot for avionic system spares demand.

Example: Step 2 – Forecast (ARMA(1,2))

Model	Model Fit Statistics		
	Stationary *R*-Squared	RMSE	MAPE
Avionic Spares	0.429	98.824	14.231

TABLE 13.26 model parameters

			Estimate	*SE*	*T*	Sig.
Avionic Spares	Constant		496.699	57.735	8.603	0.000
	AR	Lag 1	0.706	0.170	4.153	0.000
	MA	Lag 1	0.694	0.173	4.006	0.000
		Lag 2	−0.727	0.170	−4.281	0.000

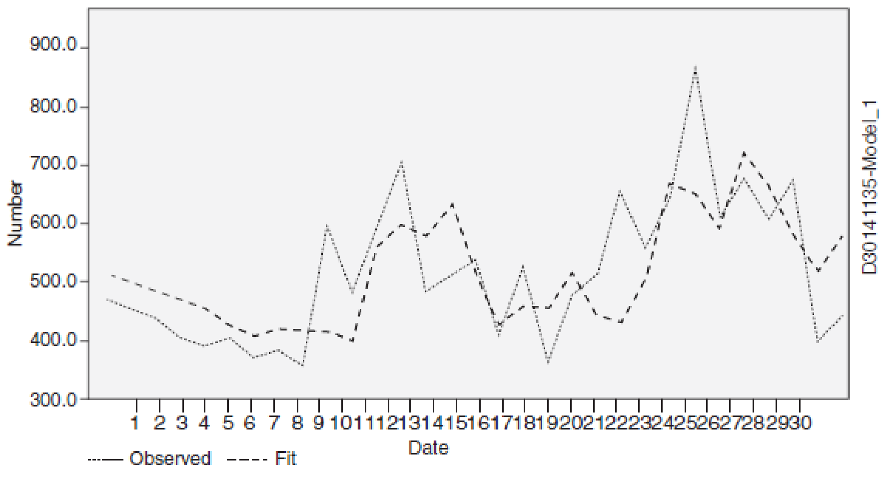


FIGURE 13.11 Observed versus forecasted demand.

All the three components in the ARMA model (AR lag 1 and MA lags 1 and 2) are statistically significant (Table 13.26). The model equation using SPSS is given by

$$Y_{t+1} - 496.669 = 0.706 \times (Y_t - 496.699) - 0.694 \times \varepsilon_t + 0.727 \times \varepsilon_{t-1} \qquad (13.45)$$

Example: Step 3 – Compute MAPE, RMSE

TABLE 13.27 ARMA(1, 2) model forecast

Month	Y_t	F_t	$(Y_t - F_t)^2$	$\lvert Y_t - F_t \rvert / Y_t$
31	503	464.8107	1458.423	0.075923
32	688	378.5341	95769.15	0.449805
33	602	444.6372	24763.04	0.2614
34	629	685.8851	3235.909	0.090437
35	823	743.5124	6318.281	0.096583
36	671	630.7183	1622.614	0.060032
37	487	649.3491	26357.22	0.333366

The RMSE and MAPE for the validation data (months 31 and 37) are 150.961 0.1953 (19.53%), respectively (Table 13.27).

The forecasted values using F_t instead of Y_t when forecasting for more than one period ahead in time are shown in Table 13.28.

TABLE 13.28 ARMA (1, 2) forecast

Month	Y_t	F_t	$(Y_t - F_t)^2$	$\lvert Y_t - F_t \rvert / Y_t$
31	503	464.4239	1488.1147	0.0767
32	688	377.8374	96200.8258	0.4508
33	602	444.5195	24800.1101	0.2616
34	629	687.2082	3388.1980	0.0925
35	823	744.9583	6090.4998	0.0948
36	671	630.5592	1635.4571	0.0603
37	487	648.3959	26048.6313	0.3314

The RMSE and MAPE for the validation data (months 31 and 37) are 151.02 and 0.1954 (19.54%), respectively.

References

Text Book:

“Business Analytics, The Science of Data-Driven Making”, U. Dinesh Kumar, Wiley 2017 (Ch. 13.10-13.13)

Additional reference for the interested reader:
Introduction to Time Series and Forecasting, Second Edition by Peter J. Brockwell, Richard A. Davis Springer 2002.

Image Courtesy

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

Lecture 6: 13.10, 13.12, 13.13 in text - AR, MA and ARMA models
(AR https://otexts.com/fpp2/AR.html) + MA (https://otexts.com/fpp2/MA.html) + ARMA Venkat Reddy's slides on ARIMA)

THANK YOU

Jyothi R

Assistant Professor, Department of Computer Science

jyothir@pes.edu

DATA ANALYTICS

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

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):

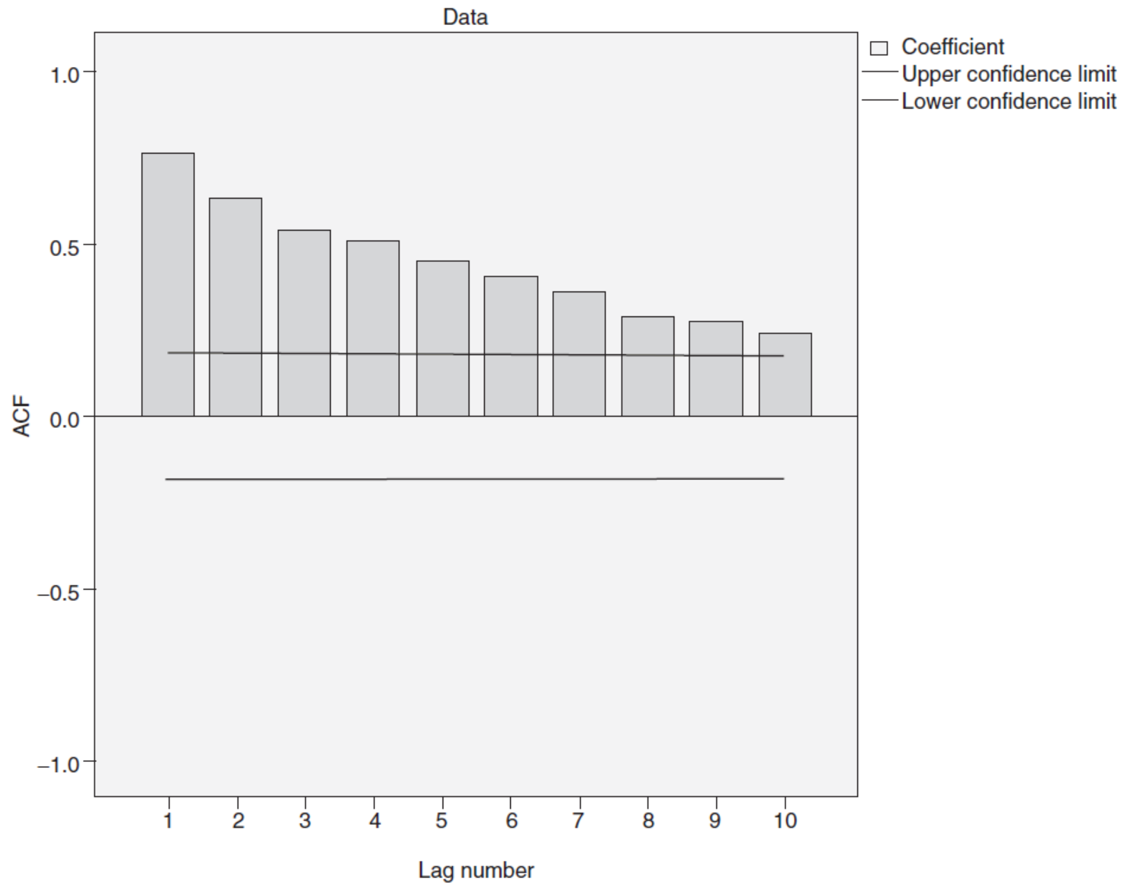
1. Auto-regressive with lag p
2. Integration component (d)
3. Moving average (q)

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

Nonstationarity could arise from deterministic or stochastic trend

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

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

Dickey Fuller Test

- ψ = 0 is same as β = 1. So, the Dickey−Fuller test can be written in terms of ψ as

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 now tested for non-stationarity.
- Again the null and alternative hypotheses are
- H_0: $\psi = 0$ (the time series is non-stationary)
- H_0: $\psi < 0$ (the time series is stationary)

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**).

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

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

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

Stationarity and differencing

- 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.

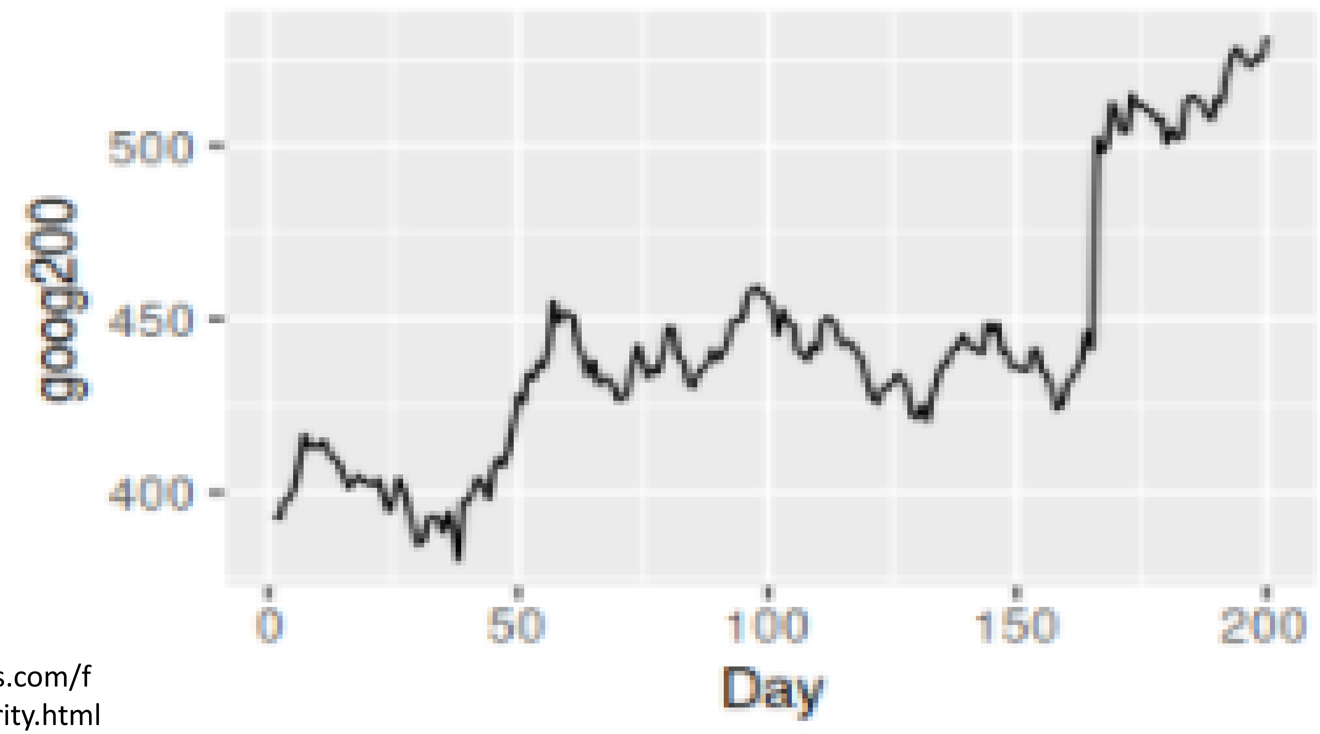
- This is because the cycles are not of a fixed length, so before we observe the series we cannot be sure where the peaks and troughs of the cycles will be.

- 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.

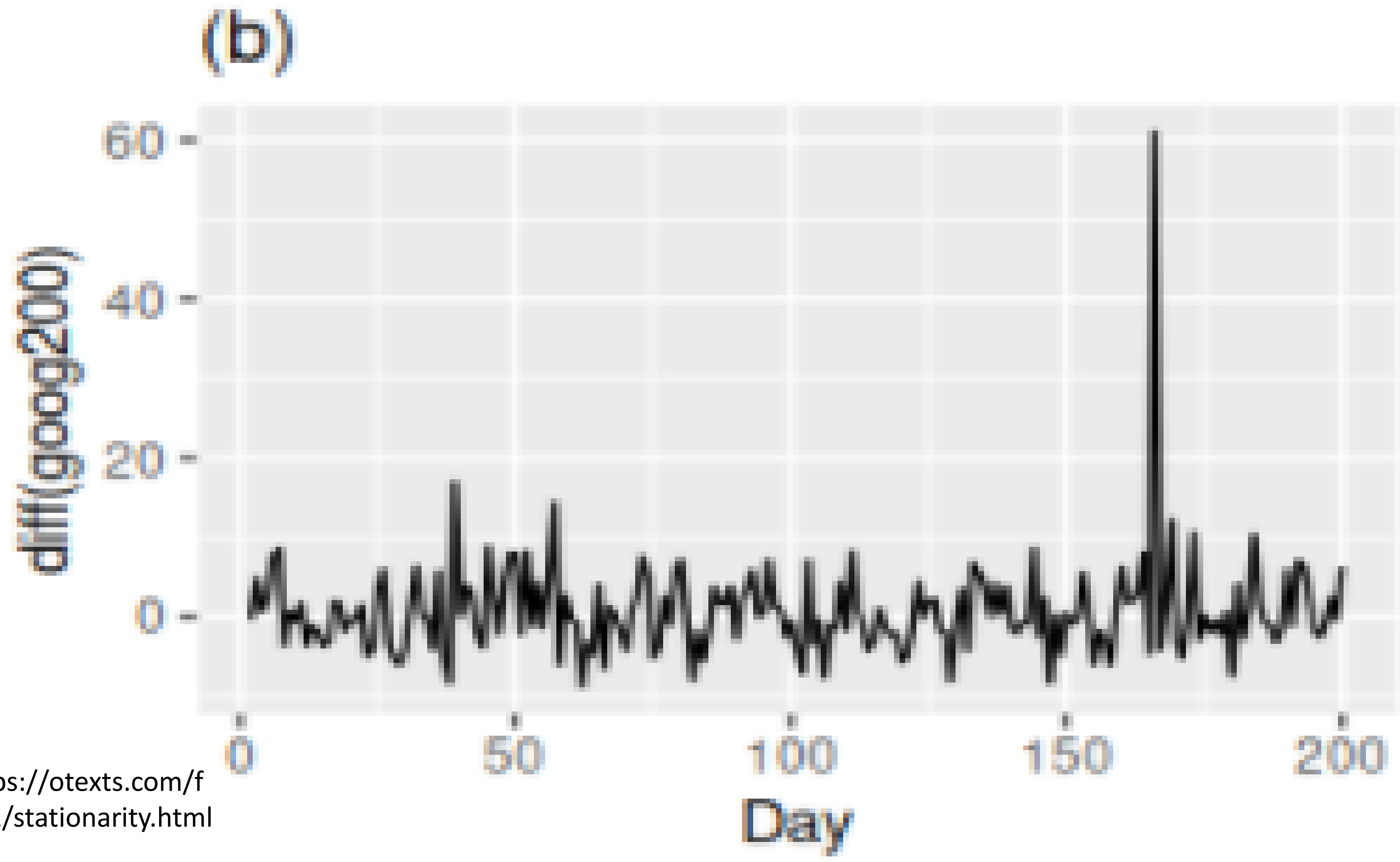
(a) Changing trends and levels



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

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

Which of these series are stationary?

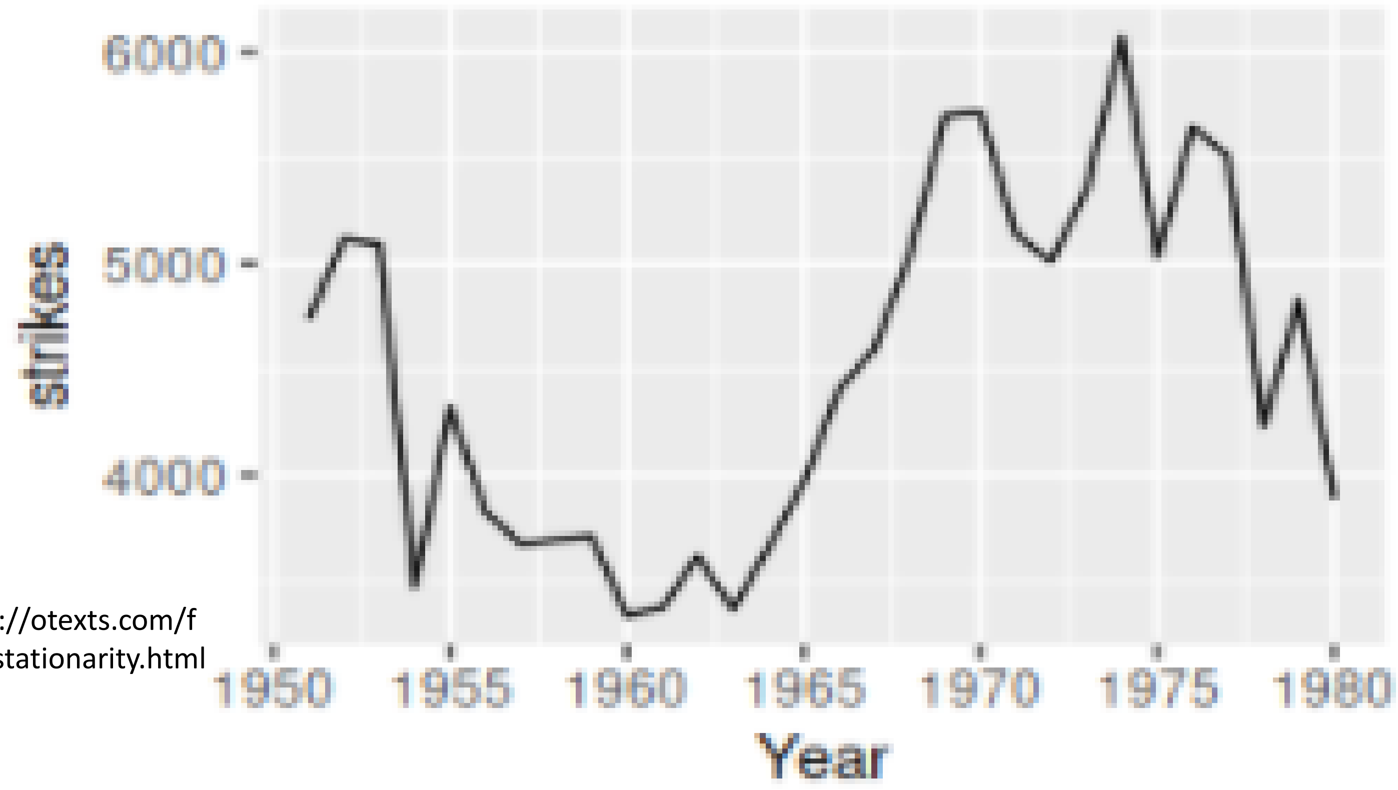


https://otexts.com/fpp2/stationarity.html

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

Which of these series are stationary?

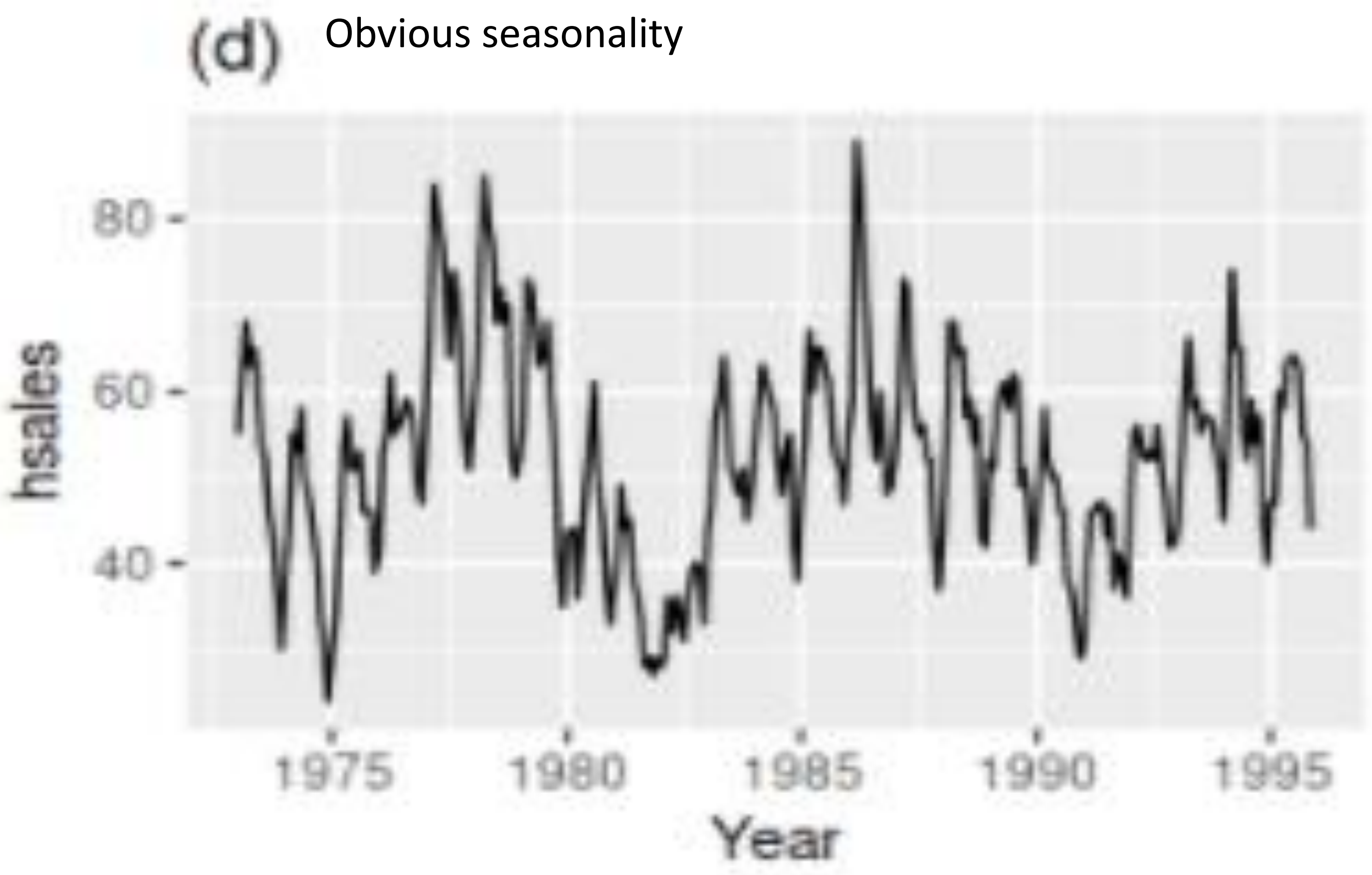
(c) Changing trends and levels



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

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

Which of these series are stationary?

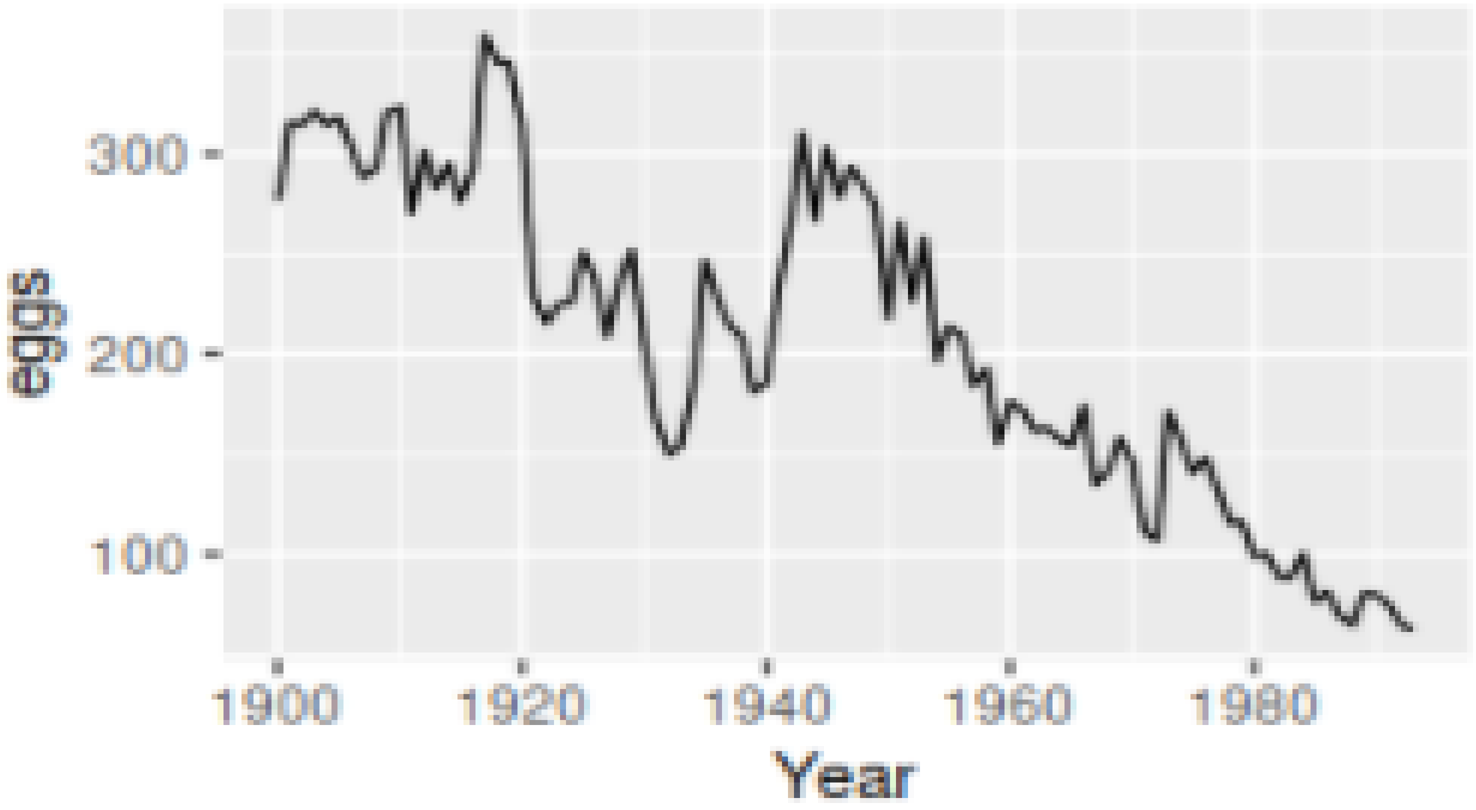


https://otexts.com/fpp2/stationarity.html

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

Which of these series are stationary?

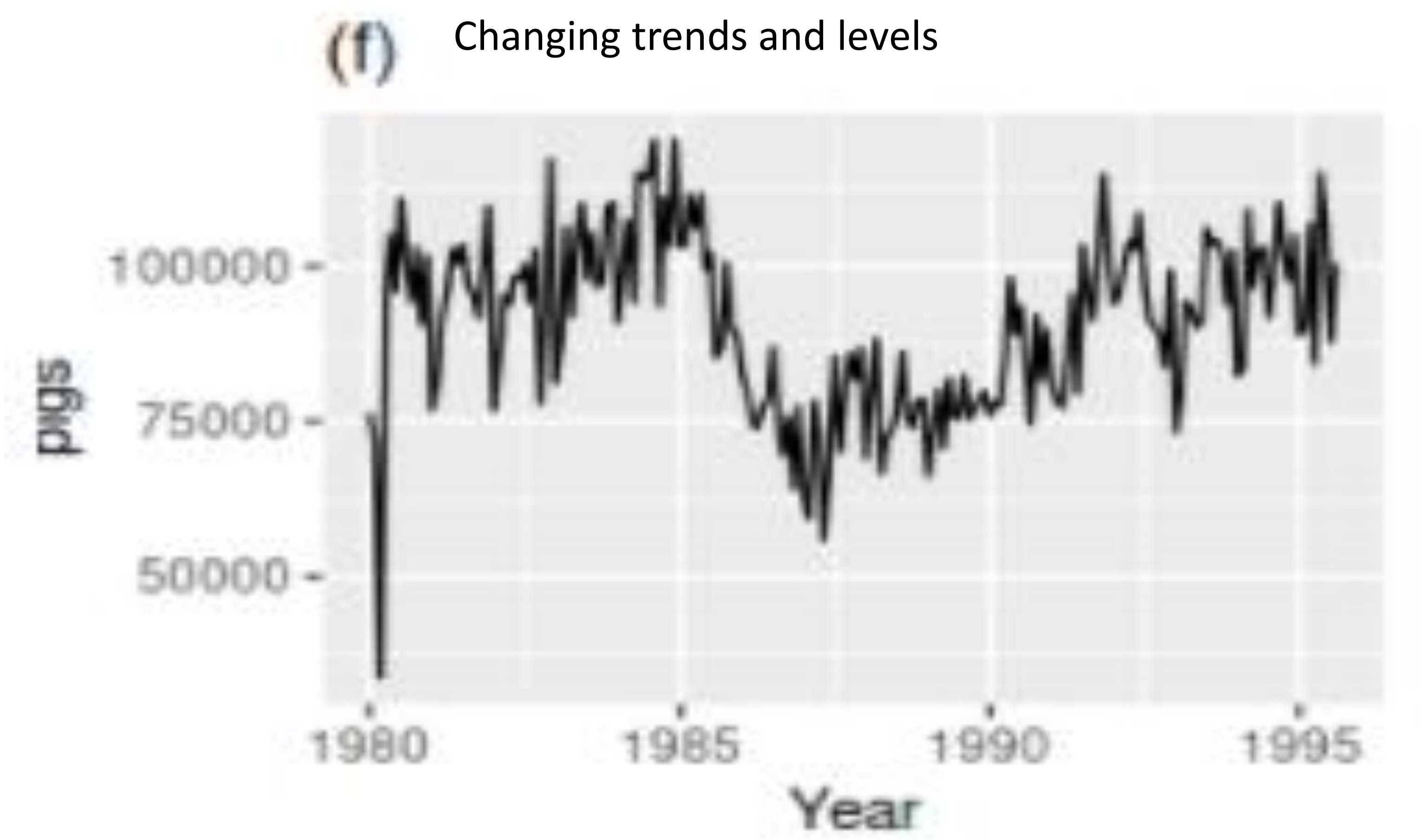
(e) Changing trends and levels



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

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

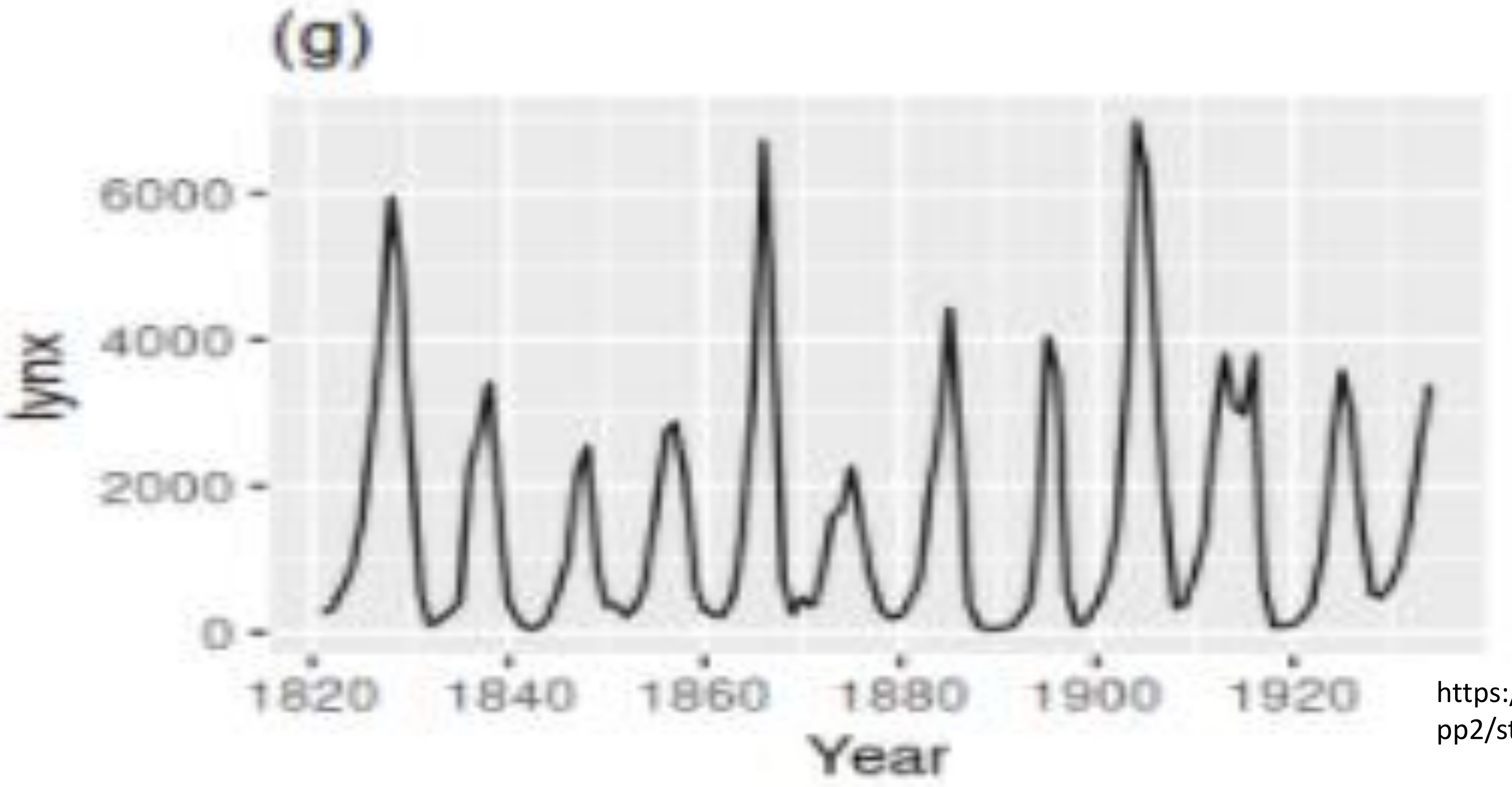
Which of these series are stationary?



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

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

Which of these series are stationary?

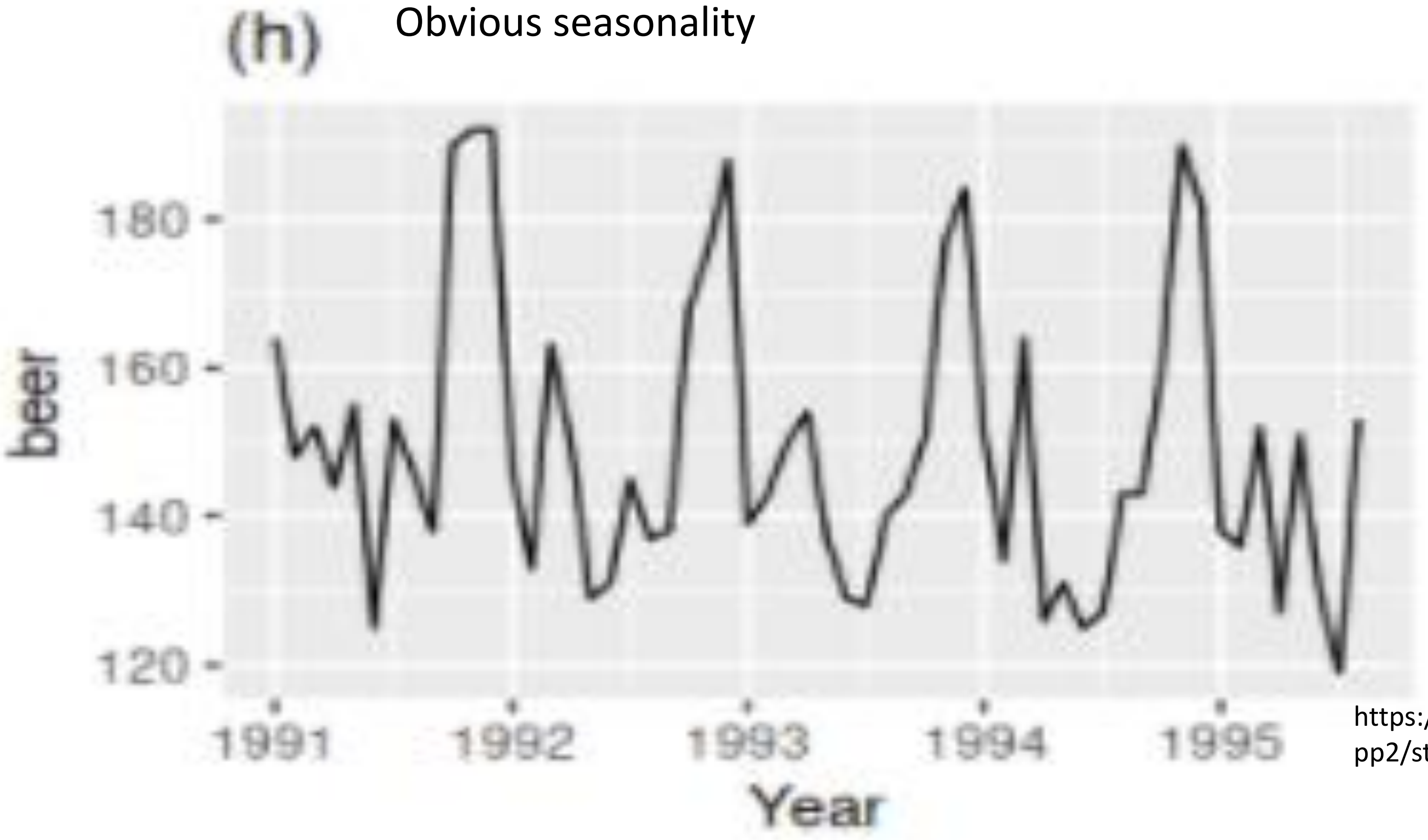


https://otexts.com/fpp2/stationarity.html

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

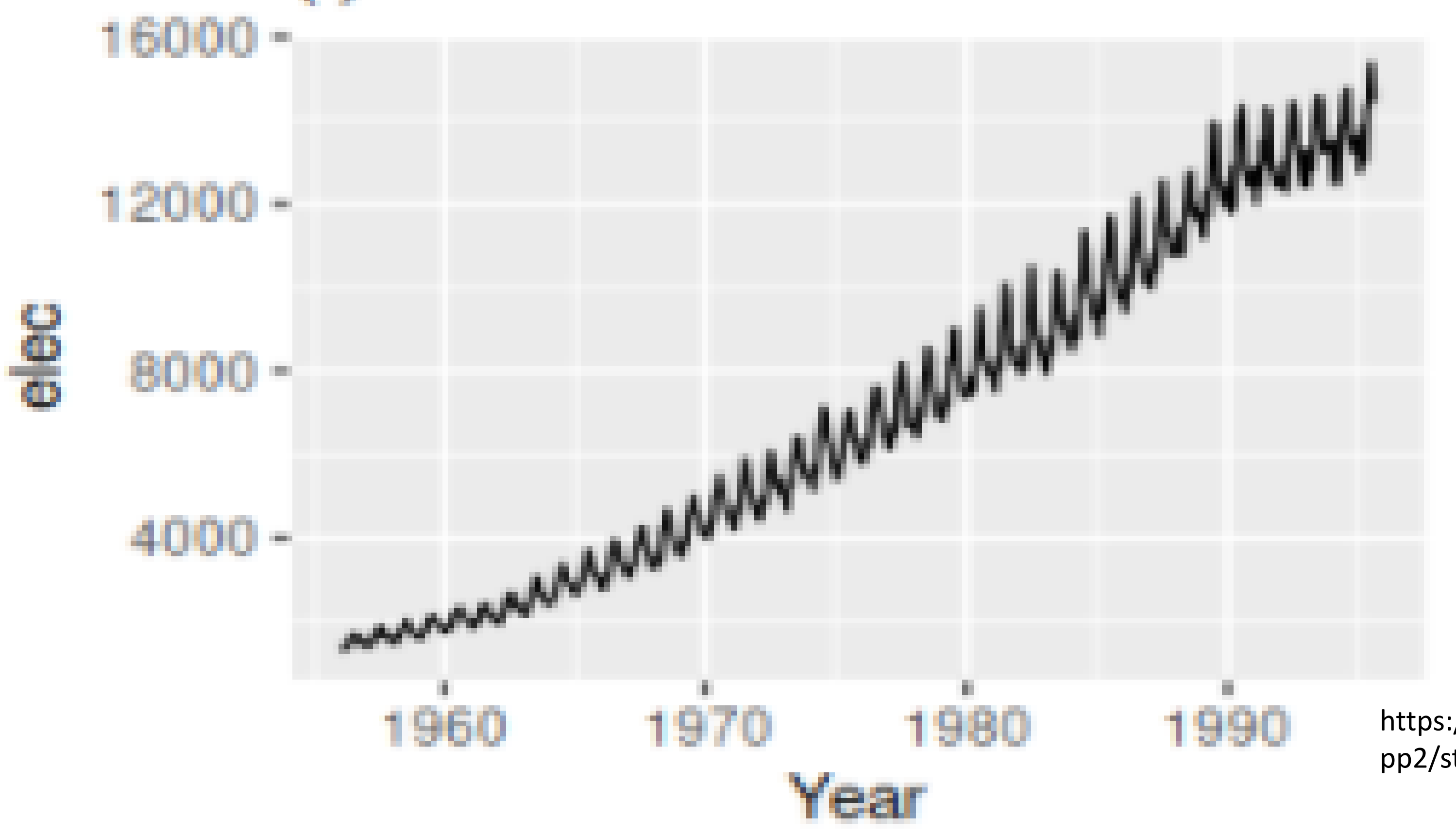


https://otexts.com/fpp2/stationarity.html

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

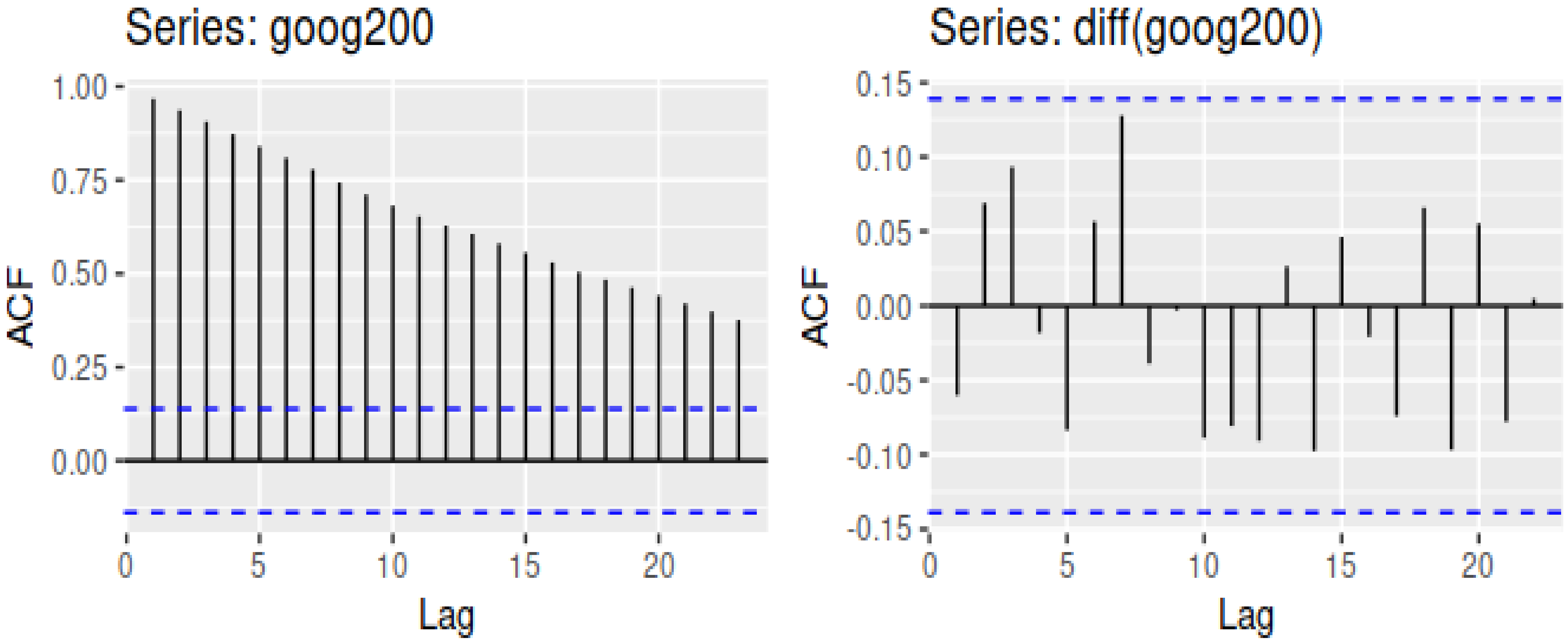


Figure 10:The ACF of the Google stock price (left) and of the daily changes in Google stock price (right).

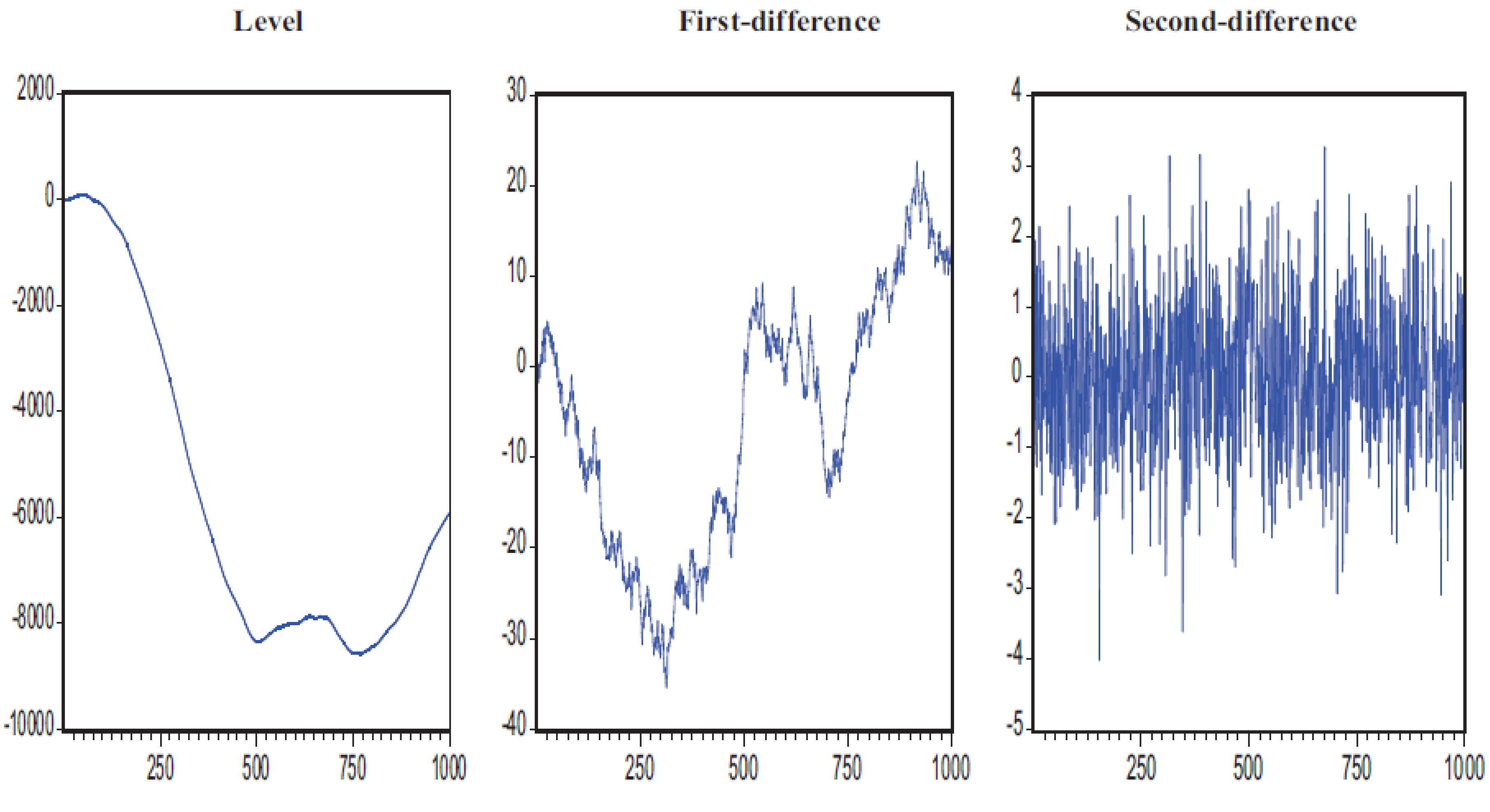
Differencing

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.

DATA ANALYTICS

Differencing



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,

Random Walk Model

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

  Then
- $y_t - y_{t-1} = c + \varepsilon_t$ or $y_t = c + y_{t-1} + \varepsilon_t$.
- $y_t - y_{t-1} = c + \varepsilon_t$ or $y_t = c + y_{t-1} + \varepsilon_t$.
- The value of cc is the average of the changes between consecutive observations.
- If cc 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

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
- The first step is to plot the ACF and PACF to identify whether the time series is stationary or not.
- If the time series is stationary then $d = 0$ and
- the model is ARIMA(p, 0, q) or ARMA(p, q) model.
- If the time series is non-stationary then it has to be converted into a stationary
- process by identifying the order of differencing.
- 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

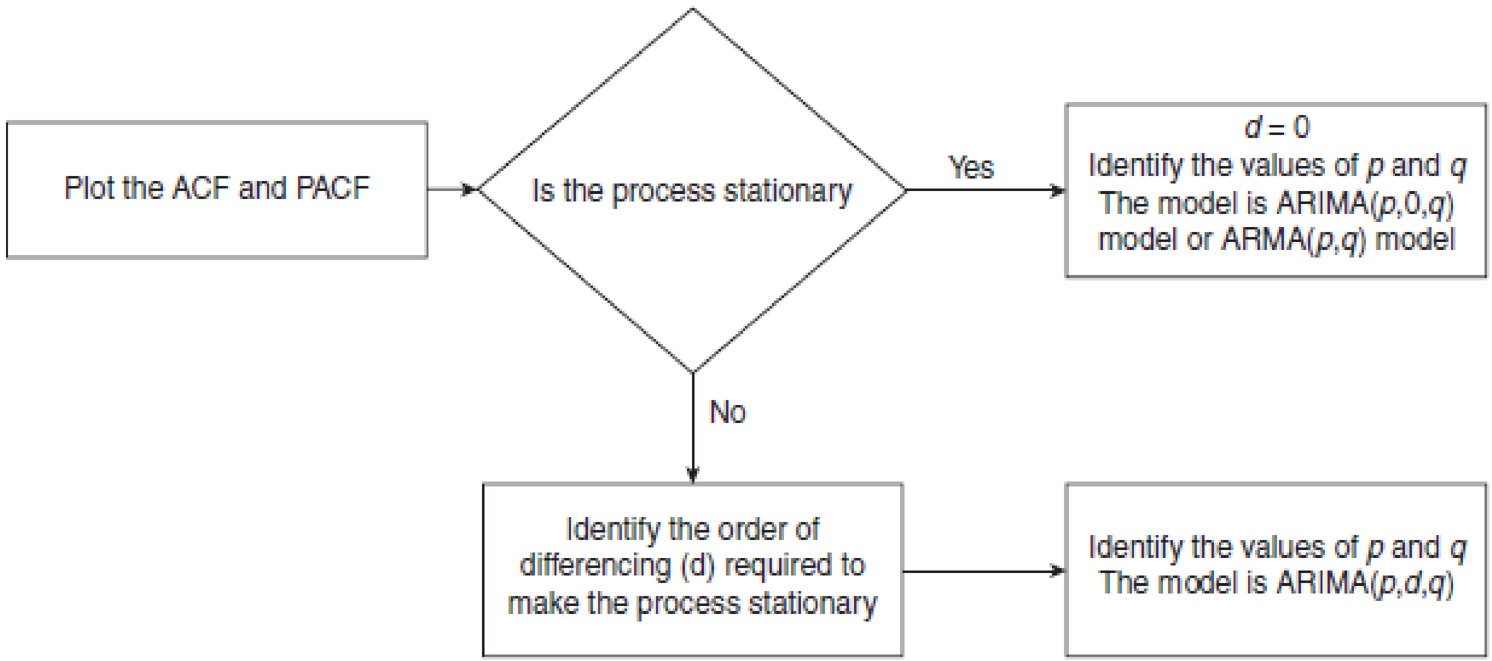


FIGURE 13.14 Model identification in ARIMA model.

Parameter Estimation and Model Selection

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

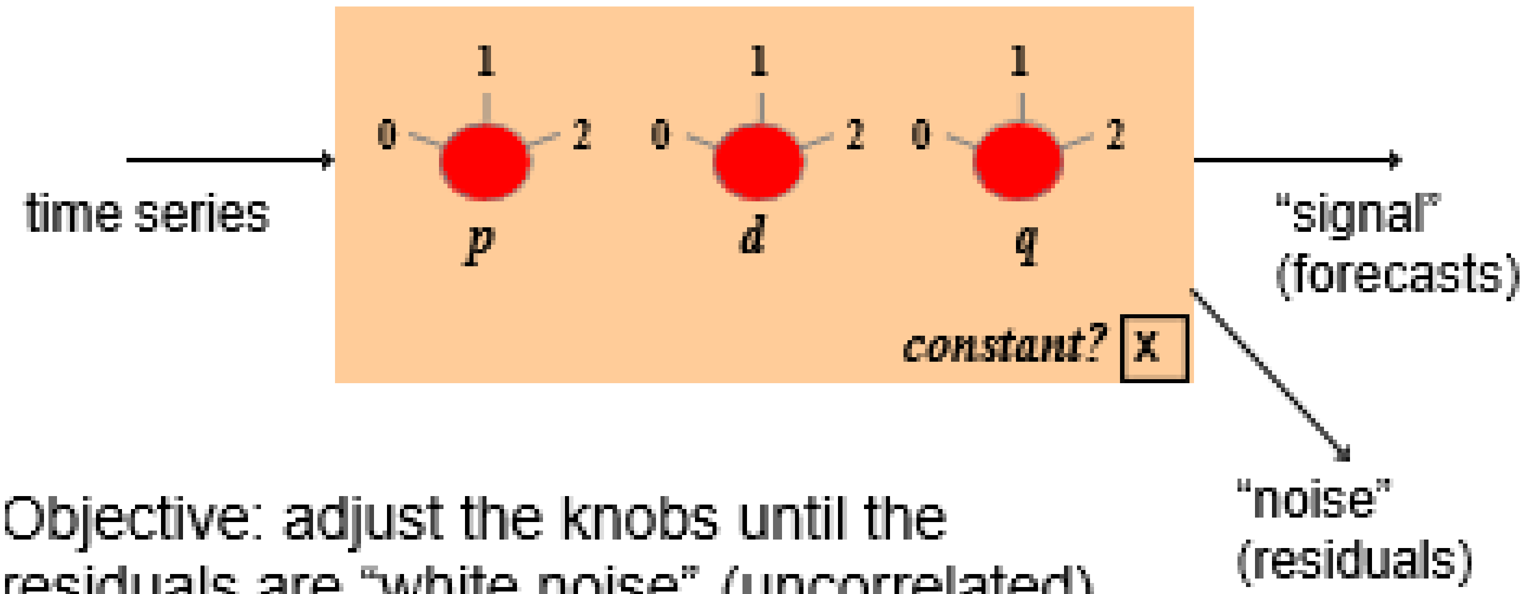
Lower values of AIC and BIC are preferred.

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”



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

ARIMA models we have already met

- ARIMA(0,0,0)+c = mean (constant) model
- ARIMA(0,1,0) = RW model
- ARIMA(0,1,0)+c = RW with drift model
- ARIMA(1,0,0)+c = regress Y on Y_LAG1
- ARIMA(1,1,0)+c = regr. Y_DIFF1 on Y_DIFF1_LAG1
- ARIMA(2,1,0)+c = ” ” plus Y_DIFF_LAG2 as well
- ARIMA(0,1,1) = SES model
- ARIMA(0,1,1)+c = SES + constant linear trend
- ARIMA(1,1,2) = LES w/ damped trend (leveling off)
- ARIMA(0,2,2) = generalized LES (including Holt’s)

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

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

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)

$$\hat{y}_t = \overbrace{\mu}^{\text{constant}} + \overbrace{\phi_1 y_{t-1} + \ldots + \phi_p y_{t-p}}^{\text{AR terms (lagged values of } y)}$$

$$\underbrace{- \theta_1 e_{t-1} \ldots - \theta_q e_{t-q}}_{\text{MA terms (lagged errors)}}$$

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

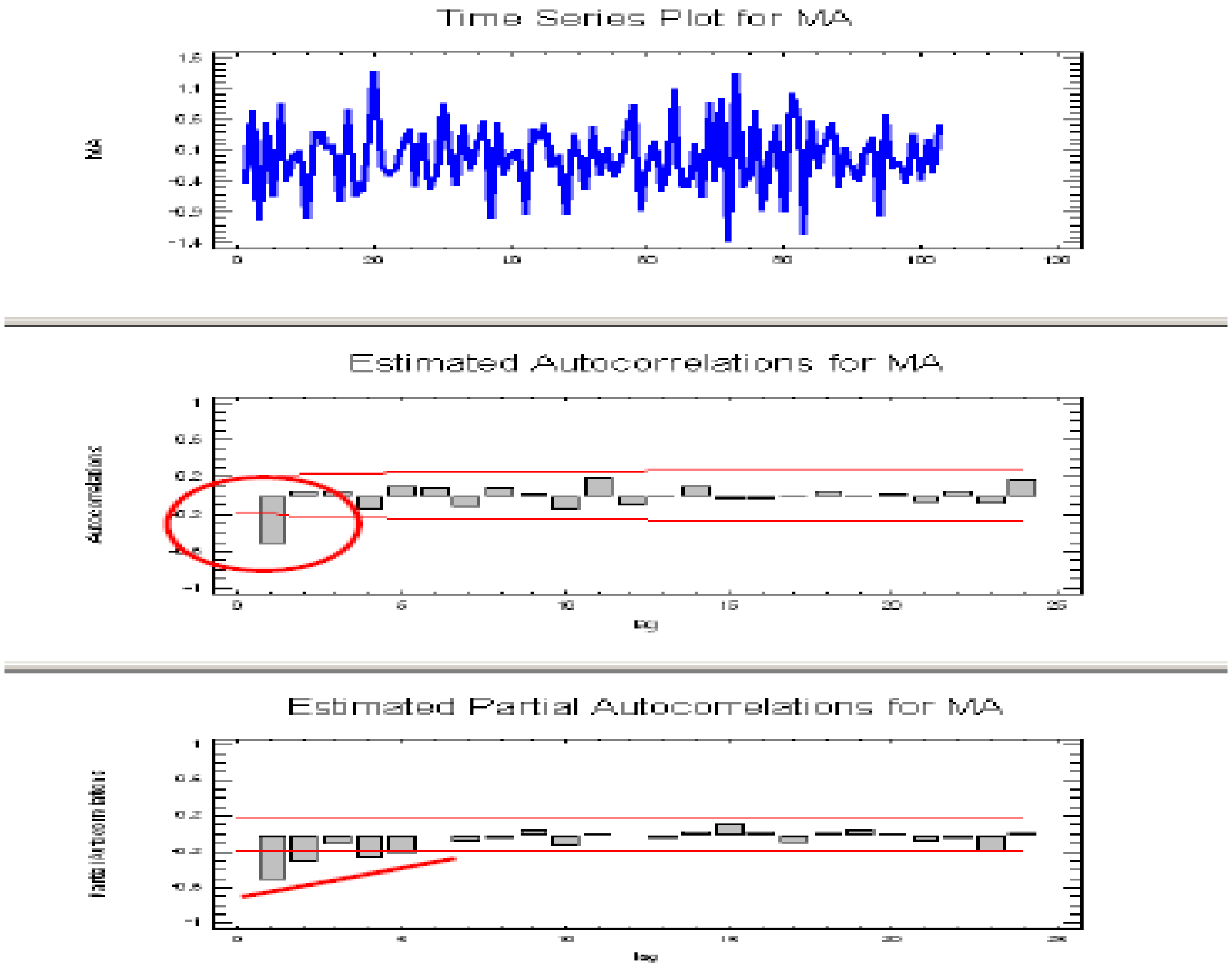
- A series displays autoregressive (AR) behavior if it apparently feels a “restoring force” that tends to pull it back toward its mean.
  - In an AR(1) model, the AR(1) coefficient determines how fast the series tends to return to its mean. If the coefficient is near zero, the series returns to its mean 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 or MA? It depends!

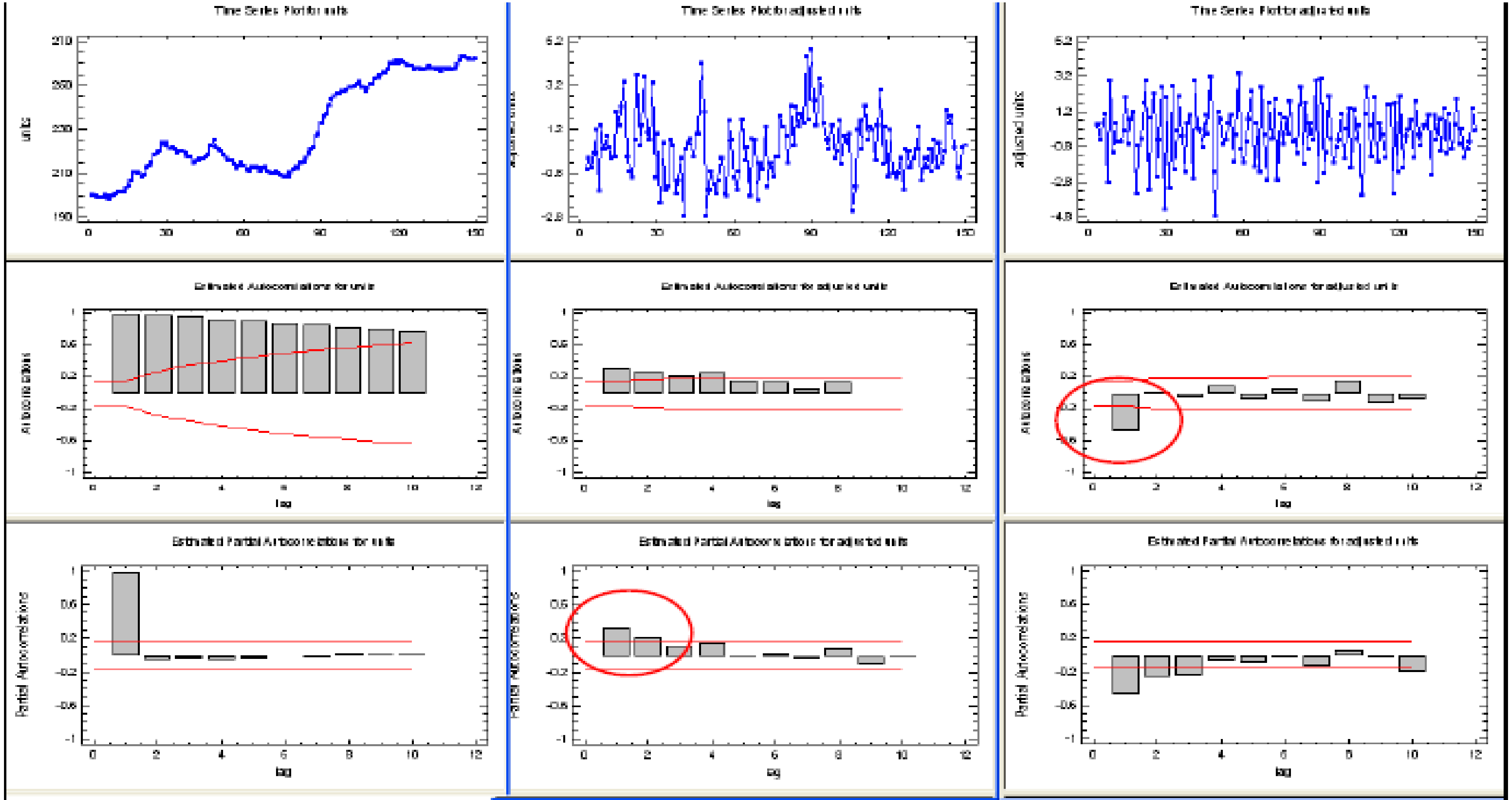
- Whether a series displays AR or MA behavior often depends on the extent to which it has been differenced.

- 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

An example



Original series: nonstationary

1st difference: AR signature

2nd difference: MA signature

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.

Tuning Parameters - Technical issues

- Backforecasting
  - Estimation algorithm begins by forecasting backward into the past to get start-up values
- Unit roots
  - Look at sum of AR coefficients and sum of MA coefficients—if they are too close to 1 you may want to consider higher or lower of differencing
- Overdifferencing
  - A series that has been differenced one too many times will show *very* strong negative autocorrelation and a strong MA signature, probably with a unit root in MA coefficients

Seasonal ARIMA terminology

- The seasonal part of an ARIMA model is summarized by three *additional* numbers:

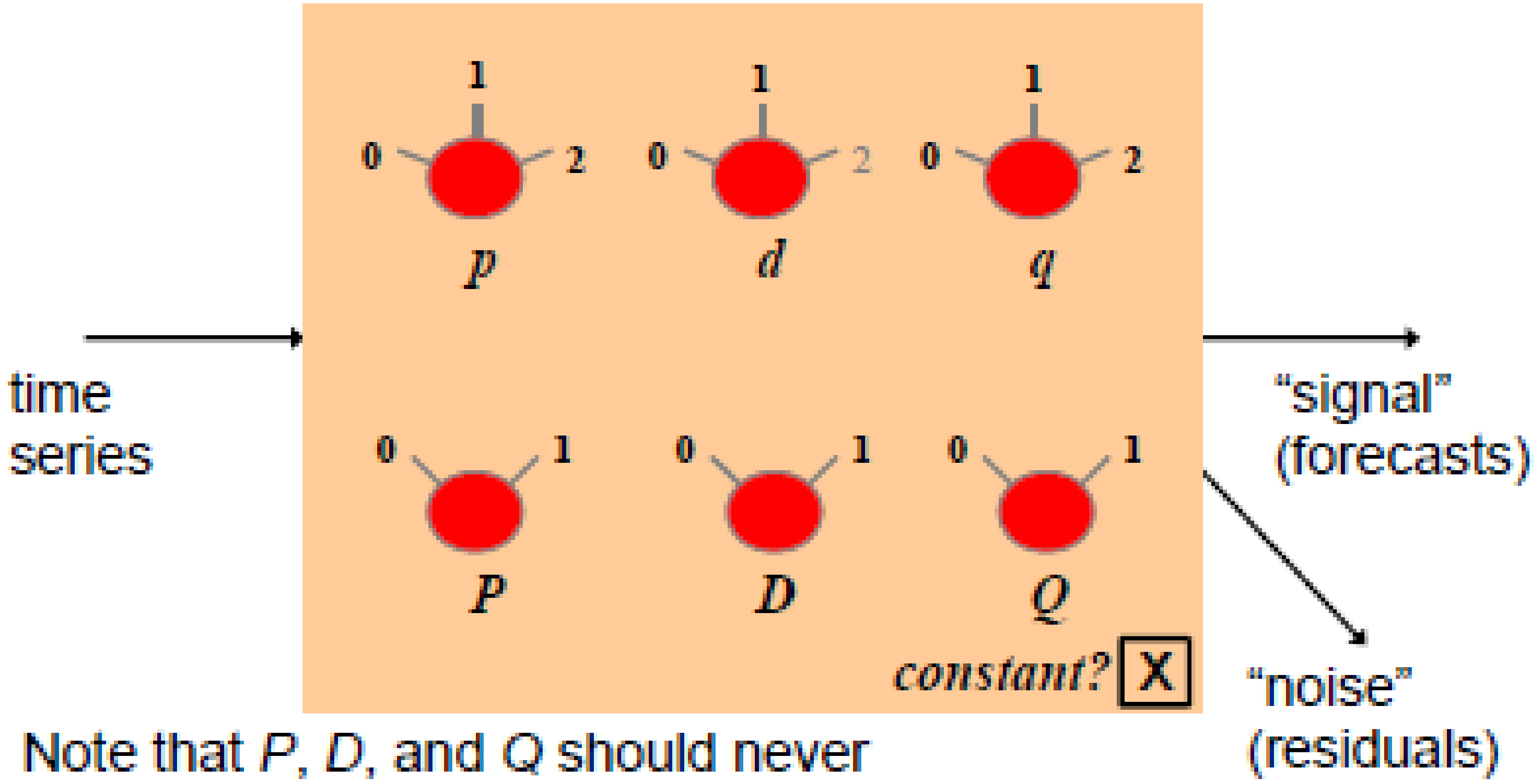
  ***P* = # of *seasonal autoregressive* terms**

  D* = # of *seasonal differences

  ***Q* = # of *seasonal moving-average* terms**

- The complete model is called an "SARIMA(*p*,*d*,*q*)(*P*,*D*,*Q*)" model

The "filtering box" now has 6 knobs:



Note that *P*, *D*, and *Q* should never be larger than 1 !!

Seasonal differences

- How non-seasonal & seasonal differences are combined to stationarize the series:

If d=0, D=1: $y_t = Y_t - Y_{t\text{-}s}$

s is the seasonal period, e.g., s=12 for monthly data

If d=1, D=1: $y_t = (Y_t - Y_{t\text{-}1}) - (Y_{t\text{-}s} - Y_{t\text{-}s\text{-}1})$

$= Y_t - Y_{t\text{-}1} - Y_{t\text{-}s} + Y_{t\text{-}s\text{-}1}$

D should never be more than 1, and *d*+*D* should never be more than 2. Also, if *d*+*D* =2, the constant term should be suppressed.

SAR and SMA terms

- How SAR and SMA terms add coefficients to the model:

- Setting *P* =1 (i.e., SAR=1) adds a multiple of
  - $y_{t\text{-}s}$ to the forecast for y_t

- Setting *Q* =1 (i.e., SMA=1) adds a multiple of
  - $e_{t\text{-}s}$ to the forecast for y_t
  - Total number of SAR and SMA factors usually should not be more than 1 (i.e., either SAR=1 or SMA=1, not both)

Model-fitting steps

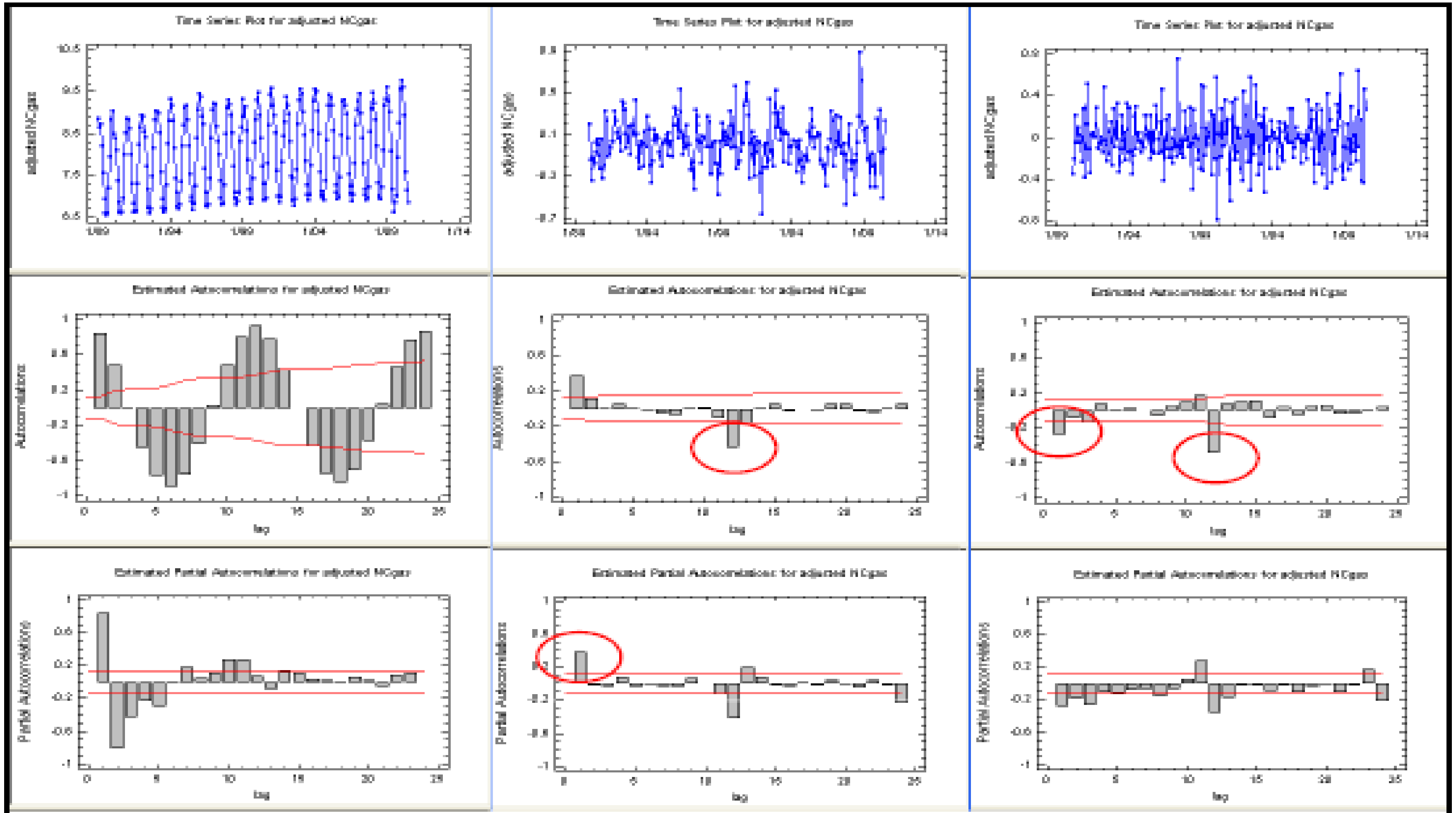
- Start by trying various combinations of one seasonal difference and/or one non-seasonal difference to stationarize the series and remove gross features of seasonal pattern.

- If the seasonal pattern is strong and stable, you MUST use a seasonal difference (otherwise it will "die out" in long-term forecasts)

Model-fitting steps, continued

- After differencing, inspect the ACF and PACF at multiples of the seasonal period (s):
  - Positive spikes in ACF at lag s, 2s, 3s…, single positive spike in PACF at lag
  - s =>SAR=1
  - Negative spike in ACF at lag s, negative spikes in PACF at lags
  - s, 2s, 3s,… => SMA=1
  - SMA=1 often works well in conjunction with a seasonal difference.
  - Same principles as for non-seasonal models, except focused on what happens at multiples of lag s in ACF and PACF.

Model-fitting steps, continued



Original series: nonstationary   Seas. diff: need AR(1) & SMA(1)   Both diff: need MA(1) & SMA(1)

A common seasonal ARIMA model

- Often you find that the “correct” order of differencing is d=1 and D=1.
- With one difference of each type, the autocorr. often negative at both lag 1 and lag s.
- This suggests an SARIMA(0,1,1)(0,1,1) model, a common seasonal ARIMA model.
- Similar to Winters’ model in estimating time-varying trend and time-varying seasonal pattern

Another common seasonal ARIMA model

- Often with D=1 (only) you see a borderline nonstationary pattern with AR(p) signature, where p=1 or 2, sometimes 3
- After adding AR=1, 2, or 3, you may find negative autocorrelation at
- lag s (=> SMA=1)
- This suggests ARIMA(p,0,0)x(0,1,1)+c, another common seasonal ARIMA model.
- Key difference from previous model: assumes a constant annual trend

Bottom-line suggestion

- When fitting a time series with a strong seasonal pattern, you generally should try

  ARIMA(0,1,q)(0,1,1) model (q=1 or 2)

  ARIMA(p,0,0)(0,1,1)+c model (p=1, 2 or 3)

  … in addition to other models (e.g., Random Walk, Single Exponential Smoothing, etc., with seasonal adjustment; or Winters Method)

- If there is a significant trend and/or the seasonal pattern is multiplicative, you should also try a natural log transformation.

Seasonal differencing

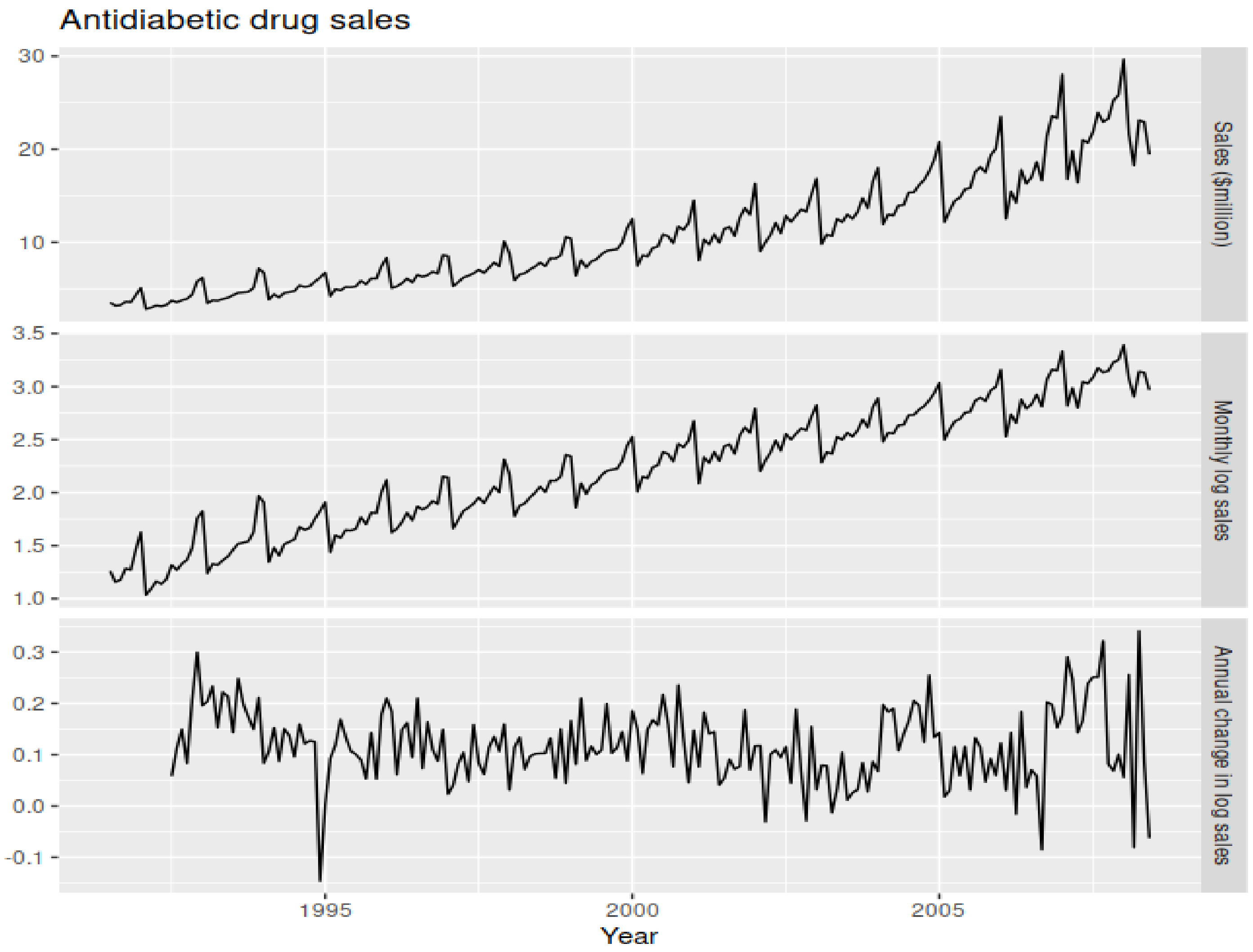


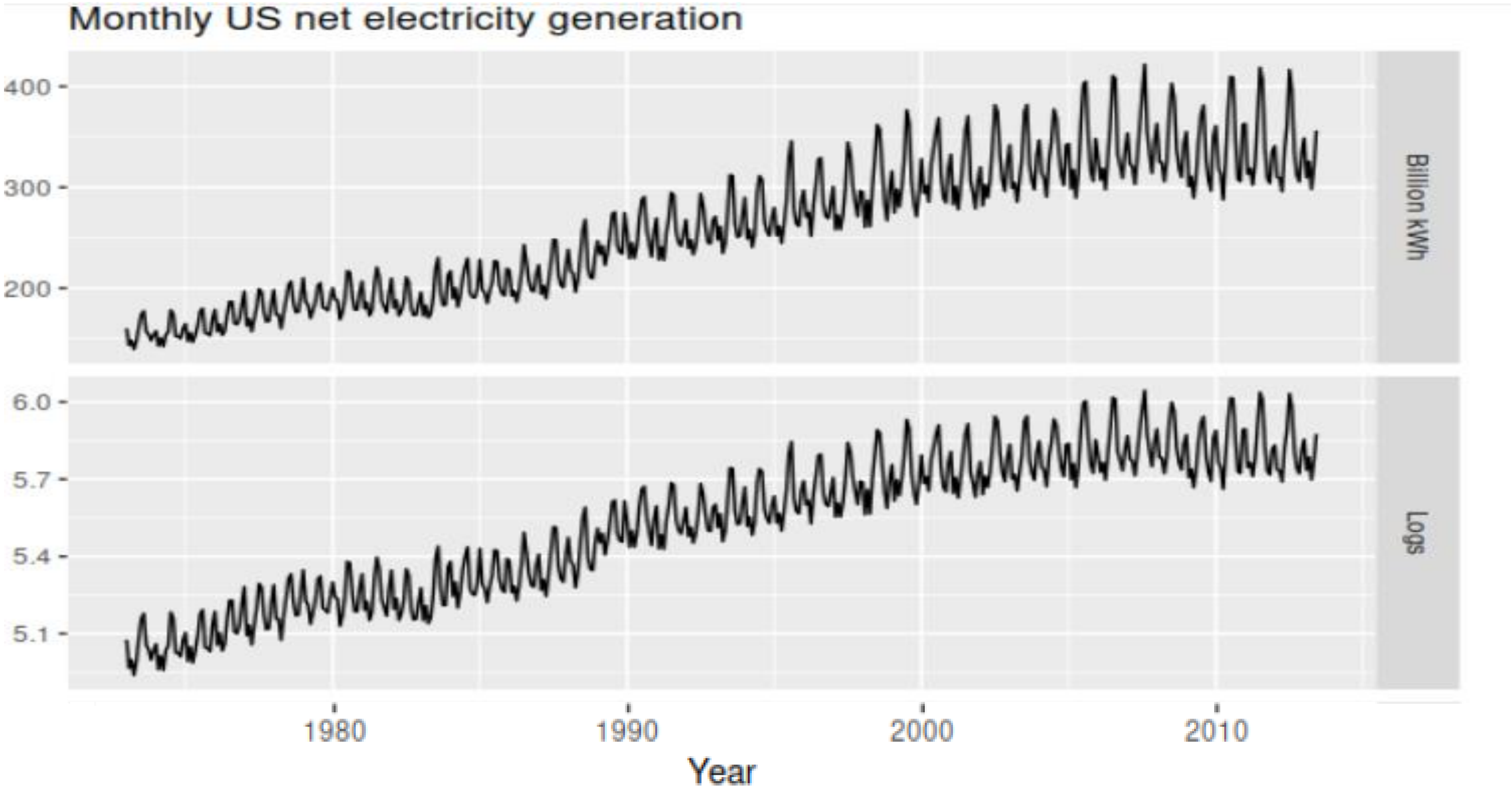
Figure 11: Logs and seasonal differences of the A10 (antidiabetic) sales data.

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

Seasonal differencing

- The bottom panel in Figure11, shows the seasonal differences of the logarithm of the monthly scripts for A10 (antidiabetic) drugs sold in Australia.
- The transformation and differencing have made the series look relatively stationary.
- The logarithms stabilise the variance, while the seasonal differences remove the seasonality and trend.
- To distinguish seasonal differences from ordinary differences, we refer to ordinary differences as “first differences”, meaning differences at lag 1.

Seasonal differencing



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

Figure 12: Monthly Net Electricity Generation in the US

Seasonal differencing

- Sometimes it is necessary to take both a seasonal difference and a first difference to obtain stationary data, as is shown in Figure 12.

- Here, the data are first transformed using logarithms (second panel), then seasonal differences are calculated (third panel).

- The data still seem somewhat non-stationary, and so a further lot of first differences are computed (bottom panel).

- Figure 12: Top panel: US net electricity generation (billion kWh).
- Other panels show the same data after transforming and differencing.

DATA ANALYTICS

Seasonal differencing

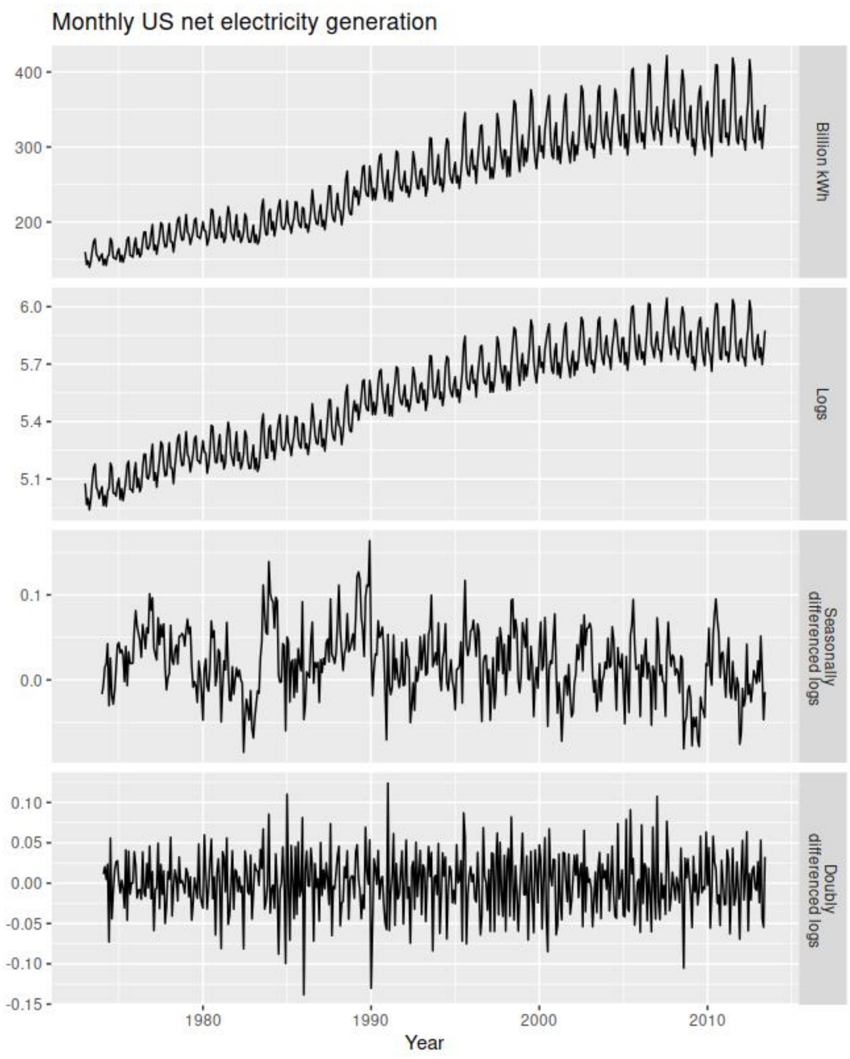


Figure 12: Monthly Net

Electricity Generation in the US

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

Take-aways

- Seasonal ARIMA models (especially the (0,1,*q*)x(0,1,1) and
- (*p*,0,0)x(0,1,1)+c models) compare favorably with other seasonal models and often yield better short-term forecasts.

- Advantages: solid underlying theory, stable estimation of time-varying trends and seasonal patterns, relatively few parameters.

- Drawbacks: no explicit seasonal indices, hard to interpret coefficients or explain “how the model works”, danger of overfitting or mis-identification if not used with care.

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

DATA ANALYTICS

Unit 3: Ljung Box and Theil's coefficient

Jyothi R.
Department of Computer Science and Engineering

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

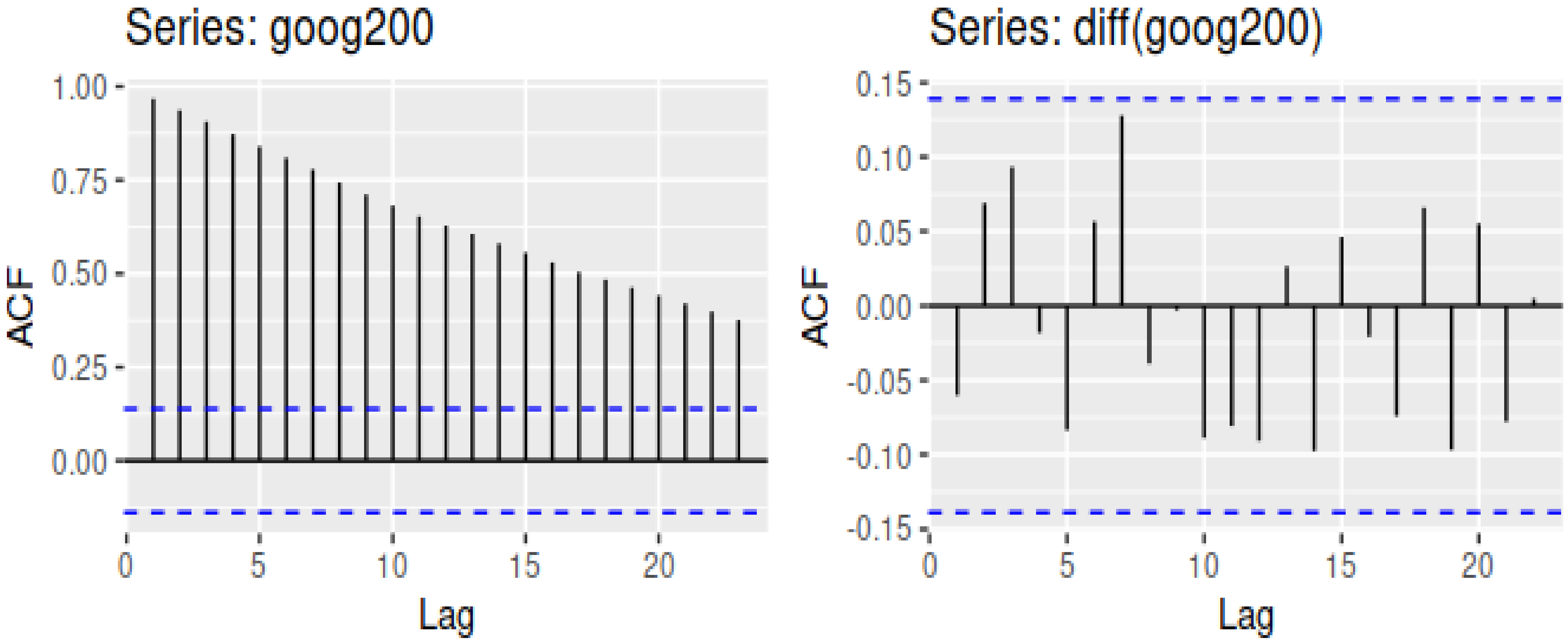


Figure 10:The ACF of the Google stock price (left) and of the daily changes in Google stock price (right).

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, and
- The Ljung -Box $Q*$statistic has a *p*-value of 0.355 (for h=10).
- This suggests that the *daily change* in the Google stock price is essentially a random amount which is uncorrelated with that of previous days.

Ljung-Box Test for Auto-Correlations

- Ljung−Box is a test of lack of fit of the forecasting model and checks whether the auto-correlations for the errors are different from zero.
- The null and alternative hypotheses are given by
- H_0: The model does not show lack of fit
- H_0: The model exhibits lack of fit

Ljung-Box Test for Auto-Correlations

- The Ljung−Box statistic (*Q*-Statistic) is given by (Ljung and Box, 1978)

$$Q(m) = n(n+2)\sum_{k=1}^{m}\frac{\rho_k^2}{n-k}$$

- where *n* is the number of observations in the time series,
- *k* is the number of lag,
- ρ_k is the auto-correlation of lag *k*, and
- *m* is the total number of lags.

Ljung-Box Test for Auto-Correlations

- *Q*-statistic is an approximate chi-square distribution with $m - p - q$ degrees of freedom where *p* and *q* are the AR and MA lags.
- The *Q*-statistic for ARIMA(1, 1, 1) is 10.216 (Table 1) and the corresponding *p*-value is 0.855 and thus we fail to reject the null hypothesis.
- Table 1: ARIMA (1, 1, 1) model summary for Omelette demand

Model	Model Fit Statistics			Ljung–Box $Q(18)$		
	R-Squared	RMSE	MAPE	Statistics	*Df*	Sig.
Omellette-Model_1	0.584	3.439	20.830	10.216	16	0.855

- $Q(m)$ measures accumulated auto-correlation up to lag *m*.

POWER OF FORECASTING MODEL: THEIL'S COEFFICIENT

- The power of forecasting model is a comparison between
- Naive forecasting model and the model developed.
- In the Naive forecasting model, the forecasted value for the next period is same as the last period's actual value
- $F_{t+1} = Y_t$.

- Theil's coefficient (*U*-statistic) is given by (Theil, 1965)

POWER OF FORECASTING MODEL: THEIL'S COEFFICIENT

- For the data shown in Table 213.14 (demand for avionic system spares),
- The *U*-statistic calculations are shown in Table 3.
- TABLE 3:U-statistic calculation

Day	Y_t	ARMA (1,2) Forecast	$(Y_t - F_t)^2$	Naïve Forecast ($F_{t+1} = Y_t$)	$(Y_t - F_t)^2$
31	503	464.8107	1458.423	443	3600
32	688	378.5341	95769.15	503	34225
33	602	444.6372	24763.04	688	7396
34	629	685.8851	3235.909	602	729
35	823	743.5124	6318.281	629	37636
36	671	630.7183	1622.614	823	23104
37	487	649.3491	26357.22	671	33856
		Total	159524.6	Total	140546

POWER OF FORECASTING MODEL: THEIL'S COEFFICIENT

- Theil's coefficient (*U*-statistic) is given by (Theil, 1965)

$$U = \frac{\sum_{t=1}^{n} (Y_{t+1} - F_{t+1})^2}{\sum_{t=1}^{n} (Y_{t+1} - Y_t)^2}$$

- Theil's coefficient is the ratio of the mean squared error of the forecasting model to the MSE of the Naïve model.
- The value of $U < 1$ indicates that forecasting model is better than the Naive forecasting model.
- $U > 1$ indicates that the forecasting model is not better than Naive model.

POWER OF FORECASTING MODEL: THEIL'S COEFFICIENT

Table 2: Monthly demand (quantity of 200 gram packets) along with average price

Period	Month	Demand in Units	Average Price	Period	Demand in Units	Average Price
1	January	10500472	37	25	10658309	36
2	February	10123572	34	26	8677622	38
3	March	7372141	36	27	7330354	37
4	April	7764303	38	28	8115471	37
5	May	6904463	40	29	8481936	34
6	June	10068862	34	30	8778999	37
7	July	6436190	40	31	10145039	32
8	August	9898436	34	32	8497839	38
9	September	6803825	39	33	8792138	34
10	October	8333787	36	34	8485358	36
11	November	7541964	39	35	8575904	36
12	December	8540662	37	36	9885156	32
13	January	10229437	37	37	11023467	35
14	February	8453201	38	38	7942451	40
15	March	7997459	35	39	12492798	32
16	April	8557825	35	40	9756258	32
17	May	7818397	36	41	8992741	32
18	June	8944499	37	42	7397807	40
19	July	8904086	36	43	9710611	32
20	August	8463682	39	44	8328379	39
21	September	7723957	37	45	11873063	32
22	October	7731422	39	46	10642507	32
23	November	8441834	35	47	10635075	32
24	December	7485122	40	48	10578547	32

POWER OF FORECASTING MODEL: THEIL'S COEFFICIENT

- For the data shown earlier on the demand for avionic system spares, the *U*-statistic calculations are shown in the Table below:

Day	Y_t	ARMA (1,2) Forecast	$(Y_t - F_t)^2$	Naïve Forecast ($F_{t+1} = Y_t$)	$(Y_t - F_t)^2$
31	503	464.8107	1458.423	443	3600
32	688	378.5341	95769.15	503	34225
33	602	444.6372	24763.04	688	7396
34	629	685.8851	3235.909	602	729
35	823	743.5124	6318.281	629	37636
36	671	630.7183	1622.614	823	23104
37	487	649.3491	26357.22	671	33856
		Total	159524.6	Total	140546

The *U*-statistic value = 159524.6 / 140546 = 1.1350.

That is, ARMA(1, 2) model is not better thanNaive forecasting.

The ‘X’ Factor (ARX, ARIMAX, etc.)

- ‘X’ = exogenous variables or explanatory variables

Important considerations:

- What other factors influence the forecast?
- How do we process this additional data to make it amenable for inclusion in our model?

Practice Quiz

1. Seasonality in time-series data is caused due to

(a) Changes in macro-economic factors such as recession, unemployment, and so on

(b) Festivals and customs in a society

(c) Random events that occur over a period of time

(d) Changes in customer behaviour driven by new products and promotions

Practice Quiz

2. In a simple exponential smoothing method, the low value of smoothing constant *α* is chosen when

(a) The data has high fluctuations around the trend line

(b) There is seasonality in the data

(c) The data is smooth with low fluctuations

(d) There are variations in the data due to cyclical component

Practice Quiz

3. White noise is

(a) Uncorrelated errors with expected value 0.

(b) Uncorrelated errors that are constant and do not change with time.

(c) Uncorrelated errors that follow normal distribution with mean 0 and constant standard deviation

(d) Errors that follow normal distribution with constant mean and standard deviation

Practice Quiz

4. A stationary process in a time series is a process for which

(a) Mean and variance are constant at different time points

(b) The time series follows normal distribution with zero mean and constant standard deviation

(c) The covariance of the time series depends only on the lag

(d) Mean and standard deviation are constant at different time points and the covariance depends only on the lag between the values and is constant for a given lag

References

Text Book:

“Business Analytics, The Science of Data-Driven Making”, U. Dinesh Kumar, Wiley 2017 Ch. 13.14.5 and 13.15

Image Courtesy

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

THANK YOU

Jyothi R
Assistant Professor, Department of Computer Science
jyothir@pes.edu

DATA ANALYTICS

Unit 3: Spectral Analysis of Time Series

Jyothi R.
Department of Computer Science and Engineering

Discrete Fourier Transform of the Time Series

- What if we are less interested in how our underlying process evolves in time and are more interested in the variance of the time series at certain frequencies?

- We may attempt to apply a Fourier transform to the data. For our time series, $x_1, \ldots, x_n$, the discrete Fourier transform would be

where $\omega_j = 0, 1/n, \ldots, (n-1)/n$.

Interpreting DFT and Another Representation

- Note that we can break up $d(\omega_j)$ into two parts

- which we can write as a cosine component and a sine component

$$d(\omega_j) = d_c(\omega_j) - i d_s(\omega_j)$$

The Periodogram

- The Periodogram is defined as

$$I(\omega_j) = |d(\omega_j)|^2 = d_c^2(\omega_j) + d_c^2(\omega_j)$$

- If there is no periodic trend in the data, then $E[d(\omega_j)] = 0$, and the Periodogram expresses the variance of x_t at frequency ω_j.

- If a periodic trend exists in the data, then $E[d(\omega_j)]$ will be the contribution to the periodic trend at the frequency ω_j.

The Periodogram

- What are we trying to estimate with the Periodogram?
- We can use the Periodogram to find periodic trends in the data.
- Is there information left in the Periodogram after the trend is removed?
- Assuming that we have a stationary time series, what does the Periodogram estimate?

The Spectral Density

- The spectral density is the Fourier transform of the auto covariance function

$$f(\omega) = \sum_{h=-\infty}^{h=\infty} e^{-2\pi i \omega h} \gamma(h)$$

- for $\omega \in (-0.5, 0.5)$. Note that this is a population quantity. (i.e. This is a constant quantity defined by the model.)

Moving Average

- A simple way to improve our estimates is to use a moving average smoothing technique

$$\hat{f}(\omega_j) = \frac{1}{2m+1} \sum_{k=-m}^{m} I(\omega_{j-k})$$

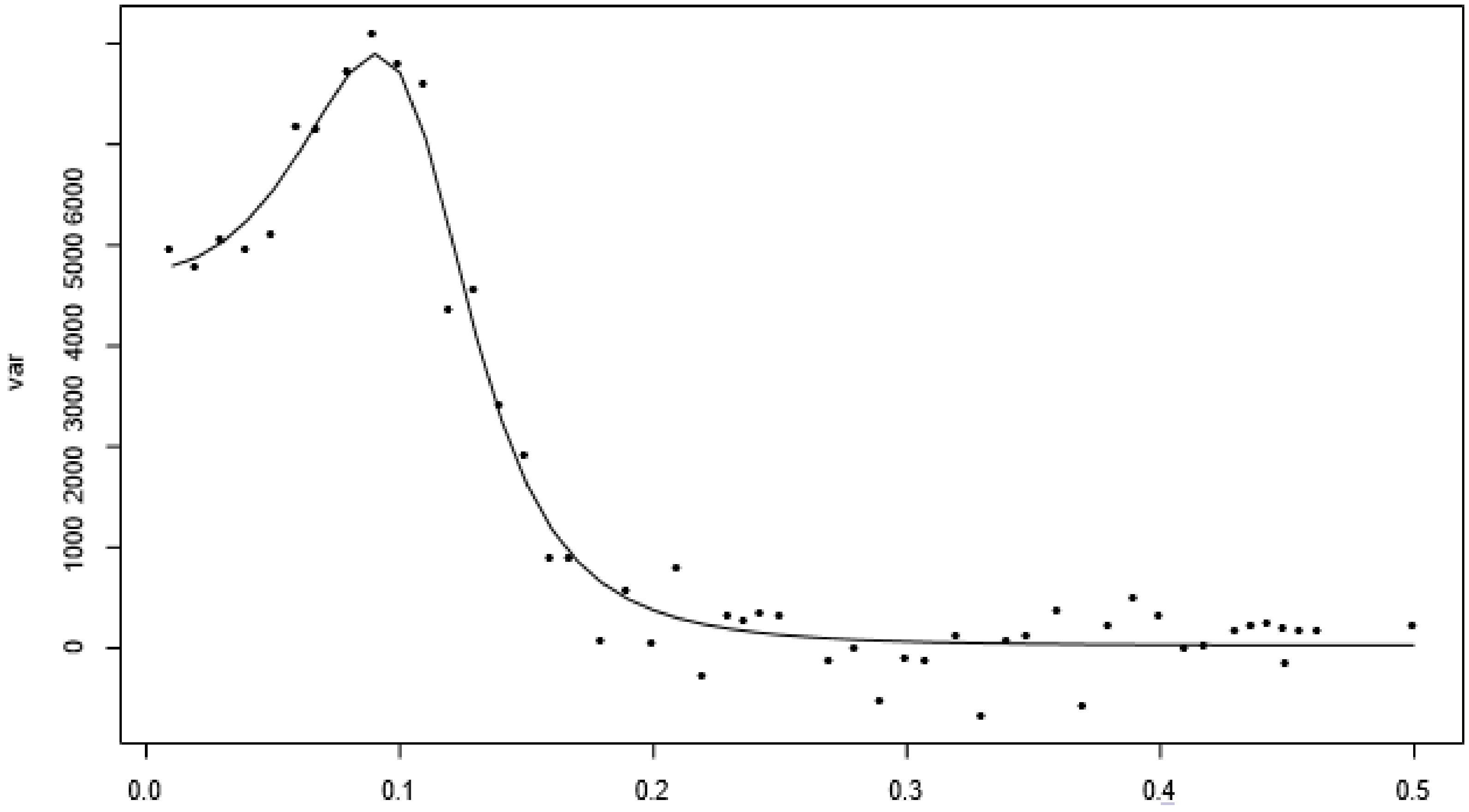
- We can also iterate this procedure of uniform weighting to be more weight on closer observations.

$$\hat{u}_t = \frac{1}{3} u_{t-1} + \frac{1}{3} u_t + \frac{1}{3} u_{t+1}$$

- Then, we iterate.
- Then, substitute to obtain better weights.

$$\hat{\hat{u}}_t = \frac{1}{3} \hat{u}_{t-1} + \frac{1}{3} \hat{u}_t + \frac{1}{3} \hat{u}_{t+1}$$

Moving Average



Smoothing Summary

- Smoothing decreases variance by averaging over the Periodogram of neighboring frequencies.
- Smoothing introduces bias because the expectation of neighboring Periodogram values are similar but not identical to the frequency of interest.
- Beware of over smoothing!

Tapering

- Tapering corrects bias introduced from the finiteness of the data.
- The expected value of the Periodogram at a certain frequency is not quite equal to the spectral density.
- It is affected by the spectral density at neighboring frequency points.
- For a spectral density which is more dynamic, more tapering is required.

Why do we need to taper?

- Our theoretical model ..., x_{-1}, x_0, x_1, ... consists of a doubly infinite time series
- We could think of our data, y_t as the following transformation of the model
- $y_t = h_t x_t$
- where $h_t = 1$ for $t = 1, ..., n$ and zero otherwise. This has repercussions on the expectation of the Periodogram of our data.

$$E[I_y(\omega_j)] = \int_{-0.5}^{0.5} W_n(\omega_j - \omega) f_x(\omega) d\omega$$

- where $W_n(\omega) = |H_n(\omega)|^2$ and $H_n(\omega)$ is the Fourier transform of the sequence h_t.

The Taper

Specifically,

$$H_n(\omega) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} h_t e^{-2\pi i \omega t}$$

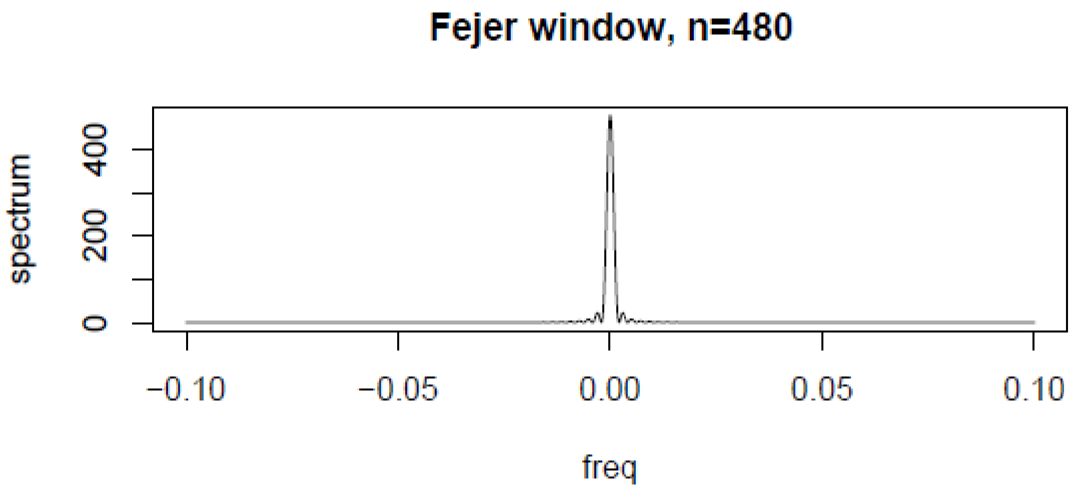
When we put in the h_t above, we obtain a spectral window of

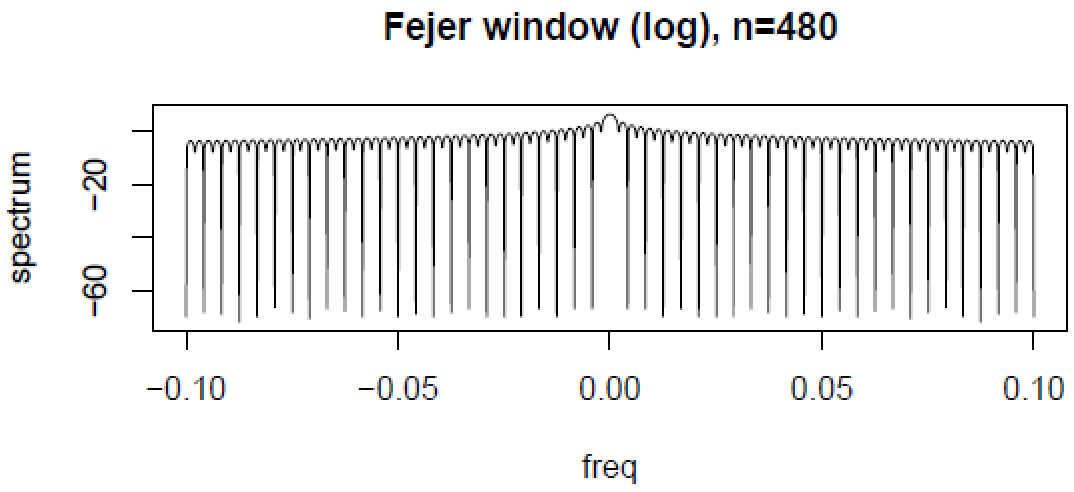
$$W_n(\omega) = \frac{\sin^2(n 2\pi\omega)}{\sin^2(\pi\omega)}.$$

We set $W_n(0) = n$.

Smoothing and Tapering

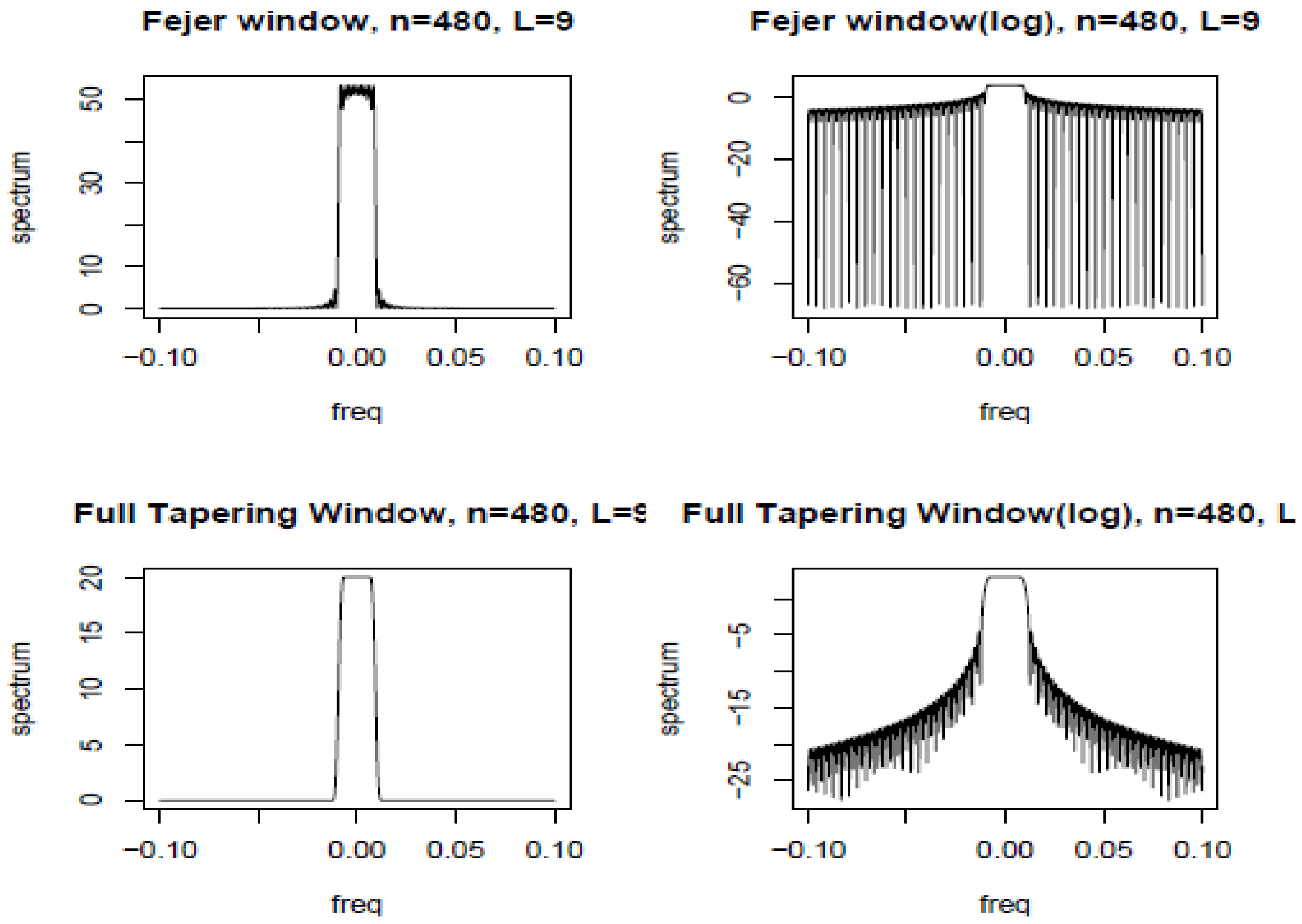
- There are problems with this spectral window, namely there is too much weight on neighboring frequencies (sidelobes).





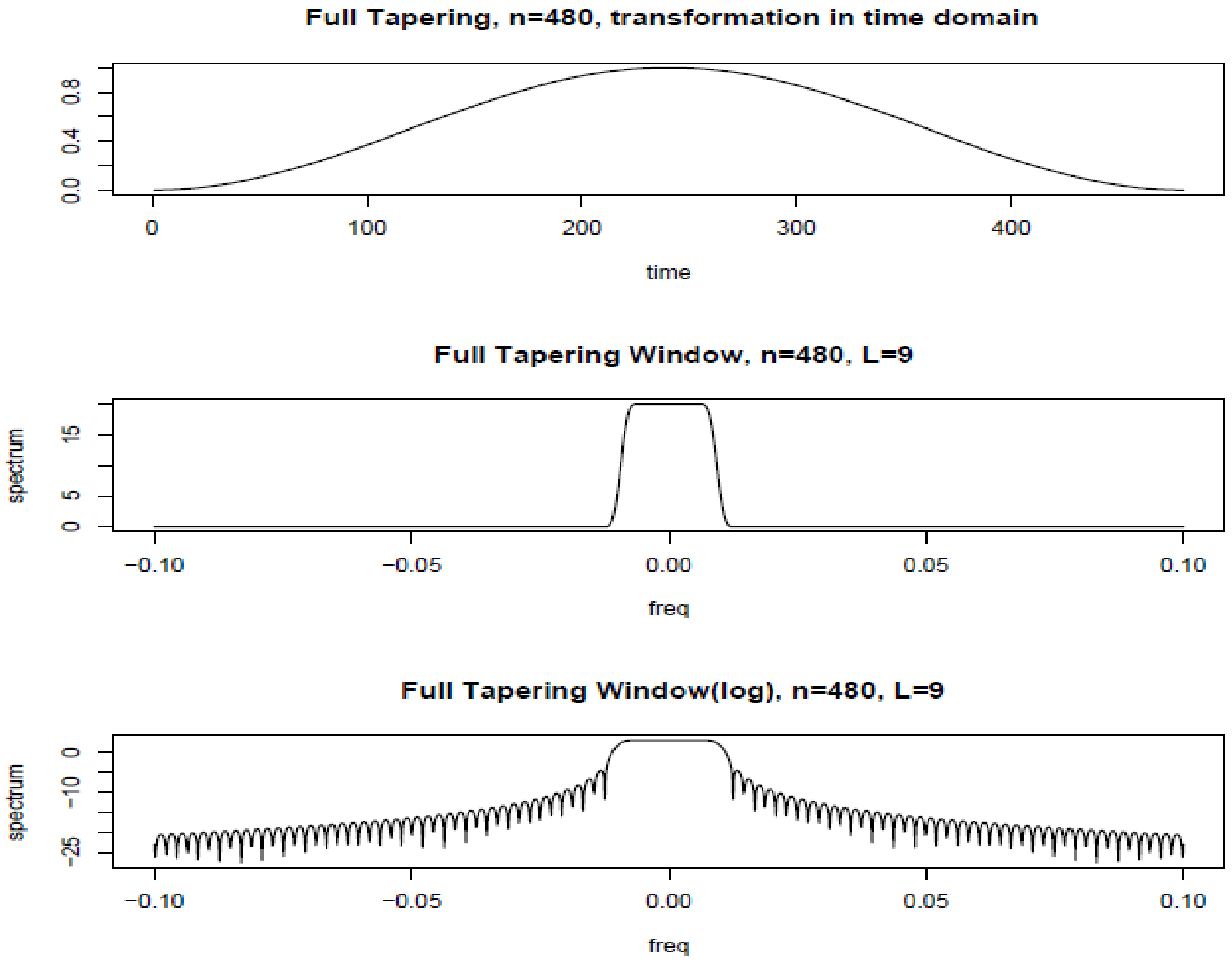
Smoothing and Tapering

One way to fix this is to use a Cosine taper. We select a transform h_t to be

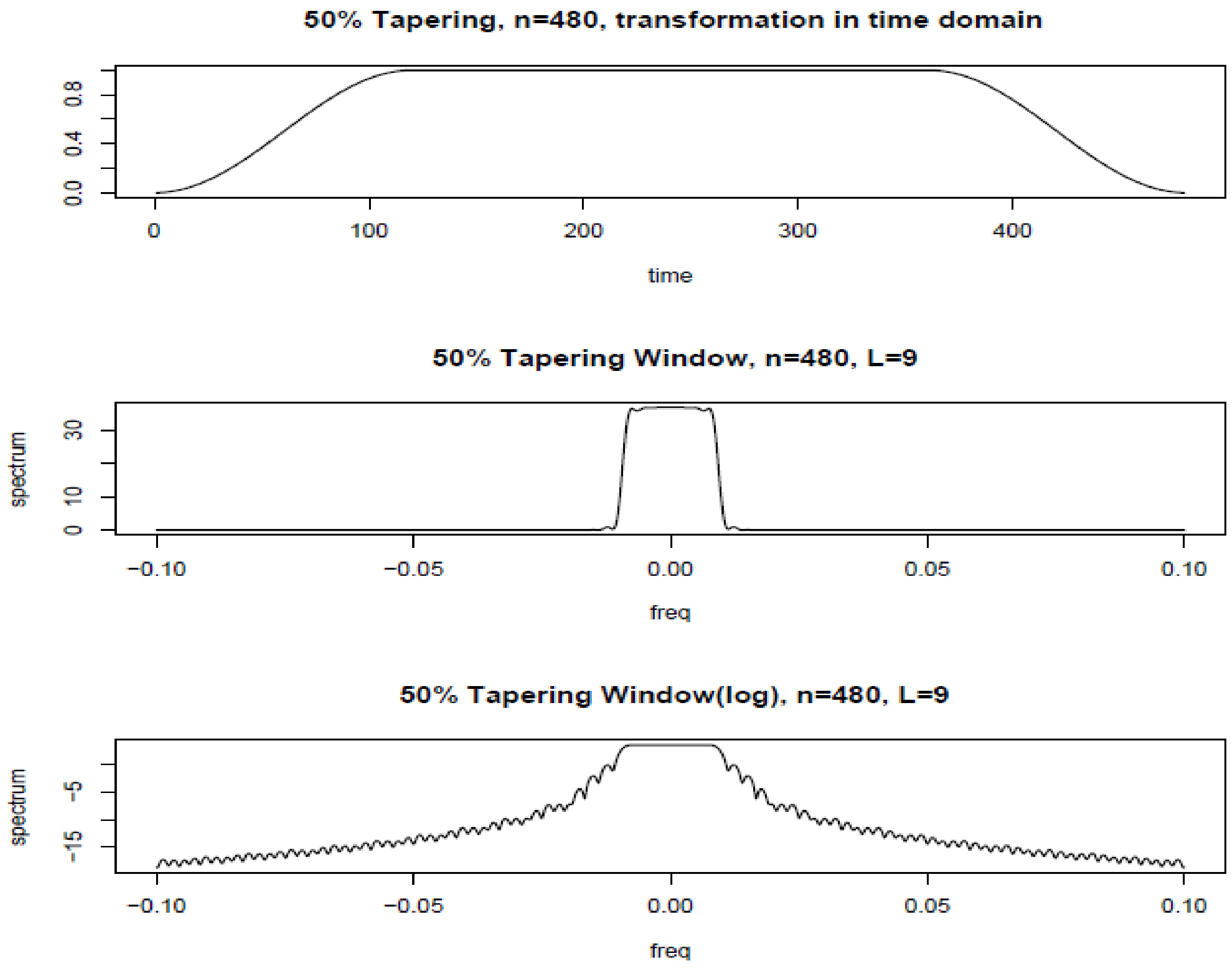


$$h_t = 0.5\left[1 + cos\left(\frac{2\pi(t - \bar{t})}{n}\right)\right]$$

Smoothing and Tapering



Smoothing and Tapering

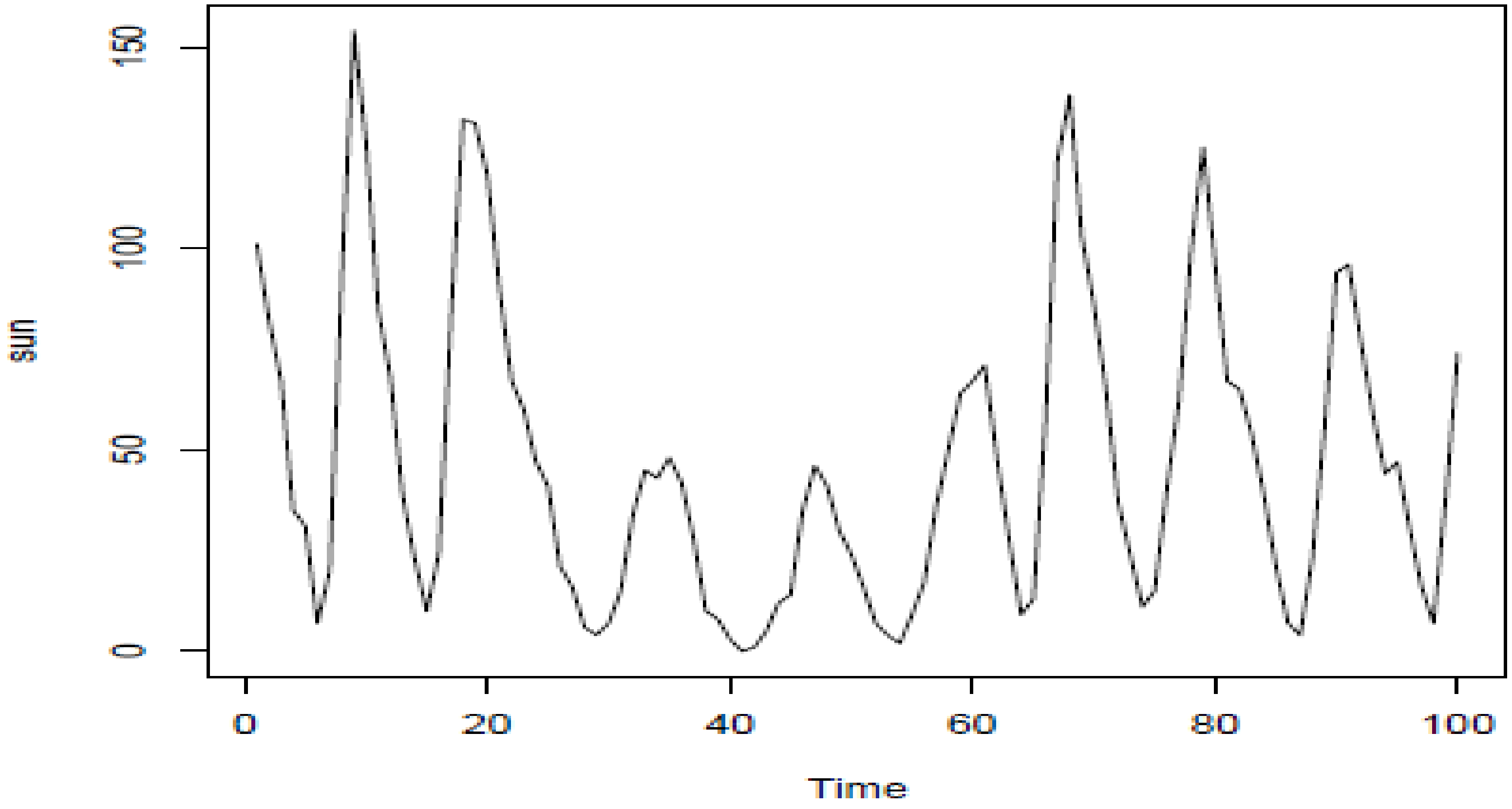


Smoothing and Tapering

- Smoothing introduces bias, but reduces variance.
- Smoothing tries to solve the problem of too many "parameters".
- Tapering decreases bias and introduces variance.
- Tapering attempts to diminish the influence of sidelobes that are introduced via the spectral window.

Examples

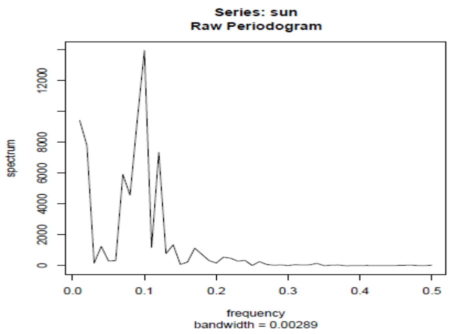
Wolfer sunspots 1770-1869

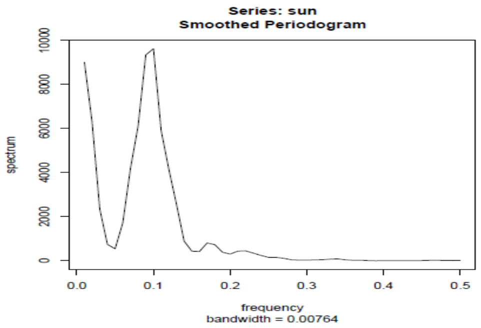


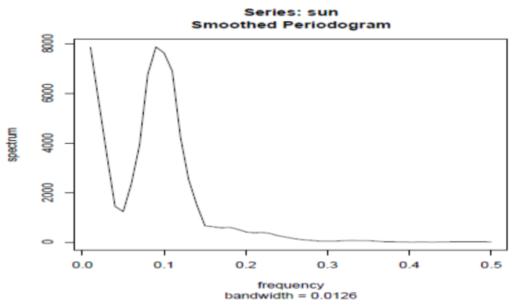
Examples: Smoothed Periodogram with ARMA Spectral Density

Raw Periodogram

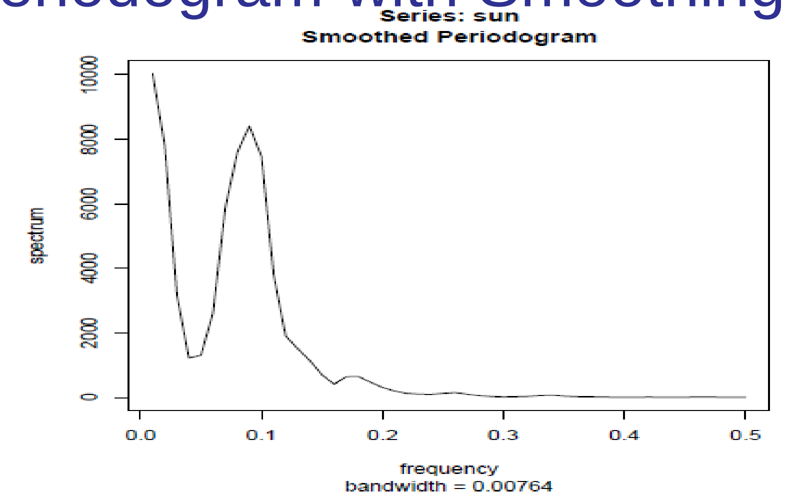
Periodogram with Smoothing Window of 3, 5

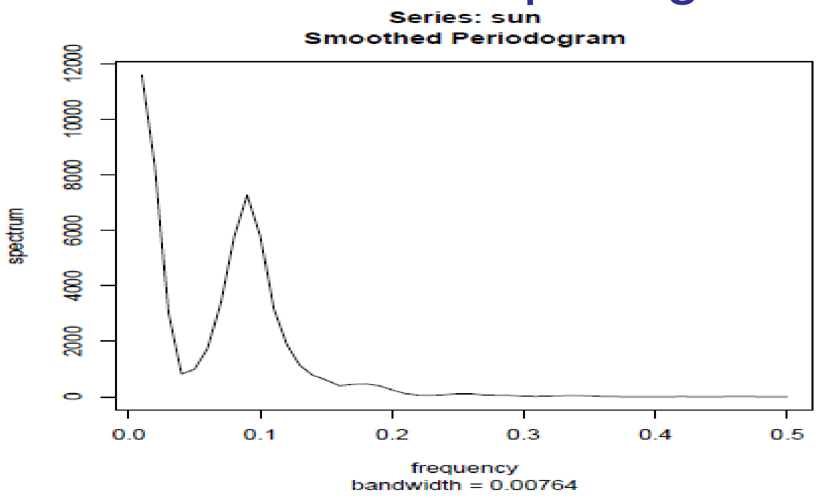






Periodogram with Smoothing Window of 3 with Tapering and more tapering





Wavelets

- We have been using Fourier components as a basis to represent stationary processes and seasonal trends.

- Since we are dealing with finite data, we must use a finite number of terms, and perhaps one could use an alternative basis.

- Wavelets are one option to accomplish this goal. They are particularly well suited to the same situation as Dynamic Fourier analysis.

References

Text Book:

“Business Analytics, The Science of Data-Driven Making”, U. Dinesh Kumar, Wiley 2017

Image Courtesy

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

https://otexts.com/fpp2/stationarity.html
https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Time+Series+Analysis:+The+Basics
https://bookdown.org/rdpeng/timeseriesbook/spectral-analysis.html
https://www.stat.berkeley.edu/~bartlett/courses/153-fall2010/lectures/15.pdf
https://astrostatistics.psu.edu/su07/fricks_2timeseries07.pdf)
https://blog.octo.com/en/time-series-features-extraction-using-fourier-and-wavelet-transforms-on-ecg-data/
https://jmread.github.io/talks/Time_Series_AI.pdf

THANK YOU

Jyothi R
Assistant Professor, Department of Computer Science
jyothir@pes.edu

Practice Quiz

5. In a pure auto-regressive process, AR(*p*), the value of *p* can be identified using

(a) Auto-correlation function

(b) Partial auto-correlation function

(c) Auto-correlation and partial auto-correlation function

(d) Ljung–Box test

Practice Quiz

6. Power of a forecasting model is calculated using

(a) Root mean square error (RMSE)

(b) Theil's coefficient

(c) Mean absolute percentage error (MAPE)

(d) Bayesian information criteria (BIC)

Practice Quiz

7. A necessary condition for accepting a time-series forecasting model is

(a) The residuals should follow a normal distribution

(b) The residuals should be white noise

(c) The residuals should be black noise

(d) The residuals should follow a normal distribution and the *R*-square should be high

Practice Quiz

8. In an ARIMA model, differencing is carried out

(a) To convert a stationary process to a non-stationary process

(b) To convert a non-stationary process to a stationary process

(c) To remove seasonal fluctuations from the data

(d) To remove cyclical fluctuations from the data

Practice Quiz

9. Overall fitness of a forecasting model is checked using

(a) Durbin−Watson Test

(a) Theil coefficient

(c) Ljung−Box test

(d) Dickey−Fuller test

10. Presence of non-stationarity is checked using

(a) Durbin−Watson Test

(b) Theil coefficient

(c) Ljung−Box test

(d) Dickey−Fuller test

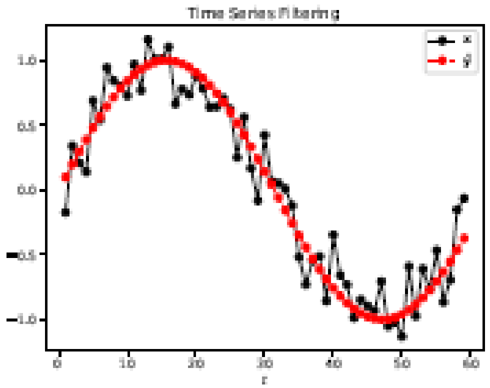
DATA ANALYTICS

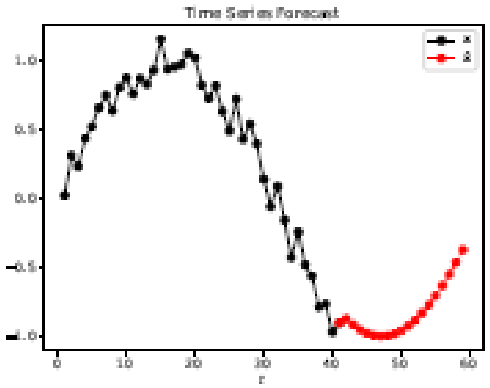
Unit 3: Time Series – Feature Extraction and Classification

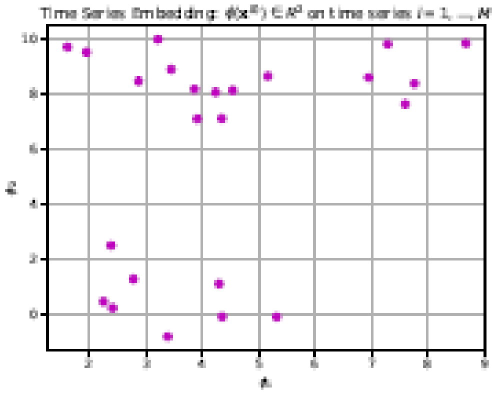
Jyothi R.
Department of Computer Science and Engineering

Time Series Tasks

- Filtering (*estimate*) $\mathbf{y}_1, \ldots, \mathbf{y}_{t-1}, \mathbf{y}_t$ from observations
- $\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}, \mathbf{x}_t$
- Forecasting (*predict*) $\mathbf{x}_{t+1}, \mathbf{x}_{t+2}, \ldots$ from time t. Embedding: Describe a time series $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ as a vector $\boldsymbol{\varphi} = [\varphi_1, \ldots, \varphi_N]$ of fixed length N.

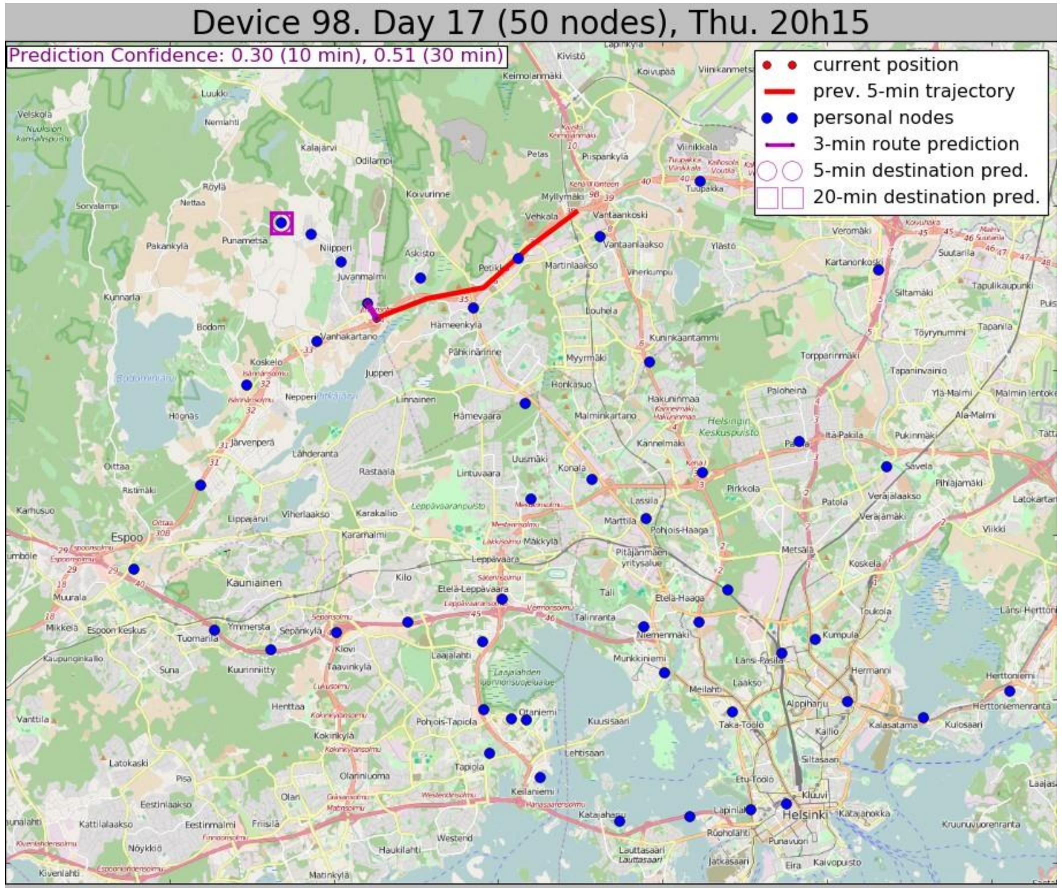






- Clustering
- Classification
- Motif extraction
- Novelty/anomaly detection
- Query by content

Machine Learning for Forecasting



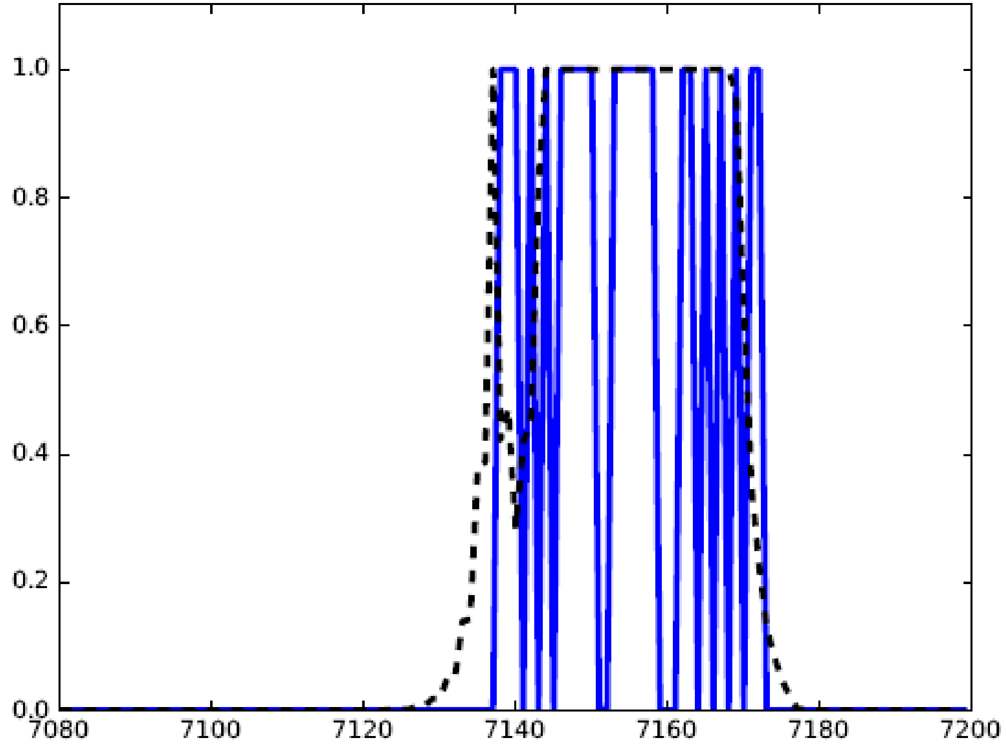
- Collected data of travellers[1]: GPS coordinates, signal strength, battery level, current time, . . .
- Predict future trajectory from current trajectory

[1] All participants volunteered to install App; share data

Work with Jaakko Hollmèn et al. @Aalto University

Example: Predictive Maintenance of Aircraft

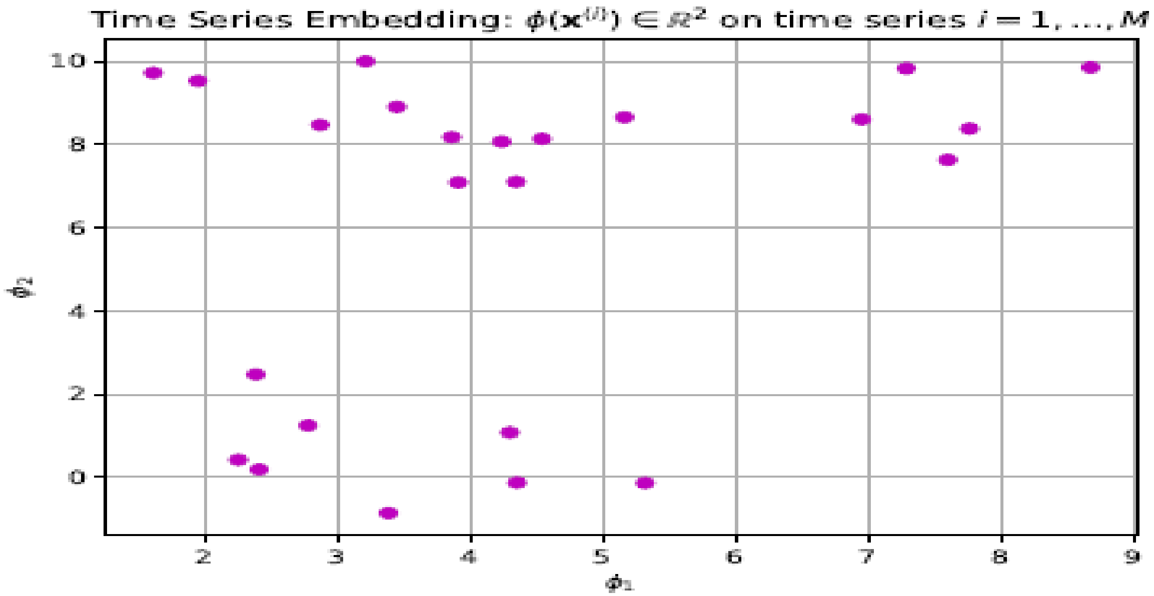
- Sensor readings from aircraft and textual description of observations
- Predict warnings/required replacement of components



Work with DaSciM team @École Polytechnique

Embedding Time Series

We seek to turn variable-length time series $\{\mathbf{x}_1^{(i)}, \ldots, \mathbf{x}_{T_i}^{(i)}\}_{i=1}^{M}$ into fixed-length vectors $\boldsymbol{\varphi}^{(i)} = [\varphi_1, \ldots, \varphi_L]$.



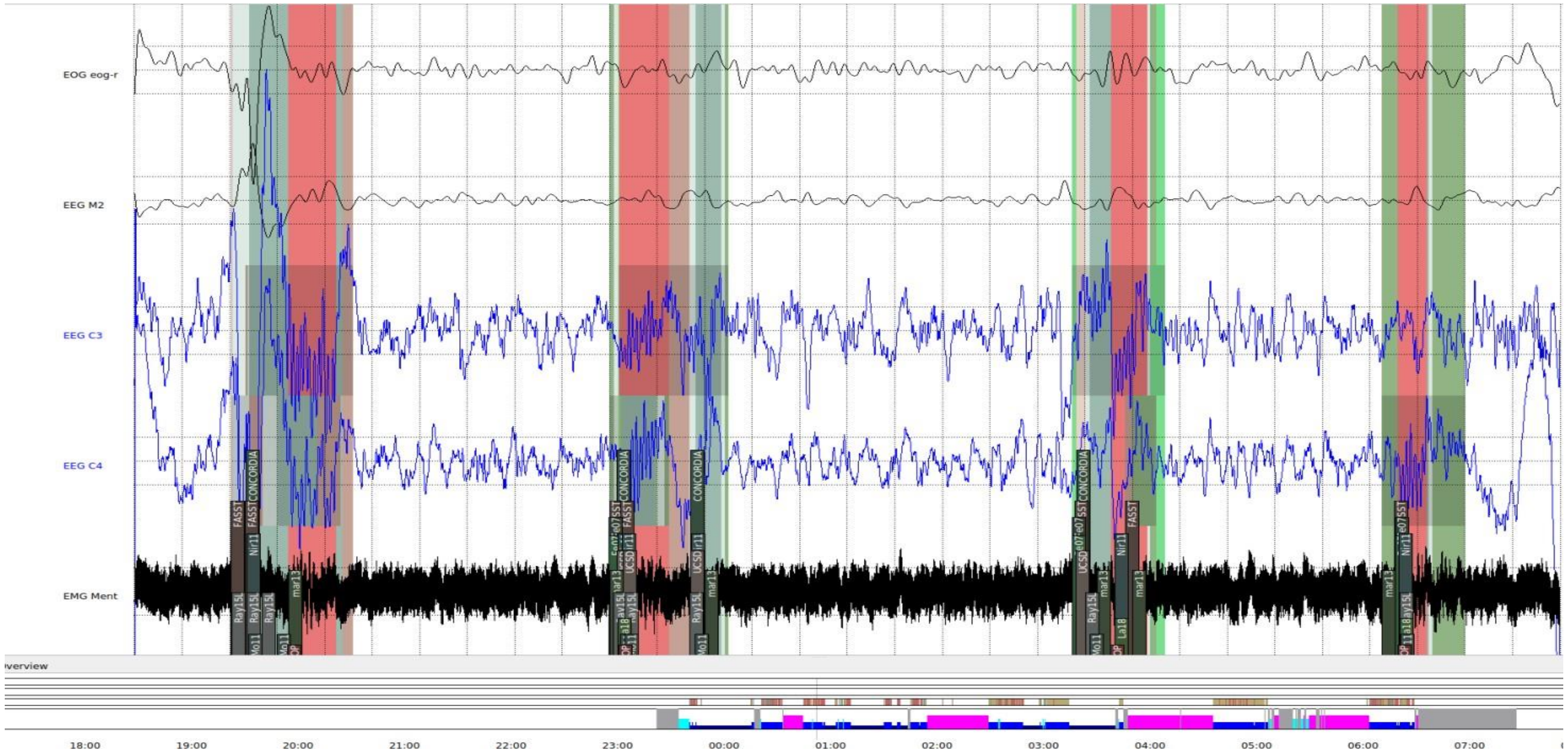
- This lets us compare and cluster time series/look for anomalies, (and classify, if we have the label): measure similarity/distance between $\varphi(\mathbf{x}^{(i)})$ and $\varphi(\mathbf{x}^{(2)})$.

Example: Modelling and Treating Chronic Insomnia

- Goal: (semi-)automate clinical assessment; what kind of insomnia + treatment recommendation.
- Data from patients:
  - Psychological questionnaires (MMPI, CAS) EEG and ECG data overnight
  - Some labels: follow-up tests/questionnaires and *biofeedback* results (some patients found success without pharmaceutical intervention, others not)

- Questionnaire data: can take ‘standard’ machine learning approach, $f : X \rightarrow Y$, and inspect feature importance, statistical correlation wrt to label variable (extent of insomnia, and improvement); cluster into groups, etc.
- Time-series data: different lengths, contains artifacts, subjects fall asleep at different times, ... How to compare?

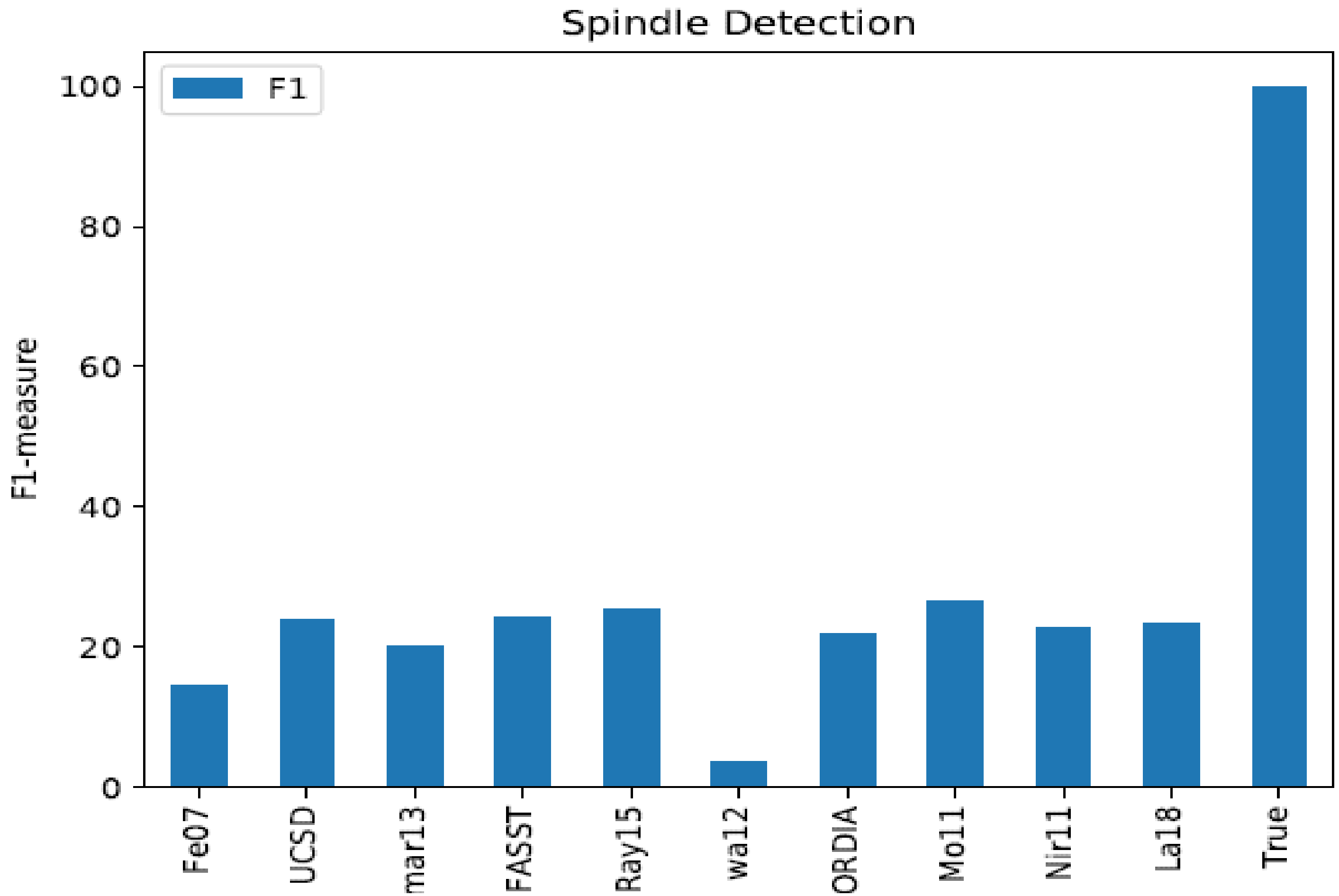
Example: Modelling and Treating Chronic Insomnia



- Certain signals are of interest: Spindles, α-waves, β-waves, . . . Simple embeddings, e.g.,
- $\varphi(\mathbf{x}^{(i)}) = [\texttt{spindles/hour, avg freq of spindle}]$. Detection and labelling by an expert is labour intensive.

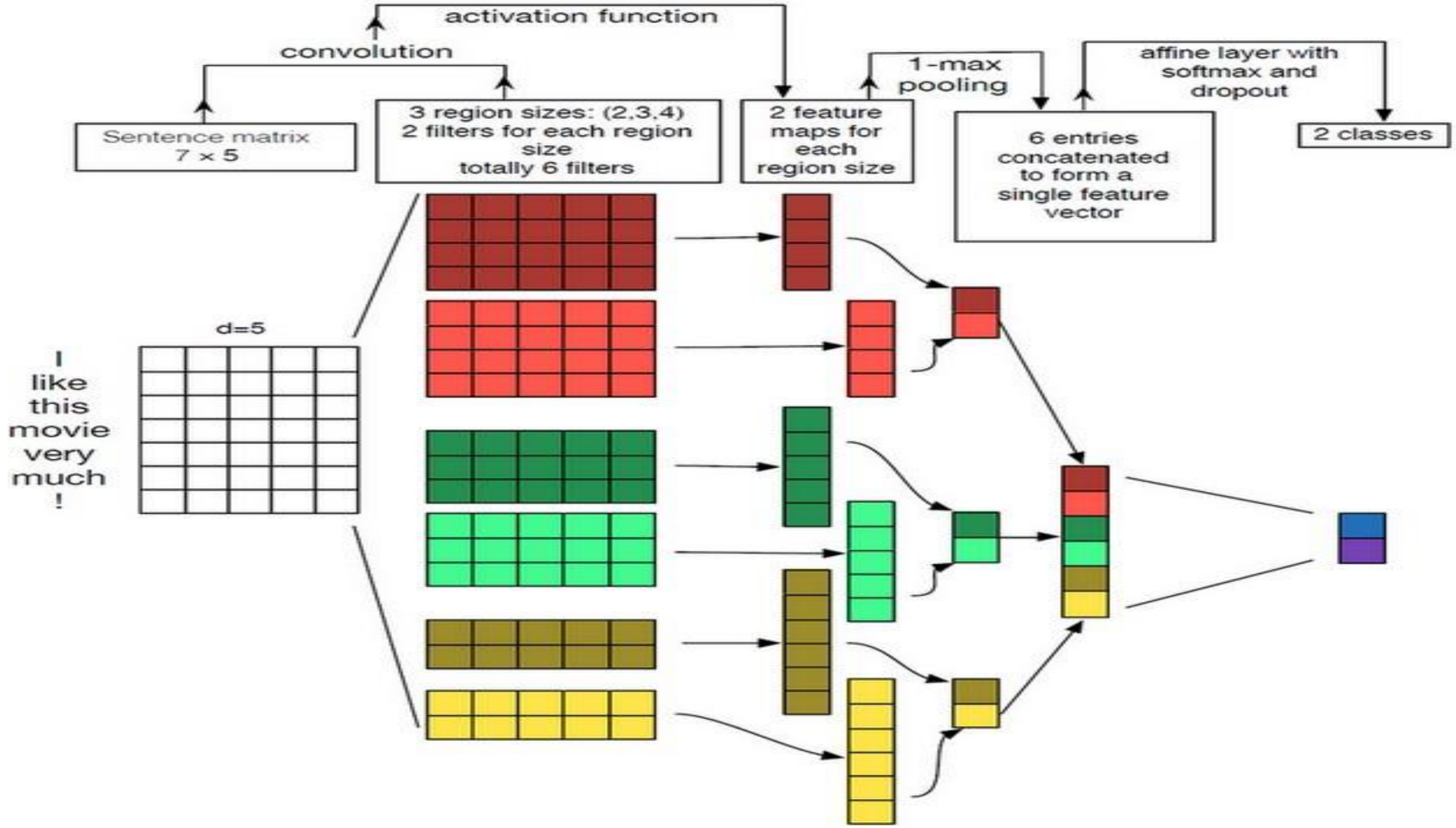
Outline

There exist many rule-based methods, e.g., wavelet analysis But predictive performance is insufficient in many practical settings

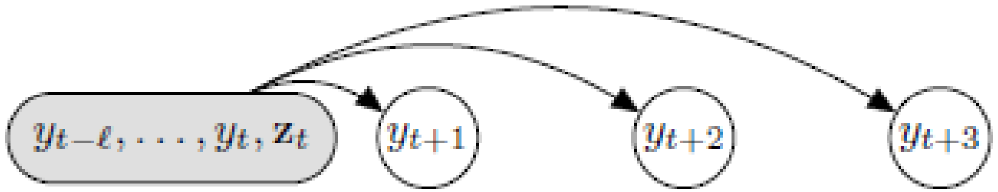


Deep learning

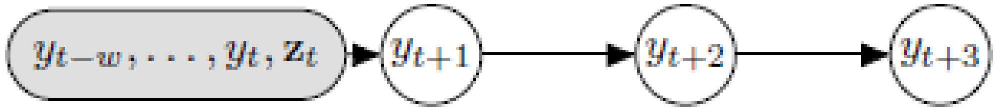
- Many current solutions are inspired by / related to NLP.
- Similar to a ‘simple’ embedding, but more data-driven.



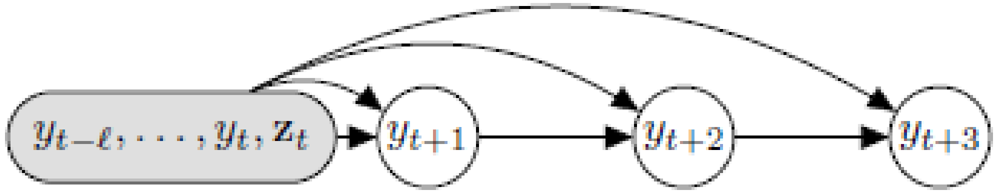
Multi-Step-Ahead Forecasting



Direct

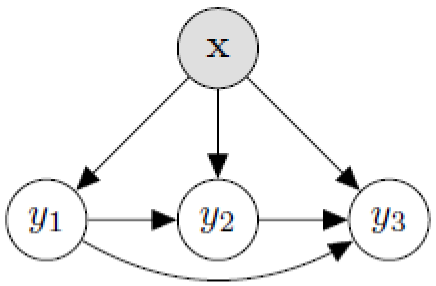


Iterated

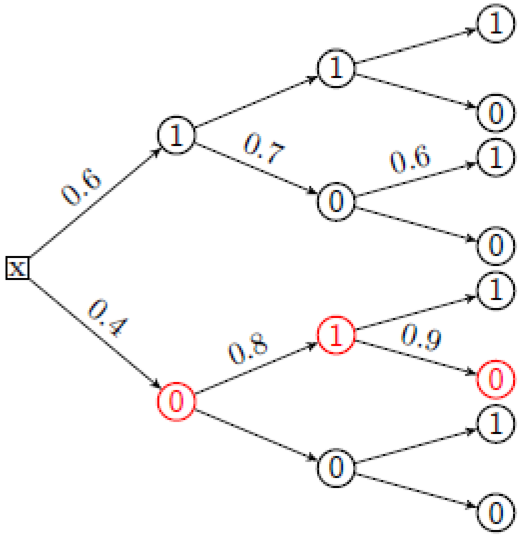


Classifier/Regressor Chain cascade

Classifier Chains



For example, where each $y_t \in \{0, 1\}$



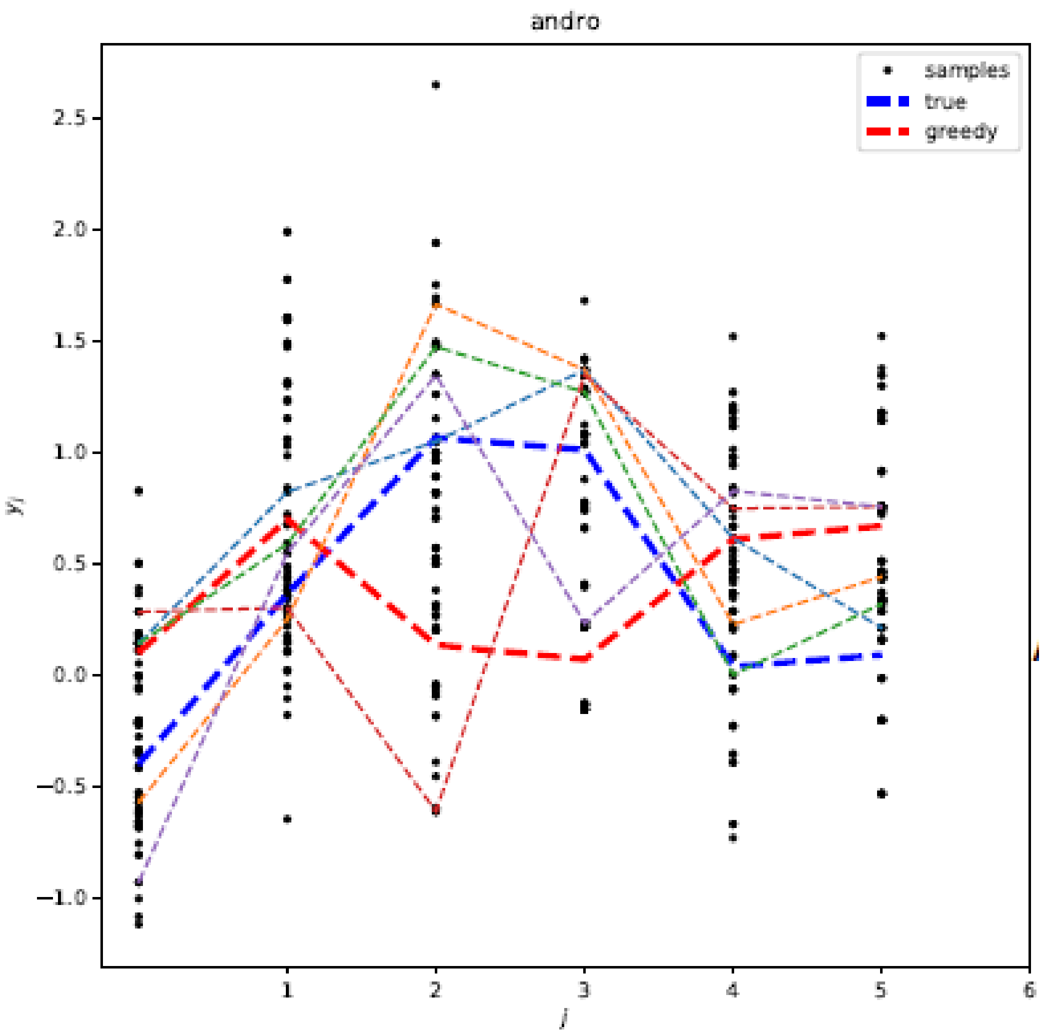
- Predictions become input, across a cascade/chain
- Efficient
- Probabilistic interpretation:

$$P(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T} P(y_t|\mathbf{x}, y_1, \ldots, y_{t-1})$$

$$\hat{\mathbf{y}} = f(\mathbf{x}) = \underset{\mathbf{y}\in\{0,1\}^3}{\operatorname{argmax}}\, P(\mathbf{y}|\mathbf{x})$$

- Search probability tree (for best prediction) with AI-search techniques (Monte-Carlo search, beam search, A* search, ...)
- Explore structure

Regressor Chains



- e.g., where $\mathbf{y} \in R^6$,
  - Sample down the chain
  - $y_{t+1} \sim p(y_{t+1}|y_1, \ldots, y_t, \mathbf{x})$
  - More samples = more hypotheses
  - Consider different loss functions
- Applications:
  - Multi-output regression
    Tracking
  - Forecasting

One-Step Decision Theory

Under uncertainty, we wish to assign $y^* = f^*(\mathbf{x})$, the best label/hypothesis, $y^* \in Y$, given $\mathbf{x} \in R^D$

.Minimizing conditional expected loss

$$f^* = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \underbrace{\sum_{y \in \mathcal{Y}} \ell(f(\mathbf{x}), y) P(y|\mathbf{x})}_{\mathbb{E}_{Y \sim P(Y|\mathbf{x})}[\ell(\hat{y}, Y)|\mathbf{x}]}$$

under loss function A, which describes our preferences. In the case of 0/1 loss (1 if $y \neq \hat{y}$, else 0),

Maximum a Posteriori

$$y^* = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, p(\mathbf{x}|y) P(y) = \underset{y \in \{0,1\}}{\operatorname{argmax}}\, P(y|\mathbf{x})$$

We can estimate P from the training data.

Expected Utility

- An intelligent agent wishes to make a decision to achieve a goal.

  The decision which involves the least risk. Another way of looking at the problem: utility.

  A rational agent maximizes their expected utility, not necessarily a simple *payoff* (e.g., amount of money):

  Expected Utility $$U(y) = \sum_{y \in \mathcal{Y}} u(y)p(y)$$

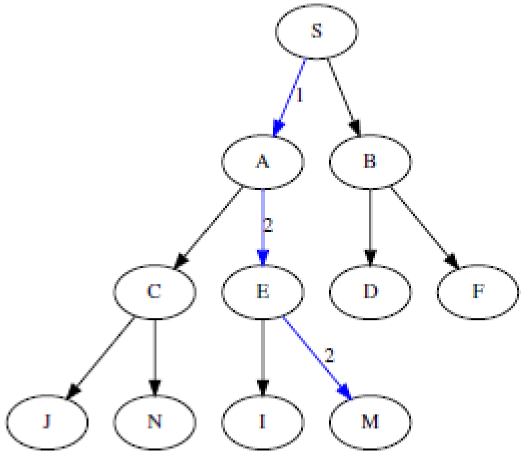
- with satisfaction/utility *u*(*y*) for outcome *y* . Different agents may have different utility functions, even when 'payoff' is the same item. Instead of labels given input, we can deal with actions given evidence and belief.

  - A risk-prone agent will tend to gamble higher stakes A conservative (risk-adverse) agent will not
  - A risk-neutral agent only cares about payoff *y* directly

What about sequential decisions?

In a Deterministic Environment

(e.g., board games – chess, etc.)



The state space, e.g., $s_t \in \{A, B, \ldots, M\}$

An initial state, e.g., $s_0 = S$

A goal state, e.g., $s_t = M$

A set of actions, e.g., $a_t \in \{1, 2\}$

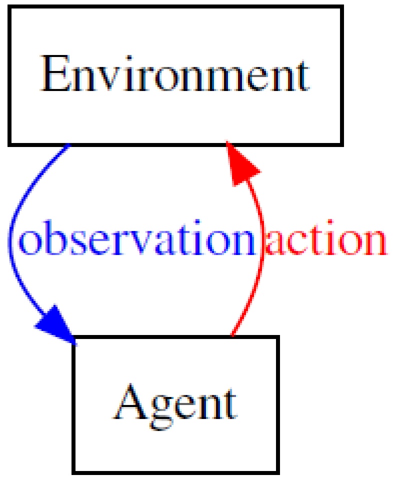
A cost for each branch, e.g., $\mathrm{Cost}(S, A) = 1$

It's just a search! AI-search techniques applicable (DFS, A^*, . . .).

Markov Decision Processes (MDP)

MDPs are models that seek to provide optimal solutions for stocastic sequential decision problems.

$$\text{MDP} = \text{Markov Chain} + \text{One-step Decision Theory}$$



Outline

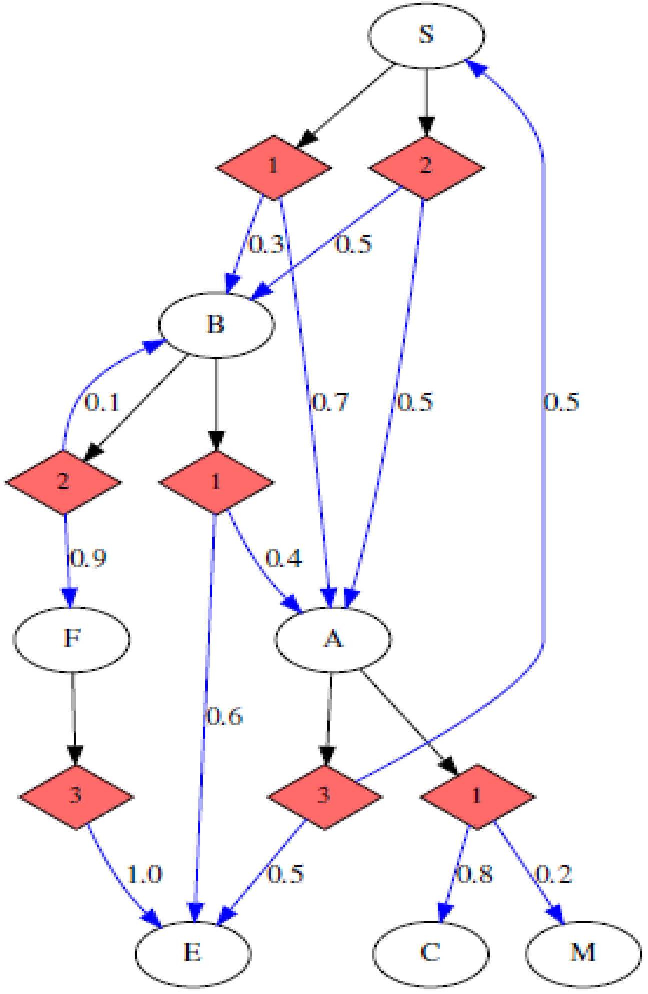
Now we have a model with

- $P(s^j|s, a)$ transition function
- $R(s^j, a, s)$ reward function
- Objective: obtain a policy

$$\pi : \mathcal{S} \mapsto \mathcal{A}$$

which maximizes expected reward:

$$\mathbb{E}[R_0|s_0 = s] = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t(s_t, a_t)\right]$$



solution can be found via dynamic programming!
Just need the model . . .

Reinforcement Learning

- We don’t have the model!

- Don’t have transition/reward functions.
- No input-output training pairs, just reward signal.
- The agent needs to experiment! Exploration vs exploitation. Deep neural net can learn a model
- . . .over millions of iterations. Emerging applications:
  - Gameplay
  - Robotics (usually trained in simulation) Parameter-tuning, etc. (as a tool)
- Transfer learning is promising

Image Courtesy

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

https://otexts.com/fpp2/stationarity.html
https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Time+Series+Analysis:+The+Basics
https://bookdown.org/rdpeng/timeseriesbook/spectral-analysis.html
https://www.stat.berkeley.edu/~bartlett/courses/153-fall2010/lectures/15.pdf
https://astrostatistics.psu.edu/su07/fricks_2timeseries07.pdf)
https://blog.octo.com/en/time-series-features-extraction-using-fourier-and-wavelet-transforms-on-ecg-data/
https://jmread.github.io/talks/Time_Series_AI.pdf