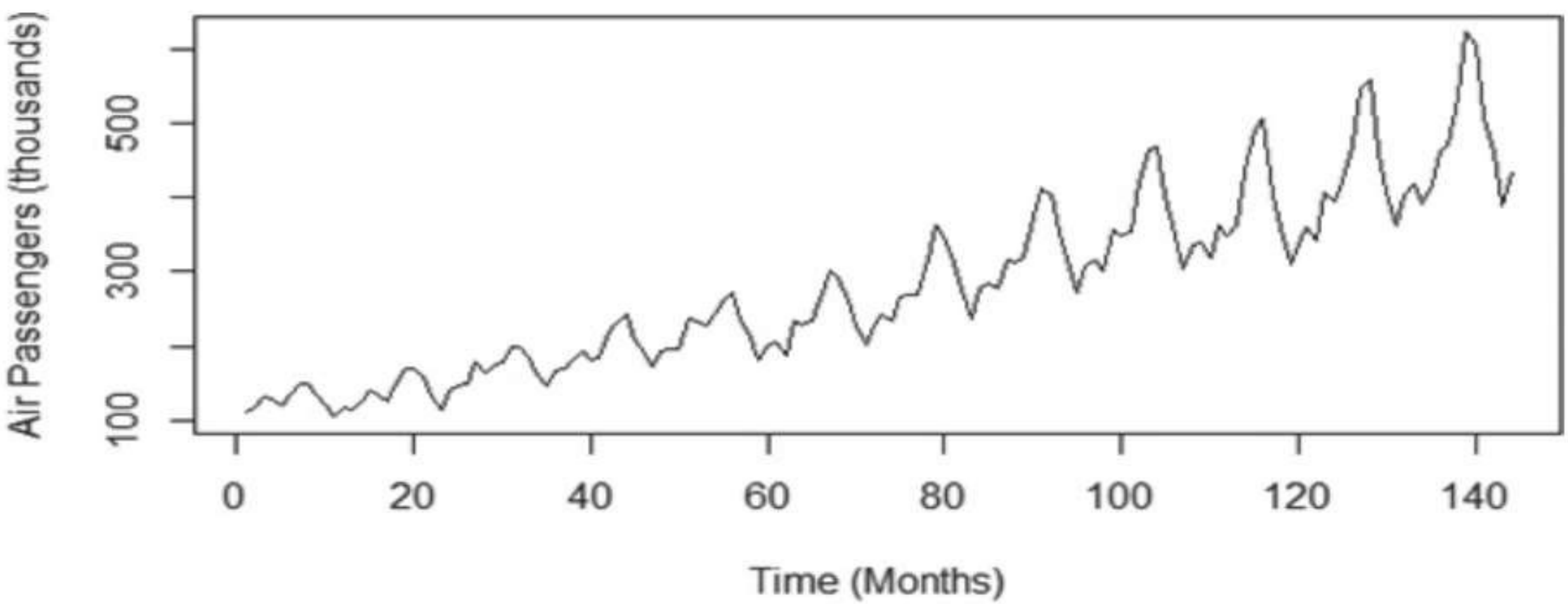
TIME SERIES ANALYSIS

Time Series Analysis: Box-Jenkins Methodology,
ARIMA (Auto Regressive Integrated Moving Average) Model,
Choice of a Model,
Overview of ARMAX,
Spectral Analysis and GARCH.

- A time series is an arrangement of statistical data in chronological order.
- Mathematically, $y = f(t)$
- Time series analysis attempts to model the underlying structure of observations taken over time.
- Time series is an ordered sequence of equally spaced values over time.
- Eg. Plot shows monthly number of international airline passengers over a period of 12-years.



Time Series Analysis

If values of a phenomenon or variable at times t1,t2,...tn are y1, y2,..., yn respectively, then the series

t : t1	t2	t3	....	tn
y : y1	y2	y3	....	yn

constitutes a *time series*.

Times series gives a **bivariate distribution.**

(Bivariate distribution are the probabilities that a certain event will occur when there are two independent random variables in scenario.
The distribution tells the probability of each possible choice of scenario.)

Depending on the frequency of observations, a time series may typically be hourly, daily, weekly, monthly, quarterly and annual. Sometimes, you might have seconds and minute-wise time series as well, like, number of clicks and user visits every minute etc.

Time Series Analysis

- Limited to equally spaced time series.
- Following are the goals of time series analysis:

✔ Identify and model the structure of the time series.

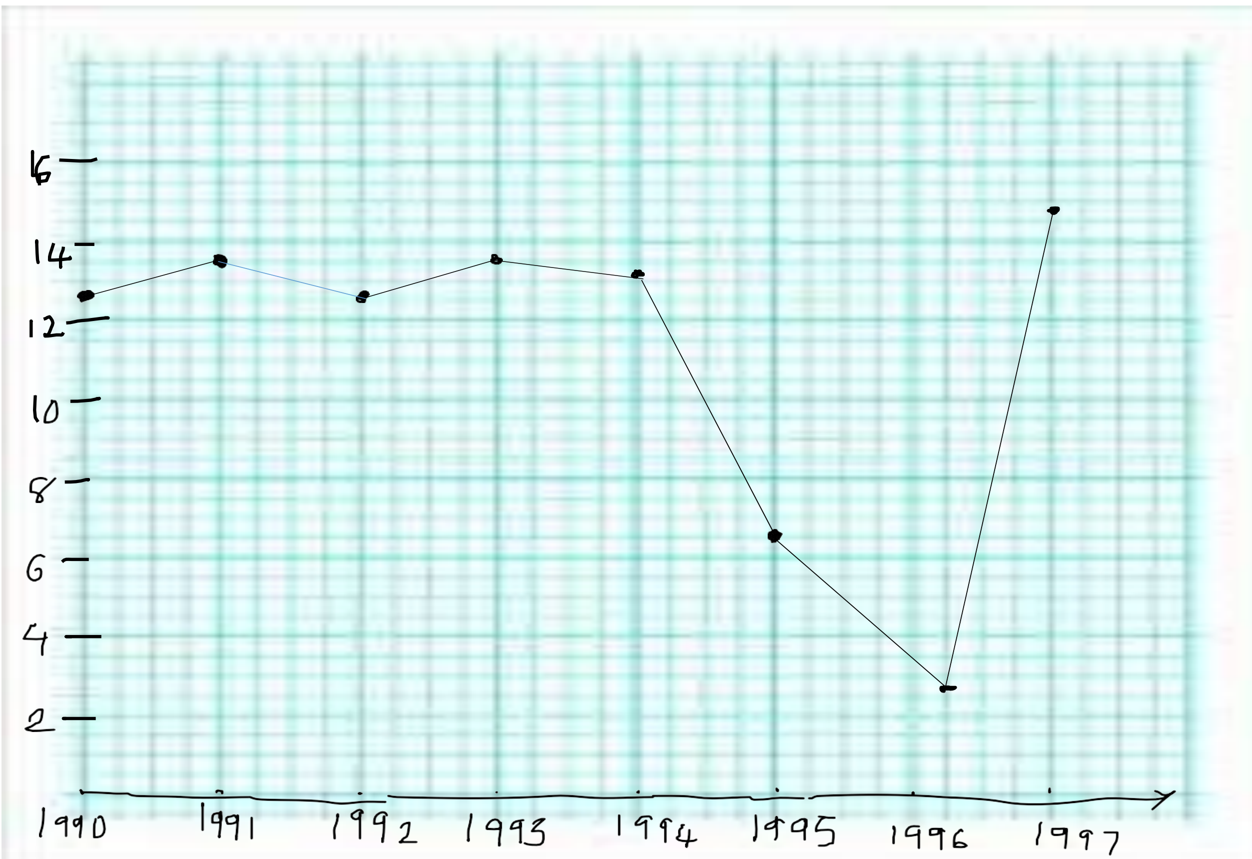
✔ Forecast future values in the time series.

- Time series analysis has many applications in **finance, economics, biology, engineering, retail, and manufacturing.**

Historigram or Horizontal Line Graphs

- Plotted by taking the independent variable time (t) along x-axis and the dependent variable (y) along y-axis.

Year	1990	1991	1992	1993	1994	1995	1996	1997
Yield (in million tons)	12.8	13.9	12.8	13.9	13.4	6.5	2.9	14.8



Scale along X-axis : 1 cm = 1 year

Scale along Y-axis : 1 cm = 2 million tons

Historigram – Two or more variables

Item	1971	1972	1973	1974	1975	1976
Cement	107.0	113.1	107.6	102.6	116.7	133.9
Iron and steel	100.6	112.0	96.1	100.2	121.3	145.0
General Index	104.2	110.2	112.0	114.3	119.3	131.2

Table showing Index numbers for industrial production of India

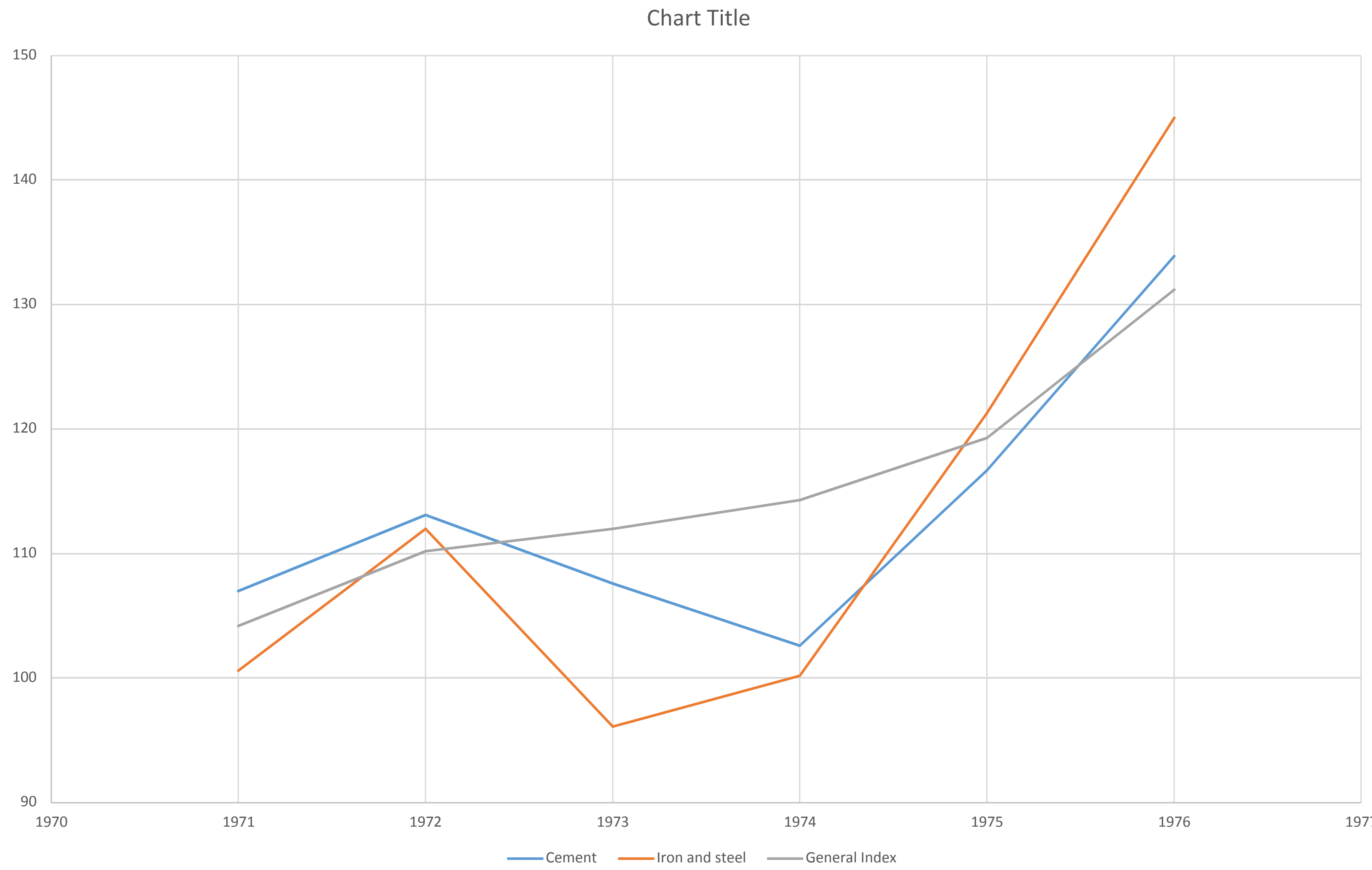


Chart Title
150
140
130
120
110
100
90
1970
1971
1972
1973
1974
1975
1976
1977
Cement
Iron and steel
General Index

Time Series vs Cross Sectional Data

More Information Online WWW.DIFFERENCEBETWEEN.COM

	Time Series Data	Cross Sectional Data
DEFINITION	A type of data consisting of observations of a single subject at multiple time intervals.	A type of data consisting of observations of many subjects at the same point in time.
MAIN FOCUS	Focuses on the same variable over a period of time.	Focuses on several variables at the same point in time.
EXAMPLES	Profit of an organization over a period of 5 years' time	Maximum temperature of several cities on a single day

Understanding Time Series

Time series data	Cross-sectional data
• One variable, one subject	• Many variables related to many subjects
– Observed over a successive equally spaced points in time	– Observed at same point of time (snapshot)
– Each observation is on same subject instance	– Each observation represents a distinct subject instance

Understanding Time Series

Time series data

- Unit of time used to order measurements is to be clearly specified
- Tasks
  - Analysis
    - Descriptive
  - Forecasting

Cross-sectional data

- Order of measurements in time does not matter
  - Time may be included as just another variable
- Tasks
  - Classification
  - Prediction

Understanding Time Series

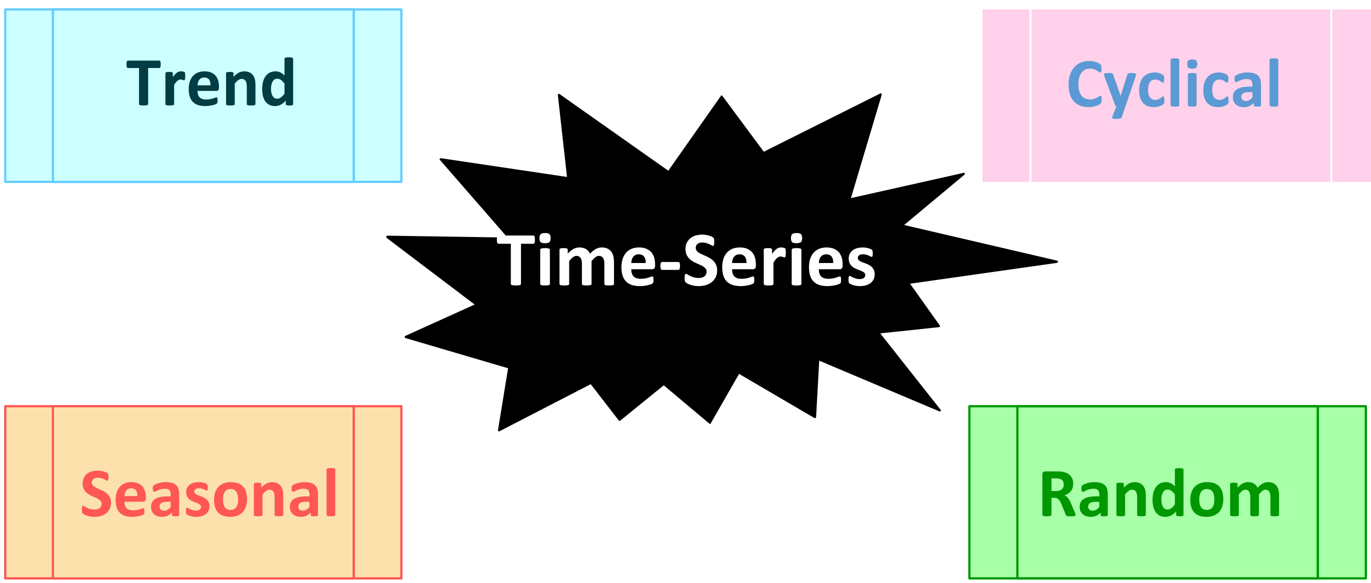
Time series data

- Main idea is to examine changes in the subject instance over time

Cross-sectional data

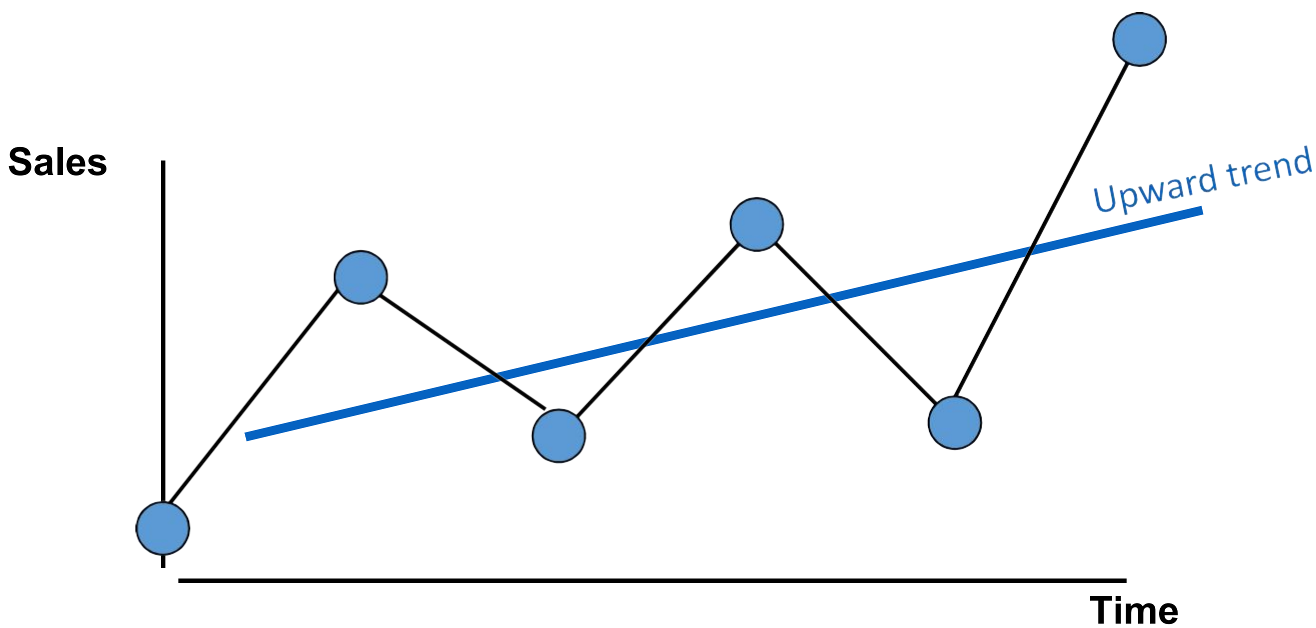
- Main idea is to compare differences among the subject instances

Time-Series Components



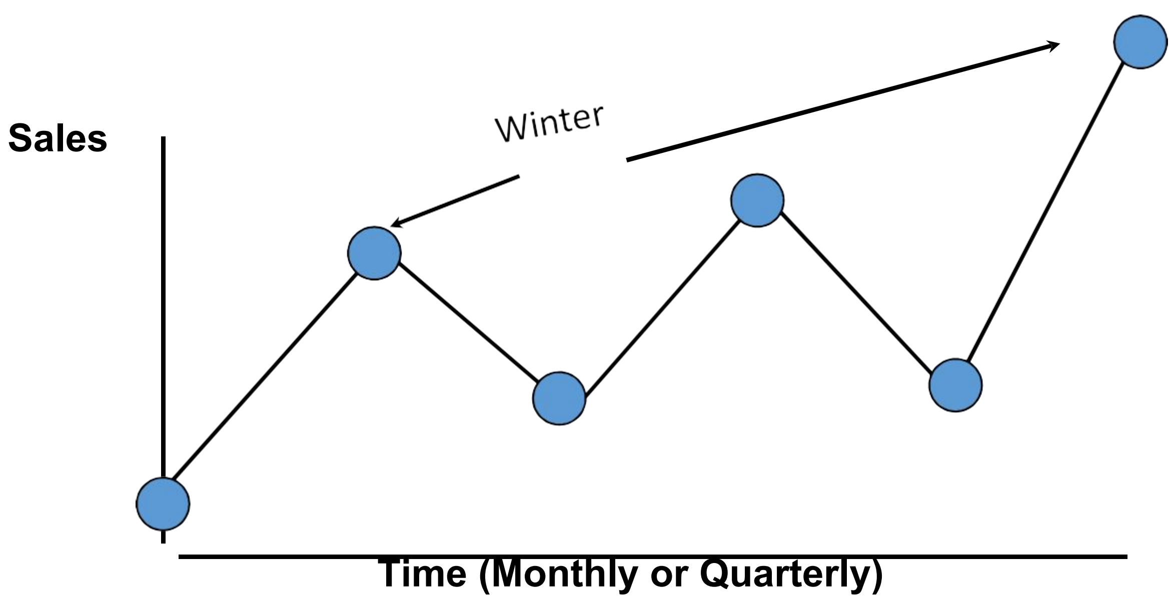
Secular Trend or Long Term Movement (T)

- Overall Upward or Downward Movement
- Data Taken Over a Period of Years



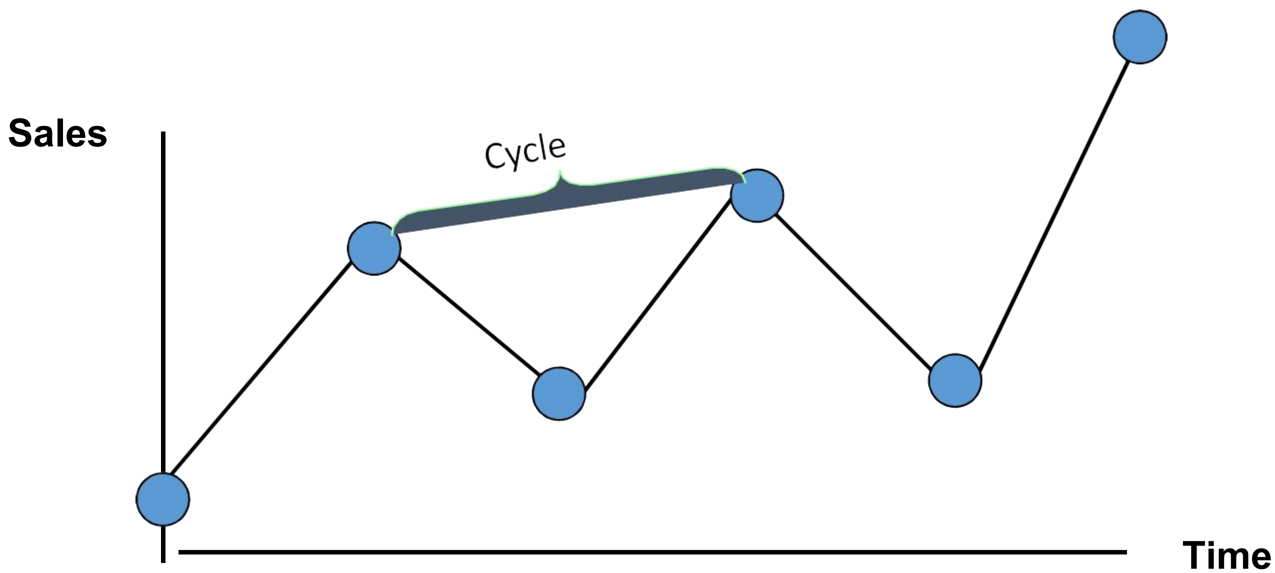
Seasonal Variations (S)

- Upward or Downward Swings
- Regular Patterns
- Observed Within 1 Year



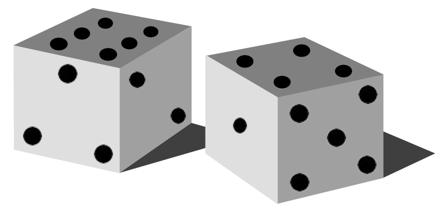
Cyclical Variations (C)

- Upward or Downward Swings
- May Vary in Length
- Usually Lasts 2 - 10 Years



Random or Irregular Component (R or I)

- Erratic, Nonsystematic, Random, ‘Residual’ Fluctuations
- Due to Random Variations of
  - Nature
  - Accidents



- Short Duration and Non-repeating

No hypothesis or trend can be used to suggest irregular or random movements in a time series.
These outcomes are unforeseen, erratic, unpredictable, and uncontrollable in nature.
Earthquakes, war, famine, and floods are some examples of random time series components.

Analysis of Time Series

Time series analysis consists of :

(i) Identifying or determining the various forces or influences whose interaction produces the variations in the time series.

(ii) Isolating, studying, analyzing, and measuring them independently i.e by holding other things constant

Mathematical Models for Time Series

Following two models are commonly used for the decomposition of a time series into its components :

(1) Additive Model or Decomposition by Additive Hypothesis

(2) Multiplicative Model or Decomposition by Multiplicative Hypothesis

Additive Time-Series Model

T_t = *Trend*

C_t = *Cyclical*

I_t = *Irregular*

S_t = *Seasonal*

$$Y_t = T_t + S_t + C_t + I_t$$

Additive model assumes that all four components of the time series operate independently of each other so that none of these components has any effect on the remaining three.

Multiplicative Time-Series Model

In the multiplicative model, the original time series is expressed as the product of trend, seasonal and irregular components.

Under this model, the trend has the same units as the original series, but the seasonal and irregular components are unitless factors, distributed around 1.

- Used Primarily for Forecasting $Y = T \times S \times C \times I,$
- Observed Value in Time Series is the product of Components
- For Annual Data: $Y_t = T_t \text{ X } C_t \text{ X } I_t$
- For Quarterly or Monthly Data:

$$Y = T \times S \times C \times I,$$

T_i = *Trend*

C_i = *Cyclical*

I_i = *Irregular*

S_i = *Seasonal*

Time Series Models

- Equations of Additive and Multiplicative models can be used to obtain a measure of one or more of the components by elimination.

Mixed Models

- $Y = TCS + I$
- $Y = TC + SI$
- $Y = T + SCI$
- $Y = T + S + CI$

Measurement of Trend

- Graphic or Free Hand Curve Fitting Method
- Method of semi-averages
- Method of Curve Fitting by the Principle of Least Squares
- Method of Moving Averages

Graphic or Free Hand Curve Fitting Method

- Take the time periods along x-axis by taking appropriate scaling

- Plot the points for observed values of the Y variable as the dependent variable against the given time periods

- Join these plotted points by line segments to get a histogram.

- Draw a free hand smooth curve (or a straight line) through the histogram
- It is generally preferred to use a curve instead of a straight line to show the secular trend.

Graphic or Free Hand Curve Fitting Method

In order to obtain proper trend line/curve :

1) The curve should be smooth

2) Number of points above the trend line/curve should be more or less equal to the number of points below it.

3) The sum of the vertical deviations of the given points above the trend line should be approximately equal to the sum of deviations of points below the trend line.

4) Sum of the squares of the vertical deviations of the given points from the trend line/curve is as minimum as possible.

Graphic or Free Hand Curve Fitting Method

5) If cycles are present in the data, then the trend line should be drawn so that:

(a) It has equal number of cycles above and below it

(b) It bisects the cycles so that the areas of the cycles above and below it are approximately the same.

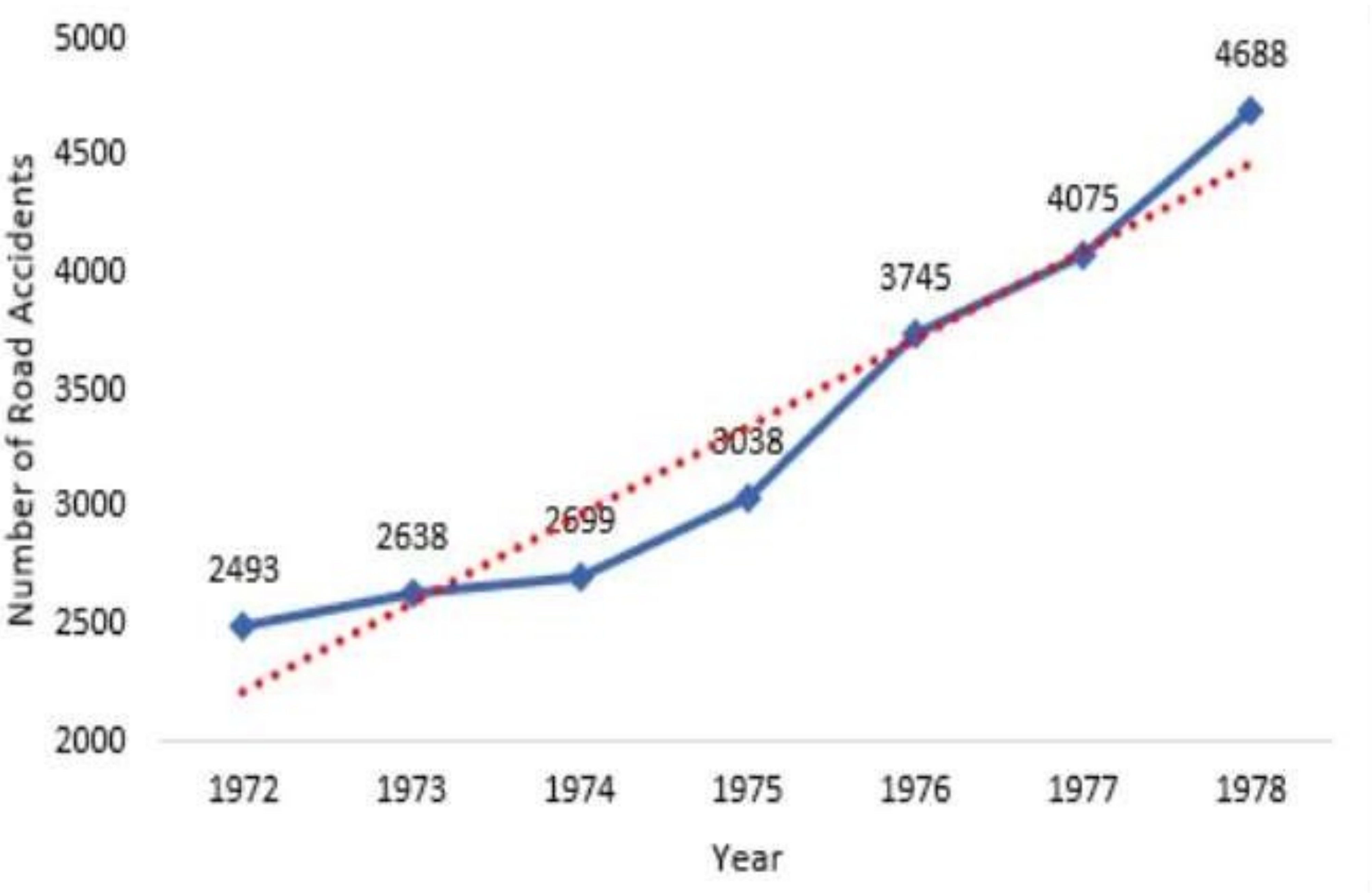
6) The minor short-term fluctuations or abrupt and sudden variations may be ignored.

Question: The following time series shows the number of road accidents in Punjab for the year 1972 to 1978.

Year	1972	1973	1974	1975	1976	1977	1978
No. of Accidents	2493	2638	2699	3038	3745	4079	4688

- Obtain the historigram showing the number of road accidents and a free hand trend line by drawing a straight line
- Find the trend values for this time series

Year	Value	Total	Mean	Trend Value
1972	2493			2200
1973	2638			2550
1974	2699			2950
1975	3038	233380	3340	3340
1976	3745			3650
1977	4079			4050
1978	4688			4499


Number of Road Accidents
5000
4500
4000
3500
3000
2500
2000
2493
2638
2699
3038
3745
4075
4688
1972
1973
1974
1975
1976
1977
1978
Year

Graphic or Free Hand Curve Fitting Method

Merits:

- The free-hand curve method is a simple, easy, and quick method for measuring secular trends.

- A well-fitted trend line (or curve) gives a close approximation to the trend based on a mathematical model.

Demerits:

- It is a rough and crude method.

- It is greatly affected by personal bias as different persons may fit different trends to the same data.

- The estimates are not reliable due to personal bias

Fit a trend line by the method of freehand method for the given data.

Year	2000	2001	2002	2003	2004	2005	2006	2007
Sales	30	46	25	59	40	60	38	65

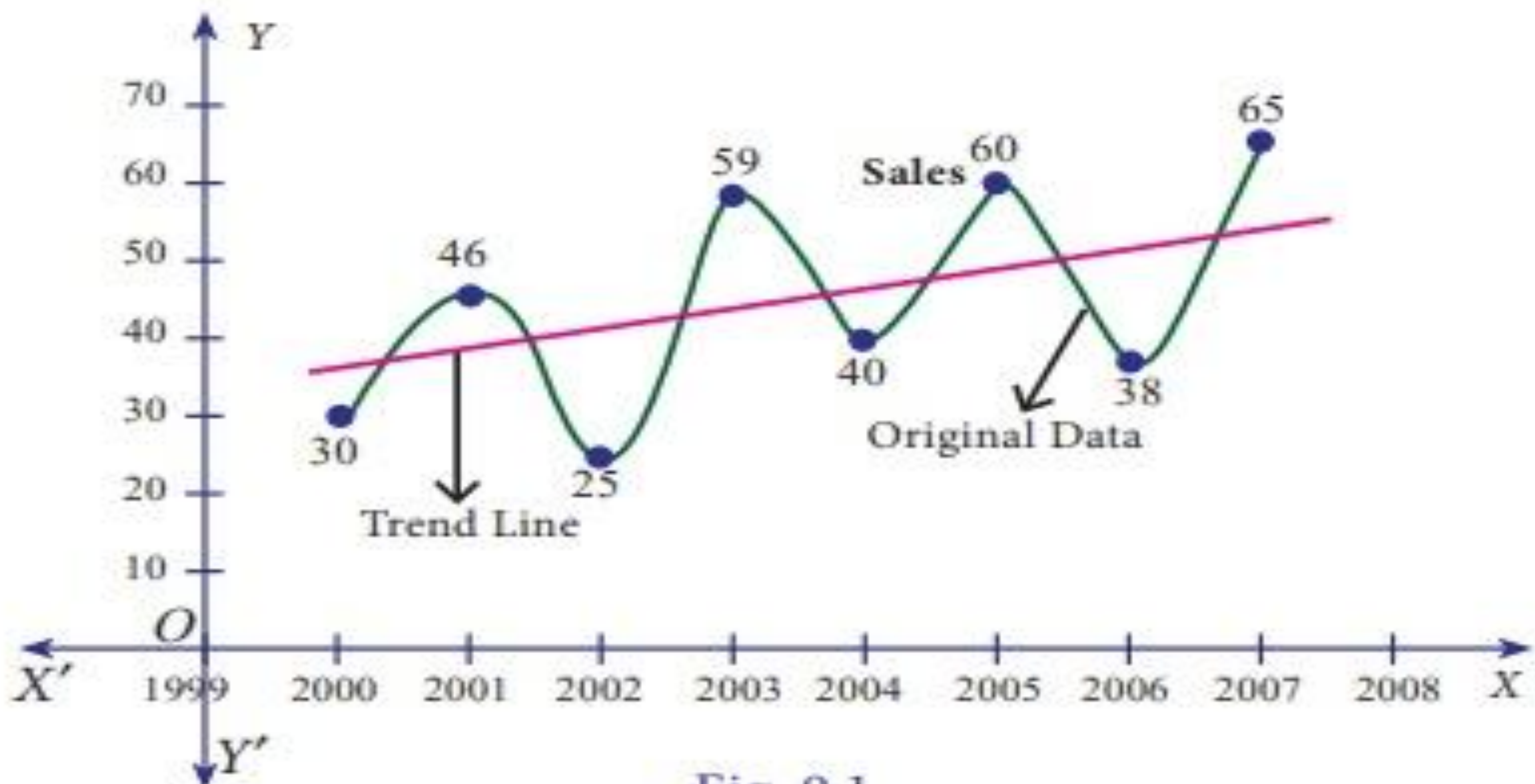


Fig: 9.1

Method of semi-averages

- Divide the whole time series data into two equal parts w.r.t time.

- Compute the arithmetic mean of time-series values for each half separately.

- These means are called **semi-averages**.

- These semi-averages are plotted as points against the middle point of the respective time periods covered by each part.

- The line joining these points gives the straight line trend fitting the given data.

Method of semi-averages

- In this method, the original data is divided into two equal parts and average are calculated for both the parts. These are averages are called Semi – averages.

- For example, we can divide the 10 years 1983 to 1992 into two equal parts;
- from 1983 to 1987 and 1988 to 1992.

- If the period is odd number of years, the values of the middle year is omitted say, from 1983 to 1993 we must omit the year 1988.
- We can draw the line by a straight line by joining the two points of averages. By extending the line downwards or upwards we can get the intermediate values or we can predict the future

Fit a trend line by the method of semi-averages for the given data.

Year	2000	2001	2002	2003	2004	2005	2006
Production	105	115	120	100	110	125	135

Year	Production	Average
2000	105	
2001	115	$\frac{105+115+120}{3}=113.33$
2002	120	
2003	100 (left out)	
2004	110	
2005	125	$\frac{110+125+135}{3}=123.33$
2006	135	

Table 9.1

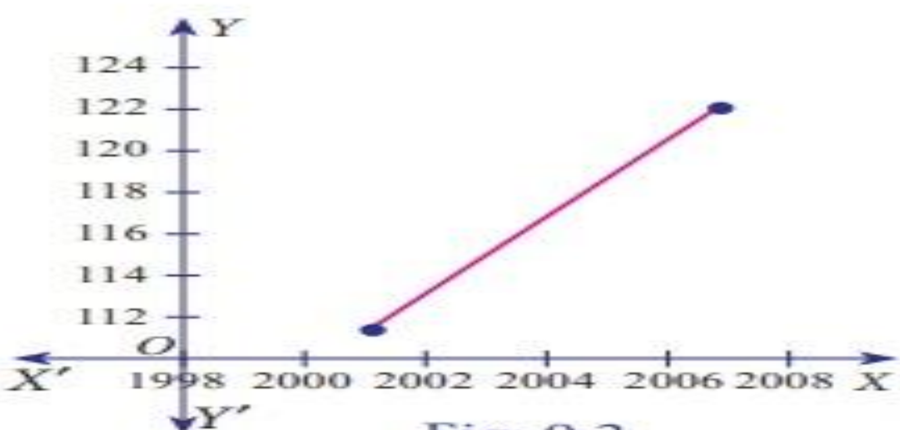


Fig: 9.2

Apply the method of semi-averages for determining trend of the following data and estimate the value for year 2000.

Part - I   Part II

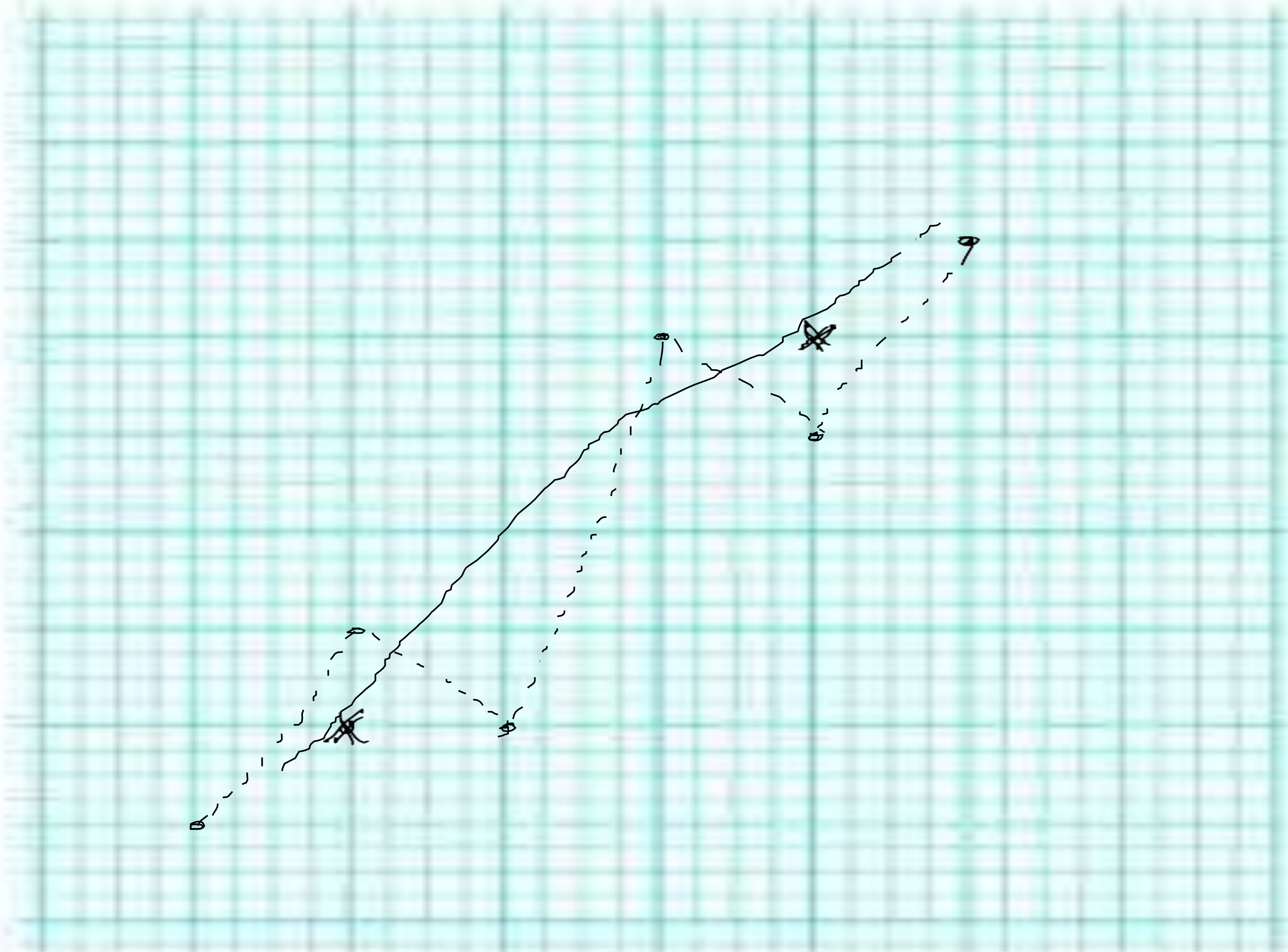
$n = 6$

Year	1993	1994	1995	1996	1997	1998
Sales (thousand units)	20	24	22	30	28	32

$\bar{x}_1 = \dfrac{20+24+22}{3} = 22$ $\bar{x}_2 = \dfrac{30+28+32}{3} = 30$

$30 - 22 = 8$

yearly increment $\left(\dfrac{8}{3}\right) = 2.667$



Apply the method of semi-averages for determining trend of the following data and estimate the value for year 1999.

Part I

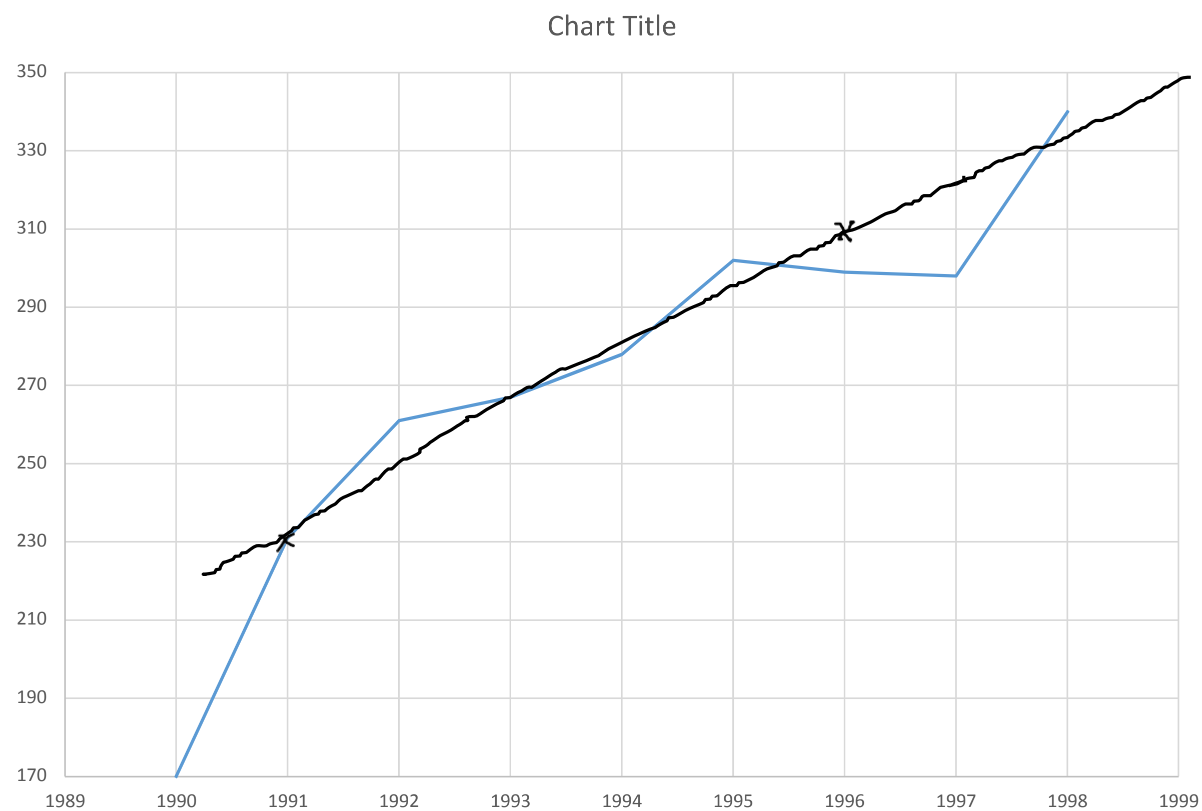
Part II

$n = 9$

Year	1990	1991	1992	1993	1994	1995	1996	1997	1998
Actual Value	170	231	261	267	278	302	299	298	340

Yea r	Actual Value	4-yearly semi totals Semi-averages	
1990			
170		929	232
1991			
231			
1992			
261		ignore	
1993			
267		1239	310
1994			
278			

1995
302
+9
1996
299
1997
298



Year	Property Damaged
1973	201
1974	238
1975	392
1976	507
1977	484
1978	549
1979	742

$$y_1' = 277, x_1 = 1, y_2' = 625, x_2 = 5$$

$$b = \frac{y_2' - y_1'}{x_2 - x_1} = \frac{625 - 277}{5 - 1} = 87$$

$$a = y_1' - bx_1 = 277 - 87(1) = 190$$

Find out the trend values for year 1980

Year	Property Damaged	Semi Total	Semi Average	Coded Year	Trend Values
1973	201			0	$y' = 190 + 87(0) = 190$
1974	238	831	277	1	$y' = 190 + 87(1) = 277$
1975	392			2	$y' = 190 + 87(2) = 364$
1976	507			3	$y' = 190 + 87(3) = 451$
1977	484			4	$y' = 190 + 87(4) = 538$
1978	549	1875	625	5	$y' = 190 + 87(5) = 625$
1979	742			6	$y' = 190 + 87(6) = 712$

$$y_1' = 277, x_1 = 1, y_2' = 625, x_2 = 5$$

$$b = \frac{y_2' - y_1'}{x_2 - x_1} = \frac{625 - 277}{5 - 1} = 87$$

$$a = y_1' - bx_1 = 277 - 87(1) = 190$$

The semi-average trend line $y' = 190 + 87x$ (with the origin at 1973)

Year	No. of books (y)	
1973	42	
1974	38	
1975	35	
1976	25	
1977	32	
1978	24	
1979	20	
1980	19	
1981	17	

$$y_1' = 35, x_1 = 1.5, y_2' = 20, x_2 = 6.5$$

$$b = \frac{y_2' - y_1'}{x_2 - x_1} = \frac{20 - 35}{6.5 - 1.5} = -3$$

$$a = y_1' - bx_1 = 35 - (-3)(1.5) = 39.5$$

$$y' = 39.5 - 3x \text{(with origin at 1973)}$$

For the year 1982, the estimated number of books sold is: $y' = 39.5 - 3(9) = 12.5$.

Year	No. of books (y)	Semi Total	Semi Average	Coded year	Trend Values
1973	42			0	$y' = 39.5–3(0) = 39.5$
1974	38	140	35	1	$y' = 39.5–3(1) = 36.5$
1975	35	2	$y' = 39.5–3(2) = 33.5$		
1976	25			3	$y' = 39.5–3(3) = 30.5$
1977	32			4	$y' = 39.5–3(4) = 27.5$
1978	24			5	$y' = 39.5–3(5) = 24.5$
1979	20	80	20	6	$y' = 39.5–3(6) = 21.5$
1980	19			7	$y' = 39.5–3(7) = 18.5$
1981	17			8	$y' = 39.5–3(8) = 15.5$

Merits of Semi – Average Method:

i. It is simple and easier to understand than moving average and least square method

ii. As the line can be extended both ways, we can get the intermediate values and predict the future values.

iii. As it does not depend upon personal judgement, everyone who applies this method will get the same trend line unlike the former method.

Demerits of Semi – Average Method:

i. Under this method, it has an assumption of linear trend whether such a relationship exists or not
ii. It is affected by the limitation of arithmetic mean
iii. This method is not enough for forecasting the future trend or for removing trend from original data

Method of Curve Fitting by Least Squares method

- The straight line method gives a line of best fit on the given data.
- Let the straight line trend between the given time-series values (y) and time (t) be given by the equation:

$$y = a + bt$$

- Then for any given time 't', the estimated value of y_e of y as given by this equation is :

$$y_e = a + bt$$

Method of Curve Fitting by the Least Squares

This method is mathematically designed to satisfy the following two conditions:

(1) Sum of squares of errors is minimum

Sum of $(y - y_e)^2$ = least

(2) Sum of errors is zero. Sum of $(y - y_e) = 0$

Method of Curve Fitting by the Principle of Least Squares

The values of constants, ‘a’ and ‘b’, are determined by the following two normal equations.

$\sum y = na + b \sum x$ ....(i)

$\sum xy = a\sum x + b \sum x^2$ ....(ii)

Change of Origin

- The process of finding values of constants a and b can be made simple by using a shortcut method, that is, by taking the origin year in such a way that it gives the total of 'x' ($\sum x$) equal to 'zero'.
- This becomes possible if we take the median year as origin period.
- Thus, the negative values in the first half of the series balance out the positive values in the second half.
- Thus, the earlier normal equation shall be changed as follows, with reference to $\sum x = 0$.

$\sum y = na$ (as $b\sum x$ becomes zero)

$\sum xy = b\sum x^2$ (as $a\sum x$ becomes zero)

Therefore, the values of two constants are obtained by the following formulae:

$$a = \frac{\Sigma y}{N}, \quad \text{and} \quad b = \frac{\Sigma xy}{\Sigma x^2}$$

Change of Origin

- The time period allotted to value 0 is known as the *period of origin.*

- This technique can be applied only if the values of *t* are equidistant, say at interval *h*.

Case (1) : if n is odd, then the transformation

is $$x = \frac{t - \text{middle value}}{\text{interval}(h)}$$

Case (2) : if n is even, then the transformation is

$$x = \frac{t - (\text{arithmetic mean of middle two values})}{\frac{1}{2}\,(\text{interval})}$$

Examples

Case(1) :

We are given sales of years 1990,1991,…1996 i.e n=7

$$x = \frac{t - \text{middle value}}{\text{interval}(h)} = t - 1993$$

Case(2) :

We are given yearly sales for years 1995,1996,…2002 i.e n=8

$$x = \frac{t - (1/2(1998+1999))}{½(1)} = 2t - 3997$$

Year	1990	1991	1992	1993	1994
Production (thousand units)	18	21	23	27	16

Here $n = 5$, hence we shift the origin to middle year i.e 1992.

Let $x = t - 1992$

Year (t)	Production (y) In Thousand tons	X = t - 1992	x^2	xy	Trend Values $y_e = 21 + 0.2x$	$y - y_e$
1990	18	-2	4	-36	$21+0.2(-2)$ $= 20.6$	$18 - 20.6 = -2$
1991	21	-1	1	-21	$21+0.2(-1)$ $= 20.8$	$21 - 20.8 = 0.2$
1992	23	0	0	0	21	$23 - 21 = 2$
1993	27	1	1	27	$21+0.2(1)$ $= 21.2$	$27 - 21.2 = 5.$
1994	16	2	4	32	$21+0.2(2)$ $= 2\ .4$	$16 - 21.4 = -5.$
	$\sum y = 105$	$\sum x = 0$	$\sum x^2 = 10$	$\sum xy = 2$	$\sum y_e$ $= 105$	$\sum 0$

Estimated production for 1995

put $x = t - 1992$

$x = 1995 - 1992 = 3$

Substituting $x = 3$ in $y = 21 + 0.2x$

$y_e = 21 + 0.2(3)$

$= 21 + 0.6$

$= 21.6$ units

Monthly increase in production

yearly production $= 0.2 \times 1000 = 200$ tons

monthly incr in prod $= \dfrac{200}{12} = 16.67$ tons

Years	Sales
1995	6.7
1996	5.3
1997	4.3
1998	6.1
1999	5.6
2000	7.9
2001	5.8
2002	6.1

Year	Sales				
1996	11				
2001	12				
2006	14				
2011	18				
2016	16				

Method of Moving Averages

- Consists of obtaining a series of moving averages (arithmetic mean) of successive overlapping groups or sections of the time series.

- The averaging process smoothens out fluctations and ups & downs in the given data.

- The moving average (MA) is characterized by a constant known as the *period* or *extent* of the moving average.

- The moving average of period 'm' is a series of successive averages of m overlapping values at a time, starting with 1^{st}, 2^{nd}, 3^{rd} value…..

Method of Moving Averages

- For the time series values y1, y2, y3, y4, ….., the moving average values of 'm' period are given by:

$$1^{st}\ MA = \frac{y_1 + y_2 + \ldots + y_m}{m}$$

$$2^{nd}\ MA = \frac{y_2 + y_3 + \ldots + y_{m+1}}{m}$$

$$3^{rd}\ MA = \frac{y_3 + y_4 + \ldots + y_{m+2}}{m} \quad \ldots\ldots\ldots\ldots\ldots\ldots$$

Year (t)	Values (y)	3-yearly MA
1	10	-
2	14	14
3	18	18
4	22	22
5	26	26
6	30	30
7	34	34
8	38	38
9	42	42
10	46	46
11	50	-

$m = 3$

$=$

Case (1) : When period is Odd

Successive values of the moving averages are placed against the middle values of the corresponding time intervals.

Second centred M.A $= \frac{1}{8}(y_2 + 2$

Case (2) : When period is Even

MA values are placed in between the two middle periods of the time intervals it covers.

Centred MA

$m = 4$

$\bar{y}_1 = \frac{1}{4}(y_1 + y_2 + y_3 + y_4)$, $\bar{y}_2 = \frac{1}{4}(y_2 + y_3 + y_4 + y_5)$, $\bar{y}_3 = \frac{1}{4}(y_3 + y_4 + y_5 +$

First Centred M.A $= \frac{1}{2}(\bar{y}_1 + \bar{y}_2)$ $= \frac{1}{2}\left[\frac{1}{4}(y_1 + y_2 + y_3 + y_4) + \frac{1}{4}(y_2 + y_3 + y_4 + y_5)\right]$

$= \frac{1}{8}\left[(y_1 + y_2 + y_3 + y_4) + (y_2 + y_3 + y_4 + y_5)\right]$

$= \frac{1}{8}\left[y_1 + 2y_2 + 2y_3 + 2y_4 + y_5\right]$

$m=5$

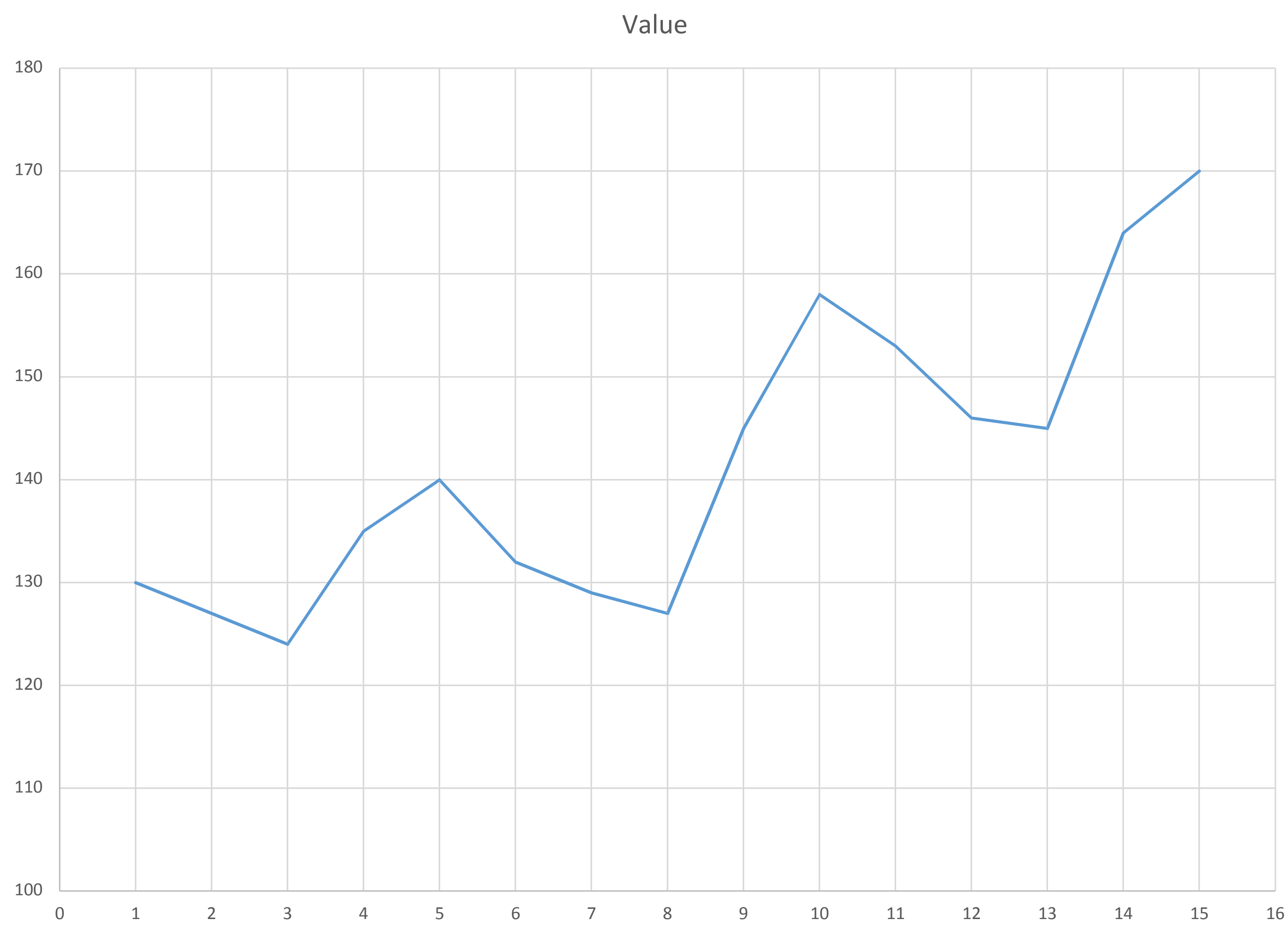
$n=11$

$m=3$

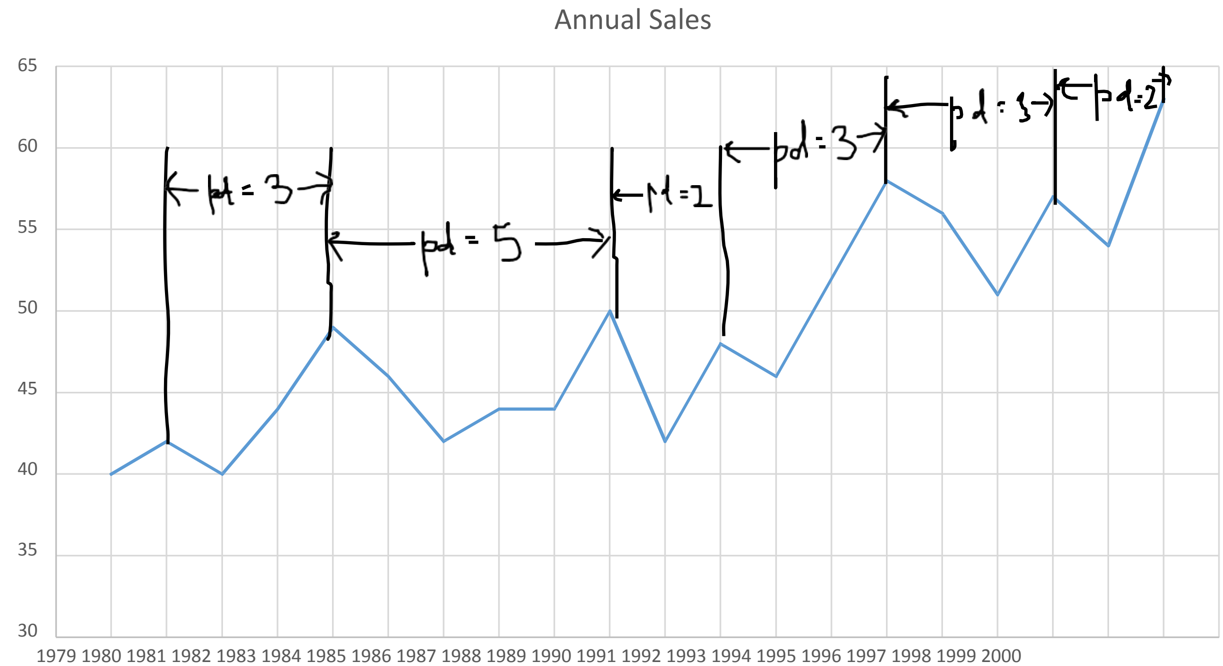
Year	y	3-year MA	y	5-year MA
1990	242		242	
1991	250	248.00	250	
1992	252	250.33	252	249.20
1993	249	251.33	249	251.80
1994	253	252.33	253	252.00
1995	255	253.00	255	253.00
1996	251	254.33	251	255.20
1997	257	256.00	257	257.60
1998	260	260.67	260	259.00
1999	265	262.33	265	
2000	262		262	

Year (1)	Sugar Prod (2)	4-yearly moving totals (3)	4-yearly moving average (4)	2-period moving total of col 4 (5)	Centered MA [Trend Values] (6) is 5%2
1971	37.4				
1972	31.1	146.7	36.675		
1973	38.7	157.2	39.3	75.975	37.98
1974	39.5	168.7	42.175	81.475	40.73
1975	47.9	178.4	44.6	86.755	43.38
1976	42.6	203.5	50.875	95.475	47.74
1977	48.4	214.0	53.5	104.375	52.18
1978	64.6	210.0	52.5	106.0	53
1979	58.4	213.0	53.25	105.75	52.875
1980	38.6	232.8	58.2	111.450	55.725
1981	51.4				
1982	84.4				

Year	Value
1	130
2	127
3	124
4	135
5	140
6	132
7	129
8	127
9	145
10	158
11	153
12	146
13	145
14	164
15	170



Year	Annual Sales
1980	40
1981	42
1982	40
1983	44
1984	49
1985	46
1986	42
1987	44
1988	44
1989	50
1990	42
1991	48
1992	46
1993	52
1994	58
1995	56
1996	51
1997	57
1998	54
1999	63



Year	1981	1984	1989	1991	1994	1997	1999
Peak Val	42	49	50	48	58	57	63
Period		3	5	2	3	3	2

$$\frac{3+5+2+3+3+2}{6} = \frac{18}{6} = 3$$

Year	Annual Sales	3-yearly MA
1980	40	
1981	42	40.67
1982	40	42.0
1983	44	44.33
1984	49	46.33
1985	46	45.67
1986	42	44.0
1987	44	43.33
1988	44	46.0
1989	50	45.33
1990	42	46.67
1991	48	45.33
1992	46	48.67
1993	52	52.0
1994	58	55.33
1995	56	55.0
1996	51	54.67
1997	57	54.0
1998	54	58.0
1999	63	

Drawbacks of Moving Average

- The main problem is to determine the extent of the moving average which completely eliminates the oscillatory fluctuations.

- This method assumes that the trend is linear but it is not always the case.

- It does not provide the trend values for all the terms.

- This method cannot be used for forecasting future trend which is the main objective of the time series analysis.

ARIMA (Auto-Regressive Integrated Moving Average)

An **autoregressive integrated moving average**, or **ARIMA,** is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends.

A statistical model is autoregressive if it predicts future values based on past values.

It's a way of modelling time series data for forecasting (i.e., for predicting future points in the series), in such a way that:

- a pattern of growth/decline in the data is accounted for (hence the "auto- regressive" part)
- the rate of change of the growth/decline in the data is accounted for (hence the "integrated" part)
- noise between consecutive time points is accounted for (hence the "moving average" part)

"time series data" = data that is made up of a sequence of data points taken at successive equally spaced points in time

ARIMA (Auto-Regressive Integrated Moving Average)

- Autoregressive Integrated Moving Average models (ARIMA models) were popularized by George Box and Gwilym Jenkins in the early 1970s.

- ARIMA models are a class of linear models that is capable of representing *stationary as well as non-stationary time series*.

- ARIMA models *do not involve independent variables* in their construction. They make use of the information in the series itself to generate forecasts.

ARIMA (Auto-Regressive Integrated Moving Average)

- ARIMA models rely heavily on autocorrelation patterns in the data.
- ARIMA methodology of forecasting is different because it does not assume any particular pattern in the historical data of the series to be forecast.
- It uses an interactive approach of identifying a possible model from a general class of models.
- The chosen model is then checked against the historical data to see if it accurately describe the series.
- a time series data is a sequence of numerical observations naturally ordered in time,

Box-Jenkins Methodology

- Box and Jenkins suggested a way of converting a non-stationary time series into a stationary time series with the help of some transformations.

- Stationarity is a property of a time series. A stationary series is one where the values of the series is not a function of time.

- That is, the statistical properties of the series like mean, variance and autocorrelation are constant over time.

- *Autocorrelation of the series is the correlation of the series with its previous values.*

- A stationary time series is devoid of seasonal effects as

Box-Jenkins Methodology

- The Box-Jenkins methodology refers to a set of procedures for *identifying*, *fitting*, and *checking* ARIMA models with time series data.

- Forecasts follow directly from the form of fitted model.

- The basis of BOX-Jenkins approach to modeling time series consists of three phases:

Box-Jenkins Methodology

1. Identification

   a) **Data preparation**
      - Transform data to stabilize variance.
      - Differencing data to obtain stationary series.

   b) **Model selection**
      - Examine data, (Autocorrection function)ACF and (Partial autocorrection function) PACF to identify potential models

Box-Jenkins Methodology

2. Estimation and testing

   a) **Estimation**
      - Estimate parameters in potential models
      - Select best model using suitable criterion

   b) **Diagnostics**
      - Check ACF/PACF of residuals
      - Do portmanteau test of residuals
      - Are the residuals white noise?

3. **Application**

   Forecasting: use model to forecast

Examining correlation in time series data

The key statistic in time series analysis is the autocorrelation coefficient (the correlation of the time series with itself, lagged 1, 2, or more periods). Recall the autocorrelation formula:

$$r_k = \frac{\sum_{t=k+1}^{n} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{n} (y_t - \bar{y})^2}$$

Examining Correlation in Time Series Data

- Recall r_1 indicates how successive values of Y relate to each other, r_2 indicates how Y values two periods apart relate to each other, and so on.

- The auto correlations at lag 1, 2, …, make up the autocorrelation function or ACF.

- Autocorrelation function is a valuable tool for investigating properties of an empirical time series.

The Partial autocorrelation coefficient

Partial autocorrelations measures the degree of association between y_t and y_{t-k}, when the effects of other time lags 1, 2, 3, …, k-1 are removed.

The partial autocorrelation coefficient of order k is evaluated by regressing y_t against $y_{t-1}, \ldots y_{t-k}$:

$$y_t = b_0 + b_1 y_{t-1} + b_2 y_{t-2} + \cdots + b_k y_{t-k}$$

α_k (partial autocorrelation coefficient of order k) is the estimated coefficient b_k.

Examining stationarity of time series data

Stationarity means no growth or decline.

Data fluctuates around a constant mean independent of time and variance of the fluctuation remains constant over time.

Stationarity can be assessed using a time series plot:

- Plot shows no change in the mean over time
- No obvious change in the variance over time.

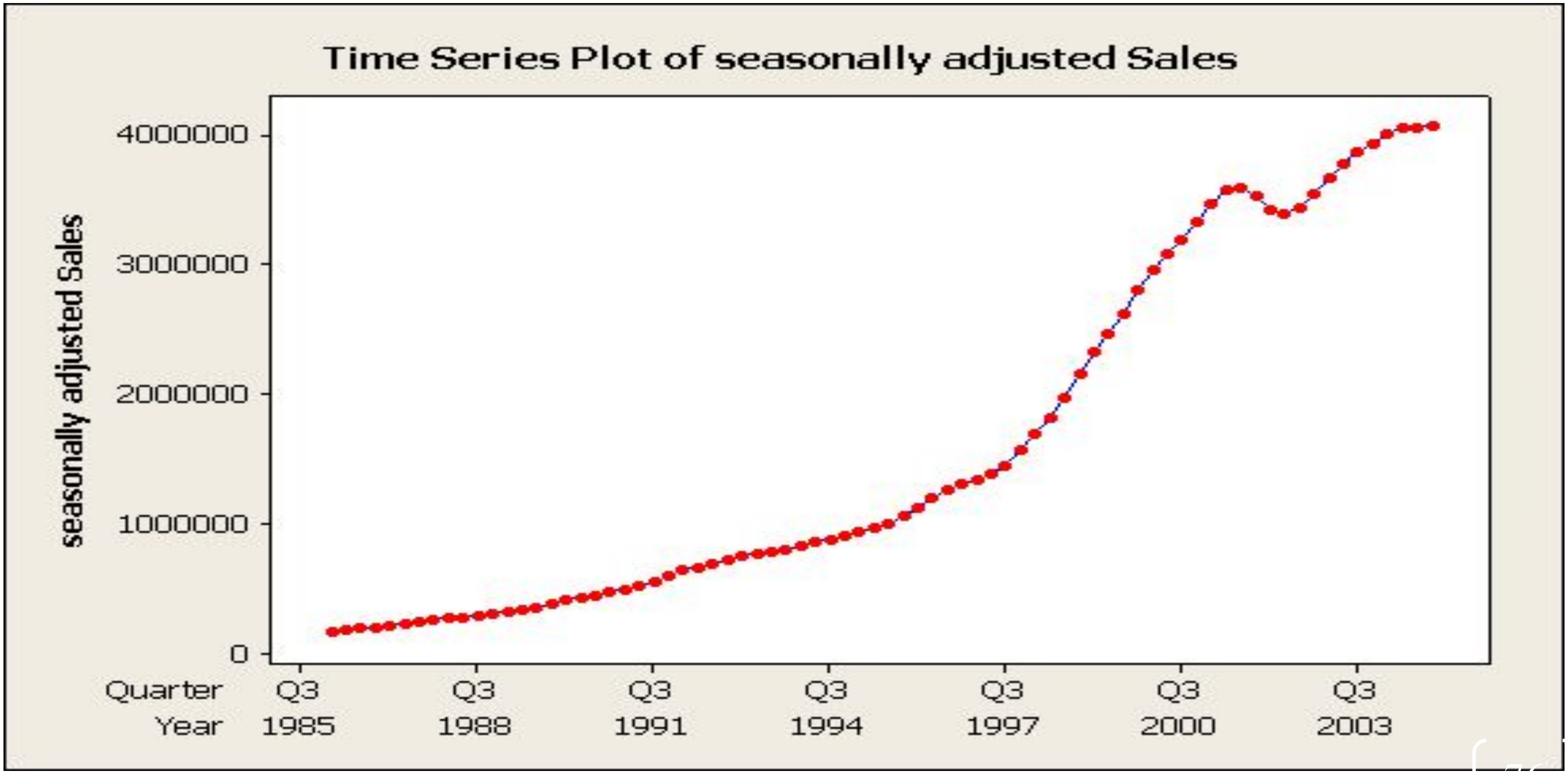
Examining stationarity of time series data

The autocorrelation plot can also show non-stationarity.

Significant autocorrelation for several time lags and slow decline in r_k indicate non-stationarity.

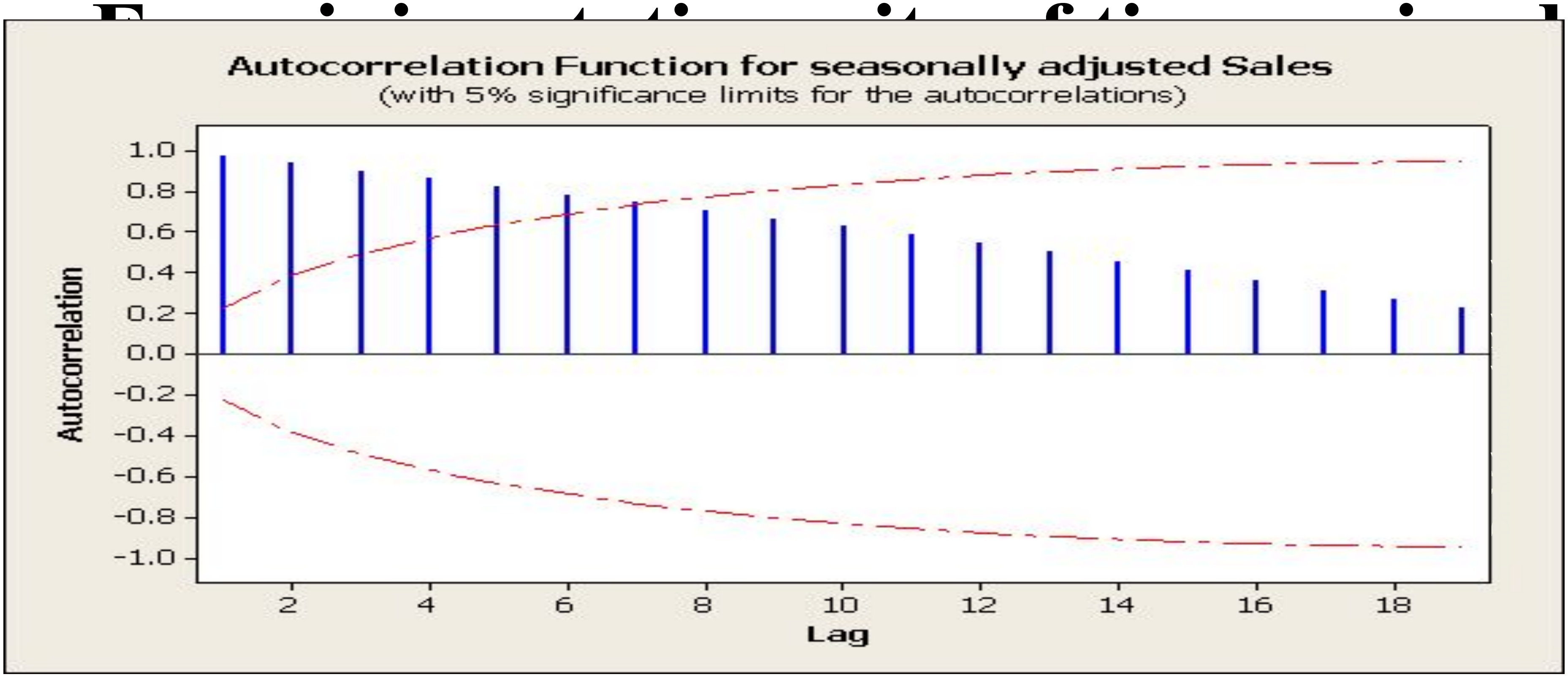
The following graph shows the seasonally adjusted sales for Gap stores from 1985 to 2003.

Examining stationarity of time series data



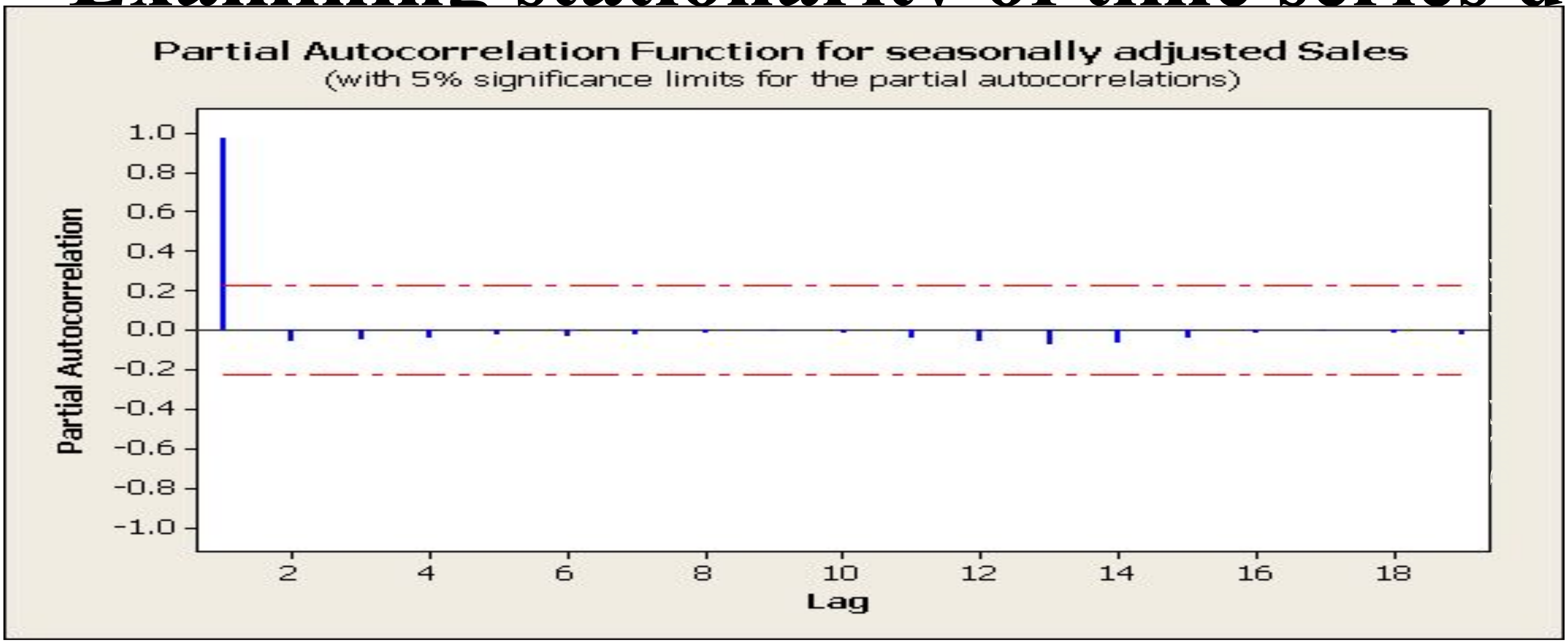
- The time series plot shows that it is **non-stationary** in the mean.
- The next slide shows the ACF plot for this data series.

ata



- The ACF also shows a pattern typical for a non-stationary series:
  - Large significant ACF for the first 7 time lag
  - Slow decrease in the size of the autocorrelations.
- The PACF is shown in the next slide.

Examining stationarity of time series data



This is also typical of a non-stationary series. Partial autocorrelation at time lag 1 is close to one and the partial autocorrelation for the time lag 2 through 18 are close to zero

Removing non-stationarity in time series

The non-stationary pattern in a time series data needs to be removed in order that other correlation structure present in the series can be seen before proceeding with model building.

One way of removing non-stationarity is through the method of differencing.

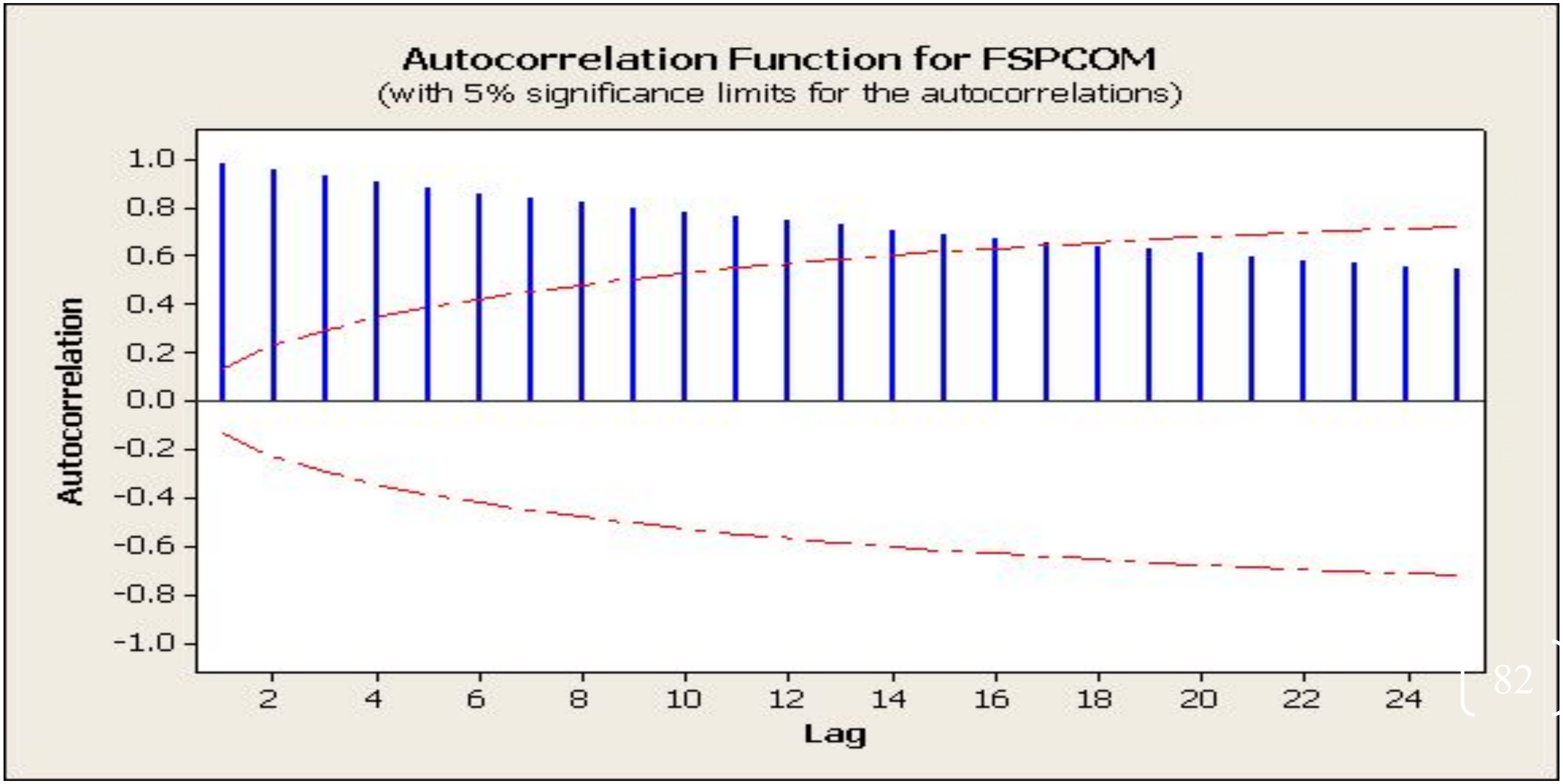
Removing non-stationarity in time series

The differenc $y'_t = y_t - y_{t-1}$:

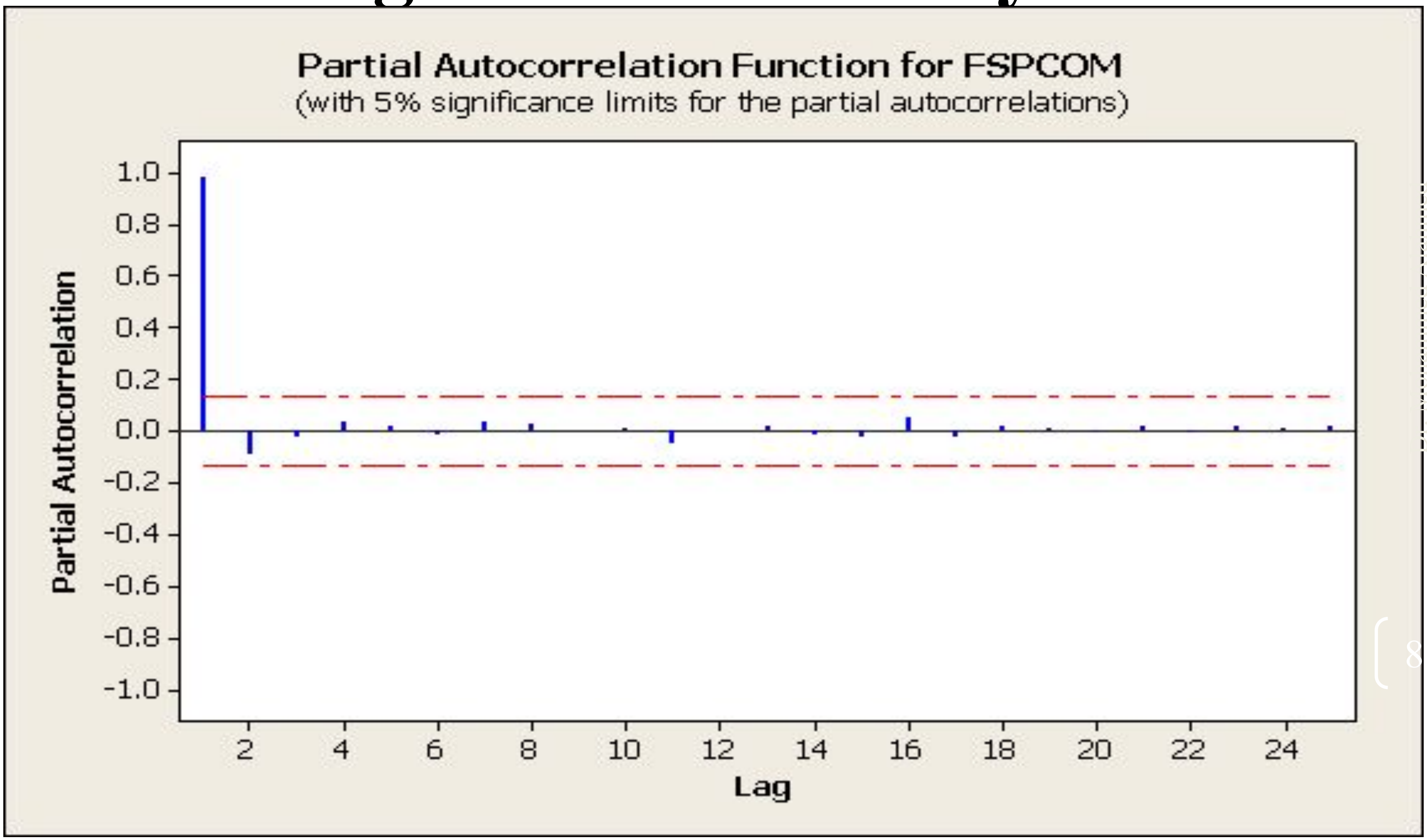
The following two slides shows the time series plot and the ACF plot of the monthly S&P 500 composite index from 1979 to 1997.

Removing non-stationarity in time series

Removing non-stationarity in time series



Removing non-stationarity in time series



Removing non-stationarity in time series

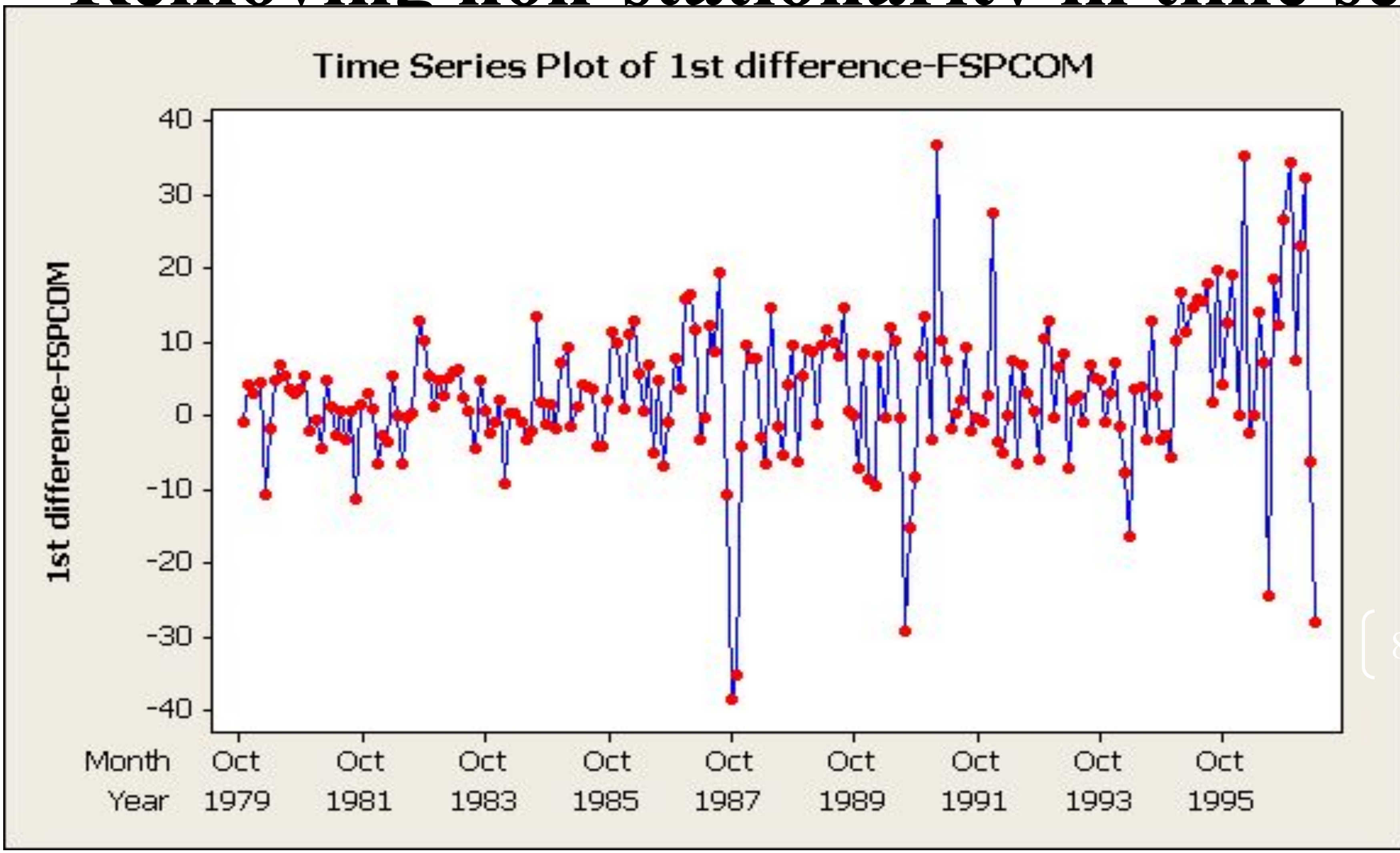
The time plot shows that it is not stationary in the mean. The ACF and PACF plot also display a pattern typical for non-stationary pattern.
Taking the first difference of the S& P 500 composite index data represents the monthly changes in the S&P 500 composite index.

Removing non-stationarity in time series

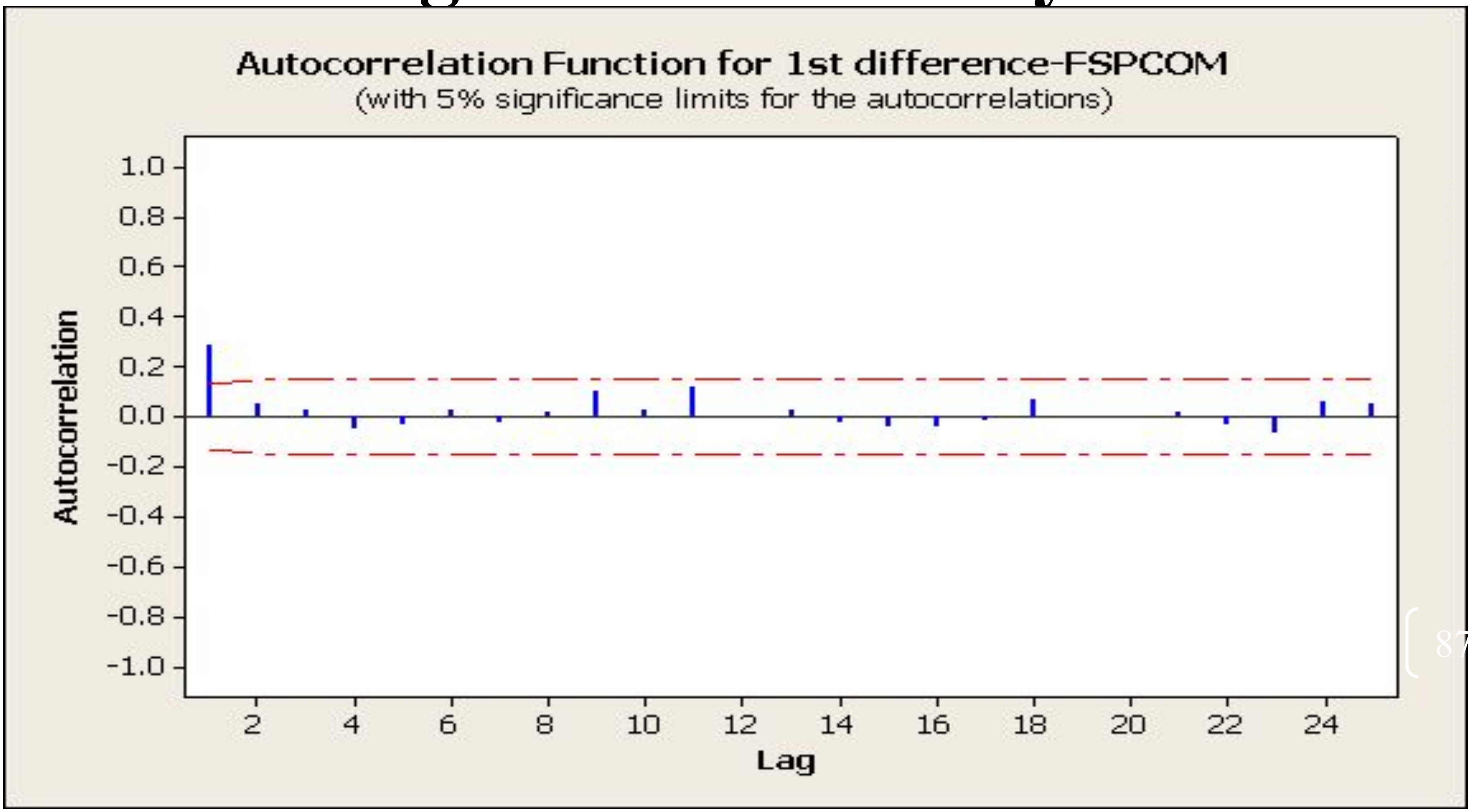
The time series plot and the ACF and PACF plots indicate that the first difference has removed the growth in the time series data.

The series looks just like a white noise with almost no autocorrelation or partial autocorrelation outside the 95% limits.

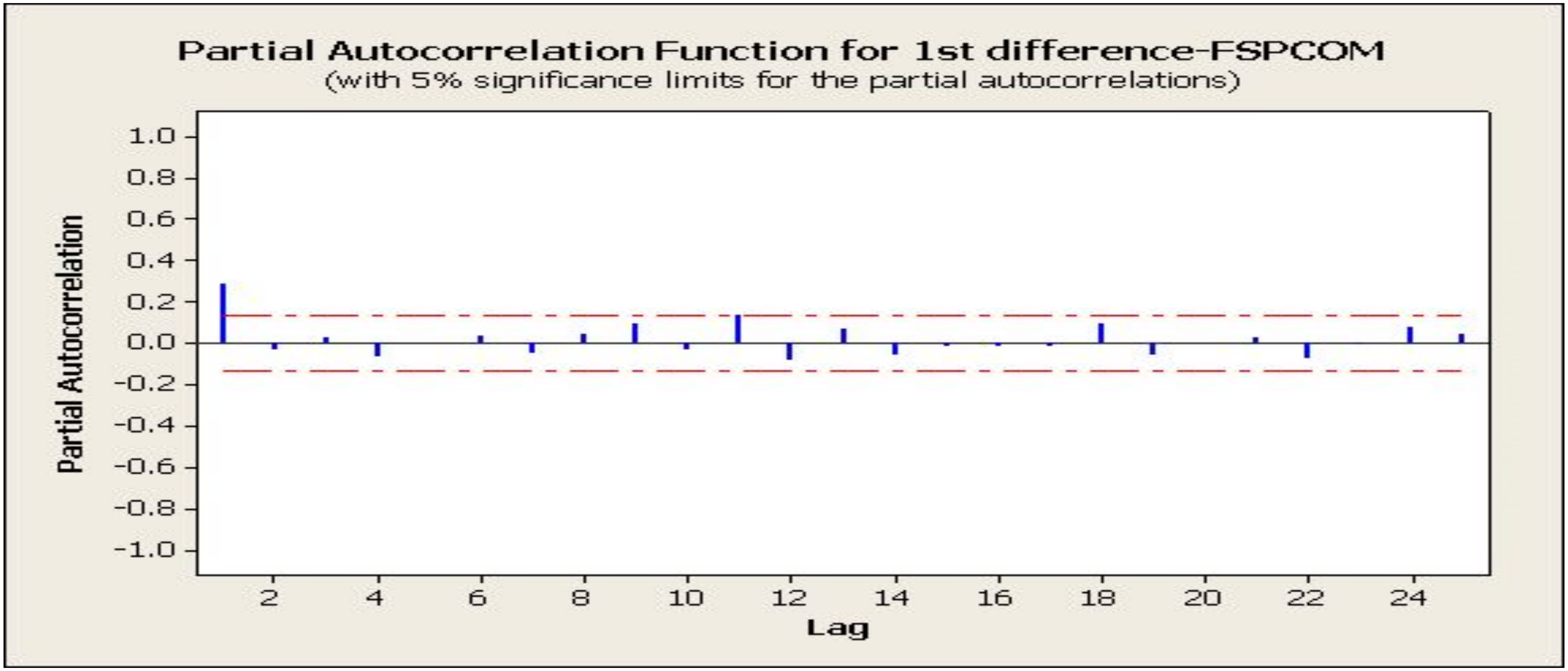
Removing non-stationarity in time series



Removing non-stationarity in time series



Removing non-stationarity in time series



Note that the ACF and PACF at lag 1 is outside the limits, but it is acceptable to have about 5% of spikes fall a short distance beyond the limit due to chance.

Random Walk

Let y_t denote the S&P 500 composite index, then the time series plot of differenced S&P 500 composite index suggests that a suitable model for the data might be

$$y_t - y_{t-1} = e_t$$

Where e_t is white noise.

Random Walk

The equation in the prev

$$y_t = y_{t-1} + e_t$$

This model is known as "random walk" model and it is widely used for *non-stationary* data. Random walks typically have long periods of apparent trends up or down which can suddenly change direction unpredictably
They are commonly used in analyzing economic and stock price series.

Removing non-stationarity in time series

Taking first differencing is a very useful tool for removing non-statioanarity, but sometimes the differenced data will not appear stationary and it may be necessary to difference the data a second time.

Th

$$y''_t = y'_t - y'_{t-1} = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2}$$

In practice, it is almost never necessary to go beyond second order differences.

Seasonal differencing

With seasonal data which is not stationary, it is appropriate to take seasonal differences.
A seasonal difference is the difference between an observation and the corresponding observation from the previous yea

$$y'_t = y_t - y_{t-s}$$

Where s is the length of the season

Seasonal differencing

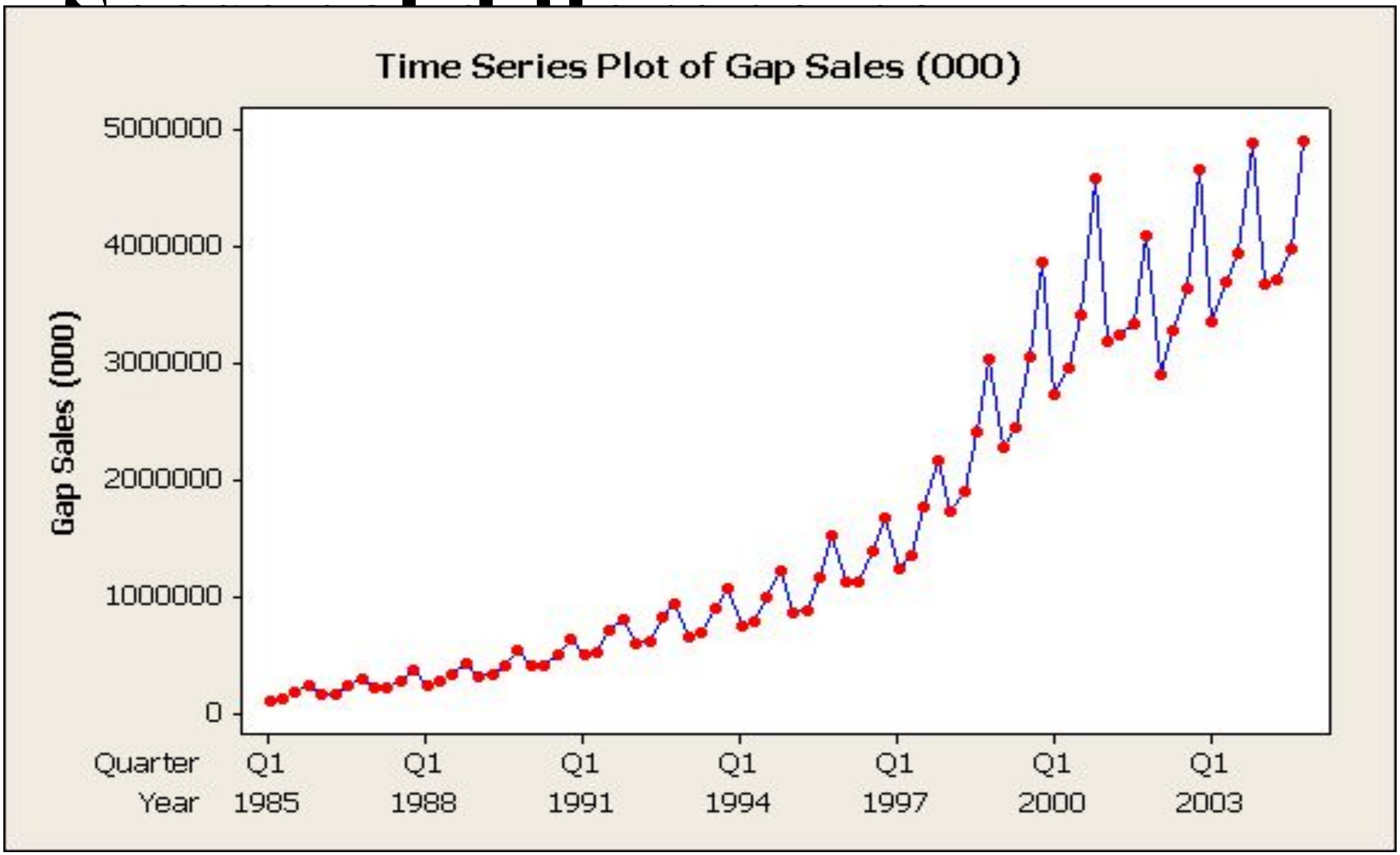
The Gap quarterly sales is an example of a non-stationary seasonal data.

The following time series plot show a trend with a pronounced seasonal component

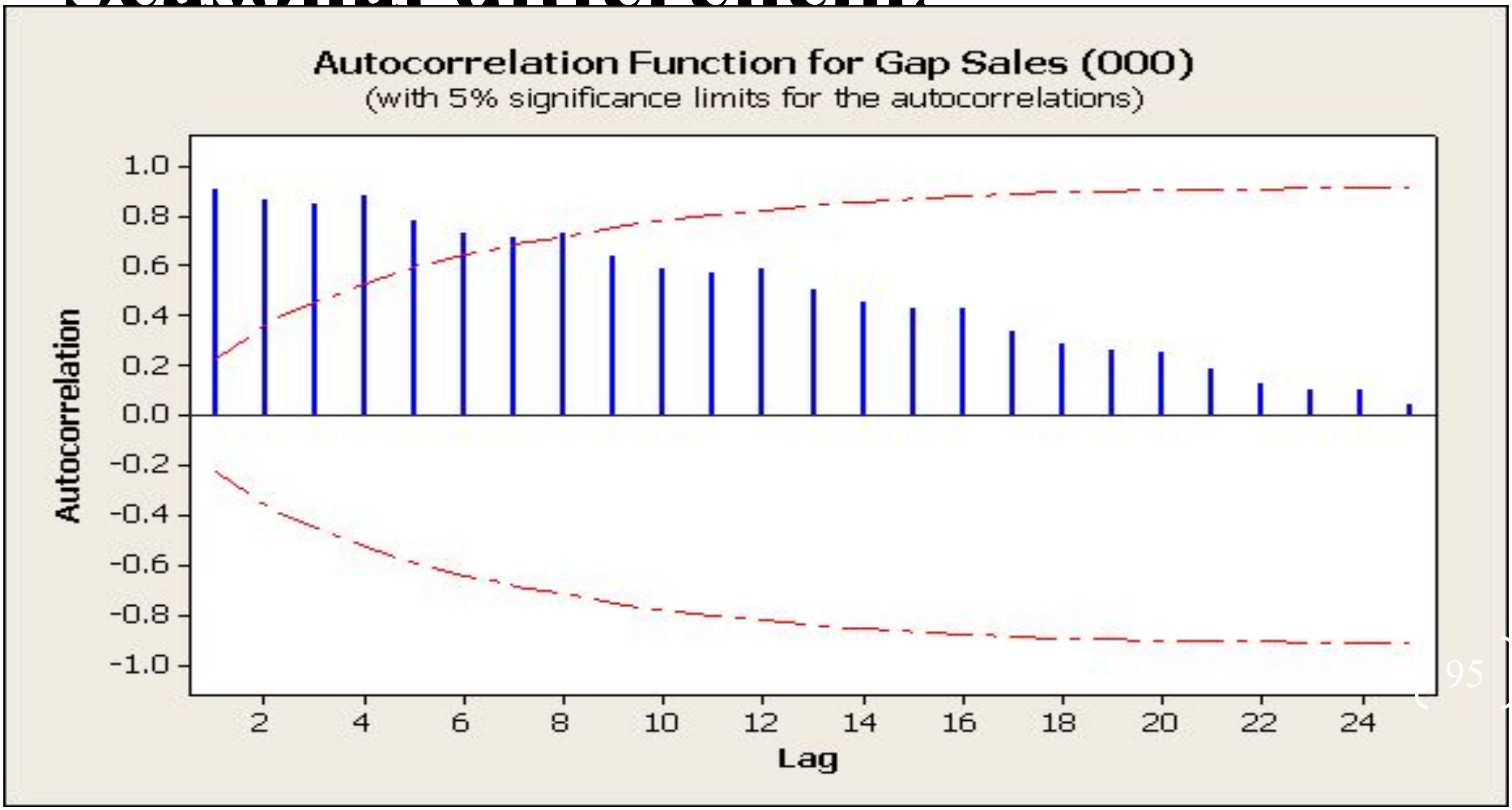
The auto correlations show that:

The series is non-stationary.

The series is seasonal.

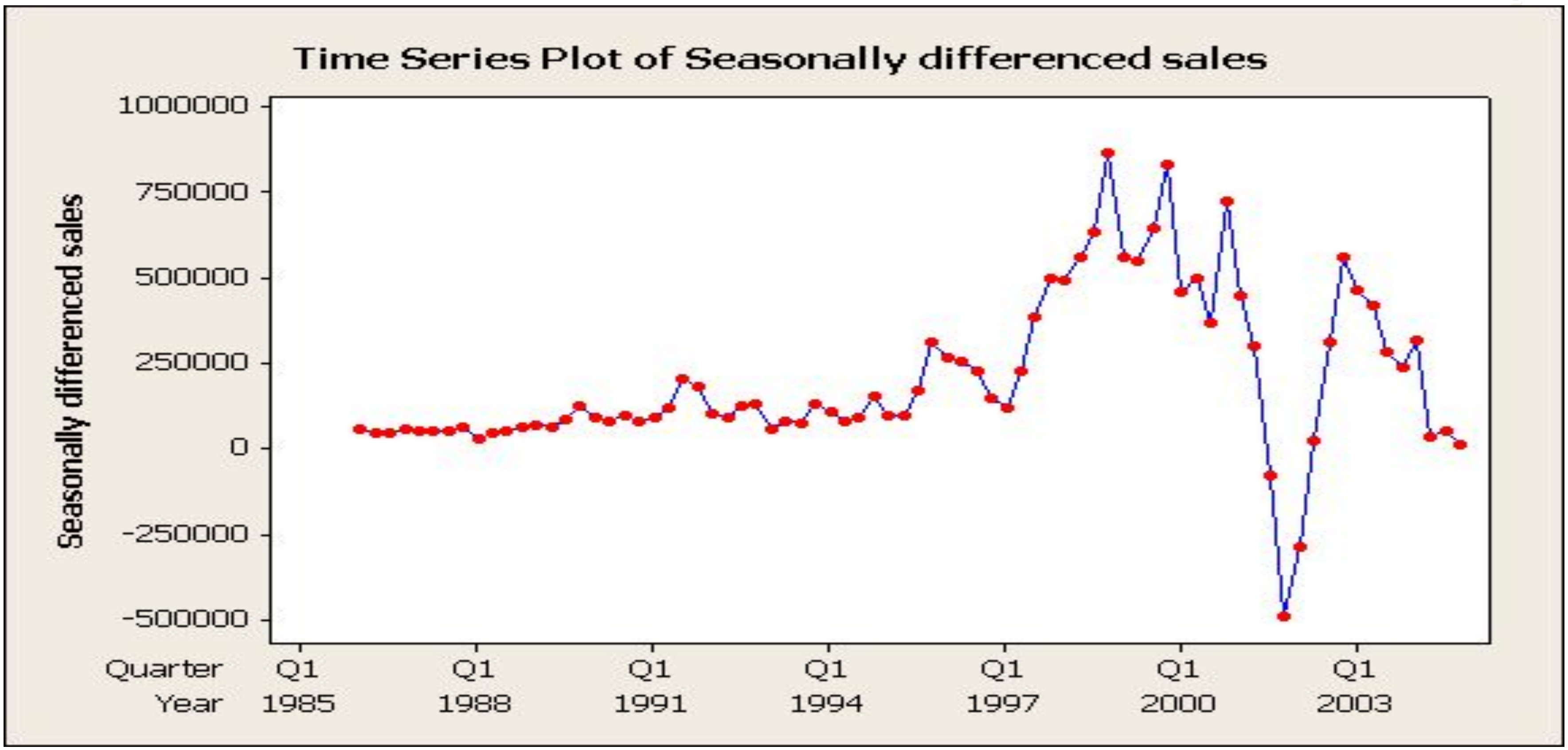

Time Series Plot of Gap Sales (000)
Gap Sales (000)
5000000
4000000
3000000
2000000
1000000
0
Quarter
Year
Q1
1985
Q1
1988
Q1
1991
Q1
1994
Q1
1997
Q1
2000
Q1
2003

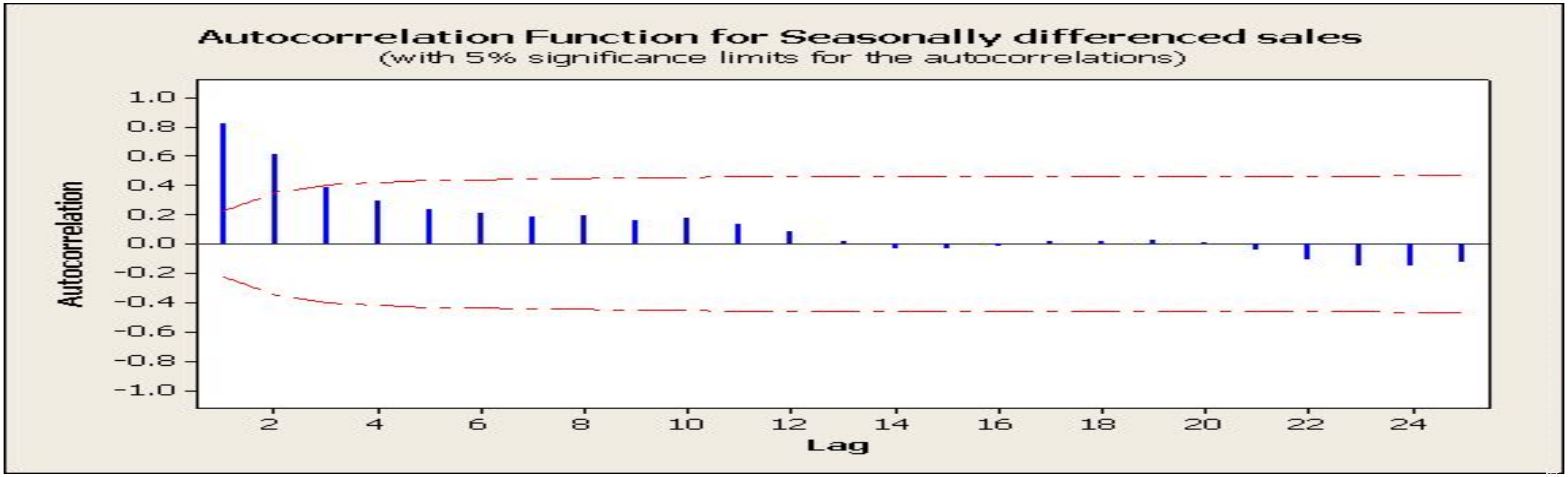
Seasonal differencing



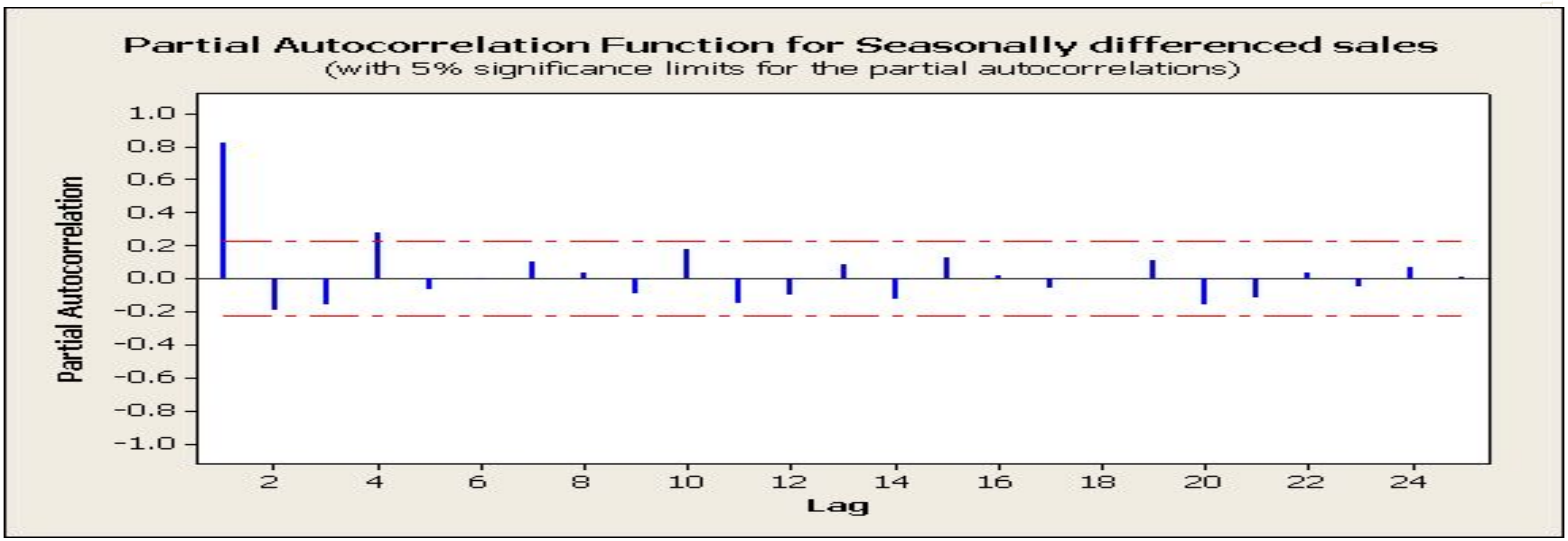
Seasonal differencing

The seasonally differenced series represents the change in sales between quarters of consecutive years.
The time series plot, ACF and PACF of the seasonally differenced Gap's quarterly sales are in the following three slides.





Autocorrelation Function for Seasonally differenced sales
(with 5% significance limits for the autocorrelations)
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
2
4
6
8
10
12
14
16
18
20
22
24
Lag



Partial Autocorrelation Function for Seasonally differenced sales
(with 5% significance limits for the partial autocorrelations)
Partial Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
2
4
6
8
10
12
14
16
18
20
22
24
Lag

Seasonal differencing

The series is now much closer to being stationary, but more than 5% of the spikes are beyond 95% critical limits and autocorrelation shov gradual decline in values.

The seasonality is still present as shown by spike at time lag 4 in the PACF.

The remaining non-stationarity in the mean can be removed with a further first difference.

When both seasonal and first differences are applied, it does not mak no difference which is done first.

Seasonal differencing

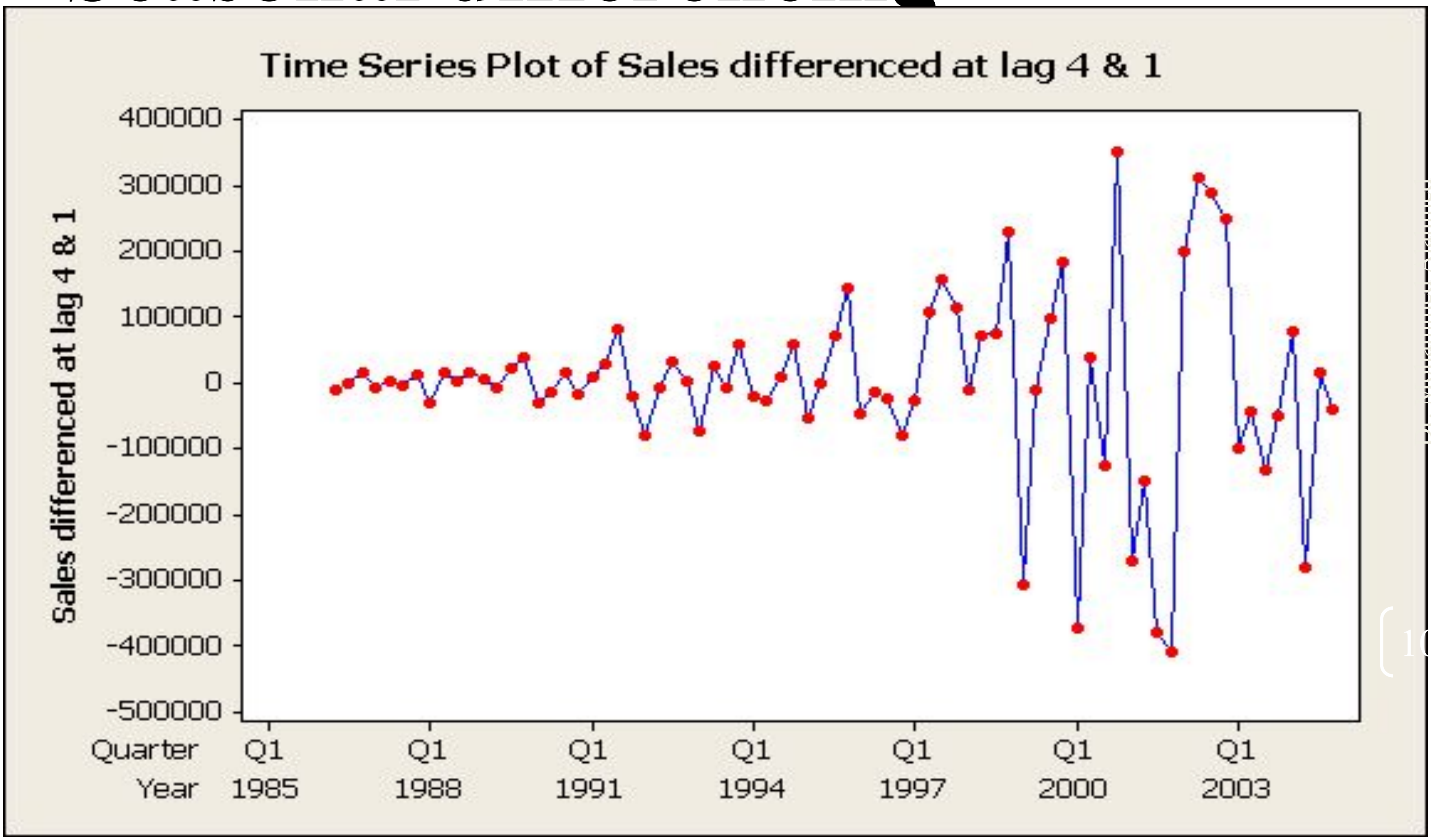
It is recommended to do the seasonal differencing first since sometimes the resulting series will be stationary and hence no need for a further first difference.

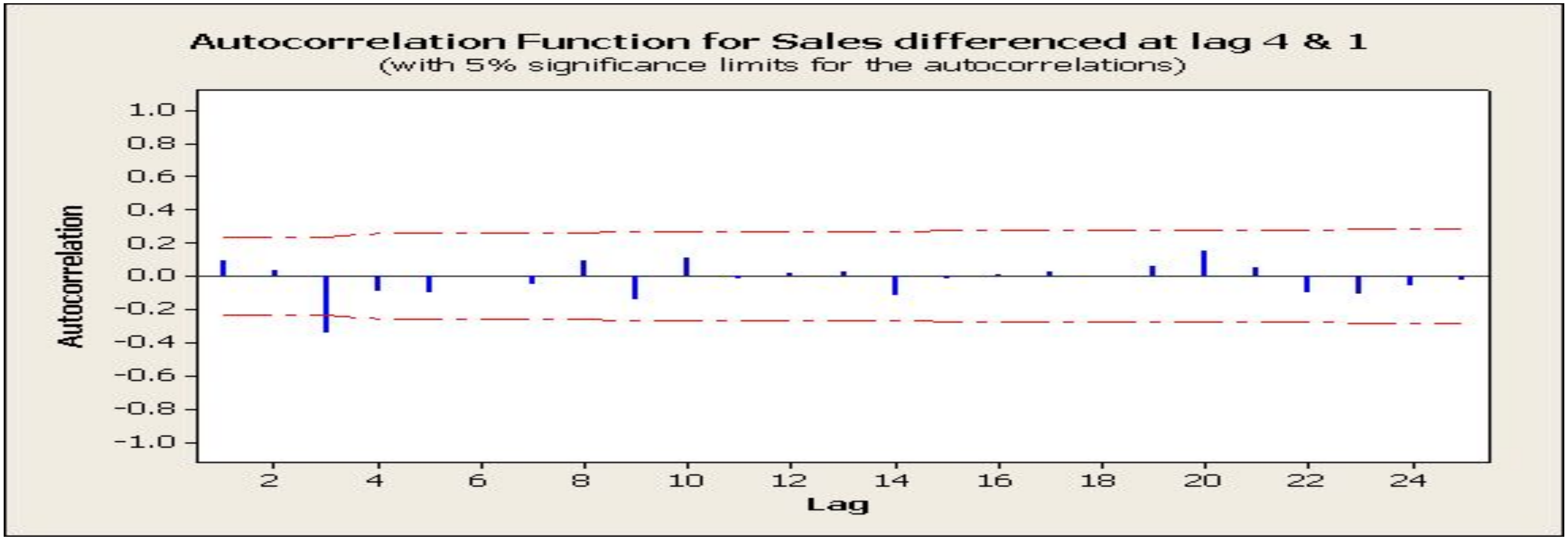
When differencing is used, it is important that the differences be interpretable.

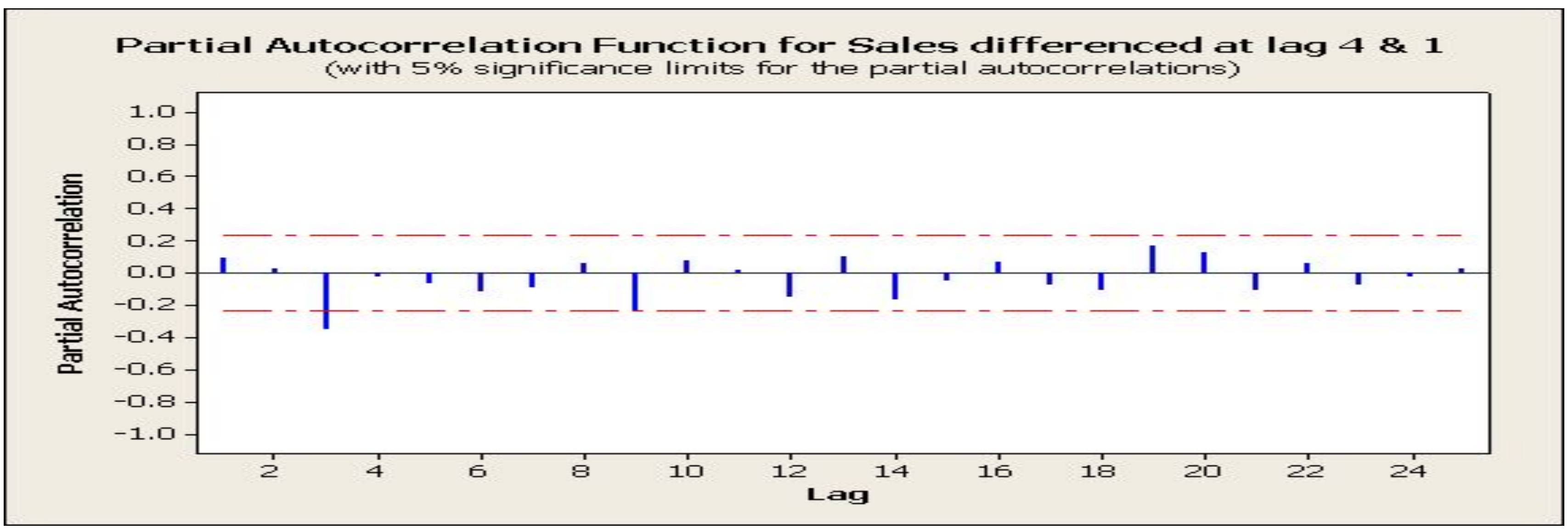
The series resulted from first difference of seasonally differenced Gap's quarterly sales data is reported in the following three slides.

Is the resulting series white noise?

Seasonal differencing




Autocorrelation Function for Sales differenced at lag 4 & 1
(with 5% significance limits for the autocorrelations)
Autocorrelation
Lag


Partial Autocorrelation Function for Sales differenced at lag 4 & 1
(with 5% significance limits for the partial autocorrelations)
Partial Autocorrelation
Lag

Tests for Stationarity

Several statistical tests has been developed to determine if a series is stationary.
These tests are also known as unit root tests.
One of the widely used such test is the Dickey-fuller test.
To carry out the test, fit the regression model

$$y'_t = \phi y_{t-1} + b_1 y'_{t-1} + b_2 y'_{t-2} + \cdots + b_p y'_{t-p}$$

Where:
The number of lagged terms p, is usually set to 3.

y'_t represents the differenced series $y_t - y_{t-1}$

Tests for Stationarity

The value of φ is estimated using ordinary least squares.
If the original series y_t needs differencing, the estimated value of φ will be close to zero.
If y_t is already stationary, the estimated value of φ will be negative.

ARIMA models for time series data

Autoregression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

Consider regression models of the form

Define

$$x_1 = y_{t-1}$$
$$x_2 = y_{t-2}$$
$$\vdots$$
$$x_p = y_{t-p}$$

ARIMA models for time series data

Then

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} + \varepsilon_t$$

The explanatory variables in this equations are time-lagged values of the variable y.
Autoregression (AR) is used to describe models of this form.

Autoregression models should be treated differently from ordinary regression models since:

The explanatory variables in the autoregression models have a built-in dependence relationship.
Determining the number of past values of y_t to include in the model is not always straight forward

ARIMA models for time series data

Moving average model

A time series model which uses past errors as explanatory variable:

$$y_t = \beta_0 + \beta_1 e_{t-1} + \beta_2 e_{t-2} + \cdots \beta_p e_{t-q} + \varepsilon_t$$

is called moving average(MA) model

Note that this model is defined as a moving average of the error series, while the moving average models we discussed previously are the moving average of the observations.

ARIMA models for time series data

Autoregressive (AR) models can be coupled with moving average (MA) models to form a general and useful class of time series model called *Autoregressive Moving Average (ARMA)* models.

These can be used when the data are stationary.

This class of models can be extended to non-stationary series by allowing the differencing of the data series.

These are called *Autoregressive Integrated Moving Average (ARIMA* models.

There are a large variety of ARIMA models.

ARIMA models for time series data

The general non-seasonal model is known as ARIMA (p, d, q):

p is the number of autoregressive terms.

d is the number of differences.

q is the number of moving average terms.

A white noise model is classified as ARIMA (0, 0, 0)

No AR part since y_t does not depend on y_{t-1}

There is no differencing involved.

No MA part since y_t does not depend on e_{t-1}

ARIMA models for time series data

A random walk model is classified as ARIMA (0, 1, 0)

There is no AR part.

There is no MA part.

There is one difference.

Note that if any of p, d, or q are equal to zero, the model can be written in a shorthand notation by dropping the unused part.

Example

ARIMA(2, 0, 0) = AR(2)

ARIMA (1, 0, 1) = ARMA(1, 1)

An autoregressive model of order one AR(1)

The basic form of an ARIMA (1, 0, 0) or AR(1) is:

$$y_t = C + \phi_1 y_{t-1} + e_t$$

Observation y_t depends on y_{t-1}

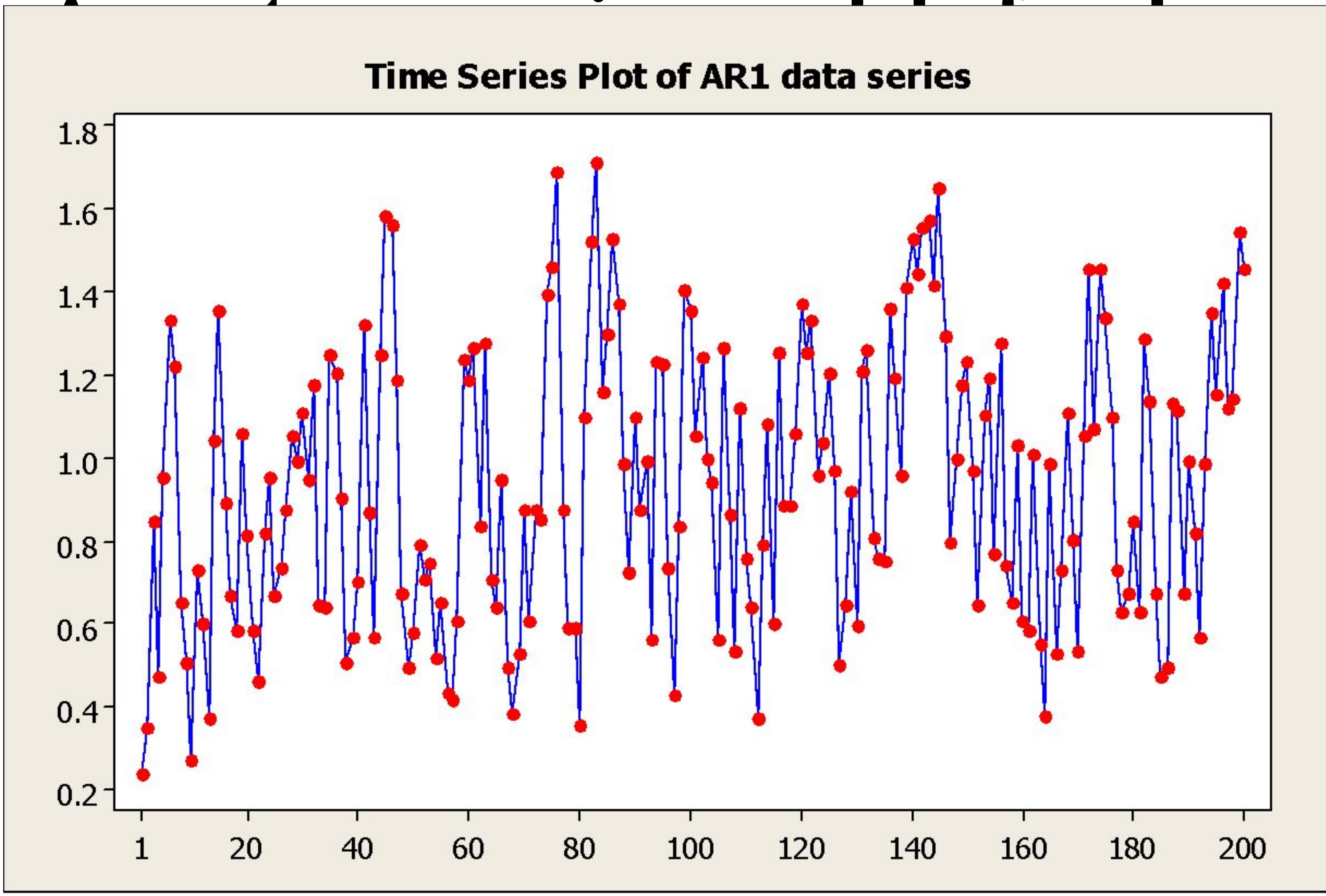
The value of autoregressive coefficient φ_1 is between –1 and 1.

An autoregressive model of order one

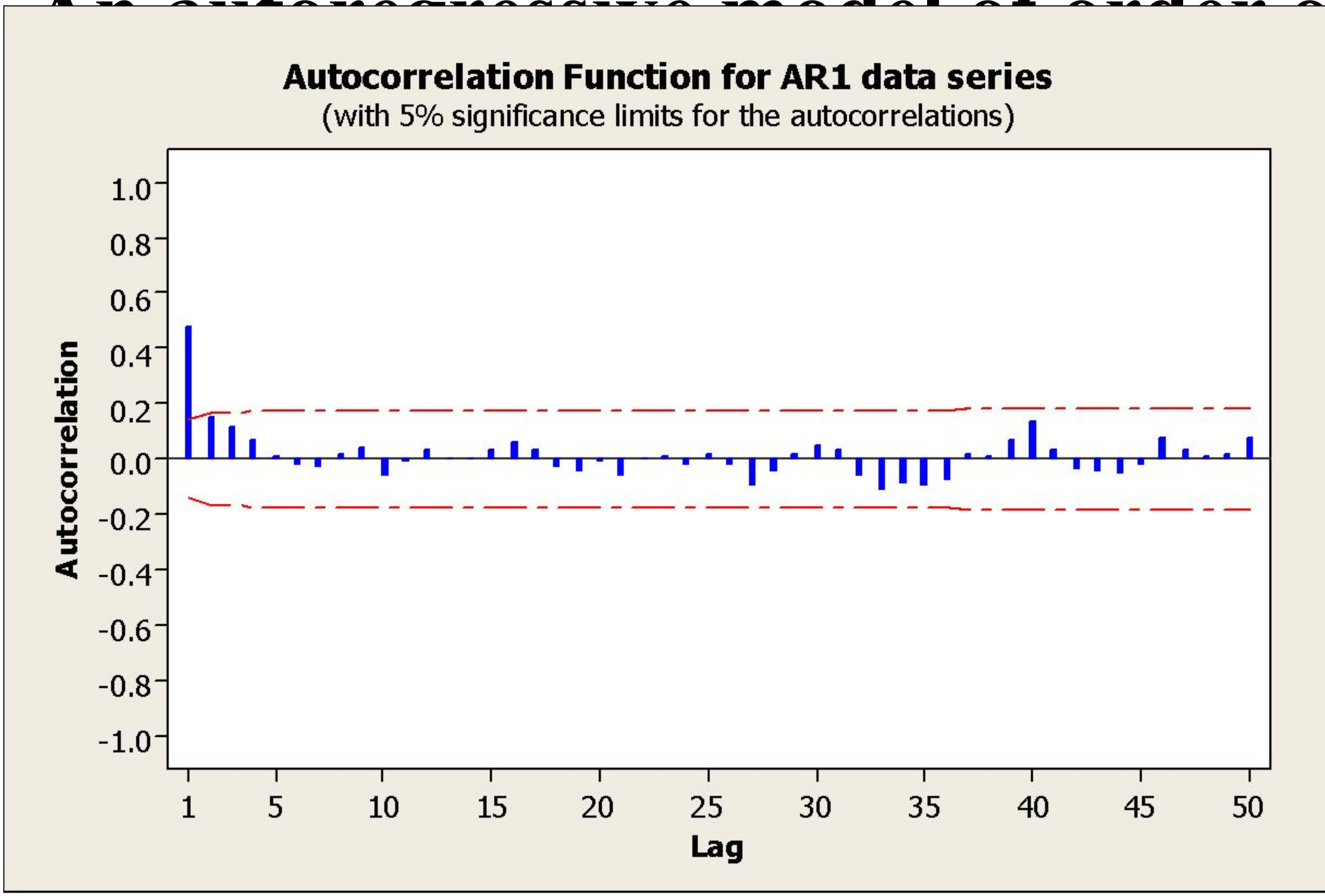
The time plot of an AR(1) model varies with the parameter φ_1:

- When $\varphi_1 = 0$, y_t is equivalent to a white noise series.
- When $\varphi_1 = 1$, y_t is equivalent to a random walk series
- For negative values of φ_1, the series tends to oscillate between positive and negative values.

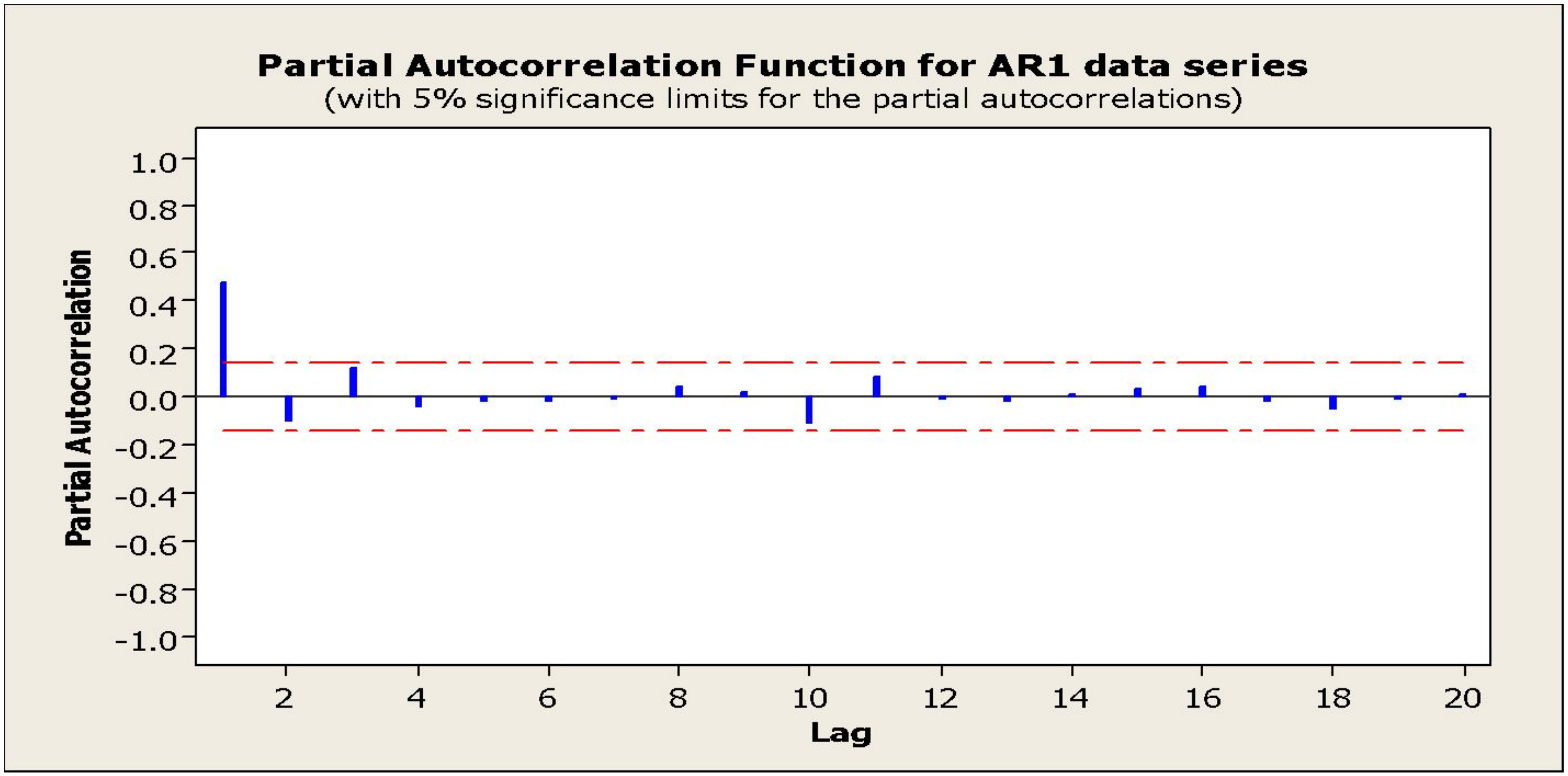
The following slides show the time series, ACF and PACF plot for an ARIMA(1, 0, 0) time series data.


Time Series Plot of AR1 data series
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
1
20
40
60
80
100
120
140
160
180
200

An autoregressive model of order one



An autoregressive model of order one



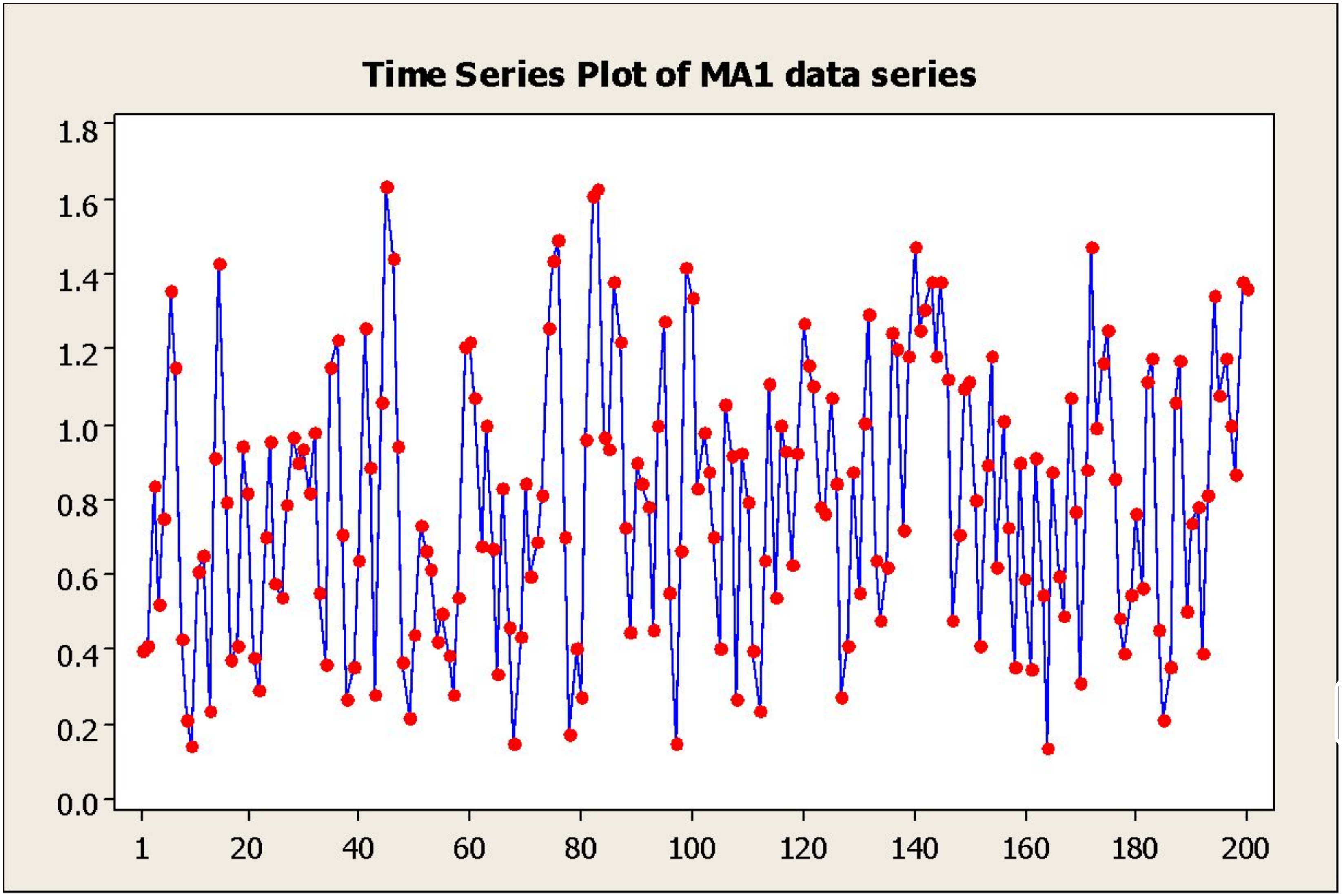
- The ACF and PACF can be used to identify an AR(1) model.
  - The autocorrelations decay exponentially.
  - There is a single significant partial autocorrelation.

Moving Average of order one MA(1)

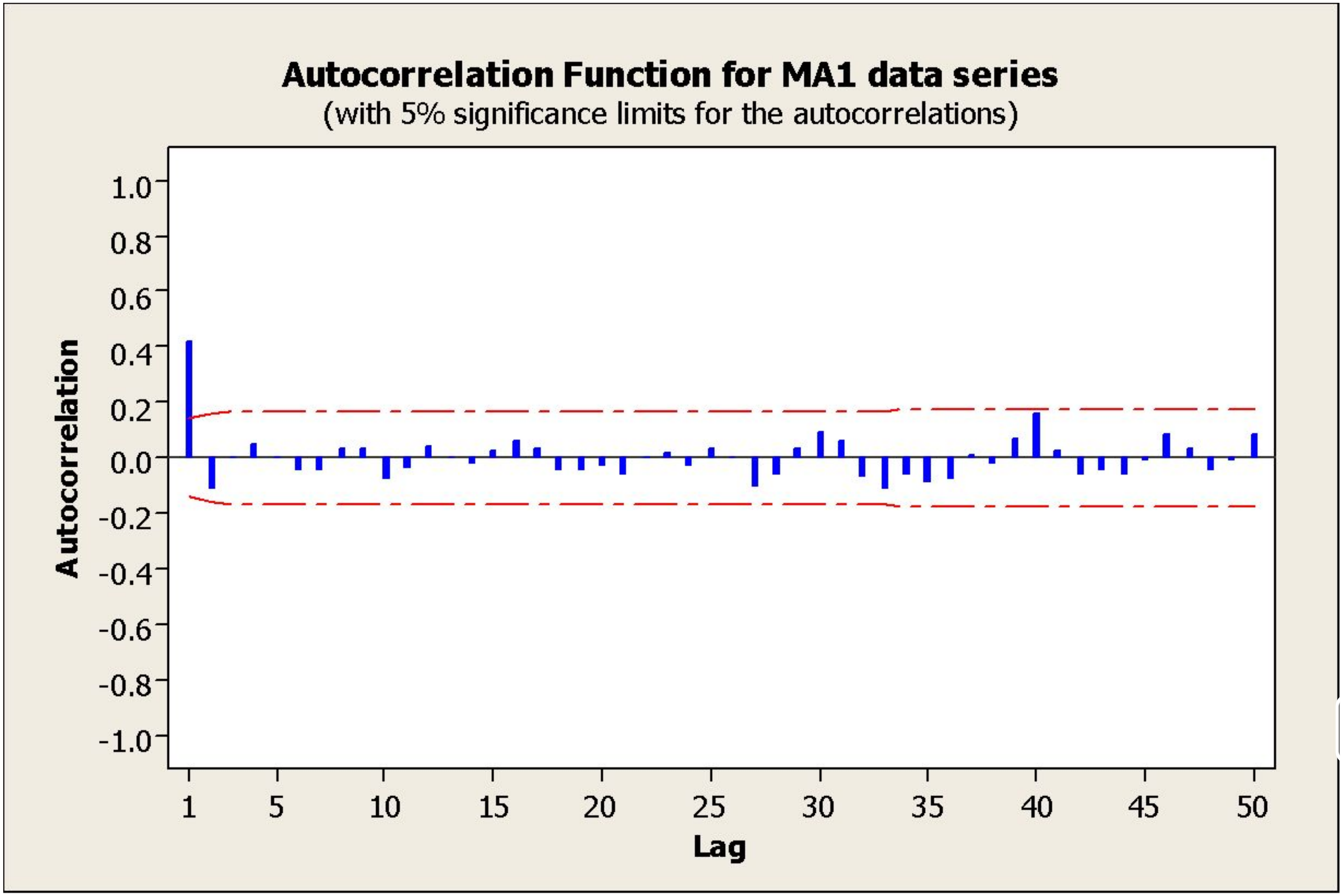
The general $y_t = C + e_t - \theta_1 e_{t-1}$ 1) or MA(1) model is

Y_t depends on the error term e_t and on the previous error term e_{t-1} with coefficient - θ_1.
The value of θ_1 is between –1 and 1.
The following slides show an example of an MA(1) data series.

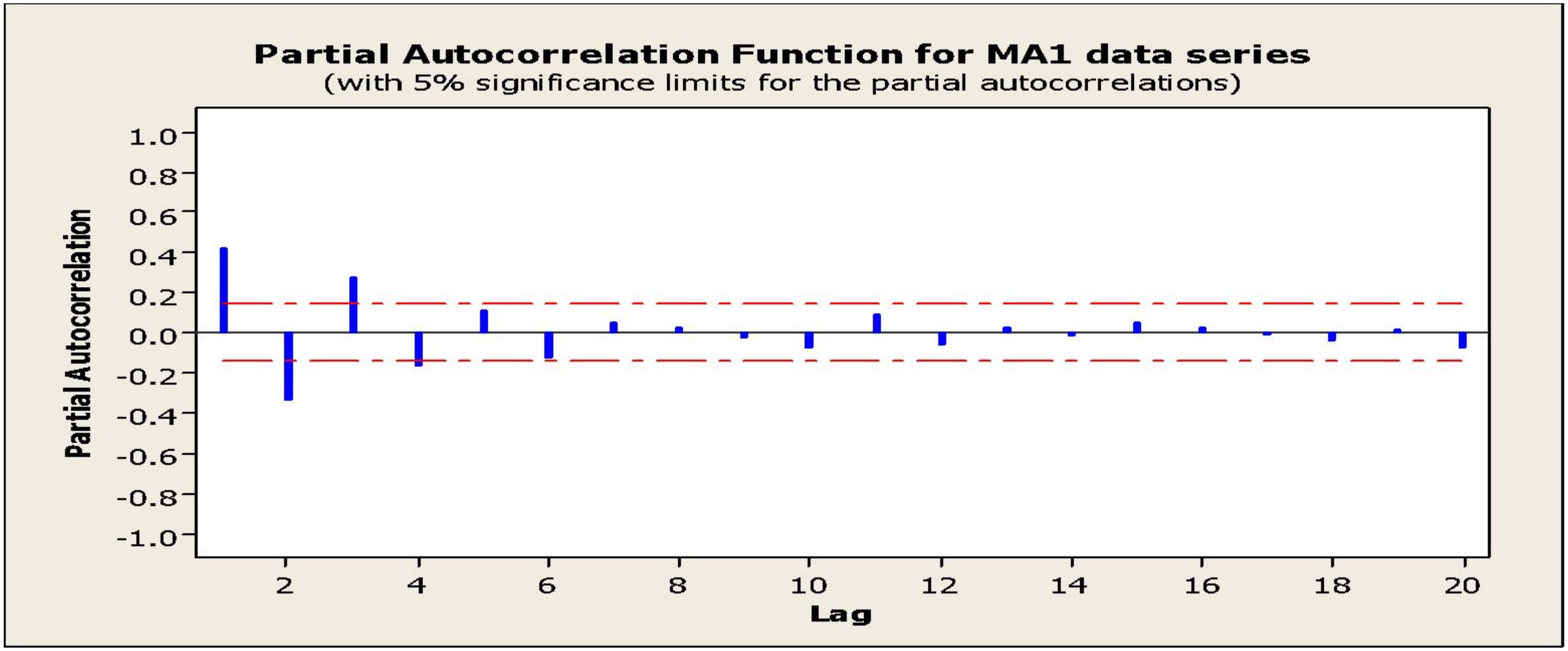
A moving average of order one MA(1)



A moving average of order one MA(1)



A moving average of order one MA(1)



- Note that there is only one significant autocorrelation at time lag 1.
- The partial autocorrelations decay exponentially, but because of random error components, they do not die out to zero as do the theoretical autocorrelation

Higher order auto regressive models

A p

$$y_t = C + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots \phi_p y_{t-p} + e_t$$

C is the constant term

φ_j is the j^{th} auto regression parameter

e_t is the error term at time t.

Higher order auto regressive models

Restrictions on the allowable values of auto regression parameters

For p =1

$-1 < \varphi_1 < 1$

For p = 2

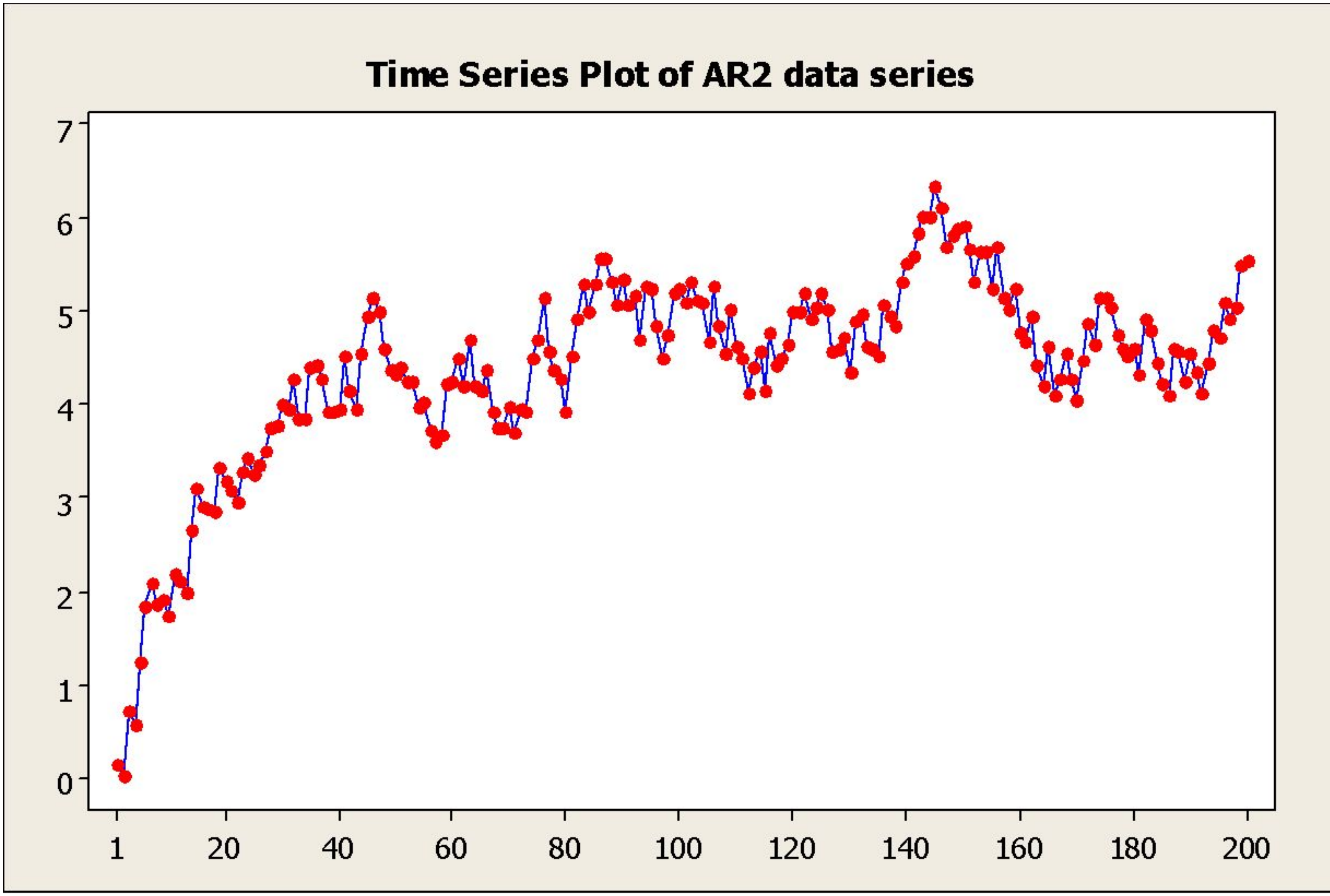
$-1 < \varphi_2 < 1$

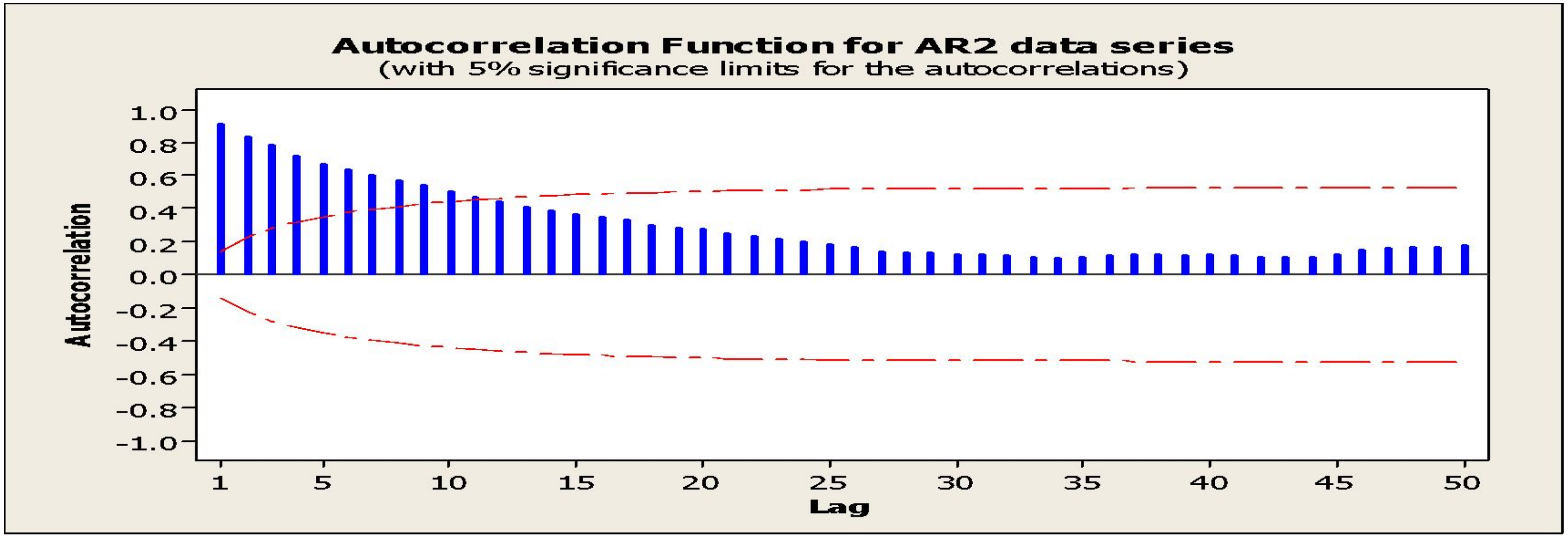
$\varphi_1 + \varphi_2 < 1$

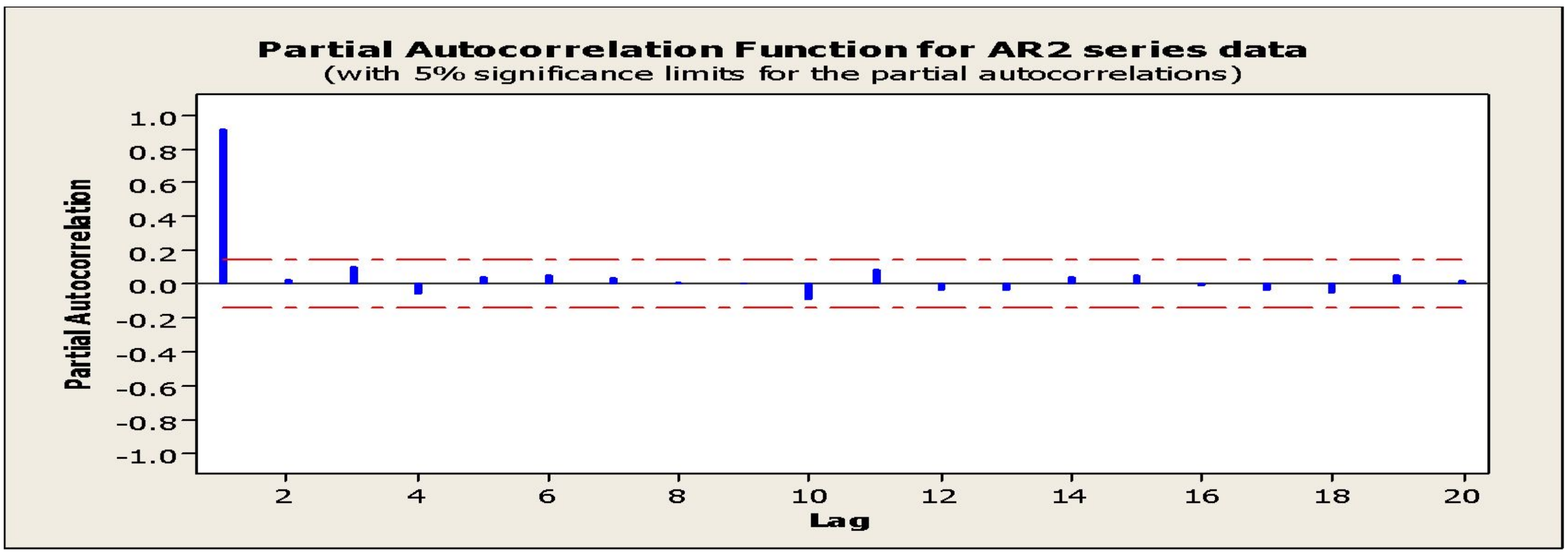
$\varphi_2 - \varphi_1 < 1$

Higher order auto regressive models

A great variety of time series are possible with autoregressive models.
The following slides shows an AR(2) model.
Note that for AR(2) models the autocorrelations die out in a damped Sine-wave patterns.
There are exactly two significant partial autocorrelations.


Time Series Plot of AR2 data series
7
6
5
4
3
2
1
0
1
20
40
60
80
100
120
140
160
180
200


Autocorrelation Function for AR2 data series
(with 5% significance limits for the autocorrelations)
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
1
5
10
15
20
25
30
35
40
45
50
Lag


Partial Autocorrelation Function for AR2 series data
(with 5% significance limits for the partial autocorrelations)
Partial Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
2
4
6
8
10
12
14
16
18
20
Lag

Higher order moving average models

The g ten as:

$$y_t = C + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q}$$

C is the constant term
θ_j is the jth moving average parameter.
e_{t-k} is the error term at time t-k

Higher order moving average models

Restrictions on the allowable values of the MA parameters.

For q =1

$-1 < \theta_1 < 1$

For q =2

$-1 < \theta_2 < 1$

$\theta_1 + \theta_2 < 1$

$\theta_2 - \theta_1 < 1$

Higher order moving average models

A wide variety of time series can be produced using moving average models.
In general, the autocorrelations of an MA(q) models are zero beyond lag q
For $q \geq 2$, the PACF can show exponential decay or damped sine-wave patterns.

Mixtures ARMA models

Basic elements of AR and MA models can be combined to produce a great variety of models.
The follo A(1) and AR(1) models

$$y_t = C + \phi_1 y_{t-1} + e_t - \theta_1 e_{t-1}$$

This is model called ARMA(1, 1) or ARIMA (1, 0, 1)
The series is assumed stationary in the mean and in the variance.

Mixtures ARIMA models

If non-stationarity is added to a mixed ARMA model, then the general ARIMA (p, d, q) is obtained.
The equation for the simplest ARIMA (1, 1, 1) is given below.

$$y_t = C + \phi_1 y_{t-1} - \phi_1 y_{t-2} + e_t - \theta_1 e_{t-1}$$

Mixtures ARIMA models

The general ARIMA (p, d, q) model gives a tremendous variety of patterns in the ACF and PACF, so it is not practical to state rules for identifying general ARIMA models.

In practice, it is seldom necessary to deal with values p, d, or q that are larger than 0, 1, or 2.

It is remarkable that such a small range of values for p, d, or q can cover such a large range of practical forecasting situations.

Seasonality and ARIMA models

The ARIMA models can be extended to handle seasonal components of a data series.

The general shorthand notation is

ARIMA (p, d, q)(P, D, Q)$_s$

Where s is the number of periods per season.

$$y_t = (1+\phi_1) y_{t-1} - \phi_1 y_{t-2} + (1+\Phi_1) y_{t-4} - (1+\phi_1+\Phi_1+\phi_1\Phi_1) y_{t-6}$$
$$- \Phi_1 y_{t-8} + (\Phi_1 + \phi_1\Phi_1) y_{t-9} - \phi_1\Phi_1 y_{t-10} + e_t - \theta_1 e_{t-1} - \Theta_1 e_{t-4} + \theta_1\Theta_1 e_{t-5}$$

Once the coefficients φ_1, Φ_1, θ_1, and Θ_1 have been estimated from the data, the above equation can be used for forecasting.

Seasonality and ARIMA models

The seasonal lags of the ACF and PACF plots show the seasonal parts of an AR or MA model.

Examples:

1. Seasonal MA model:

   $ARIMA(0,0,0)(0,0,1)_{12}$

   will show a spike at lag 12 in the ACF but no other significant spikes.

   The PACF will show exponential decay in the seasonal lags i.e. at lags 12, 24, 36,…

2. Seasonal AR model:

   $ARIMA(0,0,0)(1,0,0)_{12}$

   will show exponential decay in seasonal lags of the ACF.

   Single significant spike at lag 12 in the PACF.

Implementing the model –Building Strategy

The Box –Jenkins approach uses an iterative model-building strategy that consist of:

1. Selecting an initial model (model identification)
2. Estimating the model coefficients (parameter estimation)
3. Analyzing the residuals (model checking)

If necessary, the initial model is modified and the process is repeated until the residual indicate no further modification is necessary.
At this point the fitted model can be used for forecasting.

Model identification

The following approach outlines an approach to select an appropriate model among a large variety of ARIMA models possible.

- Plot the data
  - Identify any unusual observations
  - If necessary, transform the data to stabilize the variance

Check the time series plot, ACF, PACF of the data (possibly transformed) for stationarity.

IF

- Time plot shows the data scattered horizontally around a constant mean
- ACF and PACF to or near zero quickly

Then, the data are stationary.

Model identification

Use differencing to transform the data into a stationary series

For no-seasonal data take first differences

For seasonal data take seasonal differences

Check the plots again if they appear non-stationary, take the differences of the differenced data.

When the stationarity has been achieved, check the ACF and PACF plots for any pattern remaining.

Model identification

There are three possibilities:

AR or MA models

No significant ACF after time lag q indicates MA(q) may be appropriate.

No significant PACF after time lag p indicates that AR(p) may be appropriate.

Seasonality is present if ACF and/or PACF at the seasonal lags are large and significant.
If no clear MA or AR model is suggested, a mixture model may be appropriate

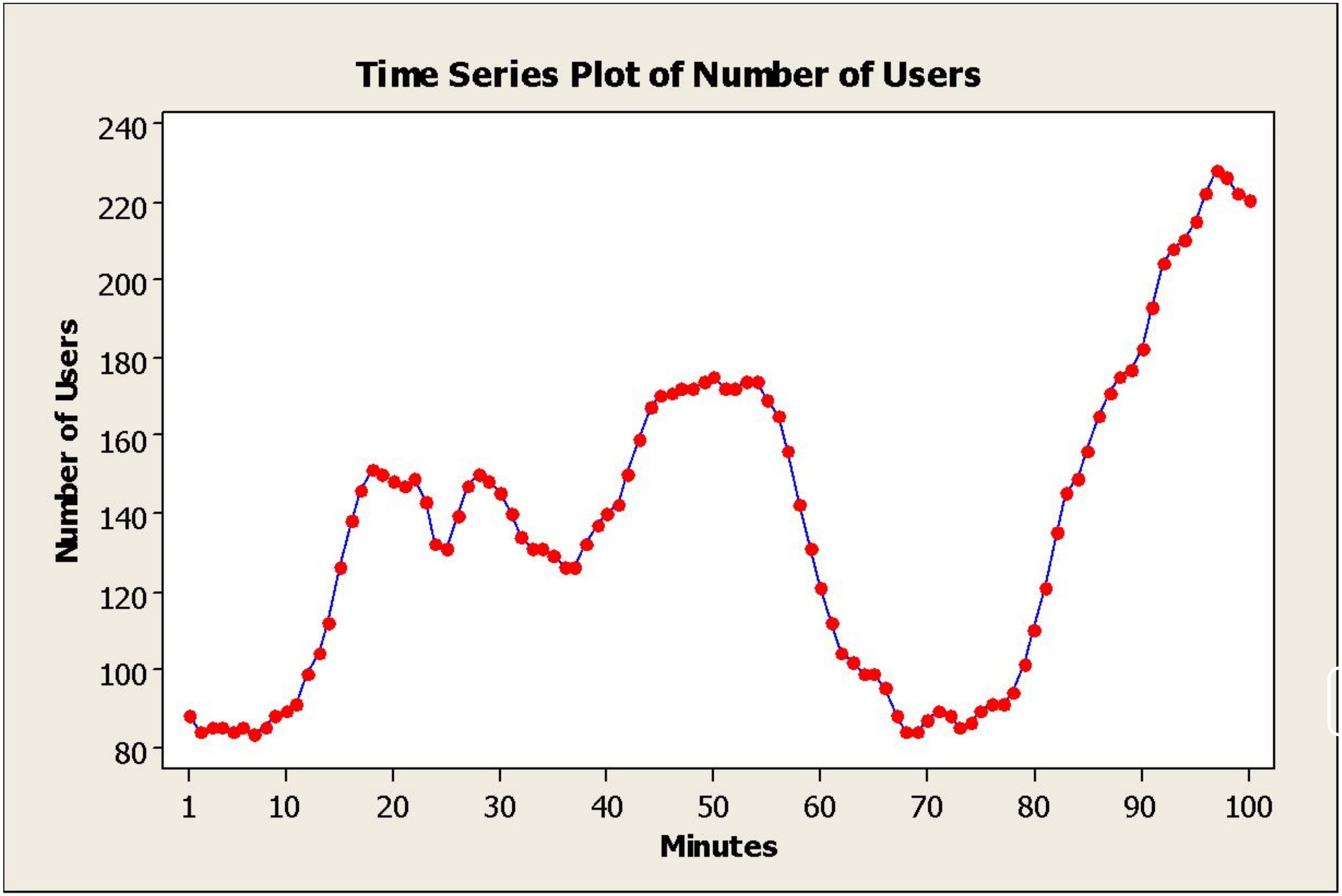
Model identification

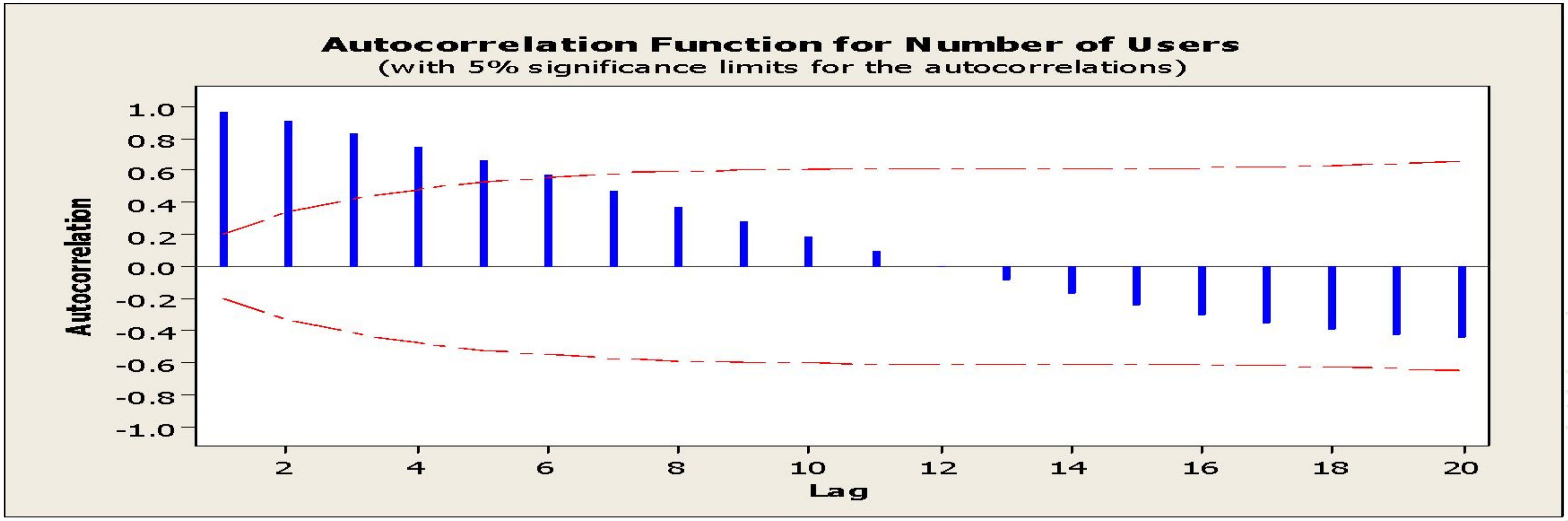
Example (1):

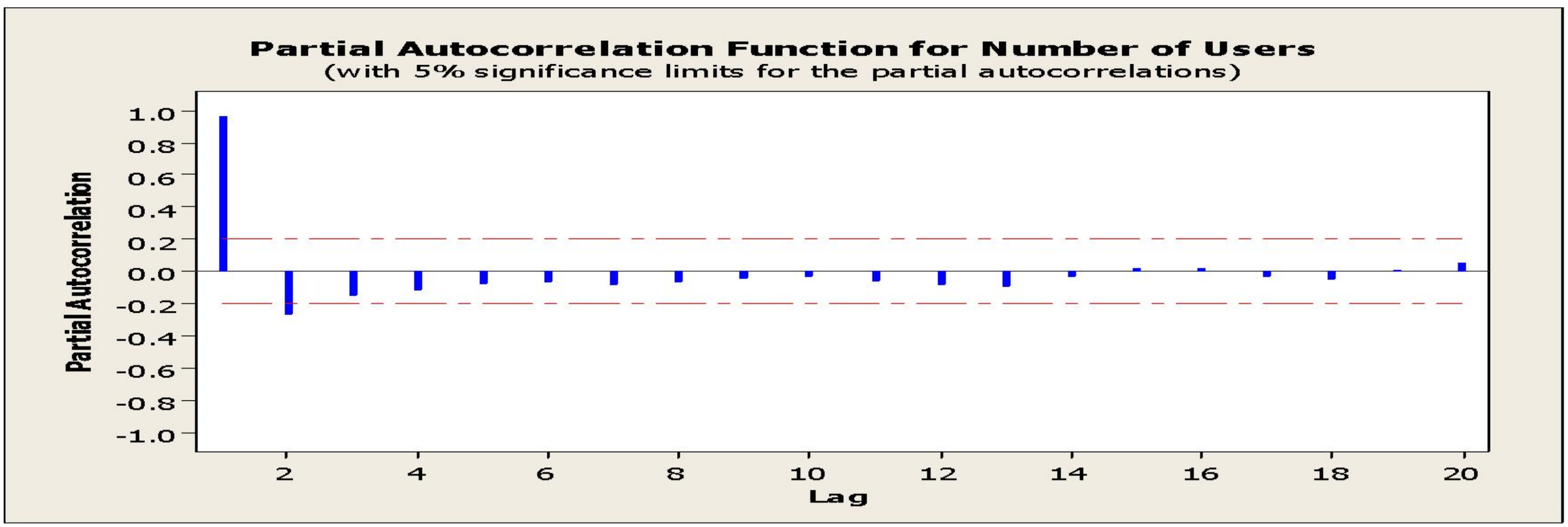
Non seasonal time series data.

The following example looks at the number of users logged onto an internet server over a 100 minutes period.

The time plot, ACF and PACF is reported in the following three slides.

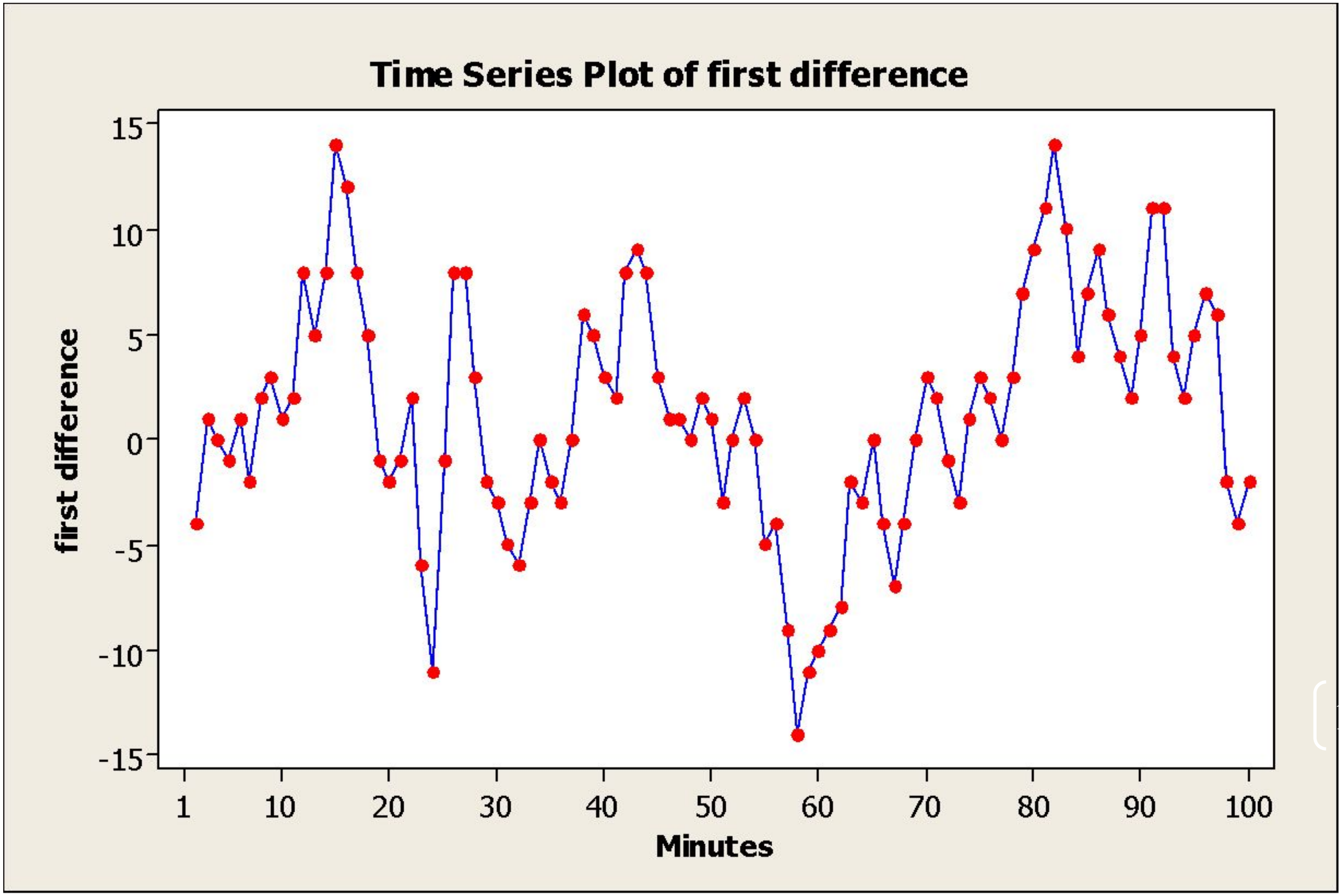

Time Series Plot of Number of Users
Number of Users
240
220
200
180
160
140
120
100
80
1
10
20
30
40
50
60
70
80
90
100
Minutes


Autocorrelation Function for Number of Users
(with 5% significance limits for the autocorrelations)
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
2
4
6
8
10
12
14
16
18
20
Lag

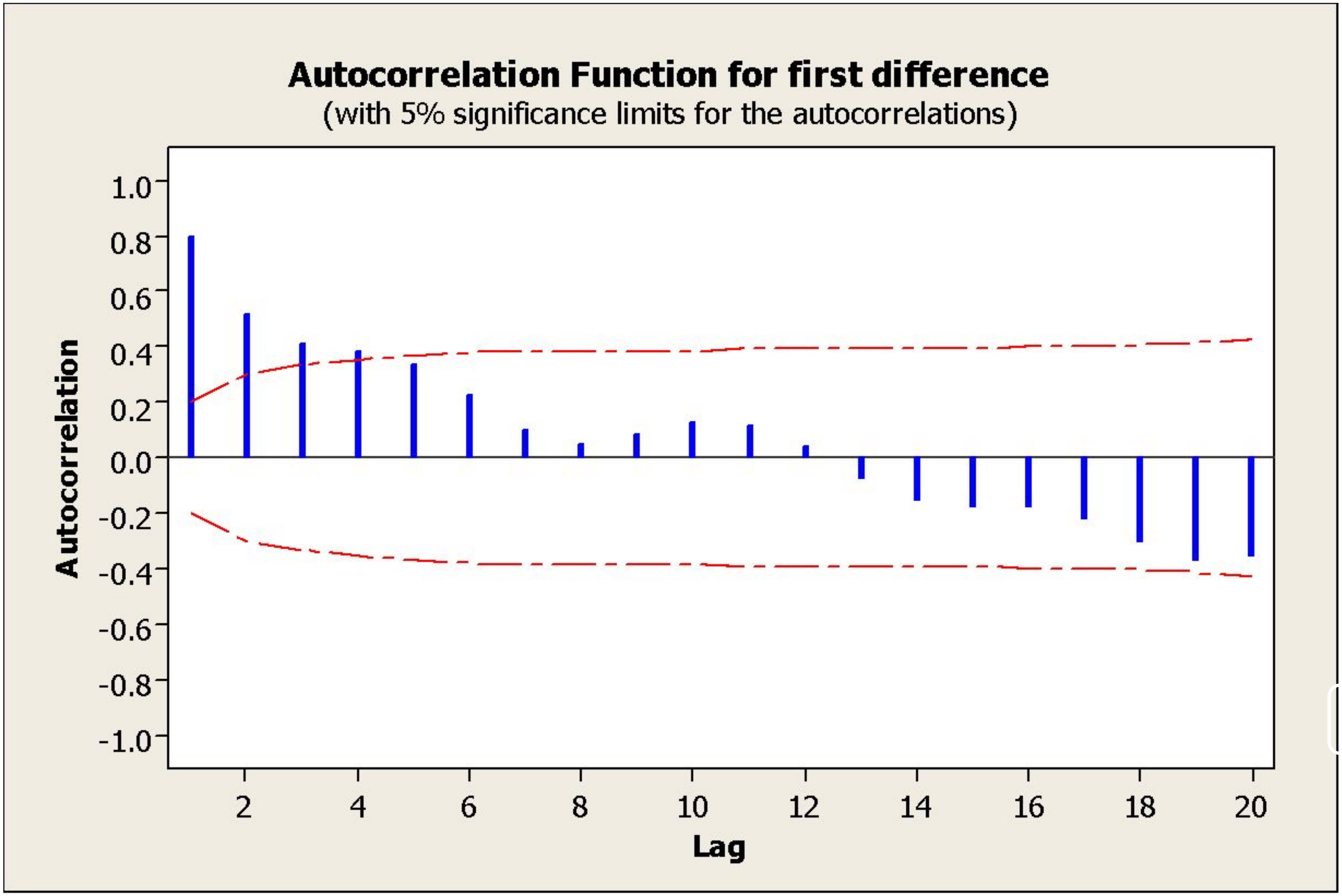

Partial Autocorrelation Function for Number of Users
(with 5% significance limits for the partial autocorrelations)
Partial Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
2
4
6
8
10
12
14
16
18
20
Lag

Model identification

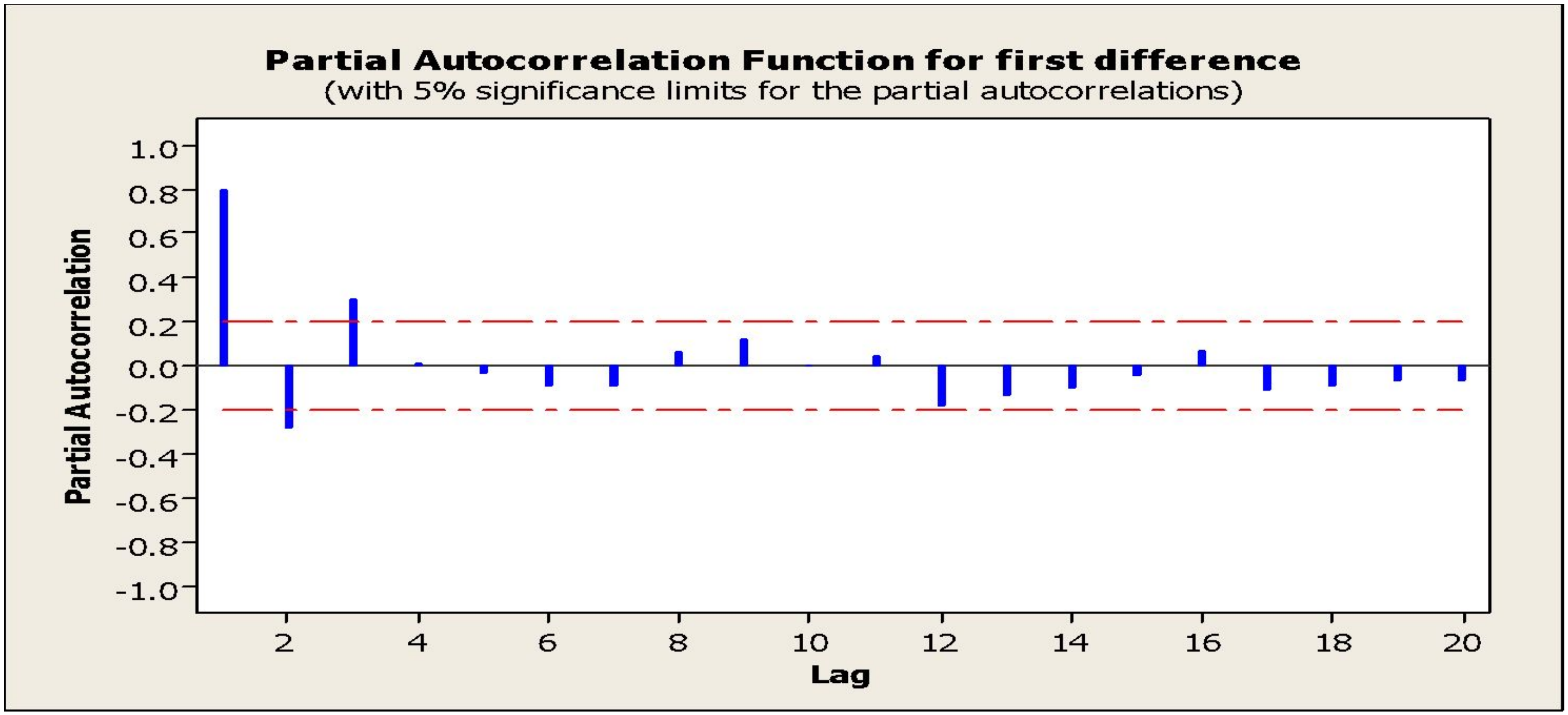
The gradual decline of ACF values indicates non-stationary series.
The first partial autocorrelation is very dominant and close to 1, indicating non-stationarity.
The time series plot clearly indicates non-stationarity.
We take the first differences of the data and reanalyze.



Time Series Plot of first difference
first difference
15
10
5
0
-5
-10
-15
1
10
20
30
40
50
60
70
80
90
100
Minutes



Autocorrelation Function for first difference
(with 5% significance limits for the autocorrelations)
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
Autocorrelation
2
4
6
8
10
12
14
16
18
20
Lag



- ACF shows a mixture of exponential decay and sine-wave pattern
- PACF shows three significant PACF values.
- This suggests an AR(3) model.
- This identifies an ARIMA(3,1,0).

Model identification

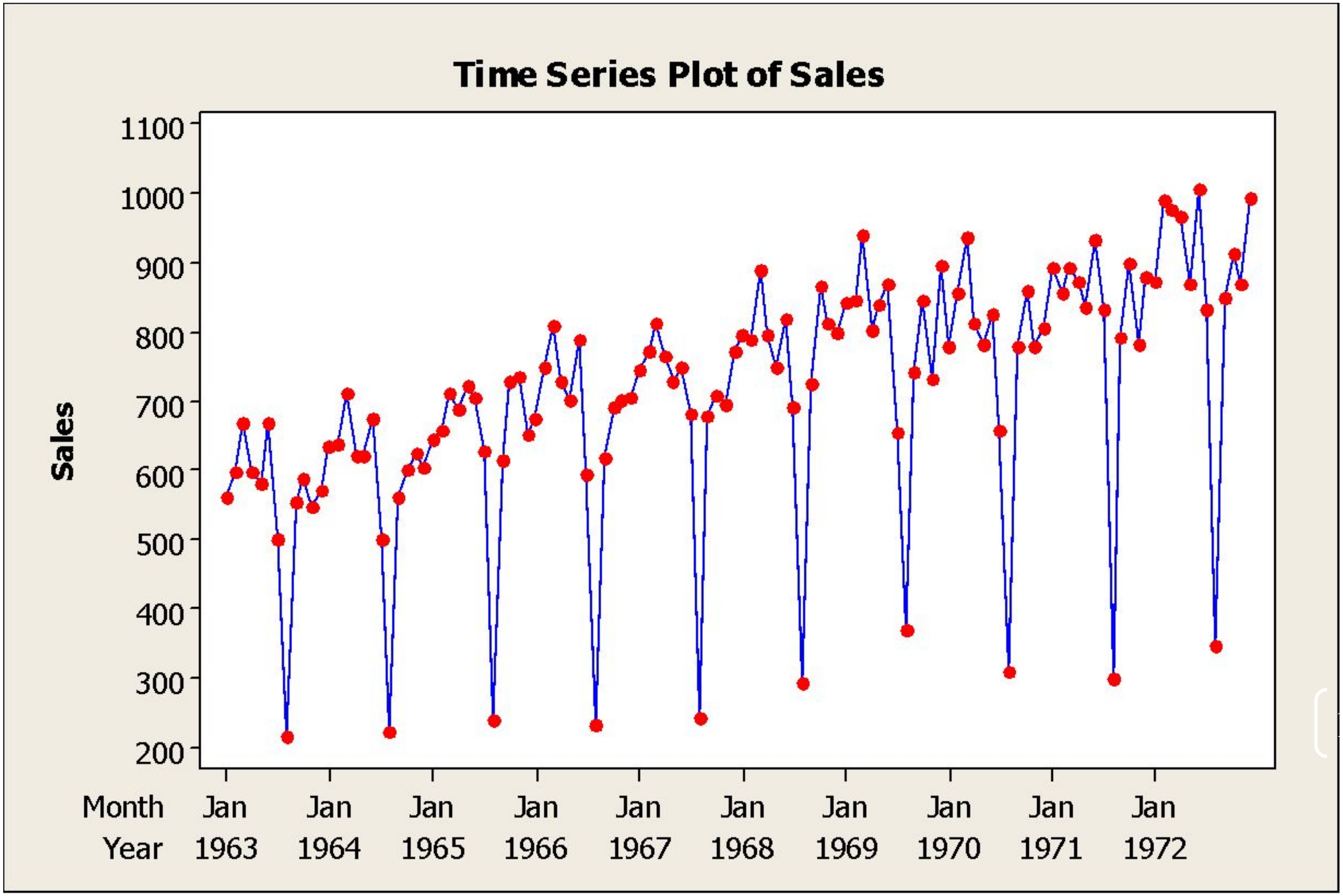
Example (2):

A seasonal time series.

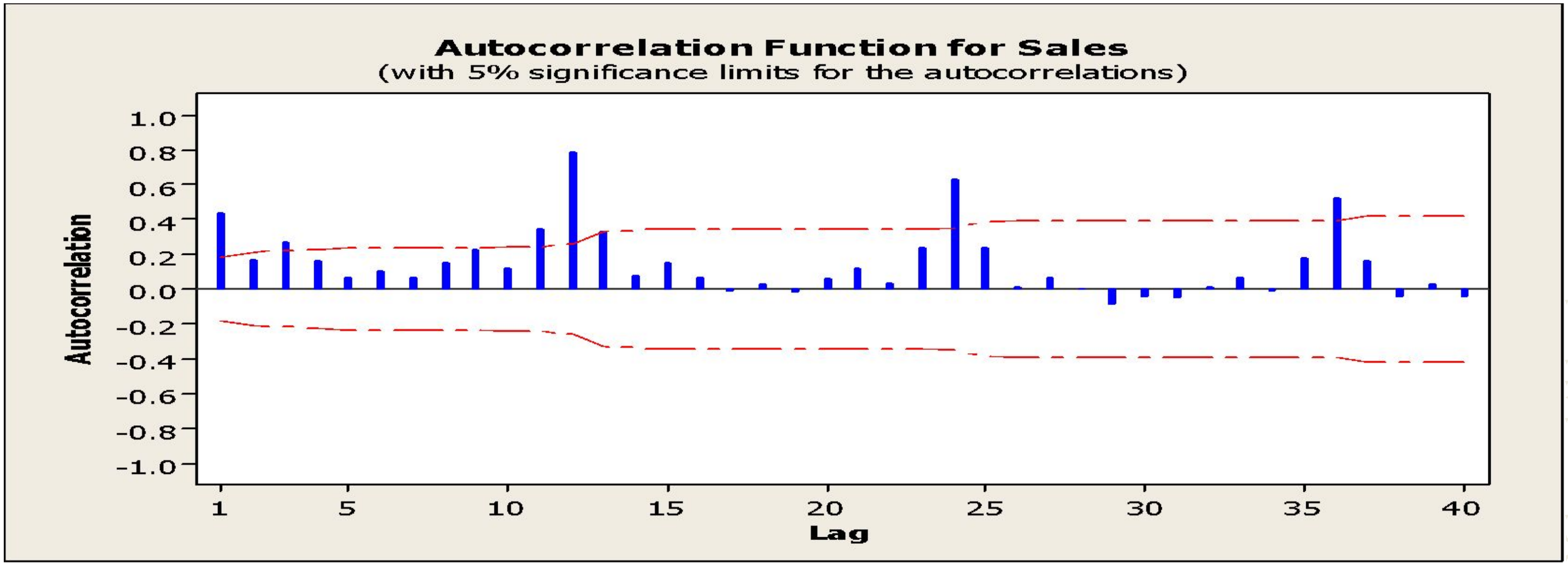
The following example looks at the monthly industry sales (in thousands of francs) for printing and writing papers between the years 1963 and 1972.

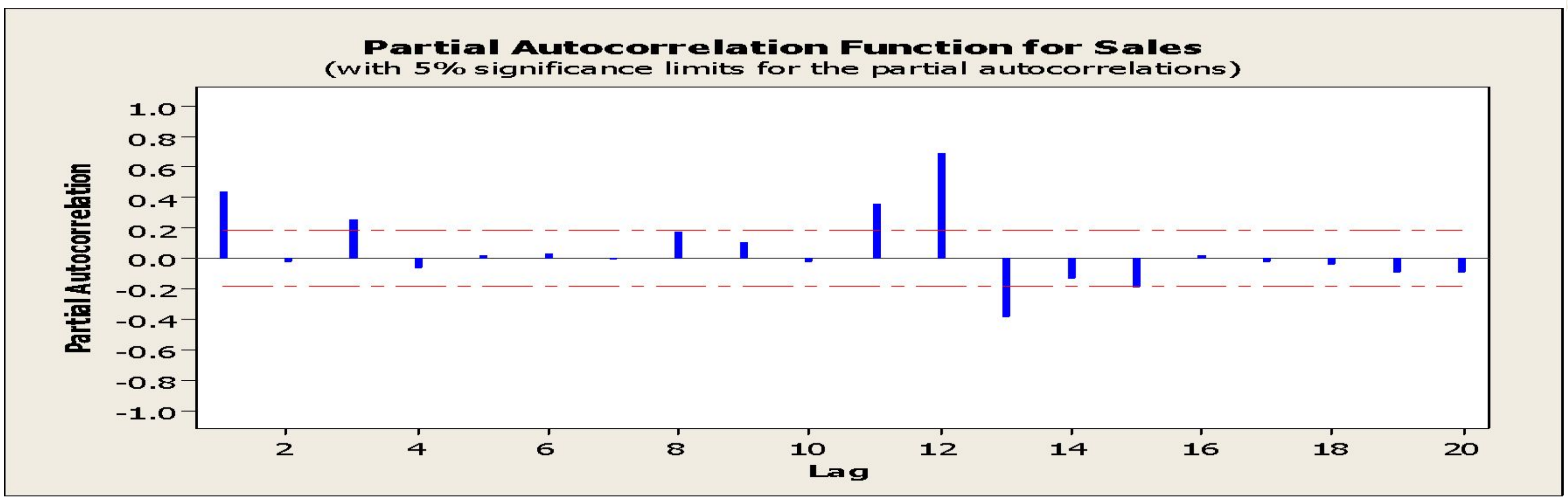
The time plot, ACF and PACF shows a clear seasonal pattern in the data.

This is clear in the large values at time lag 12, 24 and 36.



Time Series Plot of Sales
Sales
1100
1000
900
800
700
600
500
400
300
200
Month
Year
Jan 1963
Jan 1964
Jan 1965
Jan 1966
Jan 1967
Jan 1968
Jan 1969
Jan 1970
Jan 1971
Jan 1972


Autocorrelation Function for Sales
(with 5% significance limits for the autocorrelations)
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
1
5
10
15
20
25
30
35
40
Lag

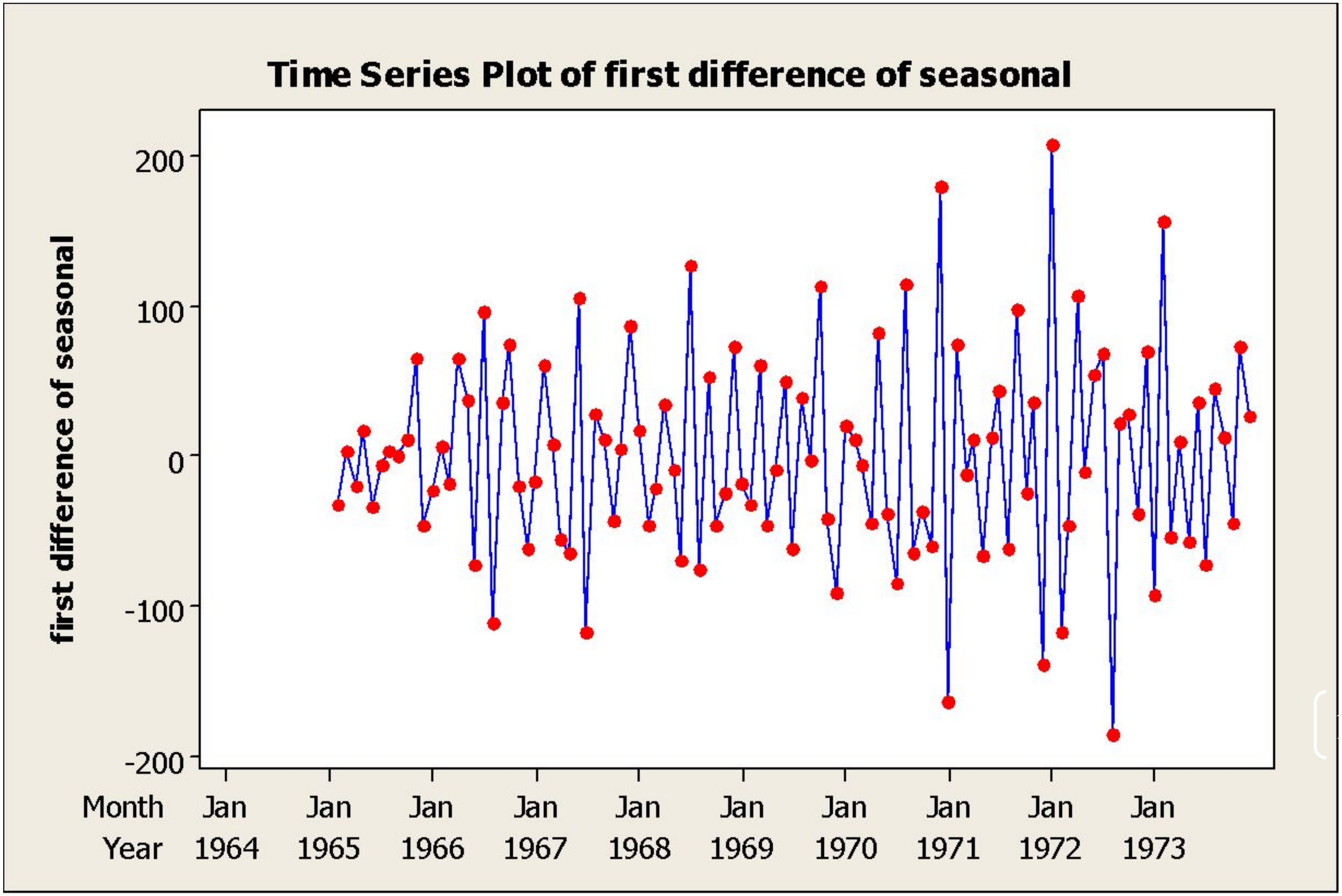

Partial Autocorrelation Function for Sales
(with 5% significance limits for the partial autocorrelations)
Partial Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
2
4
6
8
10
12
14
16
18
20
Lag

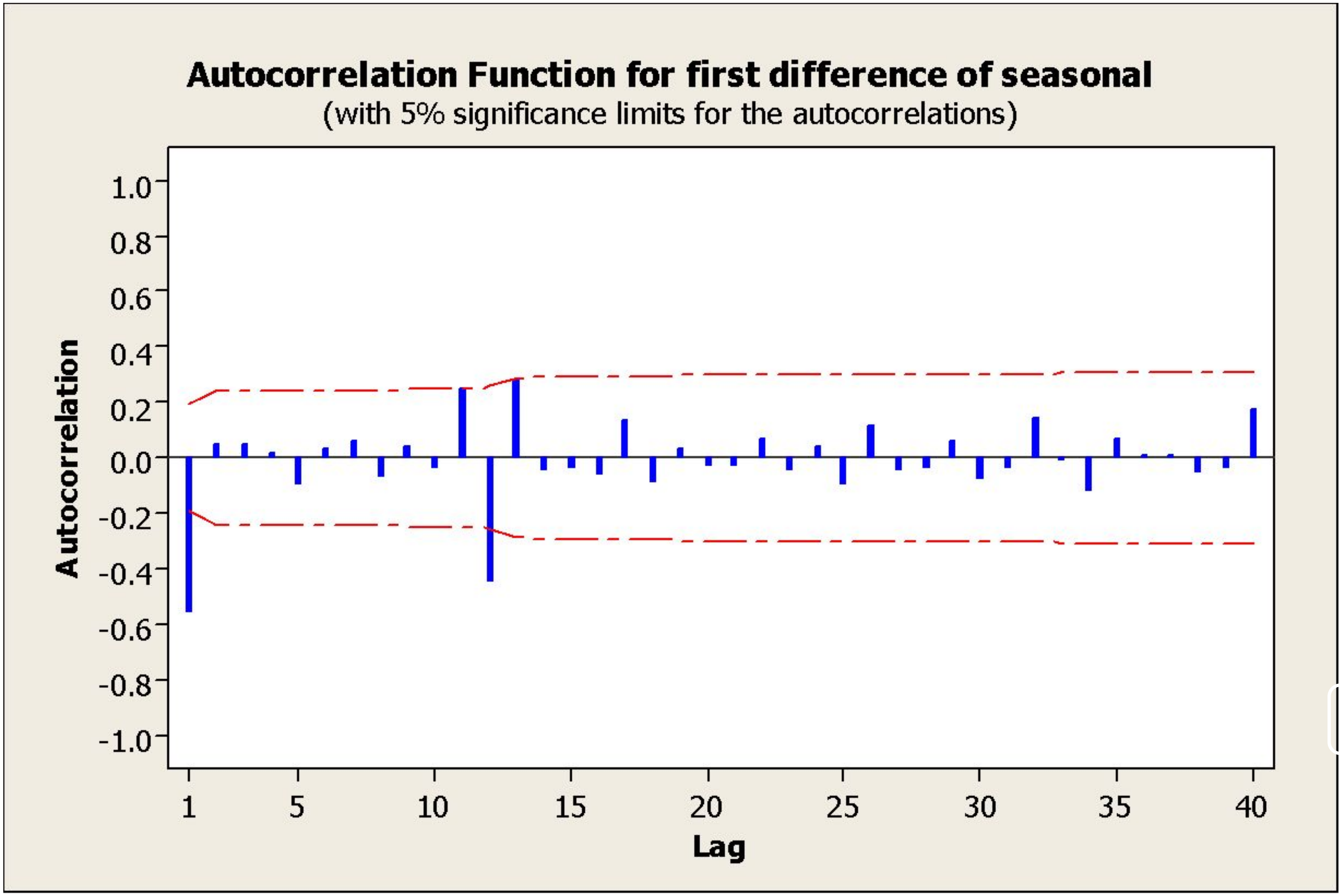
Model identification

We take a seasonal difference and check the time plot, ACF and PACF.

The seasonally differenced data appears to be non-stationary (the plots are not shown), so we difference the data again.

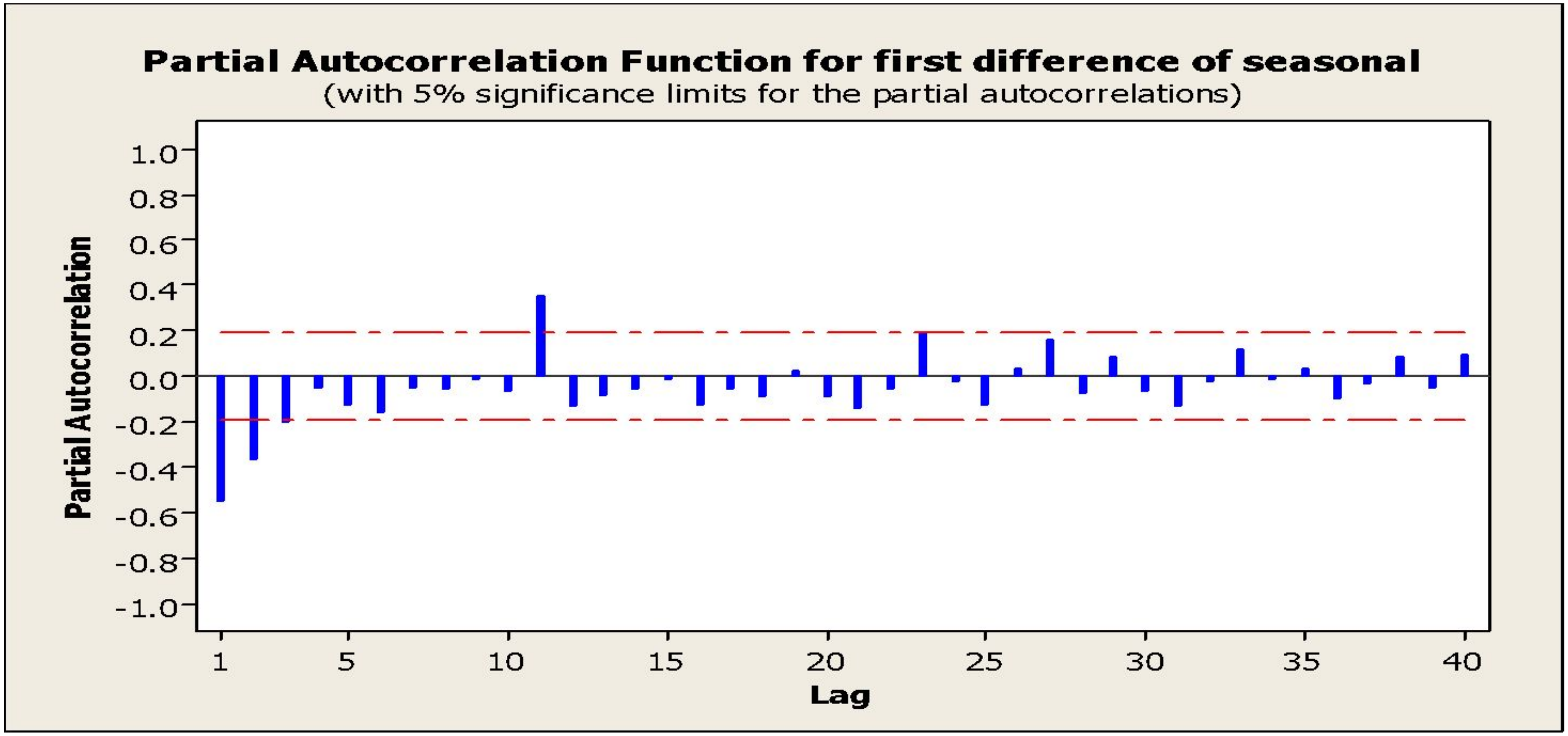
the following three slides show the twice differenced series.


Time Series Plot of first difference of seasonal
first difference of seasonal
200
100
0
-100
-200
Month
Year
Jan
1964
Jan
1965
Jan
1966
Jan
1967
Jan
1968
Jan
1969
Jan
1970
Jan
1971
Jan
1972
Jan
1973



Autocorrelation Function for first difference of seasonal
(with 5% significance limits for the autocorrelations)
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
Autocorrelation
1
5
10
15
20
25
30
35
40
Lag

Model identification



- The PACF shows the exponential decay in values.
- The ACF shows a significant value at time lag 1.
  - This suggest a MA(1) model.
- The ACF also shows a significant value at time lag 12
  - This suggest a seasonal MA(12).

Model identification

Therefore, the identifies model is

$$\text{ARIMA }(0,1,1)(0,1,1)_{12}.$$

This model is sometimes is called the “airline model” because it was applied to international airline data by Box and Jenkins.

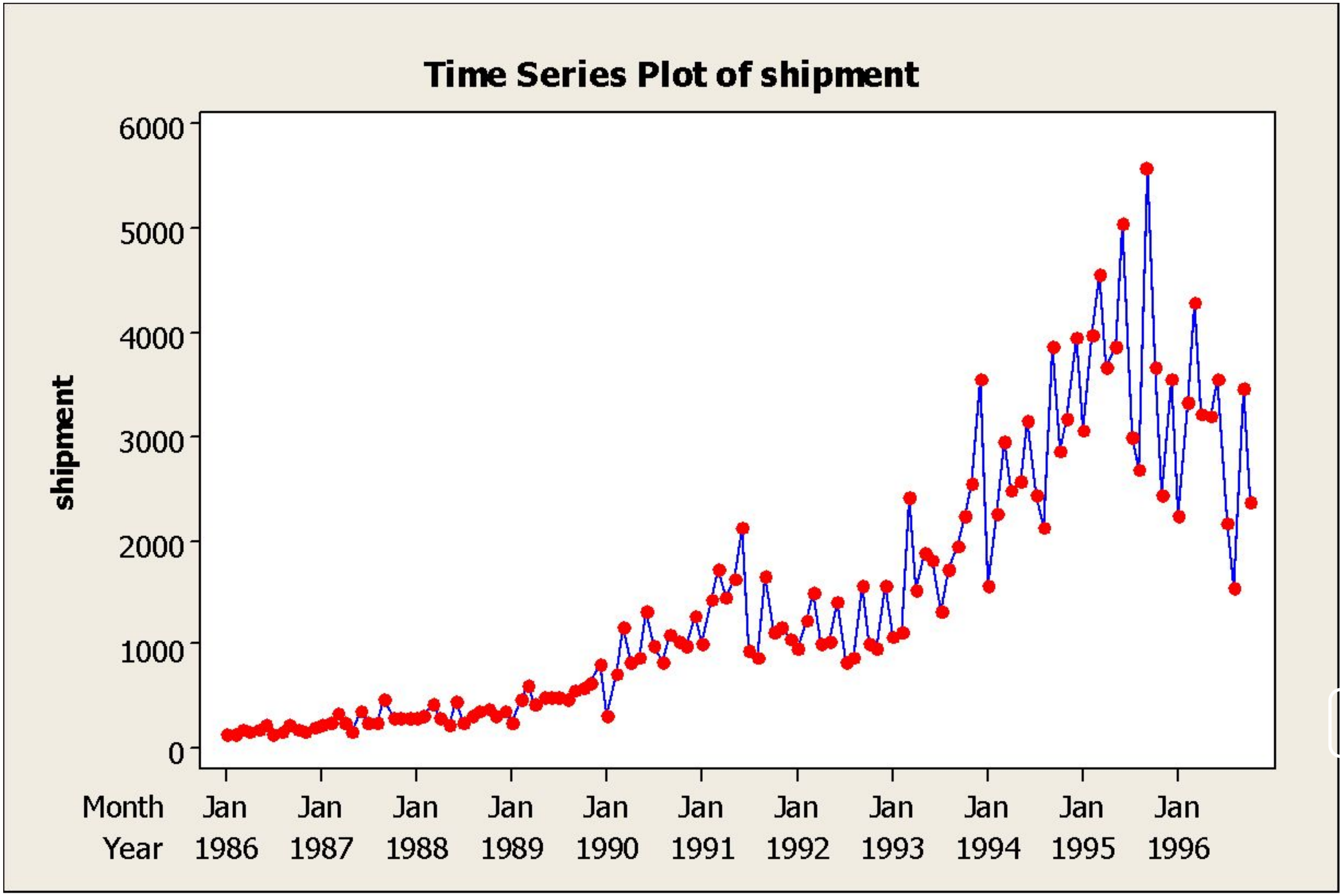
It is one of the most commonly used seasonal ARIMA model.

Model identification

Example (3):

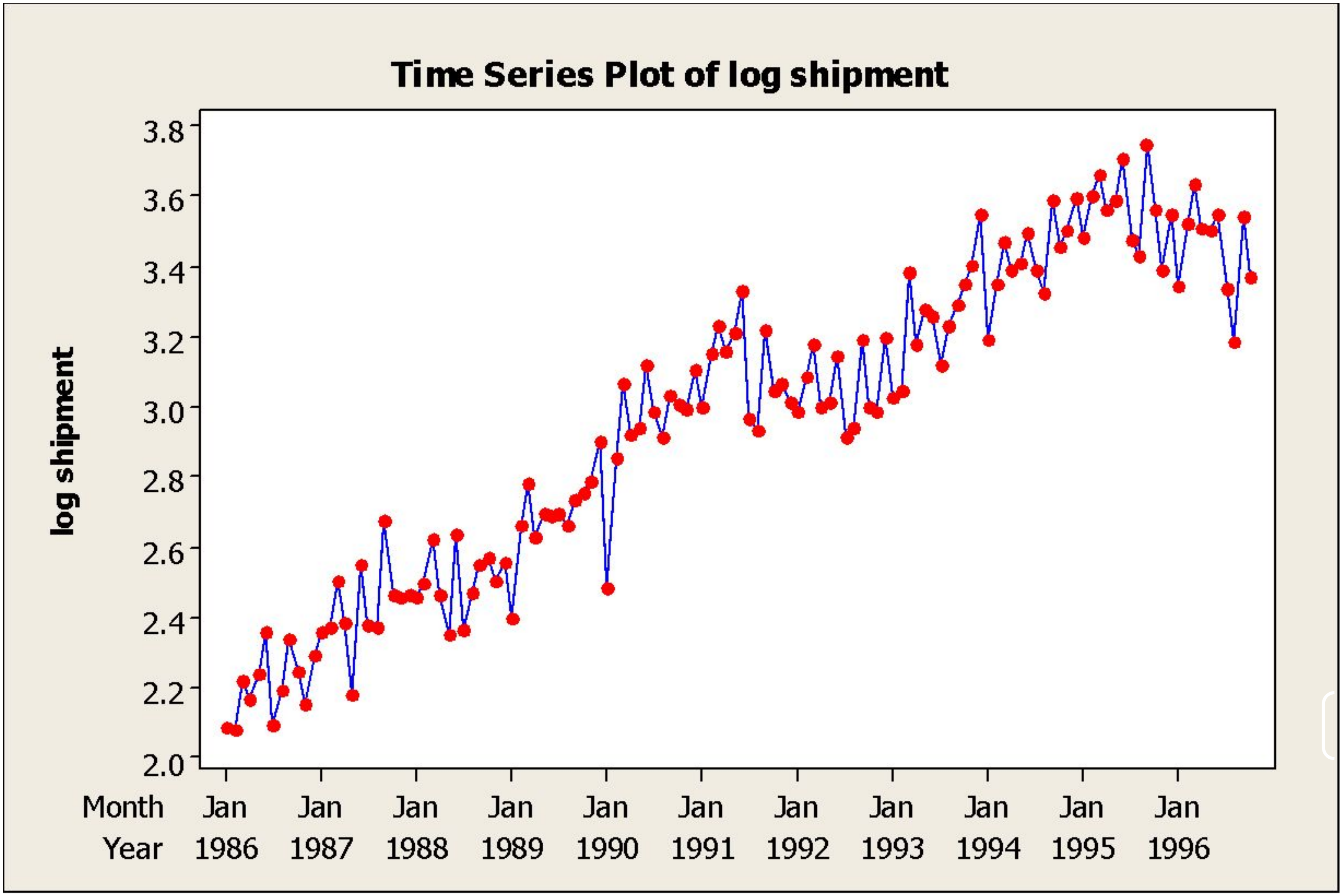
A seasonal data needing transformation

In this example we look at the monthly shipments of a company that manufactures pollution equipment The time plot shows that the variability increases as the time increases. This indicate that the data is non-stationary in the variance.

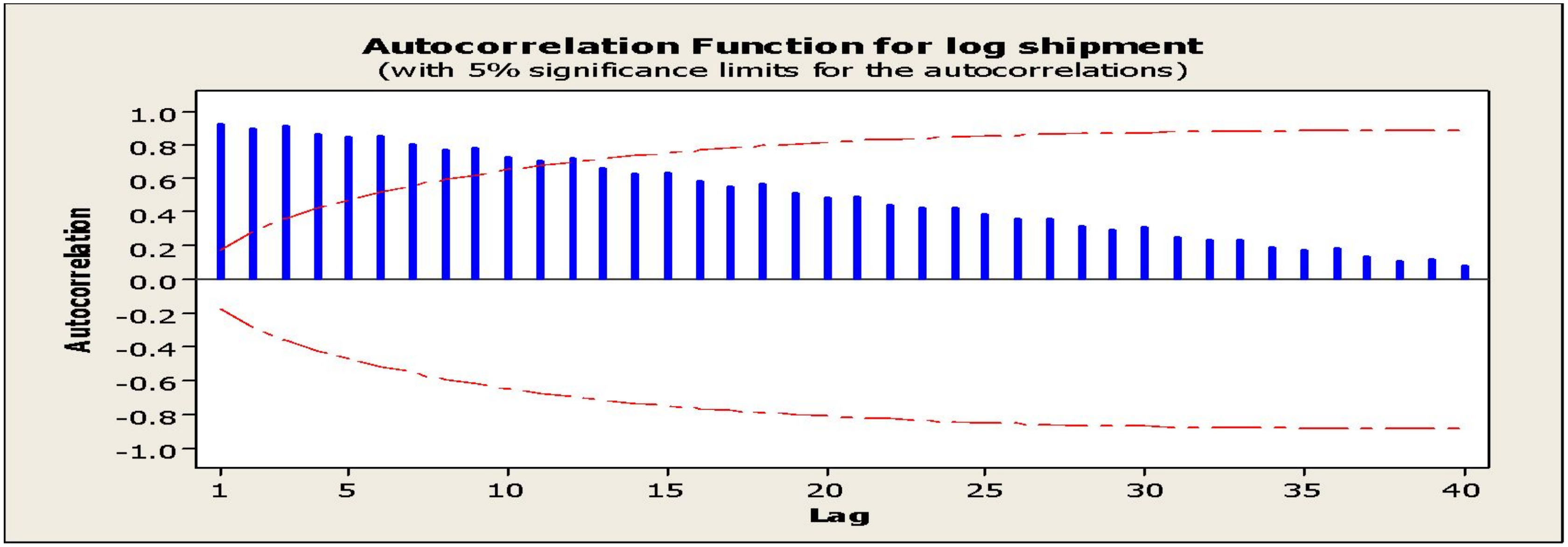

Time Series Plot of shipment
shipment
6000
5000
4000
3000
2000
1000
0
Month
Year
Jan
1986
Jan
1987
Jan
1988
Jan
1989
Jan
1990
Jan
1991
Jan
1992
Jan
1993
Jan
1994
Jan
1995
Jan
1996

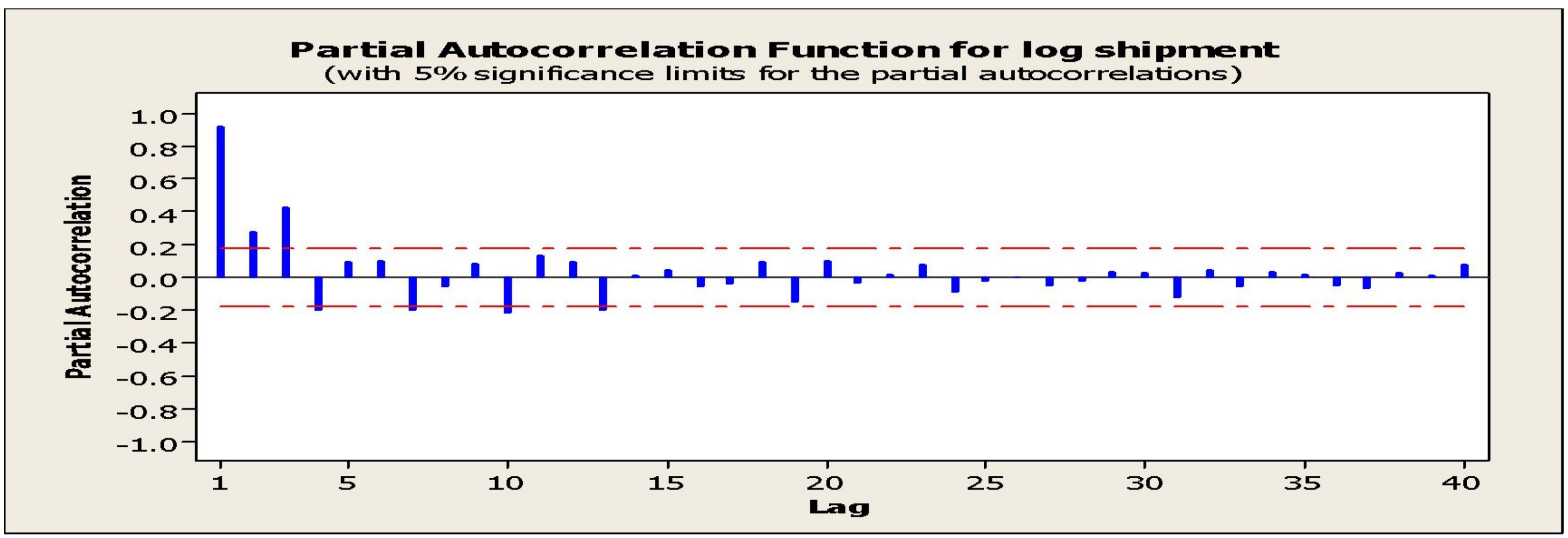
Model identification

We need to stabilize the variance before fitting an ARIMA model.
Logarithmic or power transformation of the data will make the variance stationary.
The time plot, ACF and PACF for the logged data is reported in the following three slides.



Time Series Plot of log shipment
log shipment
3.8
3.6
3.4
3.2
3.0
2.8
2.6
2.4
2.2
2.0
Month Jan Jan Jan Jan Jan Jan Jan Jan Jan Jan Jan
Year 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996


Autocorrelation Function for log shipment
(with 5% significance limits for the autocorrelations)
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
1
5
10
15
20
25
30
35
40
Lag

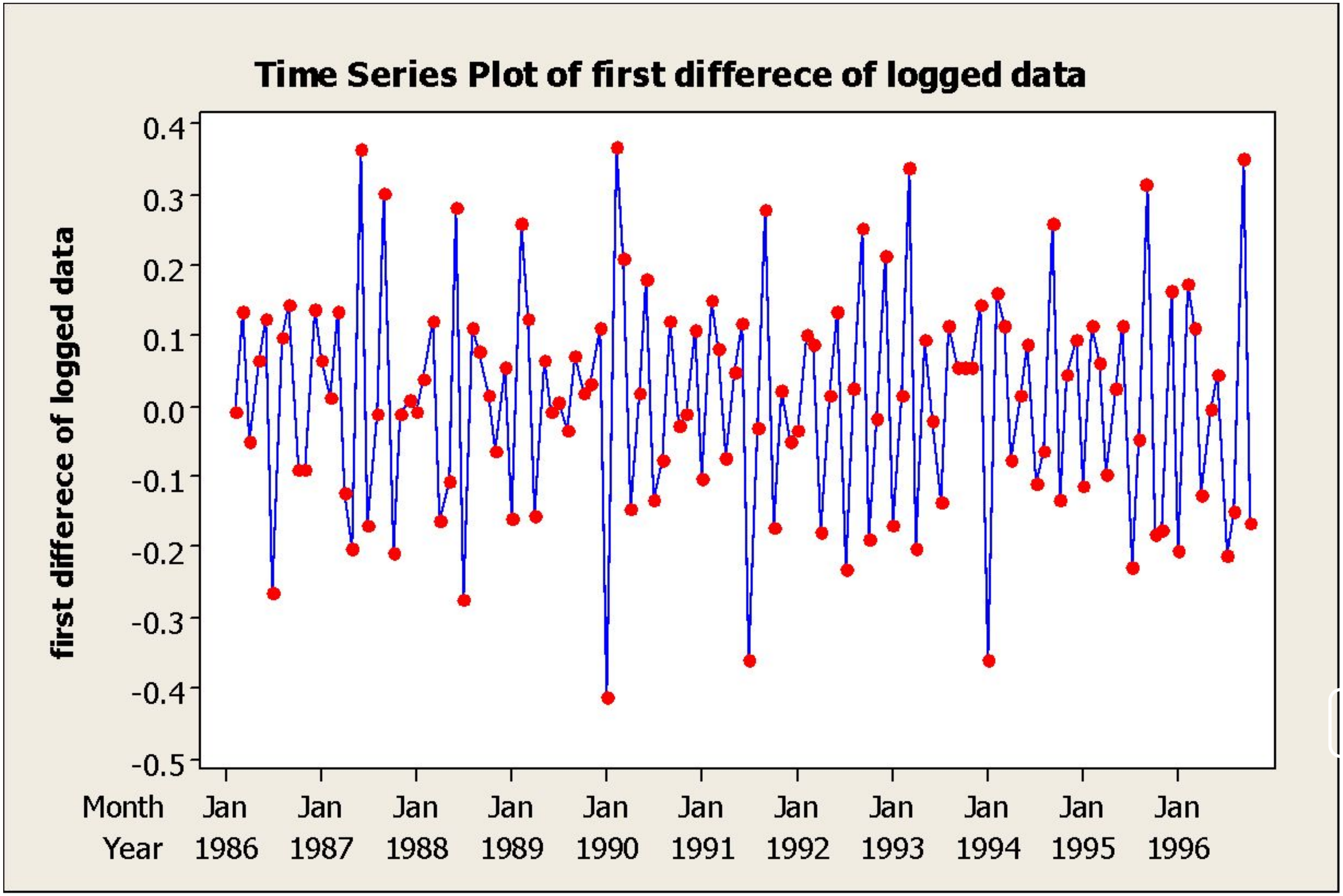

Partial Autocorrelation Function for log shipment
(with 5% significance limits for the partial autocorrelations)
Partial Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
1
5
10
15
20
25
30
35
40
Lag

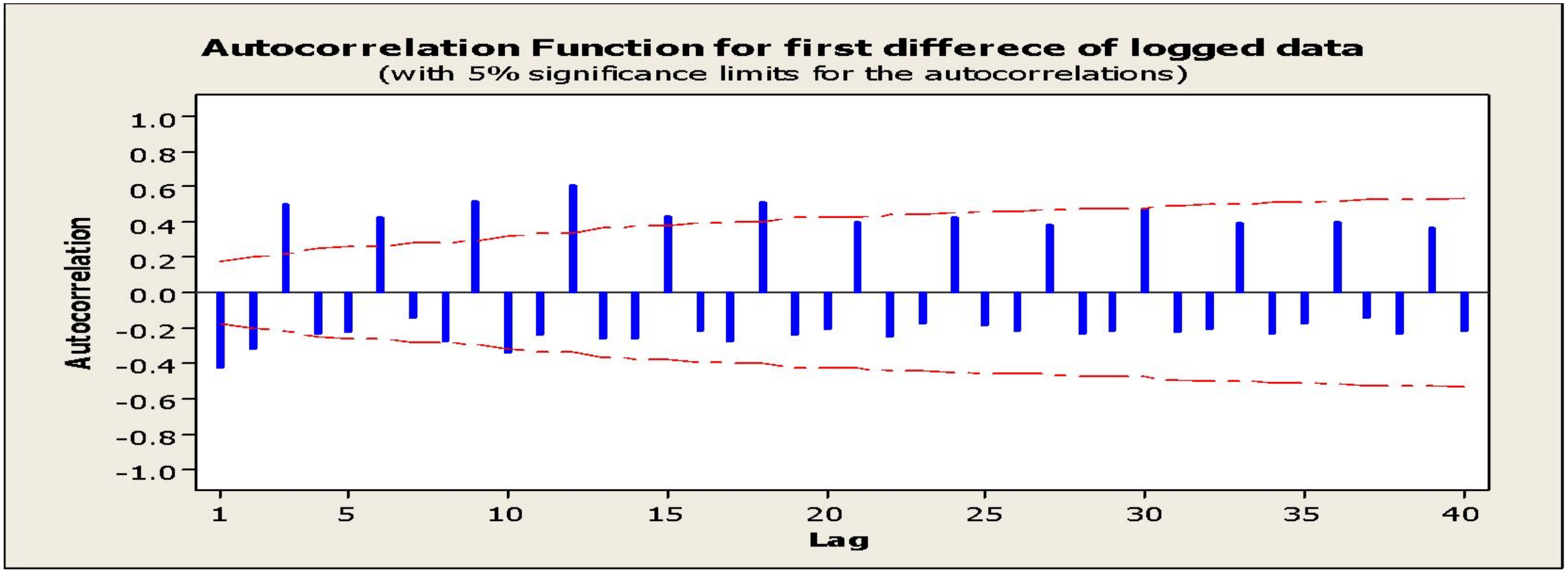
Model identification

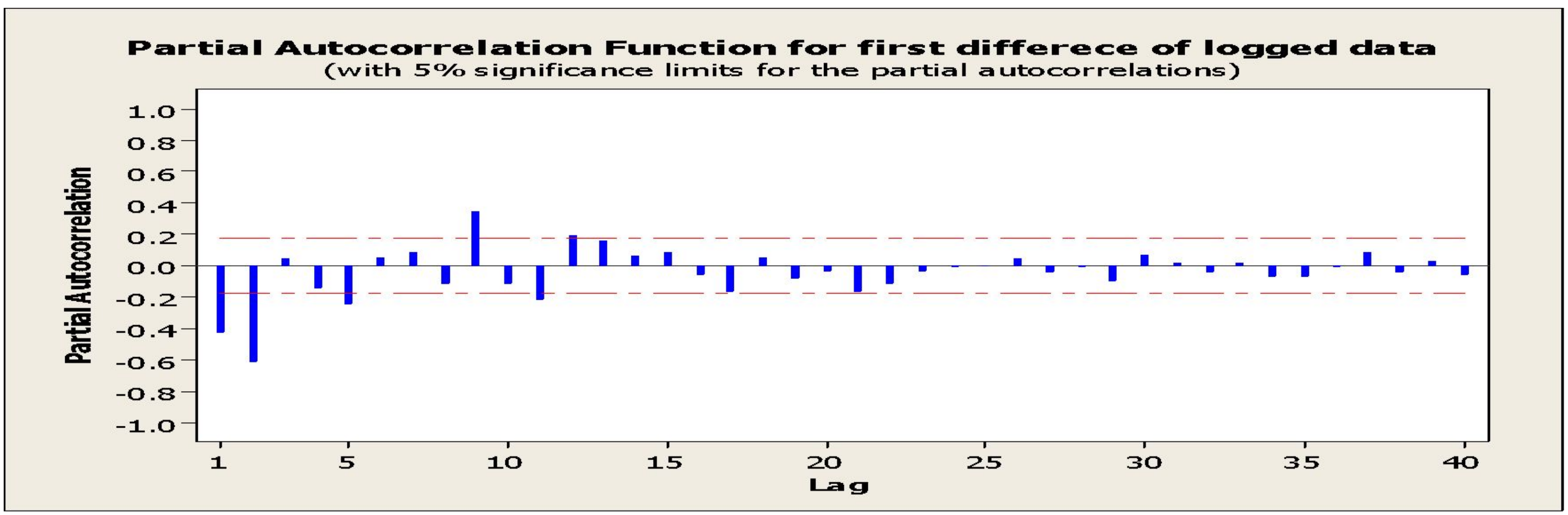
The time plot shows that the magnitude of the fluctuations in the log-transformed data does not vary with time.
But, the logged data are clearly non-stationary.

The gradual decay of the ACF values.

To achieve stationarity, we take the first differences of the logged data.
The plots are reported in the next three slides.


Time Series Plot of first differece of logged data
first differece of logged data
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
Month Jan Jan Jan Jan Jan Jan Jan Jan Jan Jan Jan
Year 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996


Autocorrelation Function for first differece of logged data
(with 5% significance limits for the autocorrelations)
Autocorrelation
Lag


Partial Autocorrelation Function for first differece of logged data
(with 5% significance limits for the partial autocorrelations)
Partial Autocorrelation
Lag

Model identification

There are significant spikes at time lag 1 and 2 in the PACF, indicating an AR(2) might be appropriate.
The single significant spike at lag 12 of the PACF indicates a seasonal AR(1) component.
Therefore for the logged data a tentative model would be

$$ARIMA(2,1,0)(1,0,0)_{12}$$

Summary

The process of identifying an ARIMA model requires experience and good judgment.

The following guidelines can be helpful:

1. Make the series stationary in mean and variance

   Differencing will take care of non-stationarity in the mean.

   Logarithmic or power transformation will often take care of non-stationarity in the variance.

Summary

2. Consider non-seasonal aspect

The ACF and PACF of the stationary data obtained from the previous step can reveal whether MA of AR is feasible.

Exponential decay or damped sine-wave. For ACF, and spikes a lags 1 to p then cut off to zero, indicate an AR(P) model.

Spikes at lag1 to q, then cut off to zero for ACF and exponential decay or damped sine-wave for PACF indicates MA(q) model.

Summary

2. Consider seasonal aspect

   Examination of ACF and PACF at the seasonal lags can help to identify AR and MA models for the seasonal aspect of the data.

   For example, for quarterly data the pattern of r_4, r_8, r_{12}, r_{16}, and so on.

Backshift notation

The backward shift operator *B* is a useful notational device when working with time series lags:

Backward shift oper … as:

$$BY_t = Y_{t-1}$$

Two applications of B to Y_t, shifts the data back two periods:

$$B(BY_t) = B^2Y_t = Y_{t-2}$$

A shift to the same quarter last year will use B^4 which is

$$B^4Y_t = Y_{t-4}$$

Backshift notation

The backward shift operator can be used to describe the differencing process.

A first

$$Y_t' = Y_t - Y_{t-1} = Y_t - BY_t = (1-B)Y_t$$

The second order differences as

$$\begin{aligned} Y_t'' &= (Y_t' - Y_{t-1}') \\ &= (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) \\ &= Y_t - 2Y_{t-1} + Y_{t-2} \\ &= (1 - 2B + B^2)Y_t \\ &= (1-B)^2 Y_t \end{aligned}$$

Backshift notation

Example:

ARMA(1,1) or ARIMA(1,0,1) model

$$Y_t = c + \phi_1 Y_{t-1} + e_t - \theta_1 e_{t-1}$$

$$(1-\phi_1 B)Y_t = c + (1-\theta_1 B)e_t$$

ARMA(p,q)

$$Y_t = c + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q}$$

$$(1-\phi_1 B - \cdots - \phi_p B^p)Y_t = c + (1-\theta_1 B - \cdots - \theta_q B^q)e_t$$

Backshift notation

ARIMA(1,1,1)

$$Y_t - Y_{t-1} = c + \phi_1 (Y_{t-1} - Y_{t-2}) + e_t - \theta_1 e_{t-1}$$

$$(1 - \phi_1 B)(1 - B) Y_t = c + (1 - \theta_1 B) e_t$$

Estimating the parameters

Once a tentative model has been selected, the parameters for the model must be estimated.
The method of least squares can be used for RIMA model.
However, for models with an MA components, there is no simple formula that can be used to estimate the parameters.
Instead, an iterative method is used. This involves starting with a preliminary estimate, and refining the estimate iteratively until the sum of the squared errors is minimized.

Estimating the parameters

Another method of estimating the parameters is the maximum likelihood procedure.
Like least squares methods, these estimates must be found iteratively.
Maximum likelihood estimation is usually favored because it has some desirable statistical properties.
After the estimates and their standard errors are determined, t values can be constructed and interpreted in the usual way.
Parameters that are judged significantly different from zero are retained in the fitted model; parameters that are not significantly different from zero are dropped from the model.

Estimating the parameters

There may have been more than one plausible model identified, and we need a method to determine which of them is preferred.

Akaike's Information Criterion (AIC)

$$AIC = -2\log L + 2m$$

L denotes the likelihood

m is the number of parameters estimated in the model: m = p+q+P+Q

Estimating the parameters

Because not all computer programs produce the AIC or the likelihoo
L, it is not always possible to find the AIC for a given model.
A useful approximation to the AIC is:

$$AIC = n(1+\log(2\pi)+n\log\sigma^2+2m$$

Diagnostic Checking

Before using the model for forecasting, it must be checked for adequacy.
A model is adequate if the residuals left over after fitting the model is simply white noise.
The pattern of ACF and PACF of the residuals may suggest how the model can be improved.
For example

- Significant spikes at the seasonal lags suggests adding seasonal component to the chosen model
- Significant spikes at small lags suggest increasing the non-seasonal AR or MA components of the model.

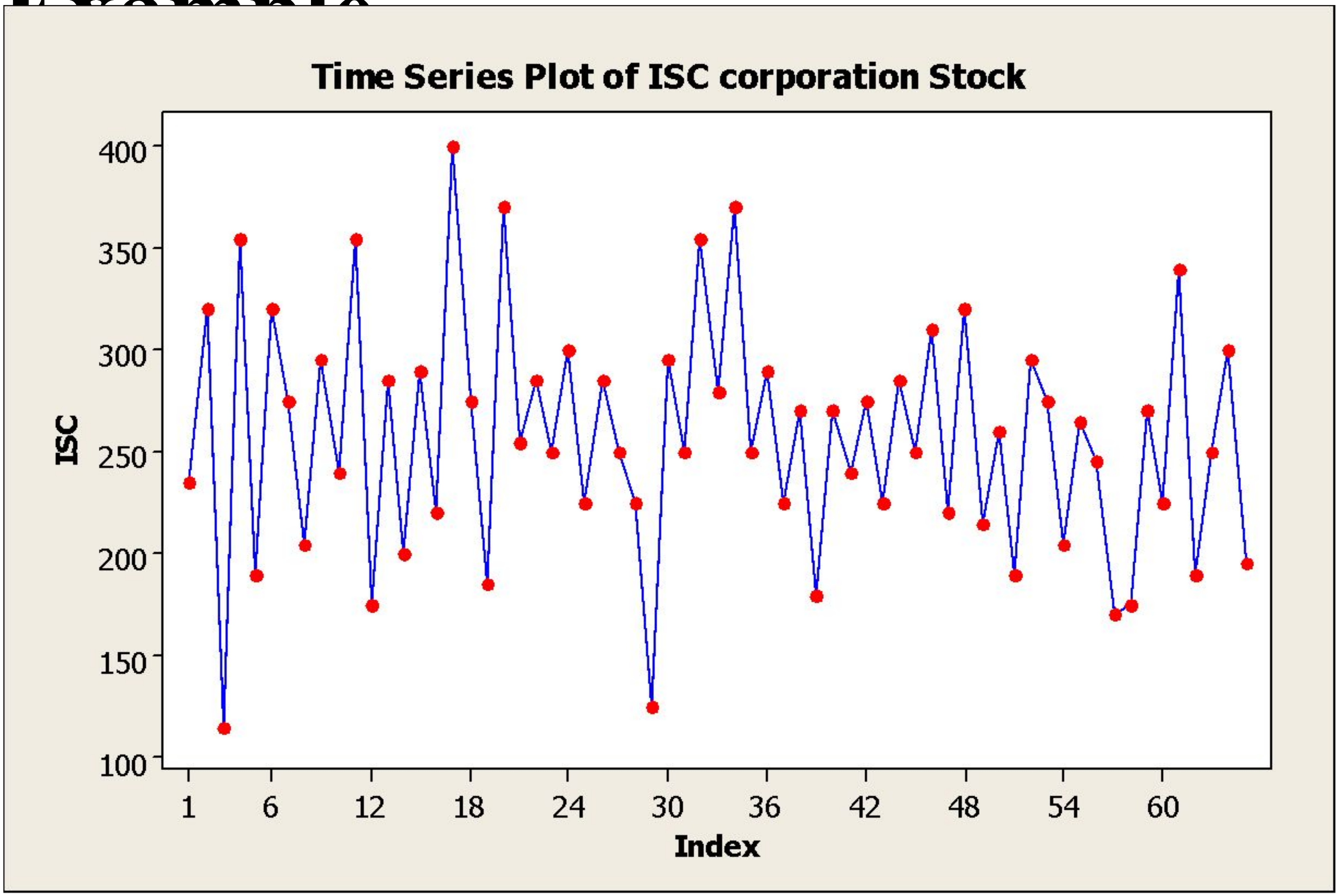
Diagnostic Checking

A portmanteau test can also be applied to the residuals as an additional test of fit.
If the portmanteau test is significant, then the model is inadequate.
In this case we need to go back and consider other ARIMA models.
Any new model will need their parameters estimated and their AIC values computed and compared with other models.
Usually, the model with the smallest AIC will have residuals which resemble white noise.
Occasionally, it might be necessary to adopt a model with not quite the smallest AIC value, but with better behaved residuals.

Example

The analyst for the ISC Corporation was asked to develop forecasts for the closing prices of ISC stock.
The stock has been languishing for some time with little growth, and senior management wanted some projections to discuss with the board of directors.
The ISC stock prices are plotted in the following slide.

Example

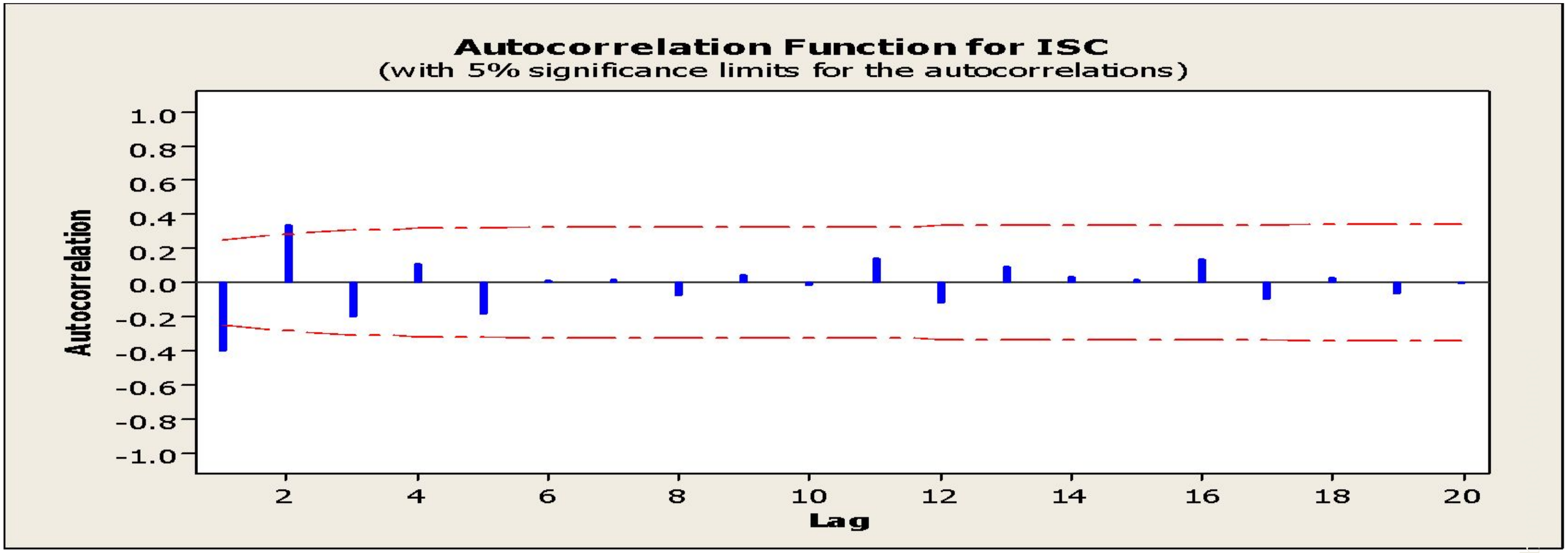


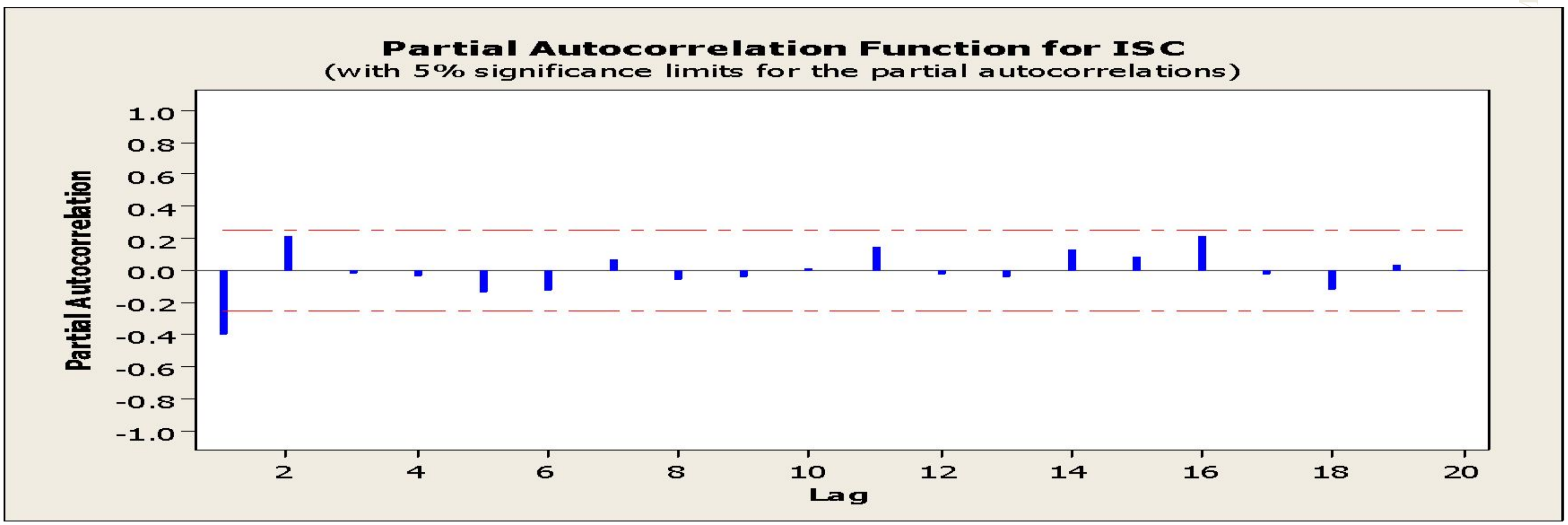
Time Series Plot of ISC corporation Stock
ISC
400
350
300
250
200
150
100
1
6
12
18
24
30
36
42
48
54
60
Index

Example

The plot of the stock prices suggests the series is stationary.
The stock prices vary about a fixed level of approximately 250.
Is the Box-Jenkins methodology appropriate for this data series?
The ACF and PACF for the stock price series are reported in the following two slides.

Example


Autocorrelation Function for ISC
(with 5% significance limits for the autocorrelations)
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
2
4
6
8
10
12
14
16
18
20
Lag


Partial Autocorrelation Function for ISC
(with 5% significance limits for the partial autocorrelations)
Partial Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
2
4
6
8
10
12
14
16
18
20
Lag

Example

The sample ACF alternate in sign and decline to zero after lag 2.
The sample PACF are similar are close to zero after time lag 2.
These are consistent with an AR(2) or ARIMA(2,0,0) model
AR(2) model is fit to the data.
WE include a constant term to allow for a nonzero level.

Example

The estimated coefficient φ_2 is not significant (t=1.75) at 5% level but is significant at the 10 % level.

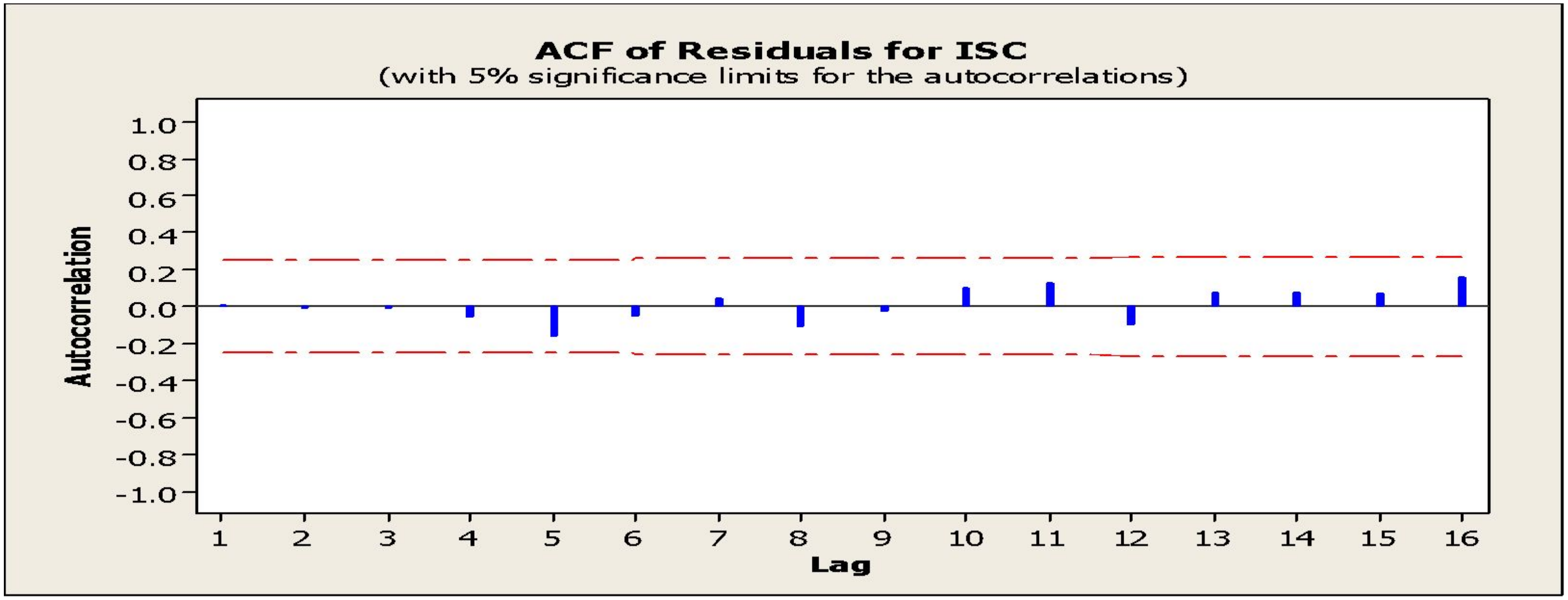
The residual ACF and PACF are given in the following two slides.

The ACF and PACF are well within their two standard error limits.

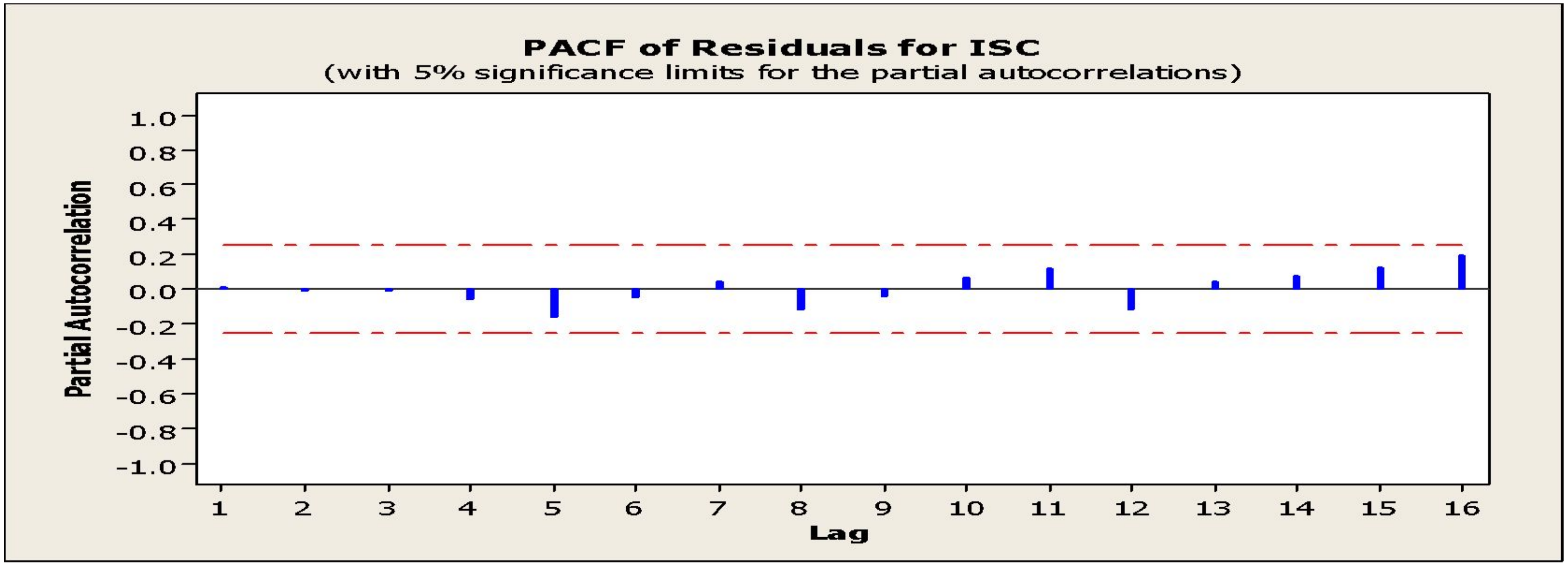
Final Estimates of Parameters

Type	Coef	SE	T	P
AR 1	-0.3243	0.1246	-2.60	0.012
AR 2	0.2192	0.1251	1.75	0.085
Constant	284.903	6.573	43.34	0.000

Example



ACF of Residuals for ISC
(with 5% significance limits for the autocorrelations)
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Lag



PACF of Residuals for ISC
(with 5% significance limits for the partial autocorrelations)
Partial Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Lag

Example

The p-value for the Ljung-Box statistics for m = 12, 24, 36, and 48 are all large (> 5%) indicating an adequate model. We use the model to generate forecasts for periods 66 and 67.

MS = 2808 DF = 62

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	6.3	13.3	18.2	29.1
DF	9	21	33	45
P-Value	0.707	0.899	0.983	0.969

Example

The forecasts are generated by the following equation:

$$\hat{Y}_t = c + \hat{\phi}_1 Y_{t-1} + \hat{\phi}_2 Y_{t-2}$$

$$\hat{Y}_{66} = 284.9 + (-.324)Y_{65} + .219Y_{64}$$
$$= 284.9 - .324(195) + .219(300) = 287.4$$

$$\hat{Y}_{67} = 284.9 + (-.324)\hat{Y}_{66} + .219Y_{65}$$
$$= 284.9 - .324(287.4) + .219(195) = 234.5$$

Example

The 95% prediction limits are approximately:

$$\hat{Y} \pm 2s$$

The 95% prediction limits for period 66 are:

$$287.4 \pm 2\sqrt{2808}$$

$$287.4 \pm 106$$

$$(181.4,\ 393.4)$$

Final Comments

- In ARIMA modeling, it is NOT good practice to include AR and MA parameters to “cover all possibilities” suggested by the sample ACF and Sample PACF.
- This means, when in doubt, start with a model containing **few** parameters rather than many parameters. The need for additional parameters will be evident from the residual ACF and PACF.
- Least square estimates of AR and MA parameters in ARIMA models tend to be highly correlated. When there are more parameters than necessary, this leads to unstable models that can produce poor forecasts.

Final Comments

- To summarize, start with a small number of clearly justifiable parameters and add one parameter at a time as needed.
- If parameters in a fitted ARIMA model are not significant, delete one parameter at a time and refit the model.
- Because of high correlation among estimated parameters, it may be the case that a previously non-significant parameter becomes significant.

Summary of rules for identifying ARIMA models

Identifying the order of differencing and the constant:

Rule 1: If the series has positive autocorrelations out to a high number of lags (say, 10 or more), then it probably needs a higher order of differencing.

Rule 2: If the lag-1 autocorrelation is zero or negative, or the autocorrelations are all small and patternless, then the series does *not* need a higher order of differencing. If the lag-1 autocorrelation is -0.5 or more negative, the series may be overdifferenced. **BEWARE OF OVERDIFFERENCING.**

Rule 3: The optimal order of differencing is often the order of differencing at which the standard deviation is lowest. (Not always, though. Slightly too much or slightly too little differencing can also be corrected with AR or MA terms. See rules 6 and 7.)

Summary of rules for identifying ARIMA models

Rule 4: A model with no orders of differencing assumes that the original series is stationary (among other things, mean-reverting). A model with one order of differencing assumes that the original series has a constant average trend (e.g. a random walk or SES-type model, with or without growth). A model with two orders of total differencing assumes that the original series has a time-varying trend (e.g. a random trend or LES-type model).

Rule 5: A model with no orders of differencing normally includes a constant term (which allows for a non-zero mean value). A model with two orders of total differencing normally does not include a constant term. In a model with one order of total differencing, a constant term should be included if the series has a non-zero average trend.

Summary of rules for identifying ARIMA models

Identifying the numbers of AR and MA terms:

Rule 6: If the partial autocorrelation function (PACF) of the differenced series displays a sharp cutoff and/or the lag-1 autocorrelation is positive--i.e., if the series appears slightly "underdifferenced"--then consider adding one or more AR terms to the model. The lag beyond which the PACF cuts off is the indicated number of AR terms.

Rule 7: If the autocorrelation function (ACF) of the differenced series displays a sharp cutoff and/or the lag-1 autocorrelation is negative--i.e., if the series appears slightly "overdifferenced"--then consider adding an MA term to the model. The lag beyond which the ACF cuts off is the indicated number of MA terms.

Summary of rules for identifying ARIMA models

Rule 8: It is possible for an AR term and an MA term to cancel each other's effects, so if a mixed AR-MA model seems to fit the data, also try a model with one fewer AR term and one fewer MA term--particularly if the parameter estimates in the original model require more than 10 iterations to converge. BEWARE OF USING MULTIPLE AR TERMS AND MULTIPLE MA TERMS IN THE SAME MODEL.

Rule 9: If there is a unit root in the AR part of the model--i.e., if the sum of the AR coefficients is almost exactly 1--you should reduce the number of AR terms by one and <u>increase</u> the order of differencing by one.

Rule 10: If there is a unit root in the MA part of the model--i.e., if the sum of the MA coefficients is almost exactly 1--you should reduce the number of MA terms by one and <u>reduce</u> the order of differencing by one.

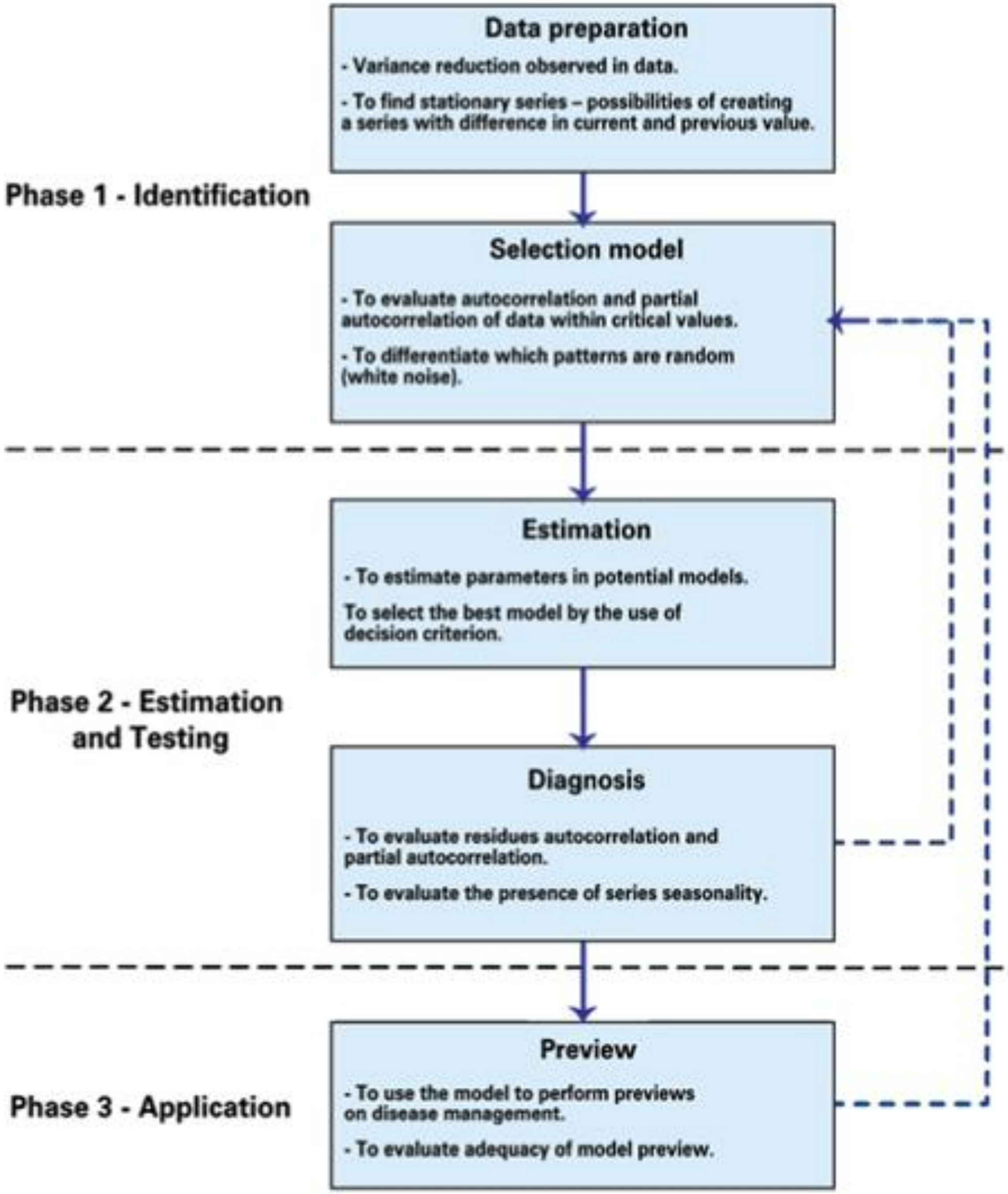
Rule 11: If the <u>long-term forecasts</u>* appear erratic or unstable, there may be a unit root in the AR or MA coefficients.

Summary of rules for identifying ARIMA models

Identifying the seasonal part of the model:

Rule 12: If the series has a strong and consistent seasonal pattern, then you must use an order of seasonal differencing (otherwise the model assumes that the seasonal pattern will fade away over time). However, never use more than one order of seasonal differencing or more than 2 orders of total differencing (seasonal+nonseasonal).

Rule 13: If the autocorrelation of the appropriately differenced series is positive at lag s, where s is the number of periods in a season, then consider adding an SAR term to the model. If the autocorrelation of the differenced series is negative at lag s, consider adding an SMA term to the model. The latter situation is likely to occur if a seasonal difference has been used, which should be done if the data has a stable and logical seasonal pattern. The former is likely to occur if a seasonal difference has not been used, which would only be appropriate if the seasonal pattern is not stable over time. You should try to avoid using more than one or two seasonal parameters (SAR+SMA) in the same model, as this is likely to lead to overfitting of the data and/or problems in estimation.


Data preparation
- Variance reduction observed in data.
- To find stationary series – possibilities of creating a series with difference in current and previous value.
Phase 1 - Identification
Selection model
- To evaluate autocorrelation and partial autocorrelation of data within critical values.
- To differentiate which patterns are random (white noise).
Estimation
- To estimate parameters in potential models.
To select the best model by the use of decision criterion.
Phase 2 - Estimation and Testing
Diagnosis
- To evaluate residues autocorrelation and partial autocorrelation.
- To evaluate the presence of series seasonality.
Preview
- To use the model to perform previews on disease management.
- To evaluate adequacy of model preview.
Phase 3 - Application

Introduction

Autoregressive Integrated Moving Average models (ARIMA models) were popularized by George Box and Gwilym Jenkins in the early 1970s.
ARIMA models are a class of linear models that is capable of representing *stationary as well as non-stationary time series*.
ARIMA models *do not involve independent variables* in their construction. They make use of the information in the series itself to generate forecasts.

Introduction

ARIMA models rely heavily on autocorrelation patterns in the data.
ARIMA methodology of forecasting is different from most methods because it does not assume any particular pattern in the historical data of the series to be forecast.
It uses an interactive approach of identifying a possible model from a general class of models.
The chosen model is then checked against the historical data to see if it accurately describe the series.

Introduction

Recall that, a time series data is a sequence of numerical observations naturally ordered in time, such as:

1. Daily closing price of IBM stock
2. Weekly automobile production by the Pontiac division of general Motors.
3. Hourly temperatures at the entrance to Grand central Station.

Introduction

Two question of paramount importance When a forecaster examines a time series data are:

1. Do the data exhibit a visible pattern?
2. Can this be used to make meaningful forecasts?

Introduction

- The Box-Jenkins methodology refers to a set of procedures for *identifying*, *fitting*, and *checking* ARIMA models with time series data. Forecasts follow directly from the form of fitted model.
- The basis of BOX-Jenkins approach to modeling time series consists of three phases:

Introduction

1. Identification
    a) Data preparation
        - Transform data to stabilize variance
        - Differencing data to obtain stationary series
    b) Model selection
        - Examine data, ACF and PACF to identify potential models

How to make a time series stationary?

You can make series stationary by:

- ✔ Differencing the Series (once or more)
- ✔ Take the log of the series
- ✔ Take the nth root of the series
- ✔ Combination of the above

The most common and convenient method to stationarize the series is by differencing the series at least once until it becomes approximately stationary.

Differencing of Series

If Y_t is the value at time 't', then the first difference is given by:

$$Y = Y_t - Y_{t-1}$$

In simpler terms, differencing the series is nothing but subtracting the next value by the current value.

If the first difference doesn't make a series stationary, you can go for the second differencing. And so on.

Differencing of Series

For example, consider the following series: [1, 5, 2, 12, 20]

First differencing gives: [5-1, 2-5, 12-2, 20-12] = [4, -3, 10, 8]

Second differencing gives: [-3-4, 10-(-3), 8-10] = [-7, 13, -2]

Why make a non-stationary series stationary before forecasting?

- Forecasting a stationary series is relatively easy and the forecasts are more reliable.
- An important reason is, autoregressive forecasting models are essentially linear regression models that utilize the lag(s) of the series itself as predictors.

ARIMA

An ARIMA model is characterized by 3 terms: p, d, q where,

- p is the order of the AR term
- q is the order of the MA term
- d is the number of differencing required to make the time series stationary

If a time series, has seasonal patterns, then you need to add seasonal terms and it becomes SARIMA, short for 'Seasonal ARIMA'.

The first step to build an ARIMA model is to make the time series stationary.

Why?

Because, term ‘Auto Regressive’ in ARIMA means it is a linear regression model that uses its own lags as predictors.

Linear regression models work best when the predictors are not correlated and are independent of each other.

d – differencing

- How to make a series stationary? – difference it!
- Subtract the previous value from the current value. Sometimes, depending on the complexity of the series, more than one differencing may be needed.
- The value of d, therefore, is the minimum number of differencing needed to make the series stationary.
- This captures the “intergrated” nature of ARIMA.
- If d=0, this means that our data does not tend to go up/down in the long term (i.e., the model is already “stationary”). In this case, then technically you are performing just ARMA, not AR-I-MA.

- ‘p’ is the order of the ‘Auto Regressive’ (AR) term.
- ‘p’ means the number of preceding (“lagged”) Y values that have to be added/subtracted to Y in the model, so as to make better predictions based on local periods of growth/decline in our data.
- This captures the “autoregressive” nature of ARIMA.
- It refers to the number of lags of Y to be used as predictors.
- ‘q’ is the order of the ‘Moving Average’ (MA) term.
- It refers to the number of lagged forecast errors that should go into the ARIMA Model.
- ‘q’ represents the number of preceding/lagged values for the error term that are added/subtracted to Y.
- This captures the “moving average” part of ARIMA

ARIM
A

In most software programs, the elements in the model are specified in the order (AR order, differencing, MA order). For examples,

- A model with (only) two AR terms would be specified as an ARIMA of order (2,0,0).
- A MA(2) model would be specified as an ARIMA of order (0,0,2).
- A model with one AR term, a first difference, and one MA term would have order (1,1,1).

AR Model

- A pure **Auto Regressive (AR only) model** is one where Yt depends only on its own lags. That is, Yt is a function of the 'lags of Yt'.

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} + \epsilon_1$$

where,

Y_{t-1} *is the lag1 of the series,*

β_1 is the coefficient of lag1 that the model estimates,

α and is the intercept term, also estimated by the model.

AR(1) Model

- Simplest ARIMA type model
- Alinear model to predict the value at the present time using the value at the previous time.
- **AR(1) model** stands for **autoregressive model of order 1**. The order of the model indicates how many previous times we use to predict the present time.

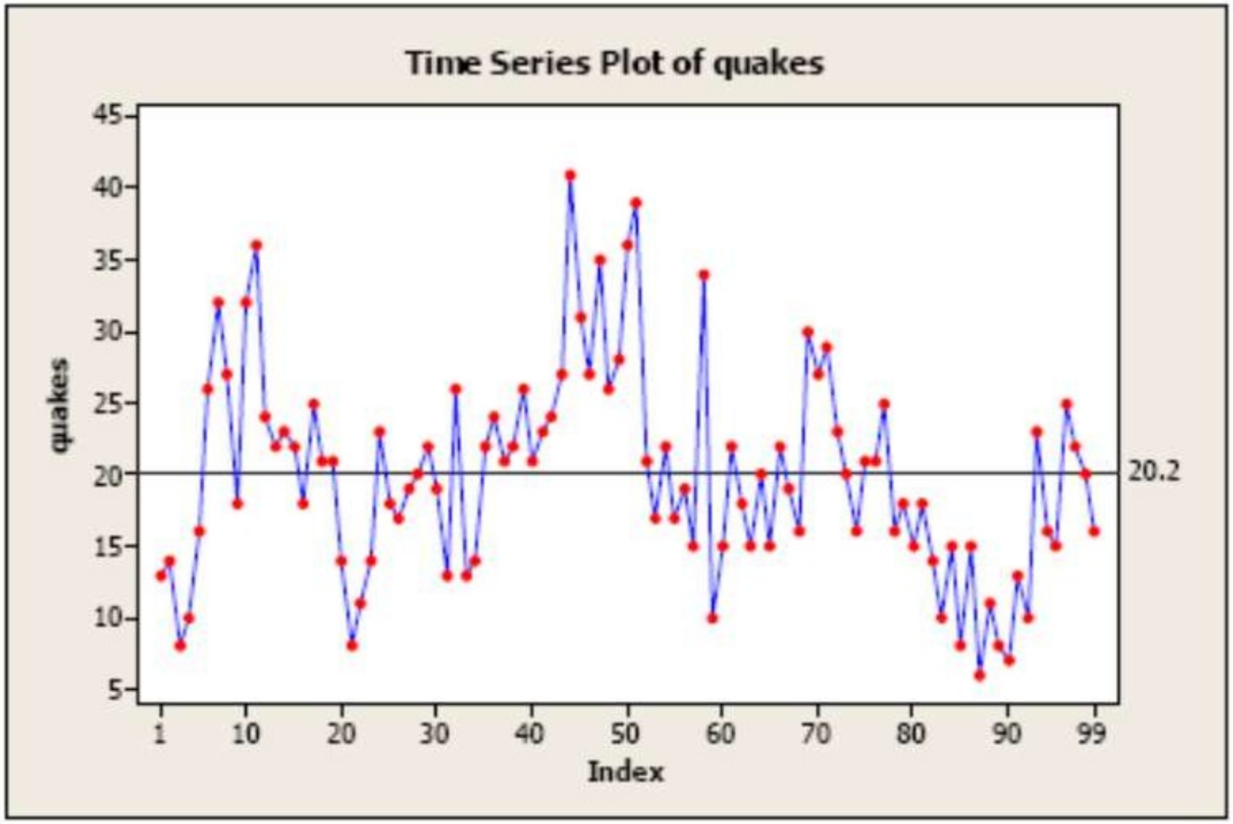
$$Y_t = \alpha + \beta_1 Y_{t-1} + \epsilon_1$$

AR(1) Model

- Plot values of the series against **lag 1 values** of the series. Let x_t denote the value of the series at any particular time t, so x_{t-1} denotes the value of the series one time before time t.

- That is, x_{t-1} is the lag 1 value of x_t.

The following plot is a **time series plot** of the annual number of earthquakes in the world with seismic magnitude over 7.0, for 99 consecutive years.



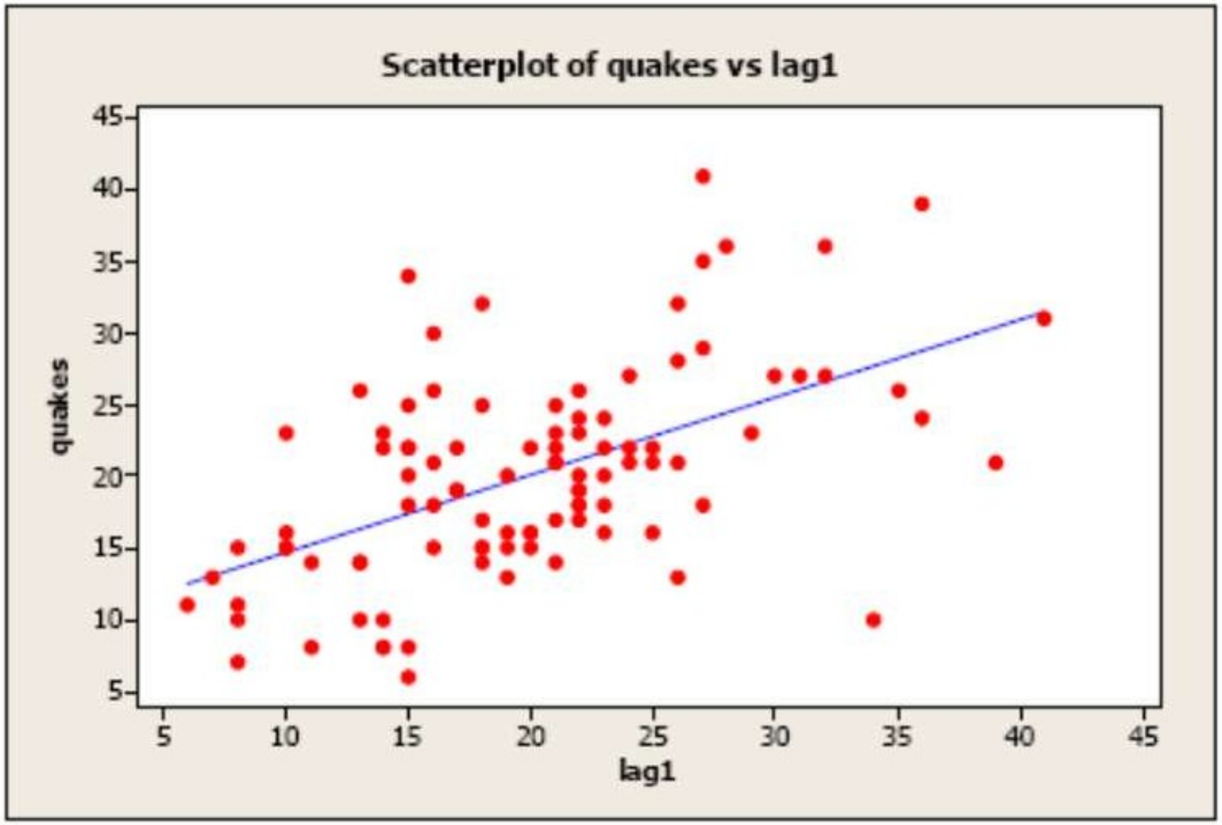
Some features of the plot:

- There is **no consistent trend** (upward or downward) over the entire time span. The series appears to slowly wander up and down. The horizontal line drawn at quakes = 20.2 indicates the mean of the series. Notice that the series tends to stay on the same side of the mean (above or below) for a while and then wanders to the other side.
- Almost by definition, there is **no seasonality** as the data are annual data.
- There are **no obvious outliers**.
- It's difficult to judge whether the variance is constant or not.

Here are the first five values in the earthquake series along with their lag 1 values:

t	x_t	x_{t-1} **(lag 1 value)**
1	13	*
2	14	13
3	8	14
4	10	8
5	16	10

For the complete earthquake dataset, here's a plot of x_t versus x_{t-1} :



Although it's only a moderately strong relationship, there is a positive linear association so an AR(1) model might be a useful model.

Autocorrelatio n

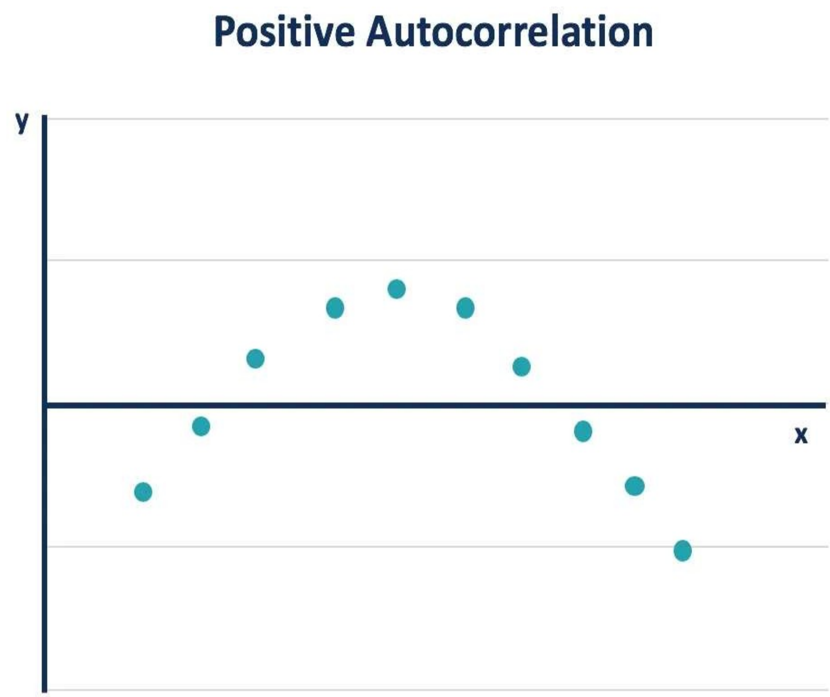
- Autocorrelation refers to the degree of correlation of the same variables between two successive time intervals.
- It measures how the lagged version of the value of a variable is related to the original version of it in a time series.
- The analysis of autocorrelation helps to find repeating periodic patterns.

How it works

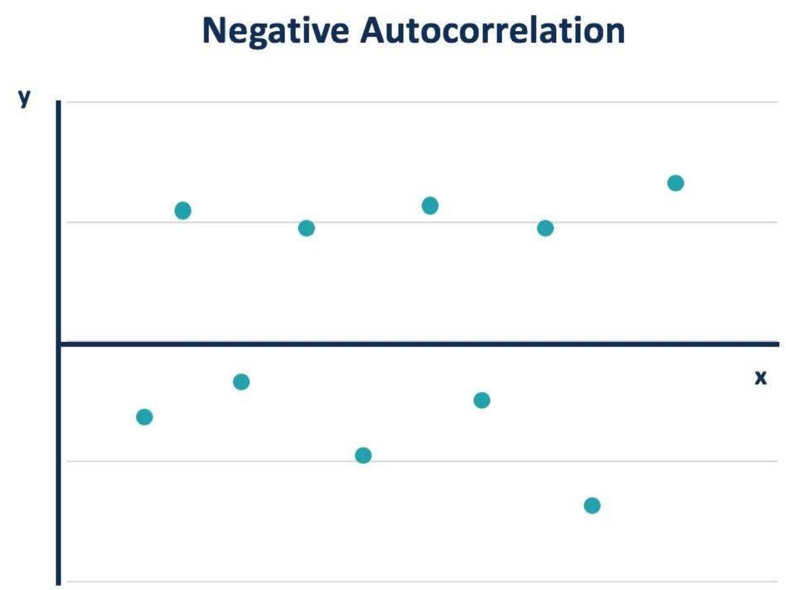
- In many cases, the value of a variable at a point in time is related to the value of it at a previous point in time.
- Autocorrelation analysis measures the relationship of the observations between the different points in time, and thus seeks for a pattern or trend over the time series.
- For example, the temperatures on different days in a month are autocorrelated.
- Similar to correlation, autocorrelation can be either positive or negative. It ranges from -1 (perfectly negative autocorrelation) to 1 (perfectly positive autocorrelation). A value between -1 and 0 represents negative autocorrelation. A value between 0 and 1 represents positive autocorrelation.
- Positive autocorrelation means that the increase observed in a time interval leads to a proportionate increase in the lagged time interval.

The example of temperature discussed above demonstrates a positive autocorrelation. The temperature the next day tends to rise when it’s been increasing and tends to drop when it’s been decreasing during the previous days.

The observations with positive autocorrelation can be plotted into a smooth curve. By adding a regression line, it can be observed that a positive error is followed by another positive one, and a negative error is followed by another negative one.



Conversely, negative autocorrelation represents that the increase observed in a time interval leads to a proportionate decrease in the lagged time interval. By plotting the observations with a regression line, it shows that a positive error will be followed by a negative one and vice versa.



- Autocorrelation can be applied to different numbers of time gaps, which is known as lag.

- A lag 1 autocorrelation measures the correlation between the observations that are a one-time gap apart.

- For example, to learn the correlation between the temperatures of one day and the corresponding day in the next month, a lag 30 autocorrelation should be used (assuming 30 days in that month).

Why autocorrelation matters

- Often, one of the first steps in any data analysis is performing regression analysis.
- However, one of the assumptions of regression analysis is that the data has no autocorrelation.
- This can be frustrating because if you try to do a regression analysis on data with autocorrelation, then your analysis will be misleading.
- And, by nature there is autocorrelation in time series data.

Autocorrelatio n

- Autocorrelation analysis also helps to uncover hidden patterns in our data and help us to select the correct forecasting methods.
- Specifically, we can use it to help identify seasonality and trend in our time series data.
- Additionally, analyzing the autocorrelation function (ACF) and partial autocorrelation function (PACF) in conjunction is necessary for selecting the appropriate ARIMA model for your time series prediction.

Testing for autocorrelation

- Any autocorrelation that may be present in time series data is determined using a correlogram, also known as an ACF plot.
- This is used to help you determine whether your series of numbers is exhibiting autocorrelation at all, at which point you can then begin to better understand the pattern that the values in the series may be predicting.
- The most common autocorrelation test is called the Durbin-Watson test.
- This test returns a value of 0 to 4. If the value returned is 2, there is no autocorrelation in your time series to speak of. If the value is between 0 and 2, it indicates positive autocorrelation. If the value is anywhere between 2 and 4, that means there is a negative correlation.

Introduction

2. Estimation and testing
   a) Estimation
      - Estimate parameters in potential models
      - Select best model using suitable criterion
   b) Diagnostics
      - Check ACF/PACF of residuals
      - Do portmanteau test of residuals
      - Are the residuals white noise?
3. Application
   - Forecasting: use model to forecast

Examining Correlation in Time Series Data

Recall r_1 indicates how successive values of Y relate to each other, r_2 indicates how Y values two periods apart relate to each other, and so on.

The auto correlations at lag 1, 2, …, make up the autocorrelation function or ACF.

Autocorrelation function is a valuable tool for investigating properties of an empirical time series.

Sampling distribution of autocorrelation

The autocorrelation coefficients of white noise data have a sampling distribution that can be approximated by a normal distribution with mean zero and standard error $1/\sqrt{n}$. where n is the number of observations in the series.

This information can be used to develop tests of hypotheses and confidence intervals for ACF.

Portmanteau Tests

This test was originally developed by Box and Pierce for testing the residuals from a forecast model.
Any good forecast model should have forecast errors which follow a white noise model.
If the series is white noise then, the Q-statistic has a chi-square distribution with (h-m) degrees of freedom, where m is the number of parameters in the model which has been fitted to the data.
The test can easily be applied to raw data, when no model has been fitted , by setting $m = 0$.

Examining stationarity of time series data

Stationarity means no growth or decline.

Data fluctuates around a constant mean independent of time and variance of the fluctuation remains constant over time.

Stationarity can be assessed using a time series plot:

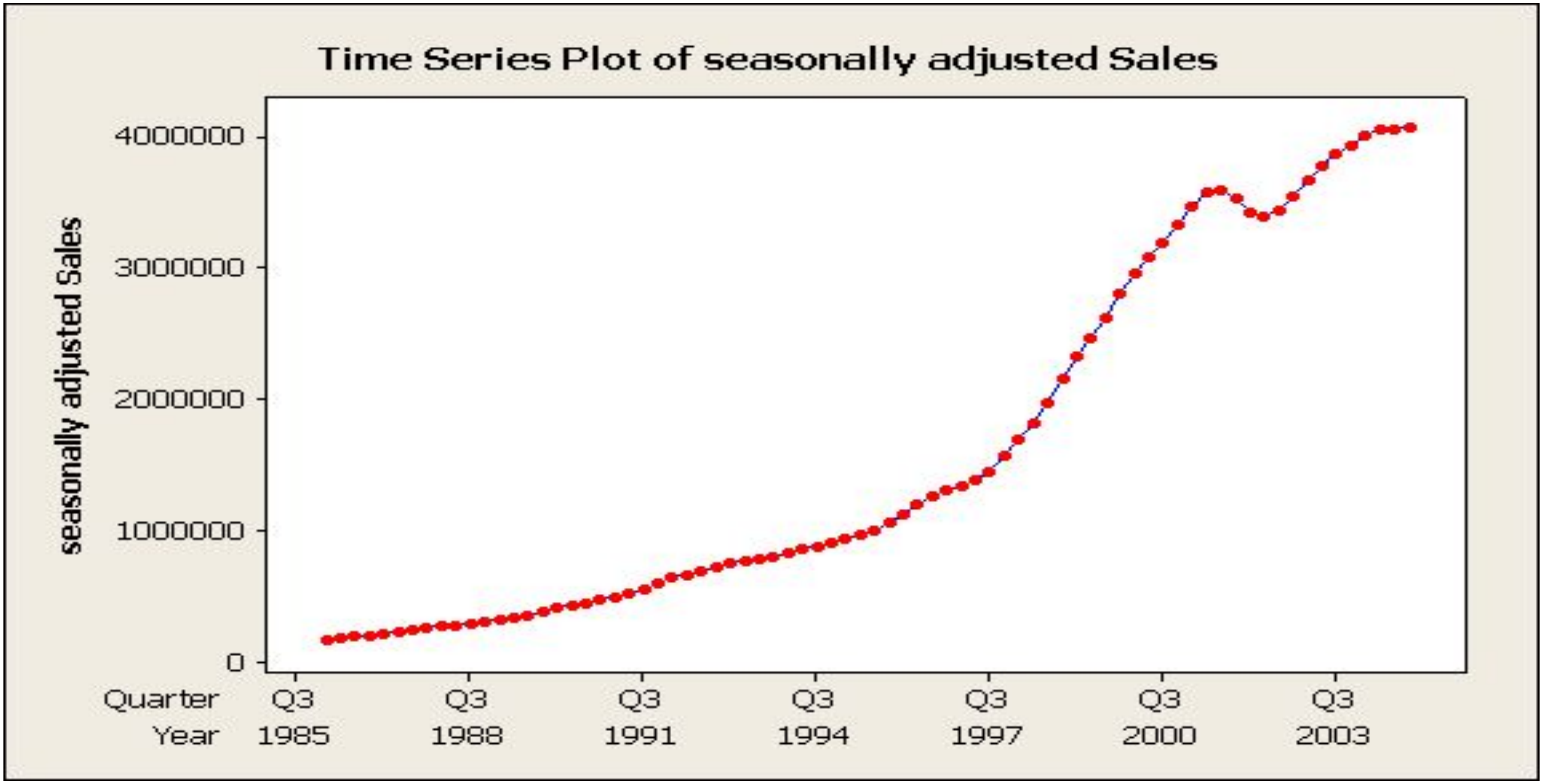
- Plot shows no change in the mean over time
- No obvious change in the variance over time.

Examining stationarity of time series data

- The autocorrelation plot can also show non-stationarity.
  - Significant autocorrelation for several time lags and slow decline in r_k indicate non-stationarity.
- The following graph shows the seasonally adjusted sales for Gap stores from 1985 to 2003.

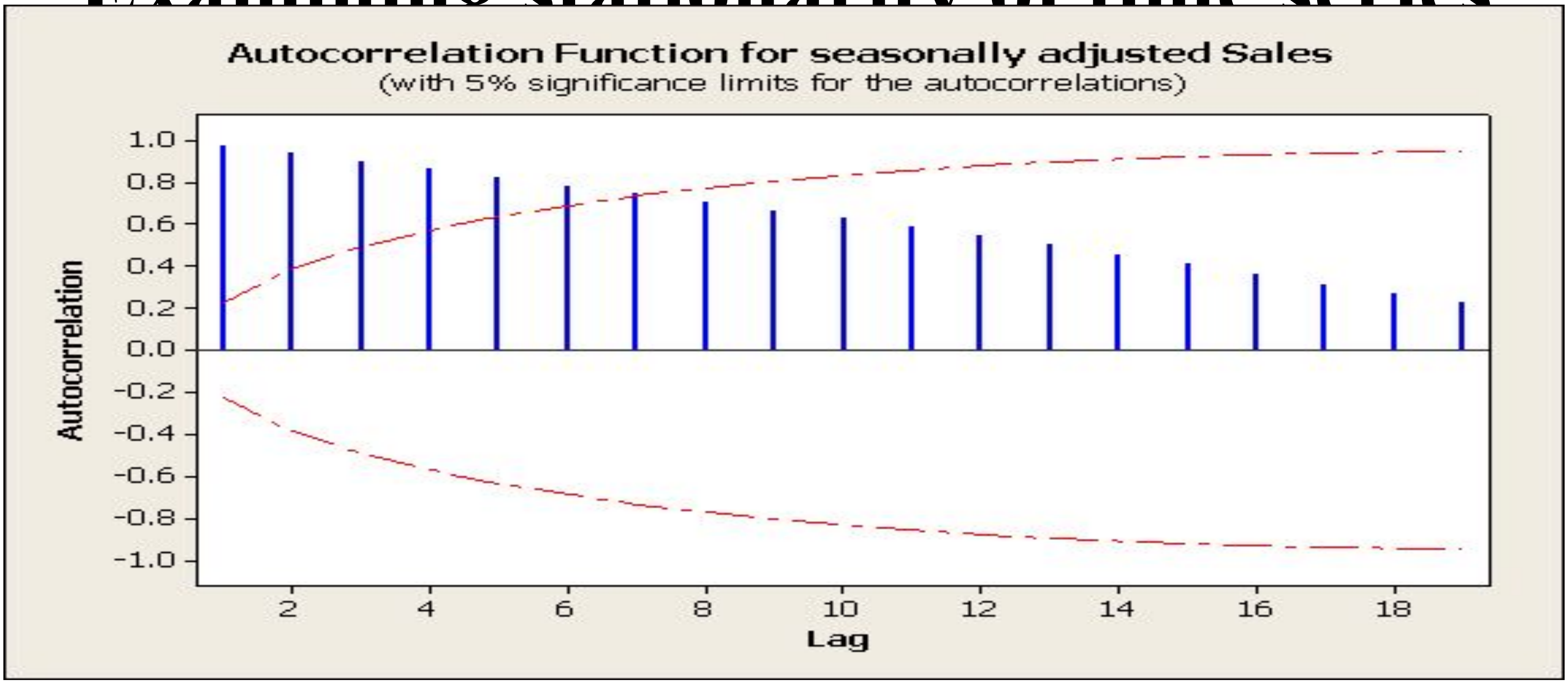
Dr. Mohammed Alahmed

Examining stationarity of time series data



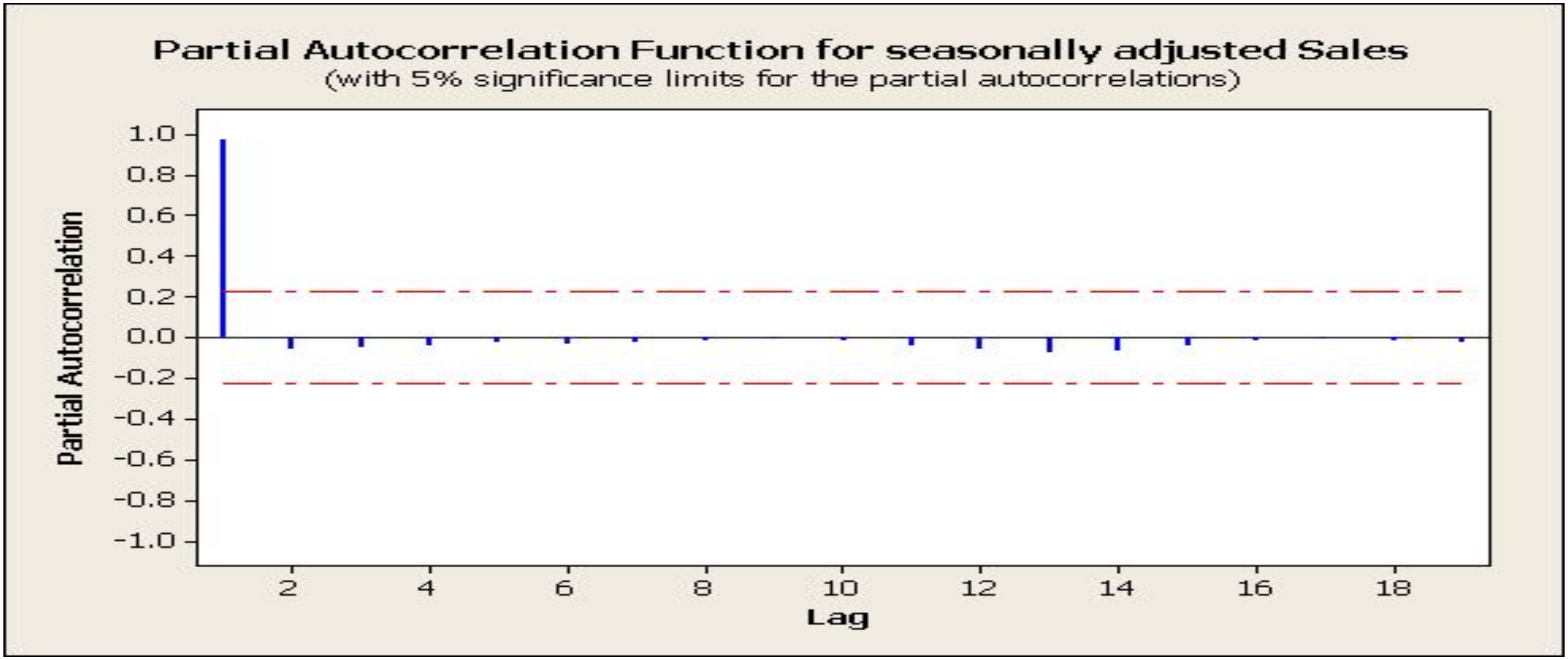
- The time series plot shows that it is **non-stationary** in the mean.
- The next slide shows the ACF plot for this data series.

Examining stationarity of time series



- The ACF also shows a pattern typical for a non-stationary series:
  - Large significant ACF for the first 7 time lag
  - Slow decrease in the size of the autocorrelations.
- The PACF is shown in the next slide.

Examining stationarity of time series



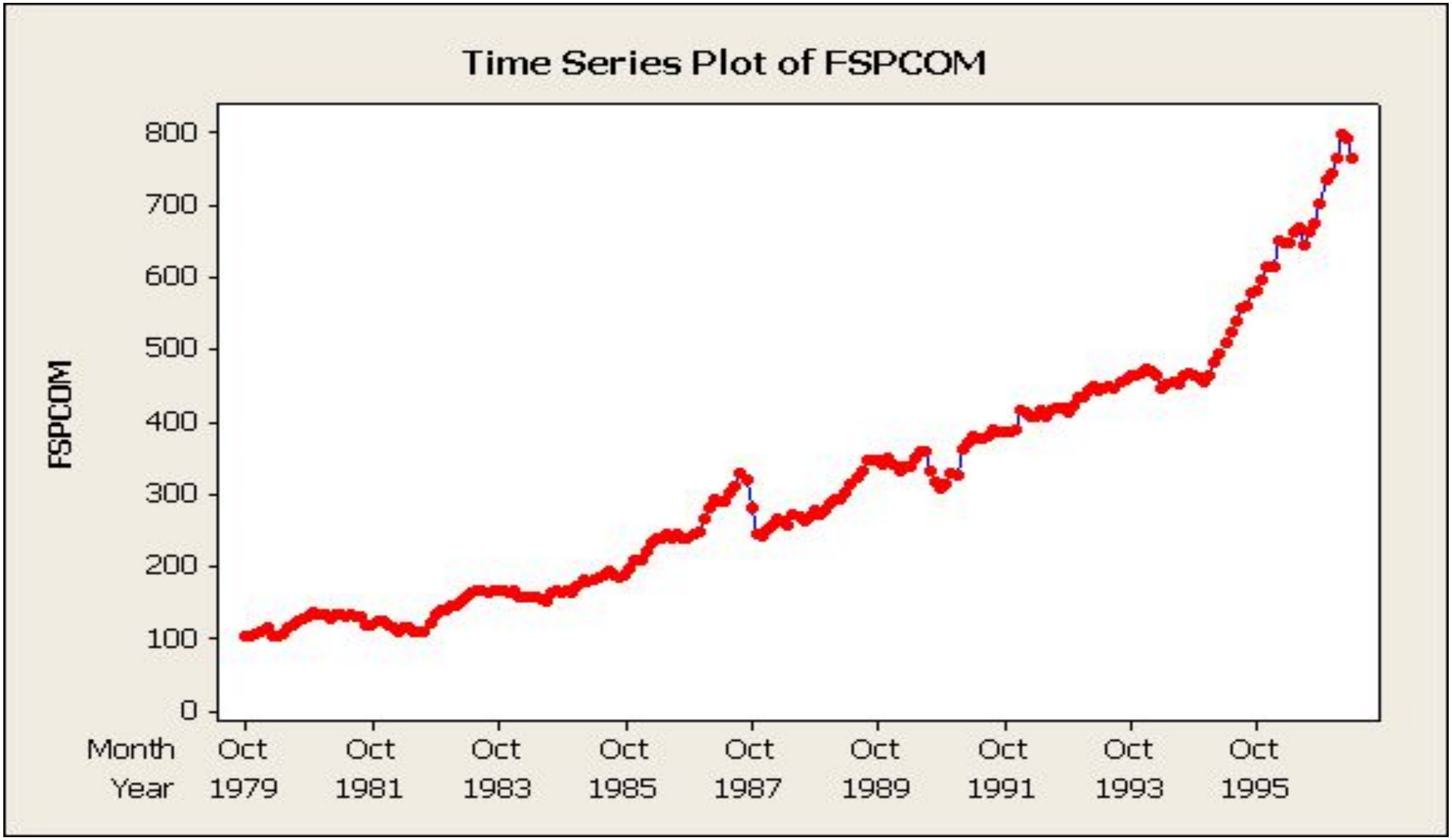
This is also typical of a non-stationary series. Partial autocorrelation at time lag 1 is close to one and the partial autocorrelation for the time lag 2 through 18 are close to zero

Removing non-stationarity in time series

The non-stationary pattern in a time series data needs to be removed in order that other correlation structure present in the series can be seen before proceeding with model building.

One way of removing non-stationarity is through the method of differencing.

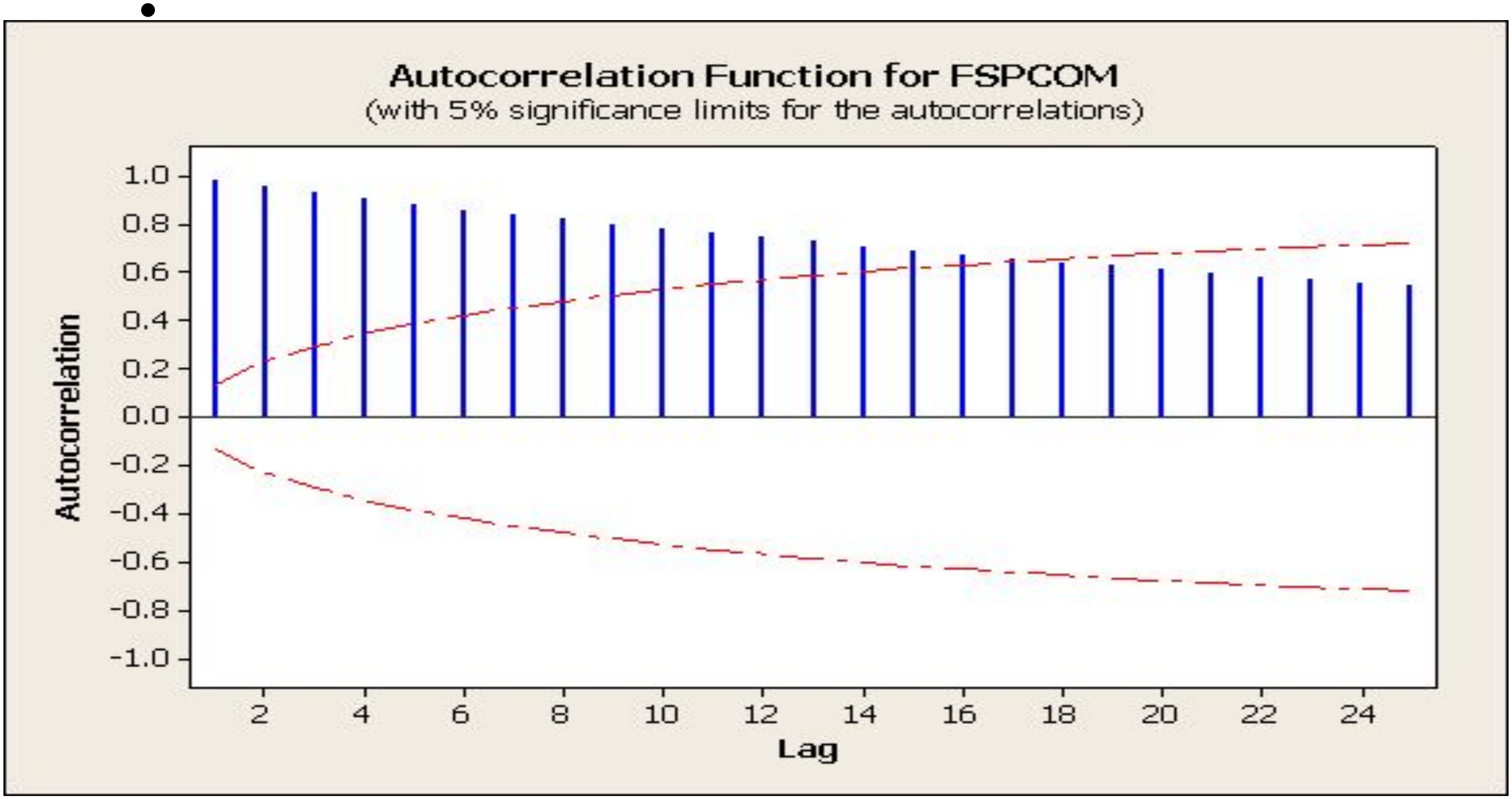
Removing non-stationarity in time



Dr. Mohammed Alahmed

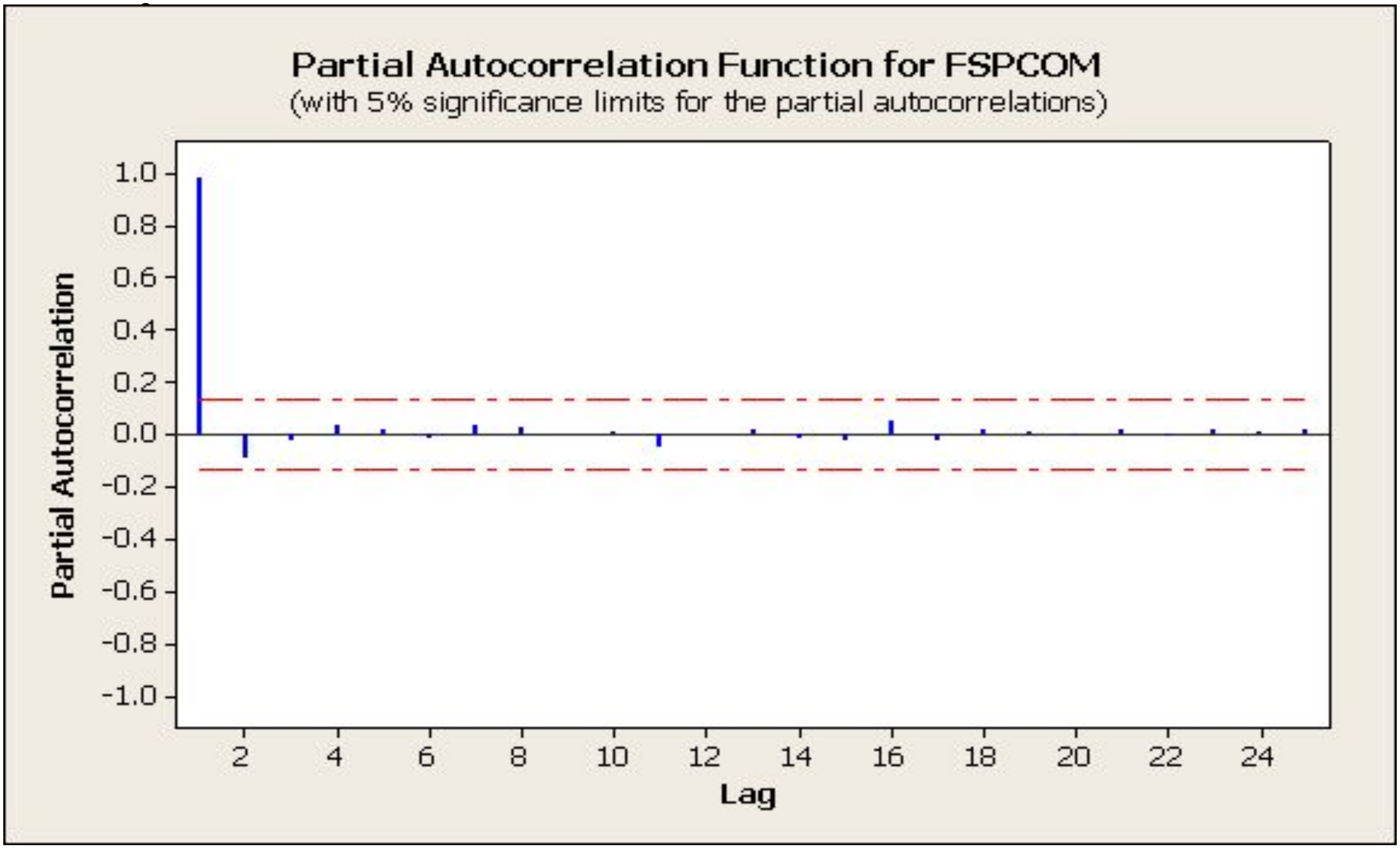
Removing non-stationarity in time

•



Dr. Mohammed Alahmed

Removing non-stationarity in time



Dr. Mohammed Alahmed

Removing non-stationarity in time series

The time plot shows that it is not stationary in the mean.

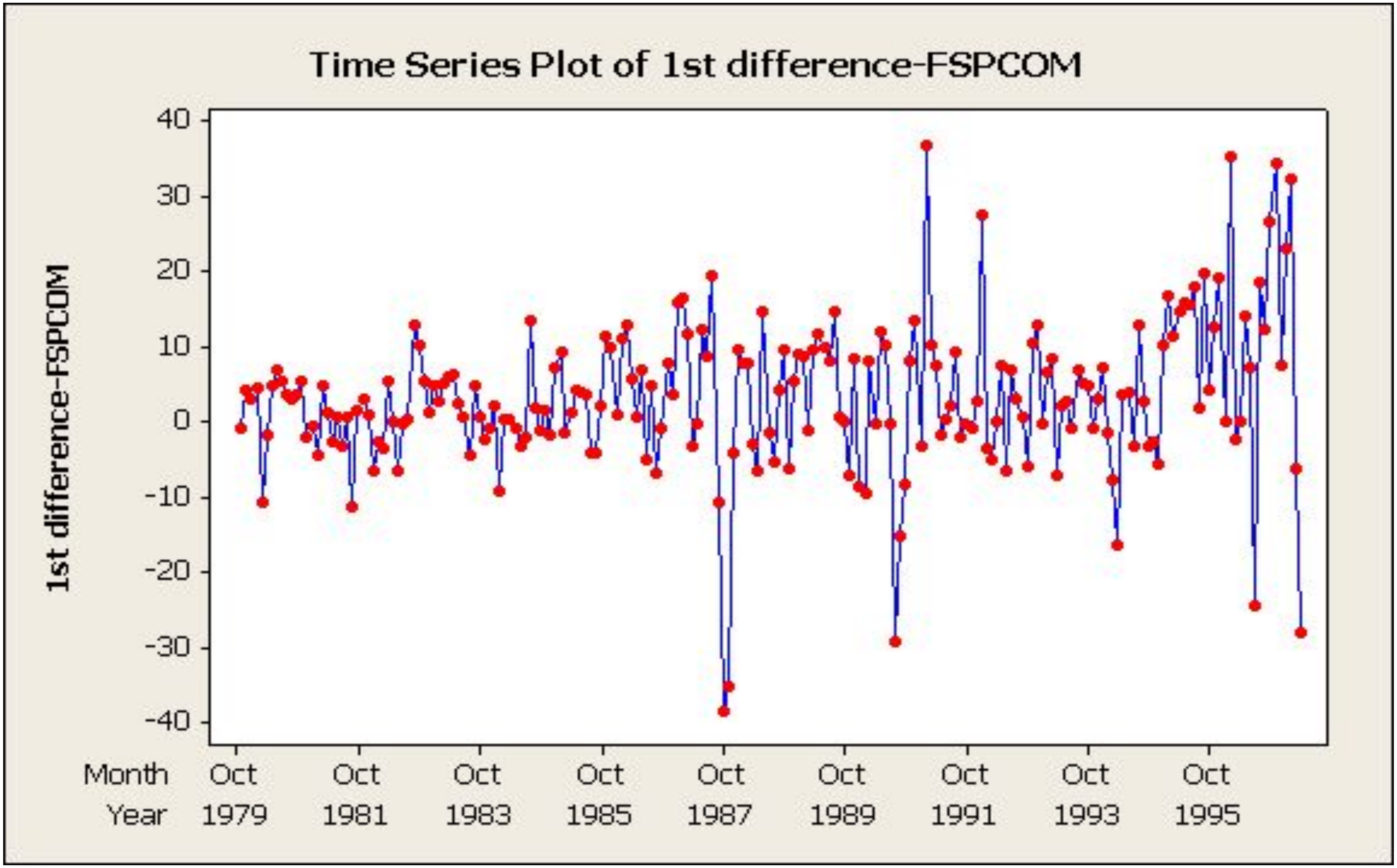
The ACF and PACF plot also display a pattern typical for non-stationary pattern.

Taking the first difference of the S& P 500 composite index data represents the monthly changes in the S&P 500 composite index.

Removing non-stationarity in time series

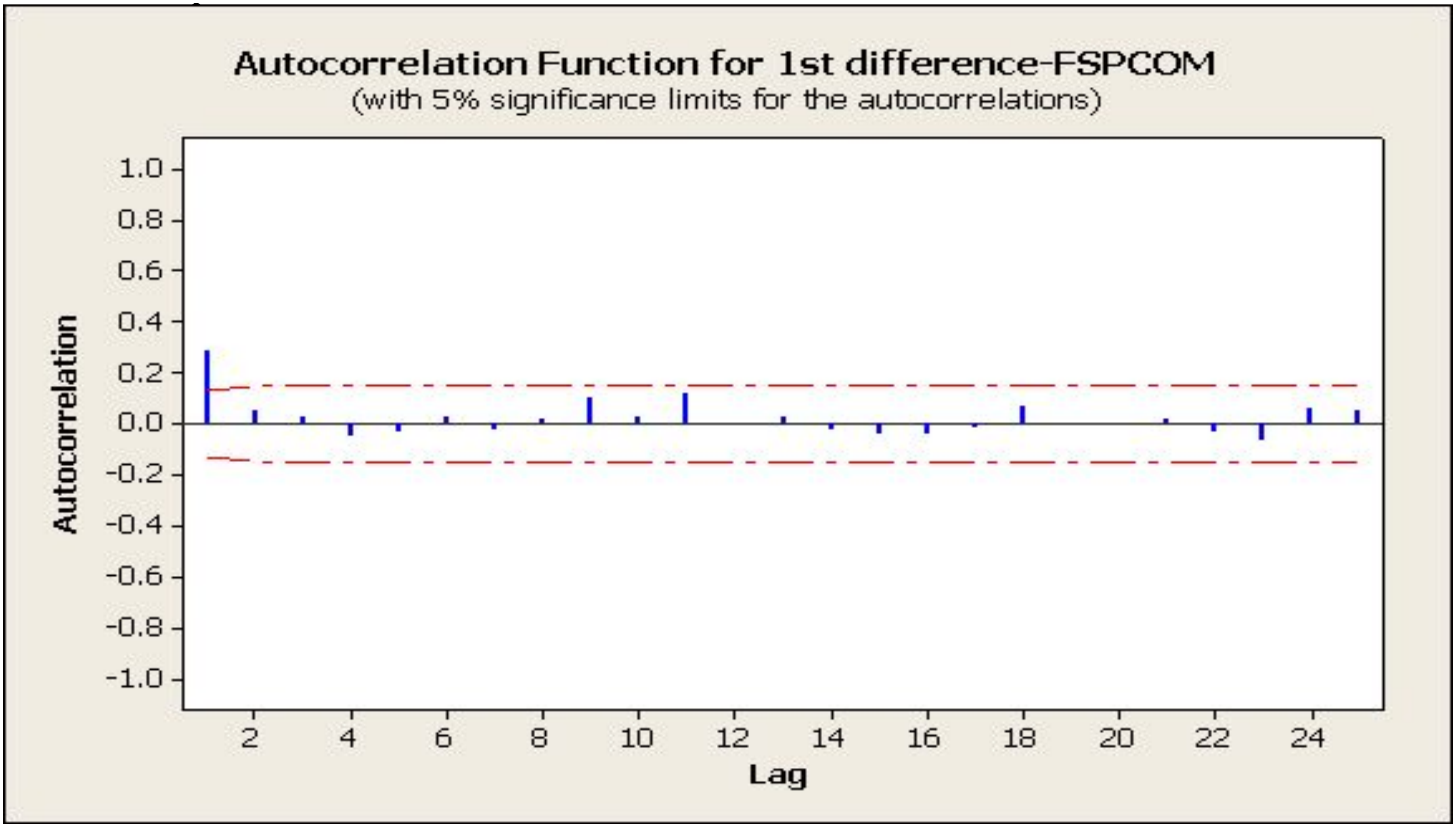
The time series plot and the ACF and PACF plots indicate that the first difference has removed the growth in the time series data.
The series looks just like a white noise with almost no autocorrelation or partial autocorrelation outside the 95% limits.

Removing non-stationarity in time



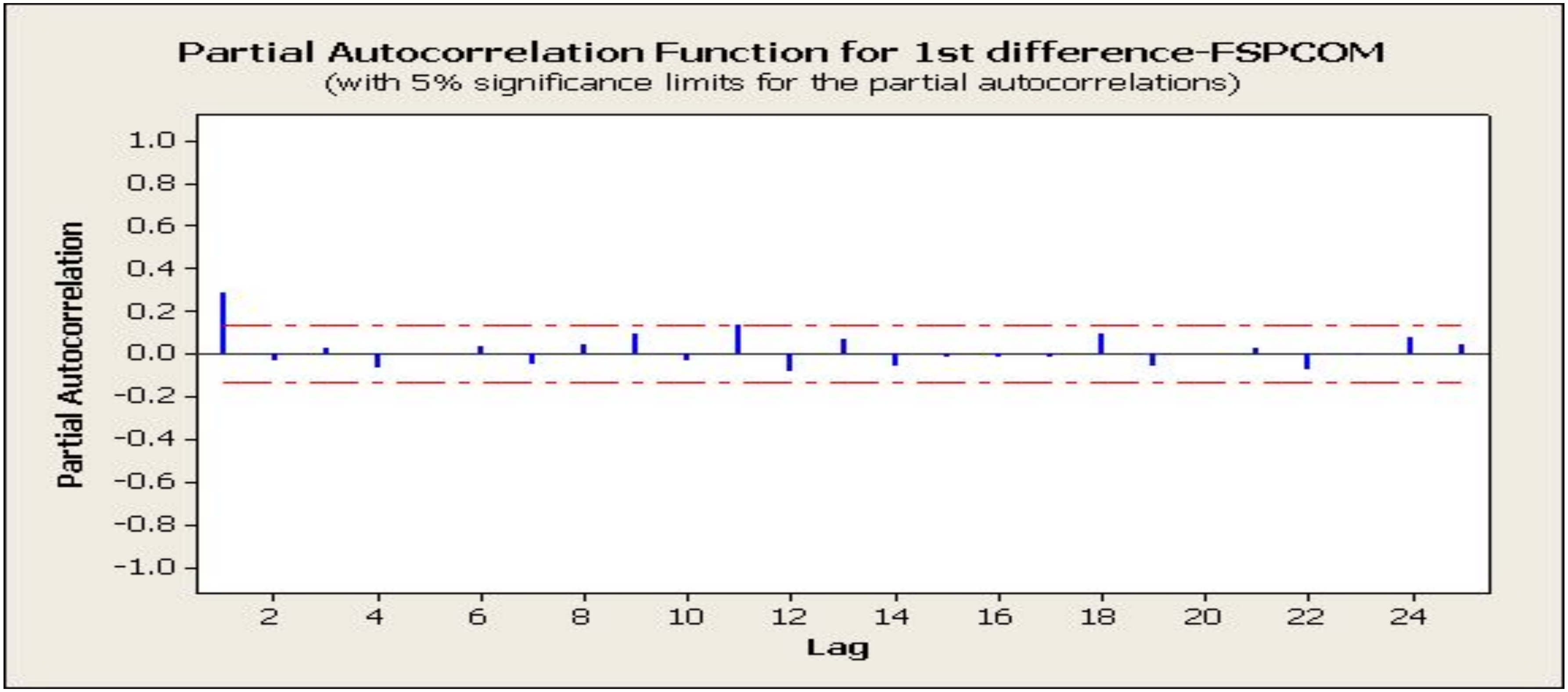
Dr. Mohammed Alahmed

Removing non-stationarity in time



Dr. Mohammed Alahmed

Removing non-stationarity in time series



Note that the ACF and PACF at lag 1 is outside the limits, but it is acceptable to have about 5% of spikes fall a short distance beyond the limit due to chance.

Seasonal differencing

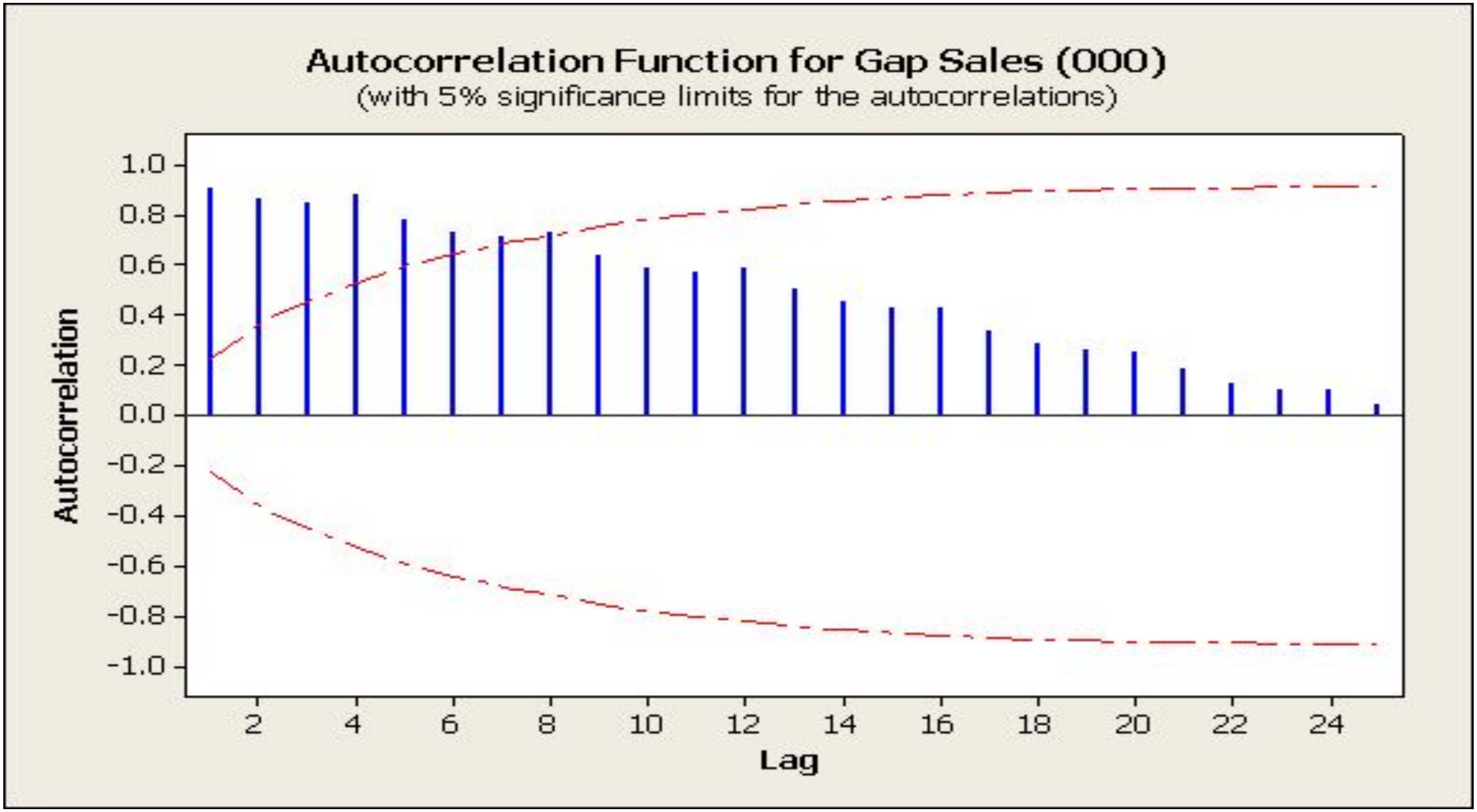
The Gap quarterly sales is an example of a non-stationary seasonal data.
The following time series plot show a trend with a pronounced seasonal component
The auto correlations show that:

- The series is non-stationary.
- The series is seasonal.

Seasonal differencing

Dr. Mohammed Alahmed

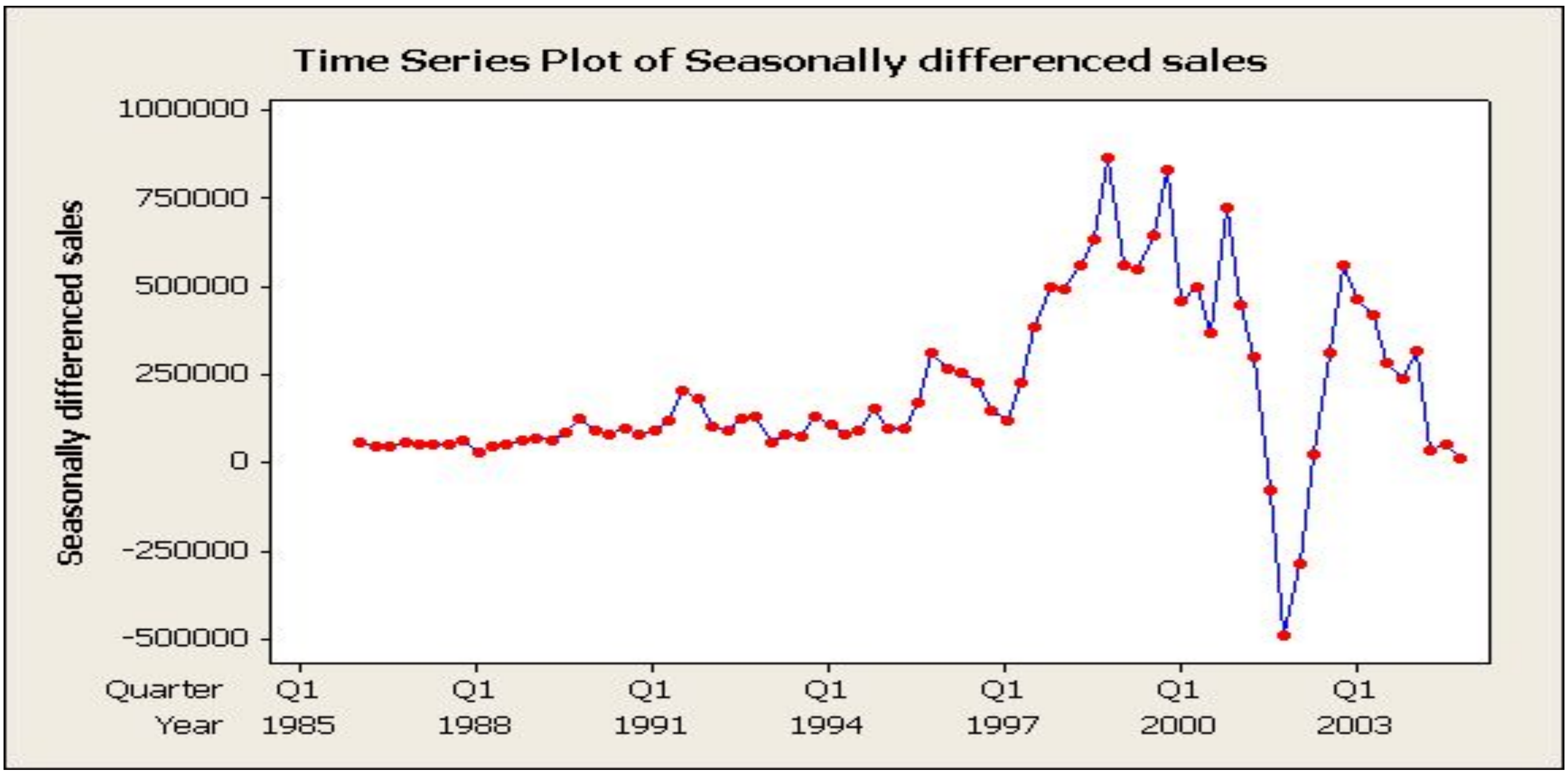
Seasonal differencing

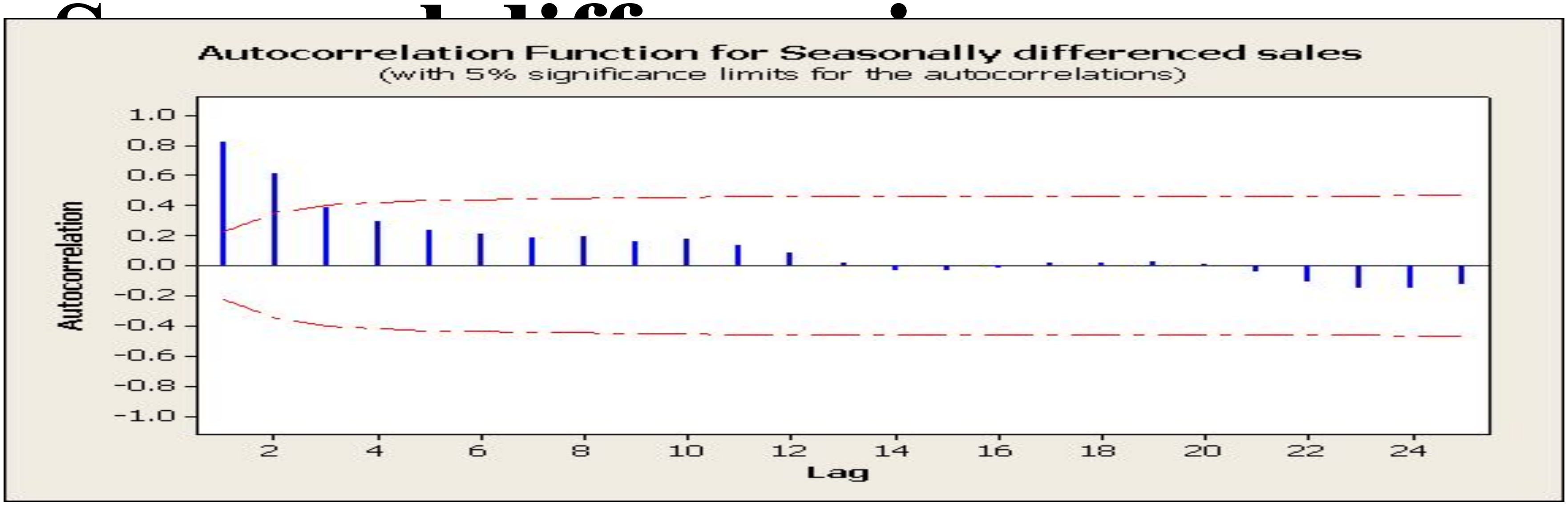


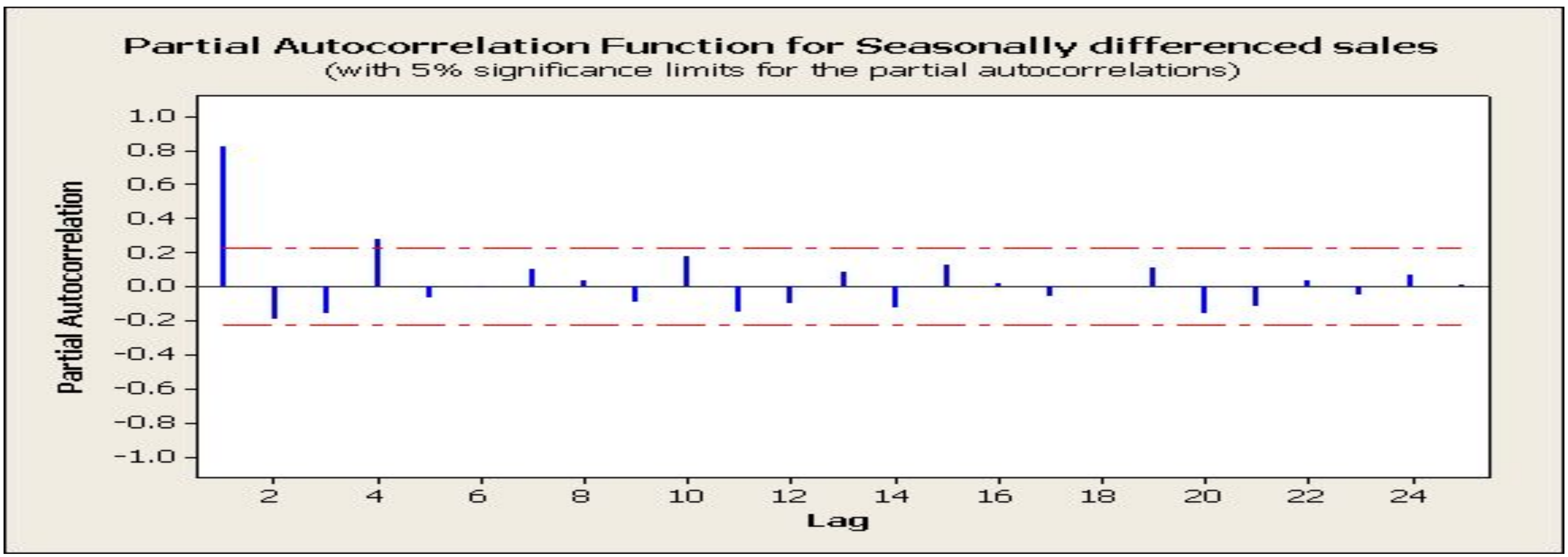
Dr. Mohammed Alahmed

Seasonal differencing

- The seasonally differenced series represents the change in sales between quarters of consecutive years.
- The time series plot, ACF and PACF of the seasonally differenced Gap’s quarterly sales are in the following three slides.




Autocorrelation Function for Seasonally differenced sales
(with 5% significance limits for the autocorrelations)
Autocorrelation
Lag


Partial Autocorrelation Function for Seasonally differenced sales
(with 5% significance limits for the partial autocorrelations)
Partial Autocorrelation
Lag

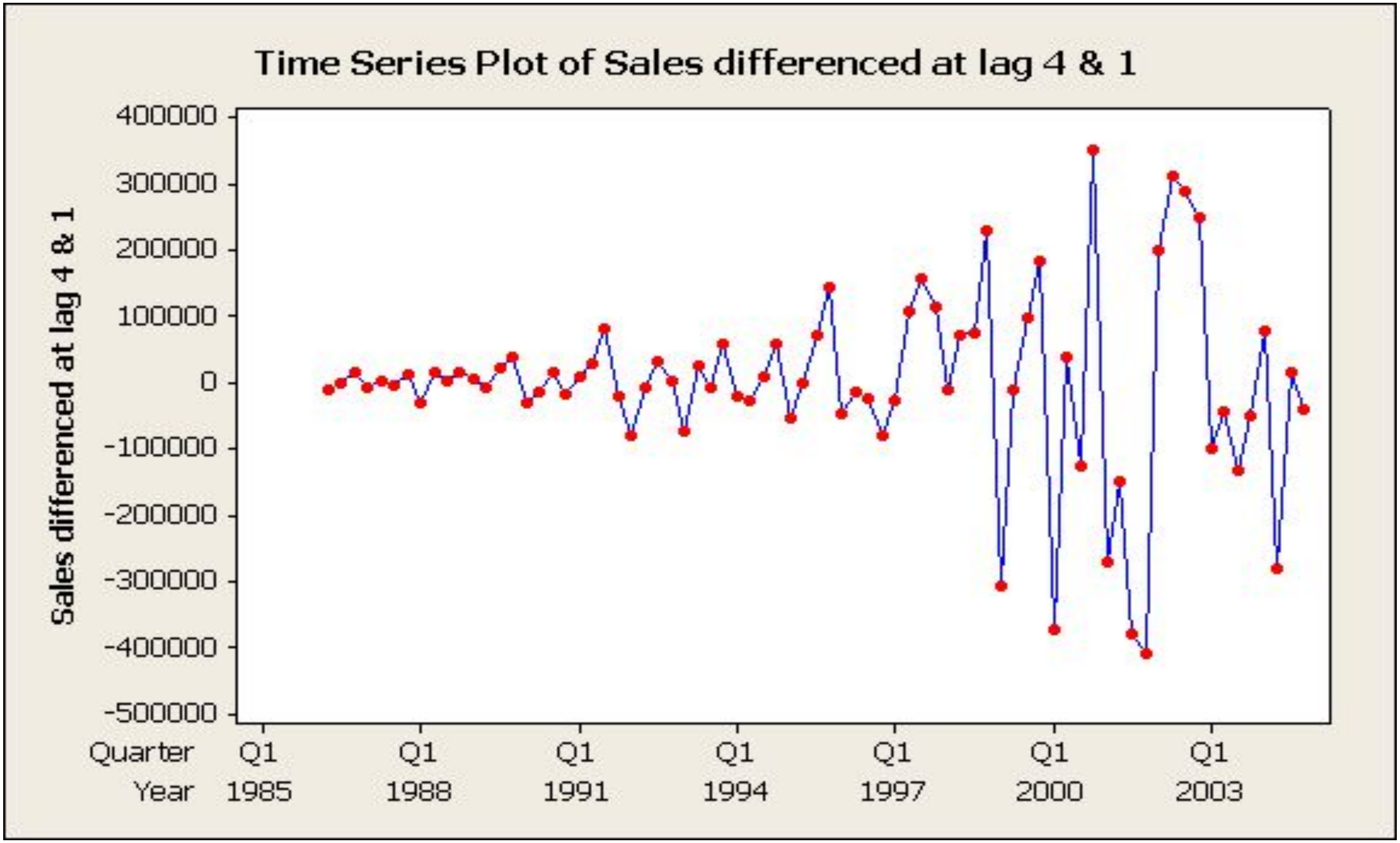
Seasonal differencing

The series is now much closer to being stationary, but more than 5% of the spikes are beyond 95% critical limits and autocorrelation show gradual decline in values.
The seasonality is still present as shown by spike at time lag 4 in the PACF.
The remaining non-stationarity in the mean can be removed with a further first difference.
When both seasonal and first differences are applied, it does not make no difference which is done first.

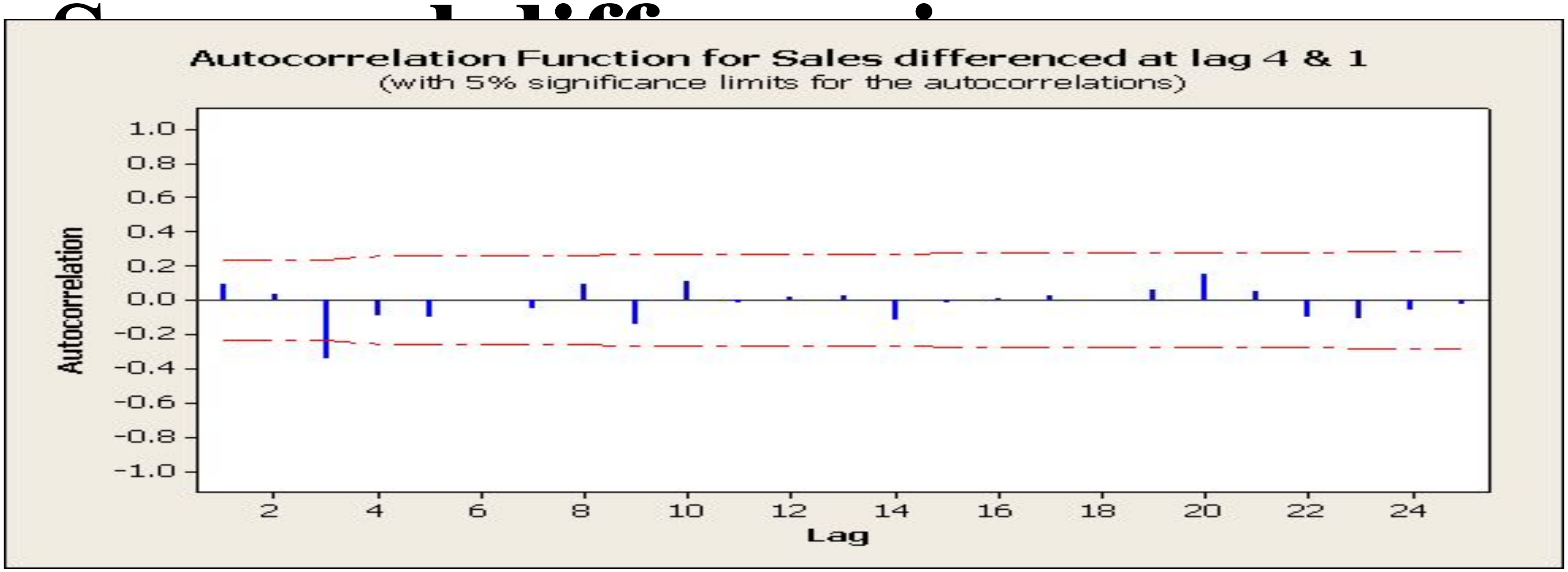
Seasonal differencing

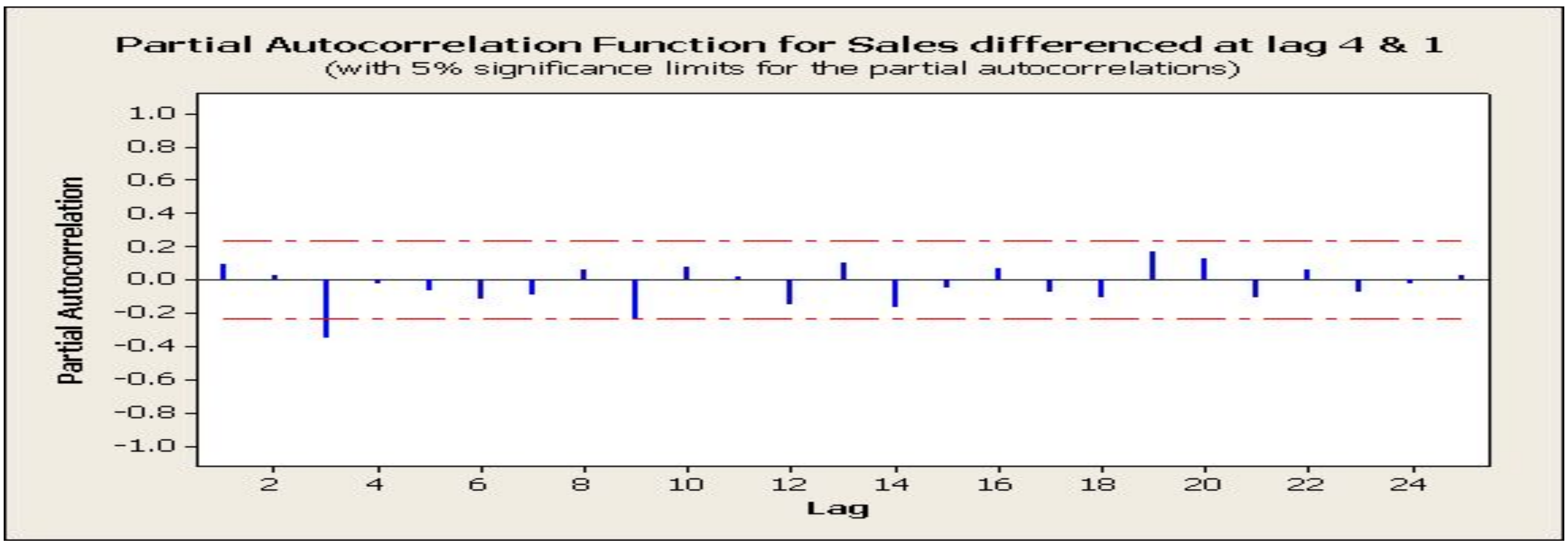
- It is recommended to do the seasonal differencing first since sometimes the resulting series will be stationary and hence no need for a further first difference.
- When differencing is used, it is important that the differences be interpretable.
- The series resulted from first difference of seasonally differenced Gap's quarterly sales data is reported in the following three slides.
- Is the resulting series white noise?

Seasonal differencing



Dr. Mohammed Alahmed


Autocorrelation Function for Sales differenced at lag 4 & 1
(with 5% significance limits for the autocorrelations)
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
2 4 6 8 10 12 14 16 18 20 22 24
Lag


Partial Autocorrelation Function for Sales differenced at lag 4 & 1
(with 5% significance limits for the partial autocorrelations)
Partial Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
2 4 6 8 10 12 14 16 18 20 22 24
Lag

Tests for Stationarity

The value of φ is estimated using ordinary least squares.
If the original series y_t needs differencing, the estimated value of φ will be close to zero.
If y_t is already stationary, the estimated value of φ will be negative.

ARIMA models for time series data

Autoregressive (AR) models can be coupled with moving average (MA) models to form a general and useful class of time series models called *Autoregressive Moving Average (ARMA)* models.

These can be used when the data are stationary.

This class of models can be extended to non-stationary series by allowing the differencing of the data series.

These are called *Autoregressive Integrated Moving Average (ARIMA)* models.

There are a large variety of ARIMA models.

ARIMA models for time series data

- The general non-seasonal model is known as ARIMA (p, d, q):
  - p is the number of autoregressive terms.
  - d is the number of differences.
  - q is the number of moving average terms.

- A white noise model is classified as ARIMA (0, 0, 0)
  - No AR part since y_t does not depend on y_{t-1}
  - There is no differencing involved.
  - No MA part since y_t does not depend on e_{t-1}

ARIMA models for time series data

A random walk model is classified as ARIMA (0, 1, 0)

- There is no AR part.
- There is no MA part.
- There is one difference.

Note that if any of p, d, or q are equal to zero, the model can be written in a shorthand notation by dropping the unused part.

Example

ARIMA(2, 0, 0) = AR(2)

ARIMA (1, 0, 1) = ARMA(1, 1)

An autoregressive model of order one

The time plot of an AR(1) model varies with the parameter φ_1:

- When $\varphi_1 = 0$, y_t is equivalent to a white noise series.
- When $\varphi_1 = 1$, y_t is equivalent to a random walk series
- For negative values of φ_1, the series tends to oscillate between positive and negative values.

The following slides show the time series, ACF and PACF plot for an ARIMA(1, 0, 0) time series data.

Higher order auto regressive models

Restrictions on the allowable values of auto regression parameters

For p =1

$-1 < \varphi_1 < 1$

For p = 2

$-1 < \varphi_2 < 1$

$\varphi_1 + \varphi_2 < 1$

$\varphi_2 - \varphi_1 < 1$

Higher order auto regressive models

A great variety of time series are possible with autoregressive models.
The following slides shows an AR(2) model.
Note that for AR(2) models the autocorrelations die out in a damped Sine-wave patterns.
There are exactly two significant partial autocorrelations.

Higher order moving average models

Restrictions on the allowable values of the MA parameters.

For q =1

$-1 < \theta_1 < 1$

For q =2

$-1 < \theta_2 < 1$

$\theta_1 + \theta_2 < 1$

$\theta_2 - \theta_1 < 1$

Higher order moving average models

A wide variety of time series can be produced using moving average models.
In general, the autocorrelations of an MA(q) models are zero beyond lag q
For $q \geq 2$, the PACF can show exponential decay or damped sine-wave patterns.

Mixtures ARIMA models

The general ARIMA (p, d, q) model gives a tremendous variety of patterns in the ACF and PACF, so it is not practical to state rules for identifying general ARIMA models.
In practice, it is seldom necessary to deal with values p, d, or q that are larger than 0, 1, or 2.
It is remarkable that such a small range of values for p, d, or q can cover such a large range of practical forecasting situations.

Seasonality and ARIMA models

- The seasonal lags of the ACF and PACF plots show the seasonal parts of an AR or MA model.
- Examples:

1. Seasonal MA model:
   - $ARIMA(0,0,0)(0,0,1)_{12}$
     - will show a spike at lag 12 in the ACF but no other significant spikes.
     - The PACF will show exponential decay in the seasonal lags i.e. at lags 12, 24, 36,…
2. Seasonal AR model:
   - $ARIMA(0,0,0)(1,0,0)_{12}$
     - will show exponential decay in seasonal lags of the ACF.
     - Single significant spike at lag 12 in the PACF.

Dr. Mohammed Alahmed

Implementing the model –Building Strategy

- The Box –Jenkins approach uses an iterative model-building strategy that consist of:
  1. Selecting an initial model (model identification)
  2. Estimating the model coefficients (parameter estimation)
  3. Analyzing the residuals (model checking)

  - If necessary, the initial model is modified and the process is repeated until the residual indicate no further modification is necessary.
  - At this point the fitted model can be used for forecasting.

Model identification

The following approach outlines an approach to select an appropriate model among a large variety of ARIMA models possible.

- Plot the data
  - Identify any unusual observations
  - If necessary, transform the data to stabilize the variance

Check the time series plot, ACF, PACF of the data (possibly transformed) for stationarity.

- IF
  - Time plot shows the data scattered horizontally around a constant mean
  - ACF and PACF to or near zero quickly
- Then, the data are stationary.

Model identification

Use differencing to transform the data into a stationary series

For no-seasonal data take first differences

For seasonal data take seasonal differences

Check the plots again if they appear non-stationary, take the differences of the differenced data.

When the stationarity has been achieved, check the ACF and PACF plots for any pattern remaining.

Model identification

- There are three possibilities:
  - AR or MA models
  - No significant ACF after time lag q indicates MA(q) may be appropriate.
  - No significant PACF after time lag p indicates that AR(p) may be appropriate.

- Seasonality is present if ACF and/or PACF at the seasonal lags are large and significant.
- If no clear MA or AR model is suggested, a mixture model may be appropriate

Model identification

Example (1):

Non seasonal time series data.

The following example looks at the number of users logged onto an internet server over a 100 minutes period.

The time plot, ACF and PACF is reported in the following three slides.

Model identification

The gradual decline of ACF values indicates non-stationary series.
The first partial autocorrelation is very dominant and close to 1, indicating non-stationarity.
The time series plot clearly indicates non-stationarity.
We take the first differences of the data and reanalyze.

Model identification

Example (2):

A seasonal time series.

The following example looks at the monthly industry sales (in thousands of francs) for printing and writing papers between the years 1963 and 1972.

The time plot, ACF and PACF shows a clear seasonal pattern in the data.

This is clear in the large values at time lag 12, 24 and 36.

Model identification

We take a seasonal difference and check the time plot, ACF and PACF.

The seasonally differenced data appears to be non-stationary (the plots are not shown), so we difference the data again.

the following three slides show the twice differenced series.

Model identification

Therefore, the identifies model is

$$\text{ARIMA } (0,1,1)(0,1,1)_{12}.$$

This model is sometimes is called the "airline model" because it was applied to international airline data by Box and Jenkins.

It is one of the most commonly used seasonal ARIMA model.

Model identification

Example (3):

A seasonal data needing transformation

In this example we look at the monthly shipments of a company that manufactures pollution equipment

The time plot shows that the variability increases as the time increases. This indicate that the data is non-stationary in the variance.

Model identification

We need to stabilize the variance before fitting an ARIMA model.
Logarithmic or power transformation of the data will make the variance stationary.
The time plot, ACF and PACF for the logged data is reported in the following three slides.

Model identification

The time plot shows that the magnitude of the fluctuations in the log-transformed data does not vary with time.

But, the logged data are clearly non-stationary.

The gradual decay of the ACF values.

To achieve stationarity, we take the first differences of the logged data.

The plots are reported in the next three slides.

Model identification

There are significant spikes at time lag 1 and 2 in the PACF, indicating an AR(2) might be appropriate.
The single significant spike at lag 12 of the PACF indicates a seasonal AR(1) component.
Therefore for the logged data a tentative model would be

$$ARIMA(2,1,0)(1,0,0)_{12}$$

Dr. Mohammed Alahmed

Summary

The process of identifying an ARIMA model requires experience and good judgment.
The following guidelines can be helpful:

1. Make the series stationary in mean and variance

   Differencing will take care of non-stationarity in the mean.
   Logarithmic or power transformation will often take care of non-stationarity in the variance.

Summary

2. Consider non-seasonal aspect

   The ACF and PACF of the stationary data obtained from the previous step can reveal whether MA of AR is feasible.

   Exponential decay or damped sine-wave. For ACF, and spikes at lags 1 to p then cut off to zero, indicate an AR(P) model.

   Spikes at lag1 to q, then cut off to zero for ACF and exponential decay or damped sine-wave for PACF indicates MA(q) model.

Summary

2. Consider seasonal aspect

   Examination of ACF and PACF at the seasonal lags can help to identify AR and MA models for the seasonal aspect of the data.

   For example, for quarterly data the pattern of r_4, r_8, r_{12}, r_{16}, and so on.

Estimating the parameters

Once a tentative model has been selected, the parameters for the model must be estimated.
The method of least squares can be used for RIMA model. However, for models with an MA components, there is no simple formula that can be used to estimate the parameters. Instead, an iterative method is used. This involves starting with a preliminary estimate, and refining the estimate iteratively until the sum of the squared errors is minimized.

Estimating the parameters

- Another method of estimating the parameters is the maximum likelihood procedure.
- Like least squares methods, these estimates must be found iteratively.
- Maximum likelihood estimation is usually favored because it has some desirable statistical properties.
- After the estimates and their standard errors are determined, t values can be constructed and interpreted in the usual way.
- Parameters that are judged significantly different from zero are retained in the fitted model; parameters that are not significantly different from zero are dropped from the model.

Diagnostic Checking

Before using the model for forecasting, it must be checked for adequacy.

A model is adequate if the residuals left over after fitting the model is simply white noise.

The pattern of ACF and PACF of the residuals may suggest how the model can be improved.

For example

- Significant spikes at the seasonal lags suggests adding seasonal component to the chosen model
- Significant spikes at small lags suggest increasing the non-seasonal AR or MA components of the model.

Diagnostic Checking

A portmanteau test can also be applied to the residuals as an additional test of fit.
If the portmanteau test is significant, then the model is inadequate.
In this case we need to go back and consider other ARIMA models.
Any new model will need their parameters estimated and their AIC values computed and compared with other models.
Usually, the model with the smallest AIC will have residuals which resemble white noise.
Occasionally, it might be necessary to adopt a model with not quite the smallest AIC value, but with better behaved residuals.

Dr. Mohammed Alahmed

Example

The analyst for the ISC Corporation was asked to develop forecasts for the closing prices of ISC stock. The stock has been languishing for some time with little growth, and senior management wanted some projections to discuss with the board of directors. The ISC stock prices are plotted in the following slide.

Example

The plot of the stock prices suggests the series is stationary.
The stock prices vary about a fixed level of approximately 250.
Is the Box-Jenkins methodology appropriate for this data series?
The ACF and PACF for the stock price series are reported in the following two slides.

Example

The sample ACF alternate in sign and decline to zero after lag 2.
The sample PACF are similar are close to zero after time lag 2.
These are consistent with an AR(2) or ARIMA(2,0,0) model
AR(2) model is fit to the data.
WE include a constant term to allow for a nonzero level.

Example

The estimated coefficient φ_2 is not significant (t=1.75) at 5% level but is significant at the 10 % level.

The residual ACF and PACF are given in the following two slides.

The ACF and PACF are well within their two standard error limits.

Final Estimates of Parameters

Type	Coef	SE	T	P
AR 1	-0.3243	0.1246	-2.60	0.012
AR 2	0.2192	0.1251	1.75	0.085
Constant	284.903	6.573	43.34	0.000

Example

The p-value for the Ljung-Box statistics for m = 12, 24, 36, and 48 are all large (> 5%) indicating an adequate model.
We use the model to generate forecasts for periods 66 and 67.

MS = 2808 DF = 62

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	6.3	13.3	18.2	29.1
DF	9	21	33	45
P-Value	0.707	0.899	0.983	0.969

Final Comments

In ARIMA modeling, it is NOT good practice to include AR and MA parameters to "cover all possibilities" suggested by the sample ACF and Sample PACF.
This means, when in doubt, start with a model containing **few** parameters rather than many parameters. The need for additional parameters will be evident from the residual ACF and PACF.
Least square estimates of AR and MA parameters in ARIMA models tend to be highly correlated. When there are more parameters than necessary, this leads to unstable models that can produce poor forecasts.

Final Comments

To summarize, start with a small number of clearly justifiable parameters and add one parameter at a time as needed.
If parameters in a fitted ARIMA model are not significant, delete one parameter at a time and refit the model.
Because of high correlation among estimated parameters, it may be the case that a previously non-significant parameter becomes significant.

Summary of rules for identifying ARIMA models

Identifying the order of differencing and the constant:

Rule 1: If the series has positive autocorrelations out to a high number of lags (say, 10 or more), then it probably needs a higher order of differencing.

Rule 2: If the lag-1 autocorrelation is zero or negative, or the autocorrelations are all small and patternless, then the series does *not* need a higher order of differencing. If the lag-1 autocorrelation is -0.5 or more negative, the series may be overdifferenced. **BEWARE OF OVERDIFFERENCING.**

Rule 3: The optimal order of differencing is often the order of differencing at which the standard deviation is lowest. (Not always, though. Slightly too much or slightly too little differencing can also be corrected with AR or MA terms. See rules 6 and 7.)

Summary of rules for identifying ARIMA models

Rule 4: A model with no orders of differencing assumes that the original series is stationary (among other things, mean-reverting). A model with one order of differencing assumes that the original series has a constant average trend (e.g. a random walk or SES-type model, with or without growth). A model with two orders of total differencing assumes that the original series has a time-varying trend (e.g. a random trend or LES-type model).

Rule 5: A model with no orders of differencing normally includes a constant term (which allows for a non-zero mean value). A model with two orders of total differencing normally does not include a constant term. In a model with one order of total differencing, a constant term should be included if the series has a non-zero average trend.

Summary of rules for identifying ARIMA models

Identifying the numbers of AR and MA terms:

Rule 6: If the partial autocorrelation function (PACF) of the differenced series displays a sharp cutoff and/or the lag-1 autocorrelation is positive--i.e., if the series appears slightly "underdifferenced"--then consider adding one or more AR terms to the model. The lag beyond which the PACF cuts off is the indicated number of AR terms.

Rule 7: If the autocorrelation function (ACF) of the differenced series displays a sharp cutoff and/or the lag-1 autocorrelation is negative--i.e., if the series appears slightly "overdifferenced"--then consider adding an MA term to the model. The lag beyond which the ACF cuts off is the indicated number of MA terms.

Summary of rules for identifying ARIMA models

Rule 8: It is possible for an AR term and an MA term to cancel each other's effects, so if a mixed AR-MA model seems to fit the data, also try a model with one fewer AR term and one fewer MA term--particularly if the parameter estimates in the original model require more than 10 iterations to converge. BEWARE OF USING MULTIPLE AR TERMS AND MULTIPLE MA TERMS IN THE SAME MODEL.

Rule 9: If there is a unit root in the AR part of the model--i.e., if the sum of the AR coefficients is almost exactly 1--you should reduce the number of AR terms by one and increase the order of differencing by one.

Rule 10: If there is a unit root in the MA part of the model--i.e., if the sum of the MA coefficients is almost exactly 1--you should reduce the number of MA terms by one and reduce the order of differencing by one.

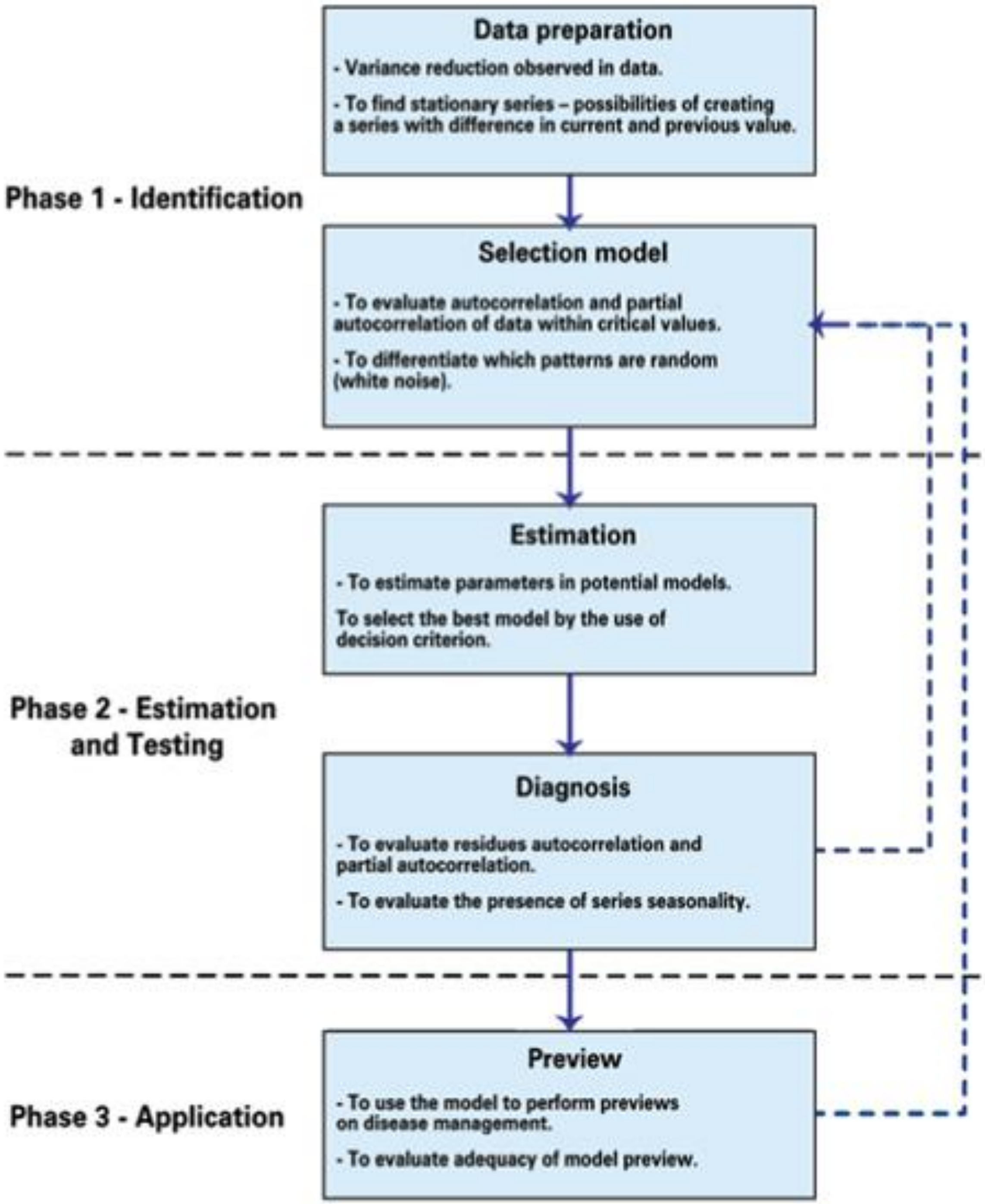
Rule 11: If the long-term forecasts* appear erratic or unstable, there may be a unit root in the AR or MA coefficients.

Summary of rules for identifying ARIMA models

Identifying the seasonal part of the model:

Rule 12: If the series has a strong and consistent seasonal pattern, then you must use an order of seasonal differencing (otherwise the model assumes that the seasonal pattern will fade away over time). However, never use more than one order of seasonal differencing or more than 2 orders of total differencing (seasonal+nonseasonal).

Rule 13: If the autocorrelation of the appropriately differenced series is positive at lag s, where s is the number of periods in a season, then consider adding an SAR term to the model. If the autocorrelation of the differenced series is negative at lag s, consider adding an SMA term to the model. The latter situation is likely to occur if a seasonal difference has been used, which should be done if the data has a stable and logical seasonal pattern. The former is likely to occur if a seasonal difference has not been used, which would only be appropriate if the seasonal pattern is not stable over time. You should try to avoid using more than one or two seasonal parameters (SAR+SMA) in the same model, as this is likely to lead to overfitting of the data and/or problems in estimation.



Dr. Mohammed Alahmed