Time Series Analysis

Agenda

Key Takeaways-

- What is Time Series analysis, Time Series data and its applications
- Stationary Series
- Time Series Components
- Decomposition Methods
- Time Series Modelling
- Simple Moving Average
- Exponential Smoothing
- ARIMA

Time Series Analysis

*"**Time series analysis** comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. **Time series forecasting** is the use of a model to predict future values based on previously observed values."* **- WikiPedia**

Time Series

1. **Time -** Having a time component (like minutes, hours, days, weeks, months, years etc.)
2. **Series** - Sequence of numeric data points

- So here the data is observed/collected at regular time intervals (or over a period of time).

- In time series, the data is dependent on previous values.

Such data may have some internal structure (such as trend, seasonality, autocorrelation etc.). This structure/phenomena should be accounted while extracting insights or making predictions. And time series analysis helps in knowing such structure/phenomena.

Time Series Analysis Applications

Time Series Analysis is used for applications such as -

- Forecasting financial market trends (sales, orders etc.)
- Stock Market Analysis and Forecasting
- Website traffic Forecasting
- Weather Forecasting
- Air Quality Forecasting
- Inventory Studies
- Census Analysis
- Budgetary Analysis

Stationary Series

Traditional/classical time series analysis/forecasting methods expects the time series to be a stationary series to perform analysis.

A stationary time series exhibits the following characteristics.

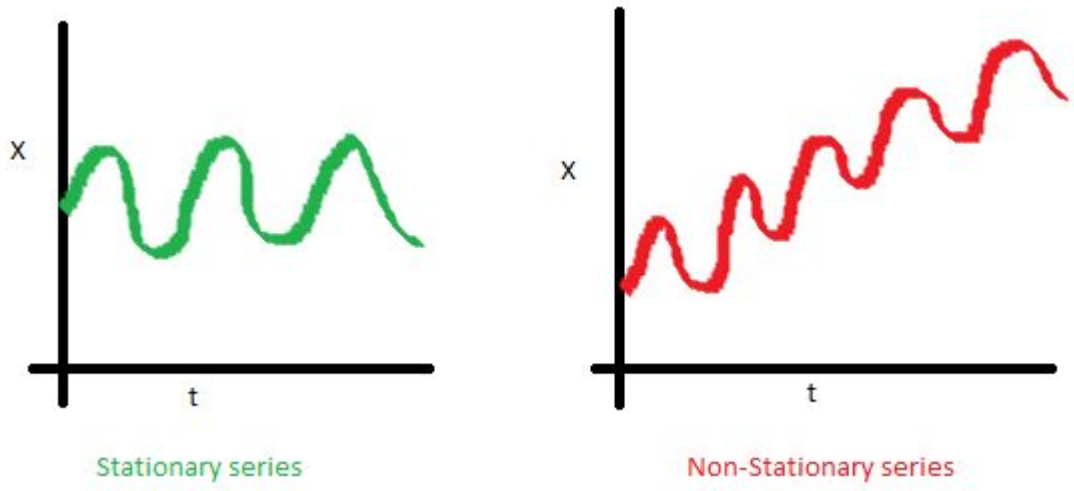
- Mean is constant
- Variance is constant
- Covariance is constant

If the time series under analysis is non-stationary then first we need to stationarize the series.

Stationary Series [Contd.]

1. Mean Constraint

The mean of the series should be constant over time. It should not be a function of time.

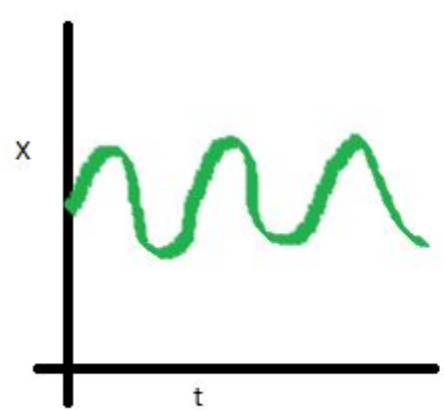


Here, in 1st graph mean is parallel to x-axis whereas in 2nd graph, the mean is upward sloping(increasing) w.r.t time.

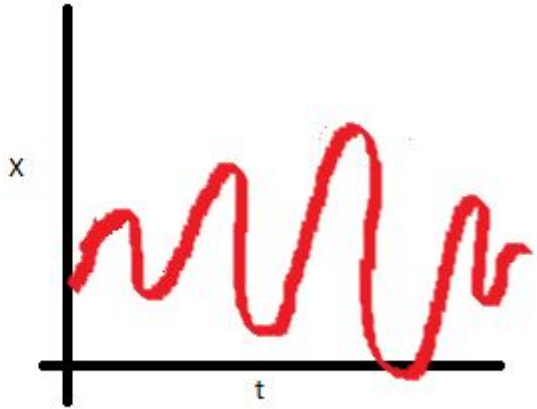
Stationary Series [Contd.]

2. Variance Constraint

- The variance of the series should be constant over time. It should not be a function of time.
- This property is also known as homoscedasticity.



Stationary series

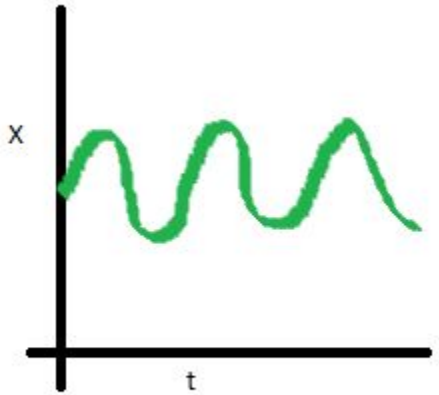


Non-Stationary series

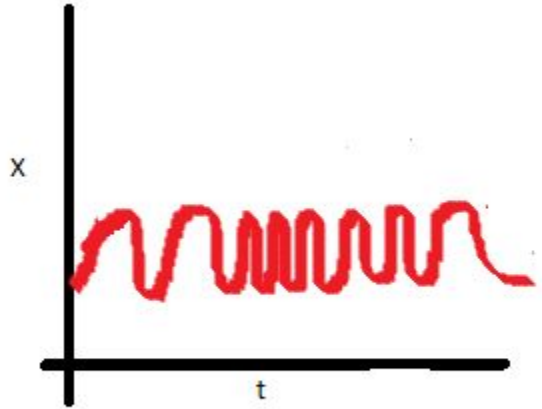
Stationary Series [Contd.]

3. Covariance Constraint

The Covariance between data points at time ‘t’ and at time ‘t+m’ should be constant over time. It should not be a function of time.



Stationary series



Non-Stationary series

Quiz 1

Choose the correct statement(s).

- In time series, data is dependent on previous values in the series.
- Heteroscedasticity indicates that mean and variance are constant.
- Heteroscedasticity indicates that mean and variance are variable.
- **1** and **2**

Time Series Components

- A time series can be further decomposed into four components and each component expresses particular aspect/property about the movement of values.

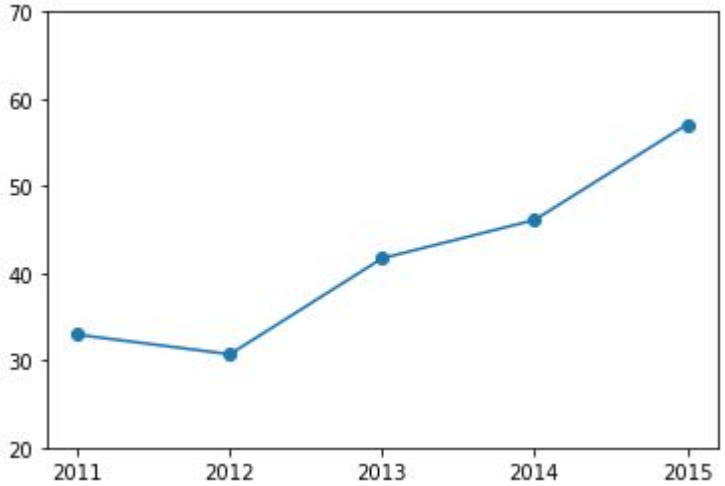
The four components are -

1. Trend
2. Seasonal
3. Cyclic
4. Random (Irregular Variations)

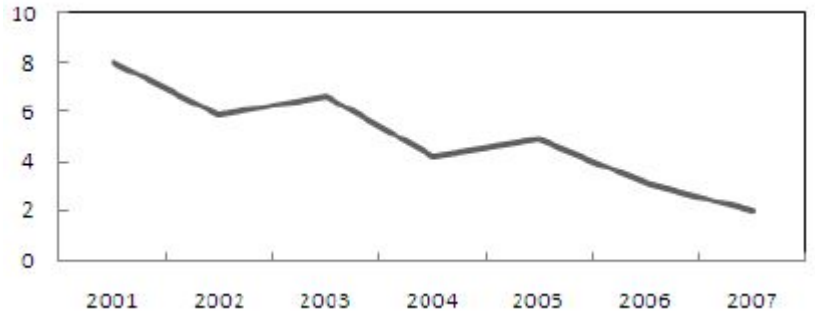
NOTE - It is not necessary for a time series to have all the components.

Trend

- A long term movement (either increase or decrease) found in the data is known as a trend.
- A time series can have overall upward or downward trend.
  - Upward Trend : E.g. Time series related to population, production etc.
  - Downward Trend : E.g. Time series related to death, epidemics etc.



Upward Trend



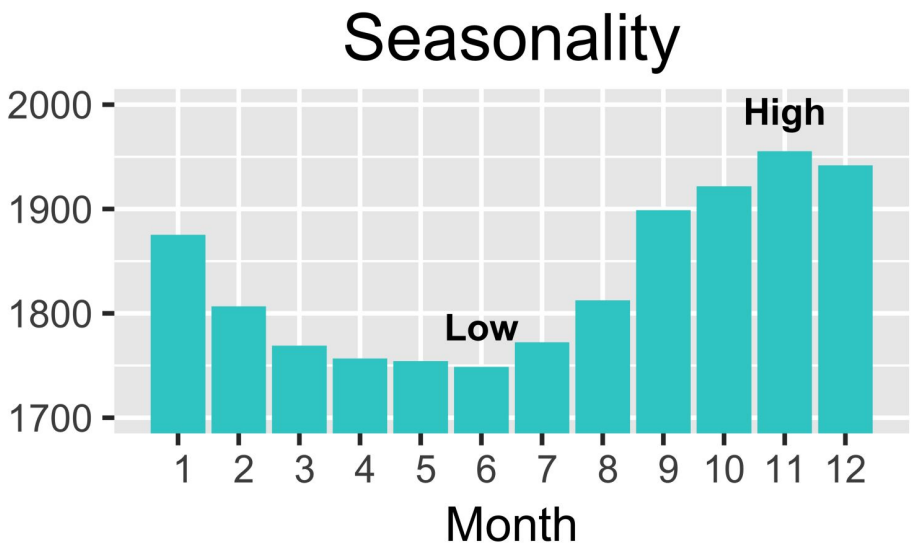
Downward Trend

Seasonal

- Cases where the data is affected by the seasonal factors.
- Seasonal Fluctuations describes any regular variation with a period of less than one year.

Seasonal factors -

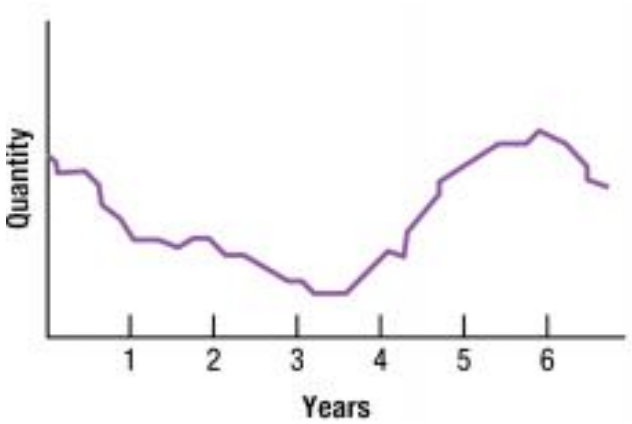
- **Natural Conditions**
  Weather fluctuations, climate changes etc.
- **Social and Cultural behaviour**
  Diwali, EID, Christmas
- **Business and Administrative procedures**
  Kids holidays, Winter vacations
- **Holiday Effects**
  New Year Eve, Independence day



Cyclic

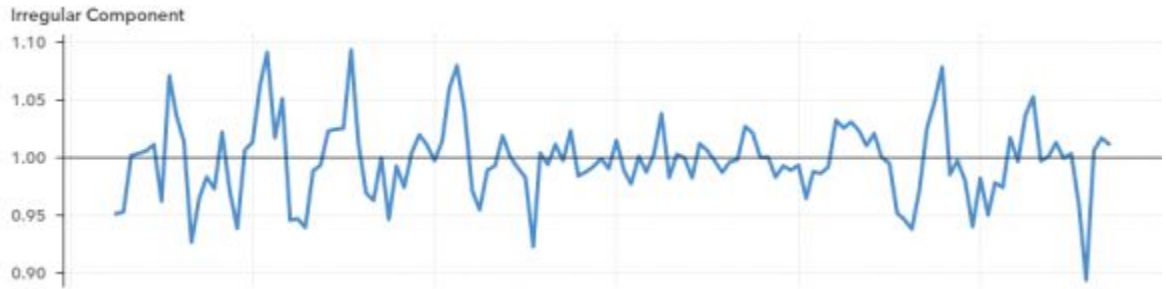
- Cyclic variations are often identified over a long period of time (more than one year).

- These variations in a time series are due to upswings and downswings recurring after a period higher than one year.



Random (irregular variations)

- These variations are neither systematic/regular nor predictable. Hence named as irregular variations.
- If all the three components from a time series are missing(or removed) then the series is termed as random.
- These fluctuations are unforeseen, uncontrollable.



Quiz 2

Choose the correct statement(s).

- Trend can be either upward or downward.
- A time series need to have all the components
- Seasonal variations usually occur in a year.
- **1** and **2**

Decomposition Methods

- The components of a time series can be extracted using decomposition methods.
- Component extraction helps in observing/exploring the impact of a component on the data which further leads to better analysis/prediction.
- There are basically two types of decomposition methods.
  - Additive
  - Multiplicative

Deseasonalized Data

- The presence of seasonal factor makes it difficult to decide whether the data has a clear upward or downward trend.

- The process of removing seasonal component from the data is known as deseasonalization and the resultant values are known as seasonally adjusted values.

- Using decomposition method first we can extract the seasonal component (if present) and then further we can remove it.

Additive

- In the time series, if the amplitude of both the seasonal and irregular variations do not change as the level of the trend rises or falls then additive model is appropriate.

- In other words, the seasonal variations are relatively constant over time.

Using additive method, the time series is decomposed as

Observed Time Series = (Trend) **+** (Seasonal) **+** (Random)

Using this,

Seasonally adjusted values = (Observed Time Series) **-** (Seasonal)

Multiplicative

- In the time series, if the amplitude of both the seasonal and irregular variations increases or decreases as the level of the trend rises or falls then multiplicative model is appropriate.

- In other words, the seasonal variations increases or decreases over time.

Using multiplicative method, the time series is decomposed as

Observed Time Series = (Trend) * (Seasonal) * (Random)

Using this,

Seasonally adjusted values = (Observed Time Series) / (Seasonal)

Quiz 3

How the Deseasonalized data is obtained -

- Adding the trend component to the time series in additive method.
- Dividing the time series by the seasonal component in multiplicative method.
- Removing the trend component to the time series in additive method.
- Multiplying the time series by the trend component in multiplicative method.

Time Series Modelling

- There are multiple ways to model a time series data in order to perform forecasting.

- In this section, we are going to cover -
  1. Moving Average
  2. Exponential Smoothing
  3. ARIMA (and its variants)

NOTE -

- Moving Average and Exponential Smoothing are also known as Smoothing techniques. They help in getting a more clear trend to forecast accurately.
- These are also known as short term forecasting techniques where we forecast immediate next value (or a set of values).

Moving Average

- It's follows a naive approach to perform time series modelling.
- As the name suggests, it forecasts the next value by taking average of past observations.
- Here, we define a window (of size k) to compute average of previous k observations.

Mathematically,

Simple Moving Average = $\frac{(y_{(t)} + y_{(t-1)} + y_{(t-2)} + \ldots\ldots y_{(t-k)})}{k}$

Window Size Selection

- A smaller window size (under smoothing case) leads to lots of variations which will disguise the trend.
- On the other end, with a larger window size (over smoothing case) some of the interesting patterns might be lost.
- Therefore, window size (k) should be selected in such a way that both the under smoothing and over smoothing cases are avoided.

Moving Average [Contd.]

- For a window size = 4, the moving average for first 3 records is not defined.

1949	# of Passengers	Window Size(k) = 4	Window Size(k) = 8
Jan	112		
Feb	118		
Mar	132		
Apr	129	(112+118+132+129)/4 = 122.75	
May	121	(118+132+129+121)/4 = 125.00	
Jun	135	(132+129+121+135)/4 = 129.25	
Jul	148	(129+121+135+148)/4 = 133.25	
Aug	148	(121+135+148+148)/4 = 138.00	130.37
Sep	136	(135+148+148+136)/4 = 141.75	133.37

Exponential Smoothing

Exponential smoothing assigns weights in decreasing order to the previous observations. For e.g., maximum weight to the most recent observation and minimum weight to the least recent observation.

Mathematically,

$$S_t = \alpha y_{(t)} + (1 - \alpha) S_{(t-1)}$$

Here, S(t) is the exponential smoothed value at time t.
y(t) is the latest observation in the series at time t.
α (alpha) is the smoothing constant.

- Exponential smoothing uses the latest observation (y(t)) and latest estimation (S(t-1)) to predict the next smoothed value (S(t)).

- α (alpha) ranges between **0** and **1**.
  - Usually α is not set to **0** because if we set to **0** then smoothed value will depend only on the latest estimation (S(t-1)).
  - Similarly α is not set to **1** because if we set to **1** then smoothed value will depend only on the latest observation (y(t)).

Exponential Smoothing [Contd.]

- For α (alpha) = 0.2

1949	# of Passengers	Mathematical Calculation	Smoothed Value
Jan	112		112
Feb	118	(0.2)*118 + (1-0.2)*112	113.2
Mar	132	(0.2)*132 + (1-0.2)*113.2	116.96
Apr	129	(0.2)*129 + (1-0.2)*116.96	119.36
May	121	(0.2)*121 + (1-0.2)*119.36	119.69
Jun	135	(0.2)*135 + (1-0.2)*119.69	122.75
Jul	148	(0.2)*148 + (1-0.2)*122.75	127.80
Aug	148	(0.2)*148 + (1-0.2)*127.80	131.84
Sep	136	(0.2)*136 + (1-0.2)*131.84	132.67

Quiz 4

Choose the correct statement(s).

- If window size (k) is set to larger value in moving average then there are less fluctuations in the overall trend.
- Exponential smoothing assigns low weight to most recent observation and high weight to least recent observation.
- Moving Average and Exponential Smoothing are short-term forecasting techniques.
- All of the above

ARIMA

- ARIMA models are the most commonly used models in time series modelling.

- Here, AR stands for Auto Regression
  I stands for Integration
  MA stands for Moving average

- So ARIMA is a combination of simpler models (Auto Regression and Moving Average) to make a complex model.

- In order to use ARIMA model we first need to ensure that our time series is a stationary series.

Dickey-Fuller test for Stationarity Check

- It is a hypothesis testing based approach to check for stationarity of a series.

- Here, the null and alternate hypothesis are defined as -
  - Null Hypothesis (H0) : Series is non-stationary.
  - Alternate Hypothesis (H1) : Series is stationary.

p-value ≤ 0.05	Reject the null hypothesis, means the series is assumed to be stationary.
p-value > 0.05	Failed to reject the null hypothesis, means the series is assumed to be non-stationary.

Dealing with Non-Stationary data

- A non-stationarity series can be made stationary by differencing (or using other methods like log).

- A series which becomes stationary after being differenced by **1**, is denoted as **I(1)** i.e. Integration of order **1**.

- So in general, a series which becomes stationary after being differenced **d** times, is denoted as **I(d)** i.e. Integrated of order **d**.

NOTE - The term **I** in ARIMA signifies this above discussed Integration order.

Autoregressive Model

- As the name suggests, this is basically a regression of the time series onto itself.
- It is a linear regression model that can be used to predict the future values in a series based on previous values.

Mathematically,

AR(p) : $$y_{(t)} = \beta_0 + \beta_1 y_{(t-1)} + \beta_2 y_{(t-2)} + \ldots\ldots\ldots\ldots + \beta_p y_{(t-p)}$$

Here, **p** denotes the number of previous observations to be considered. Also, represents the maximum lag to be considered.

For **p** = **1** and **3** respectively -

AR(1) : $$y_{(t)} = \beta_0 + \beta_1 y_{(t-1)}$$

AR(3) : $$y_{(t)} = \beta_0 + \beta_1 y_{(t-1)} + \beta_2 y_{(t-2)} + \beta_3 y_{(t-3)}$$

Moving Average

- Rather than using past values of the time series in a autoregression, a moving average model uses past errors to forecast future values.

Mathematically,

MA(q) : $y_{(t)} = \beta_0 + \alpha_0 \epsilon_{(t)} + \alpha_1 \epsilon_{(t-1)} + \ldots\ldots\ldots\ldots\ldots + \alpha_p \epsilon_{(t-p)}$

Here, **q** denotes order of the moving average model i.e. how many previous errors to be considered.

For **q** = **1** and **3** respectively -

MA(1) : $y_{(t)} = \beta_0 + \alpha_0 \epsilon_{(t)} + \alpha_1 \epsilon_{(t-1)}$

MA(3) : $y_{(t)} = \beta_0 + \alpha_0 \epsilon_{(t)} + \alpha_1 \epsilon_{(t-1)} + \alpha_2 \epsilon_{(t-2)} + \alpha_3 \epsilon_{(t-3)}$

Autocorrelation and Partial Autocorrelation Function

- To estimate **p** and **q** parameter of ARIMA, autocorrelation function (ACF) and partial autocorrelation function (PACF) is used.
  - PACF is used to estimate **p** parameter (order of auto regression)
  - ACF is used to estimate **q** parameter (order of moving average)

- ACF is the correlation between the current observation (y(k)) and the observation k periods away from the current one (y(t-k)).

- PACF is used to measure the degree of association between y(k) and y(t-k), when the effects at other time lags 1,2,3,....., (k-1) are removed.

ARIMA other variants

- Non-seasonal ARIMA (p, d, q)
  - p = Autoregressive order
  - d = Integration order
  - q = Moving Average order

- Seasonal ARIMA (p, d, q) x (P, D, Q) S
  - P = number of seasonal autoregressive (SAR) terms
  - D = number of seasonal differences
  - Q = number of seasonal moving average (SAM) terms
  - S = Seasonality period

Quiz 5

Which is most appropriate to non-seasonal time series data?

- ARIMA (p, d, q) x (P, D, Q) S
- ARIMA (p, d, q)
- ARMA (p, d, q)
- ARIMA (P, D, Q)