ISM 6136 – DATA MINING/PREDICTIVE ANALYTICS

Presenter: Dr. B. Sharma

LECTURE 12

FORECASTING TIME SERIES

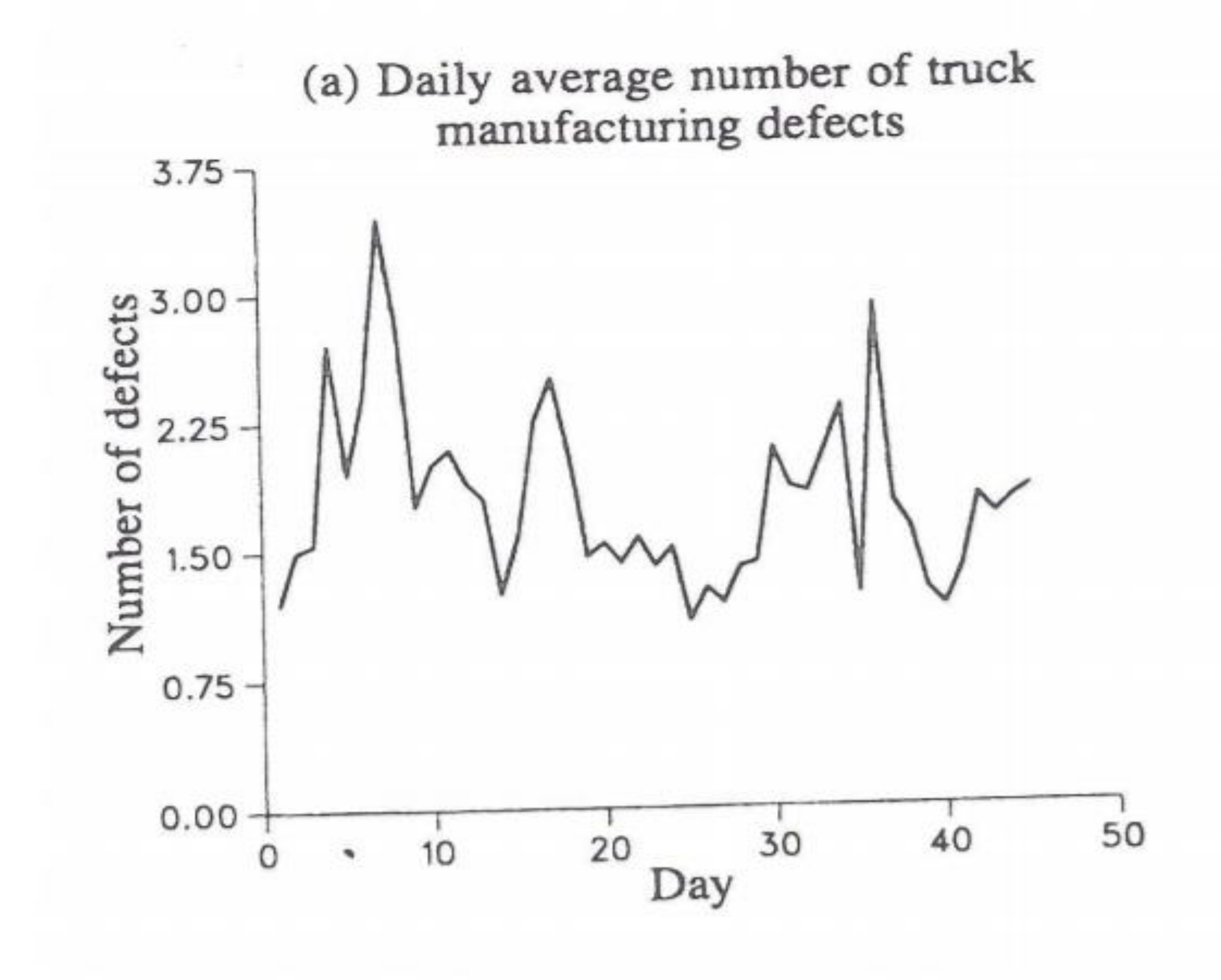
LEARNING OBJECTIVES

- **Time series concept**
- **Time series forecasting**
- **Autocorrelation concept**
- **Using XLMiner for forecasting**

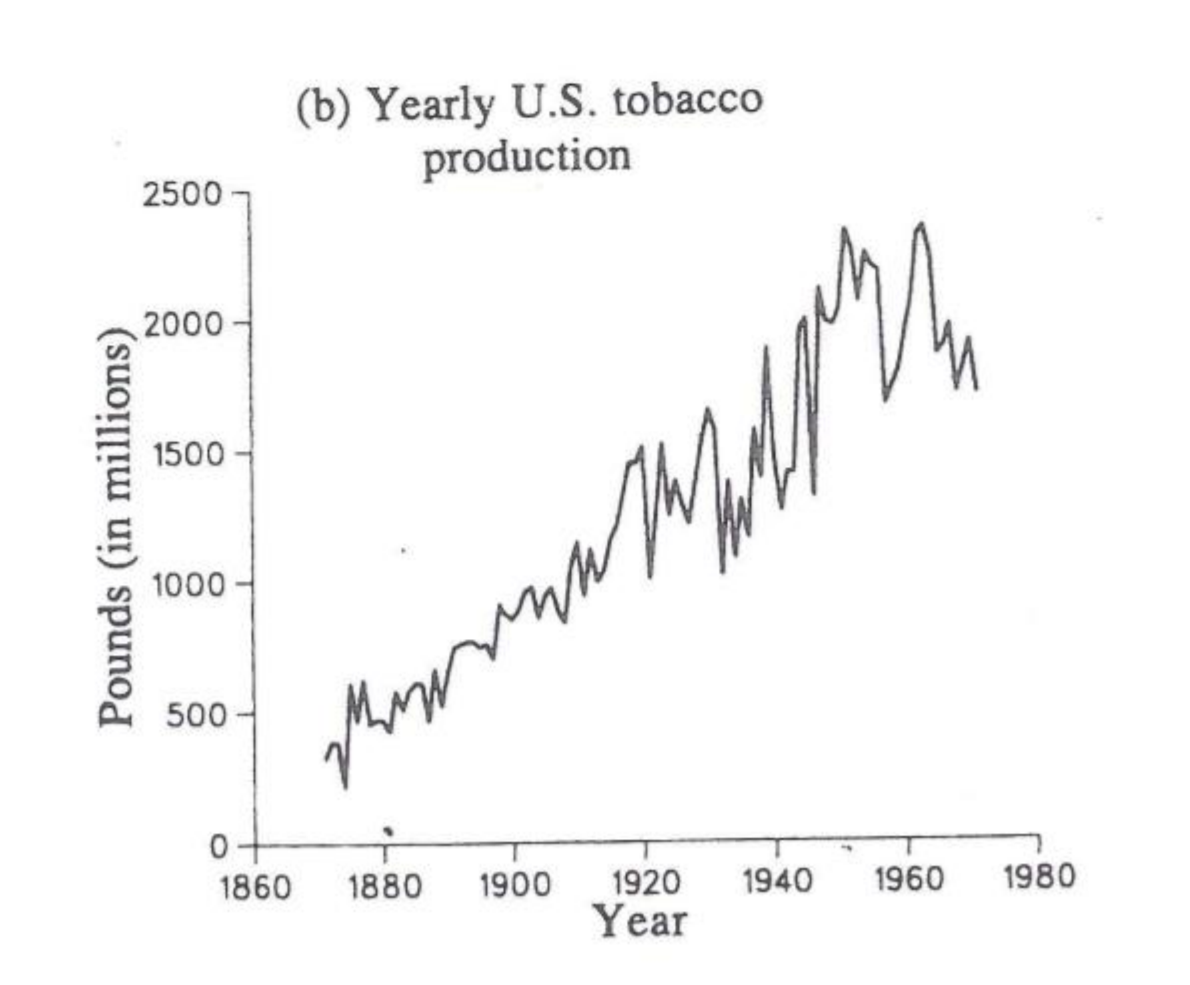
TIME SERIES CONCEPT

- Quantifiable data → Time series forecasting performed
- Forecasting sales, production, demand, prices, enrollment, inflation etc
- So far dealt with cross-sectional data (sequence of measurements over time - does not matter)
- Continuous time series data – data recorded on frequent time scales at equal time interval
  - Stock data at ticker level
  - Purchases recorded in real time
  - Daily closing value of the Dow Jones Index
  - Annual flow volume of the River Nile
  - Precipitation in a specific location
  - Size of an organism, measured daily
  - Annual U.S. population data
  - Other examples ?

TIME SERIES CONCEPT

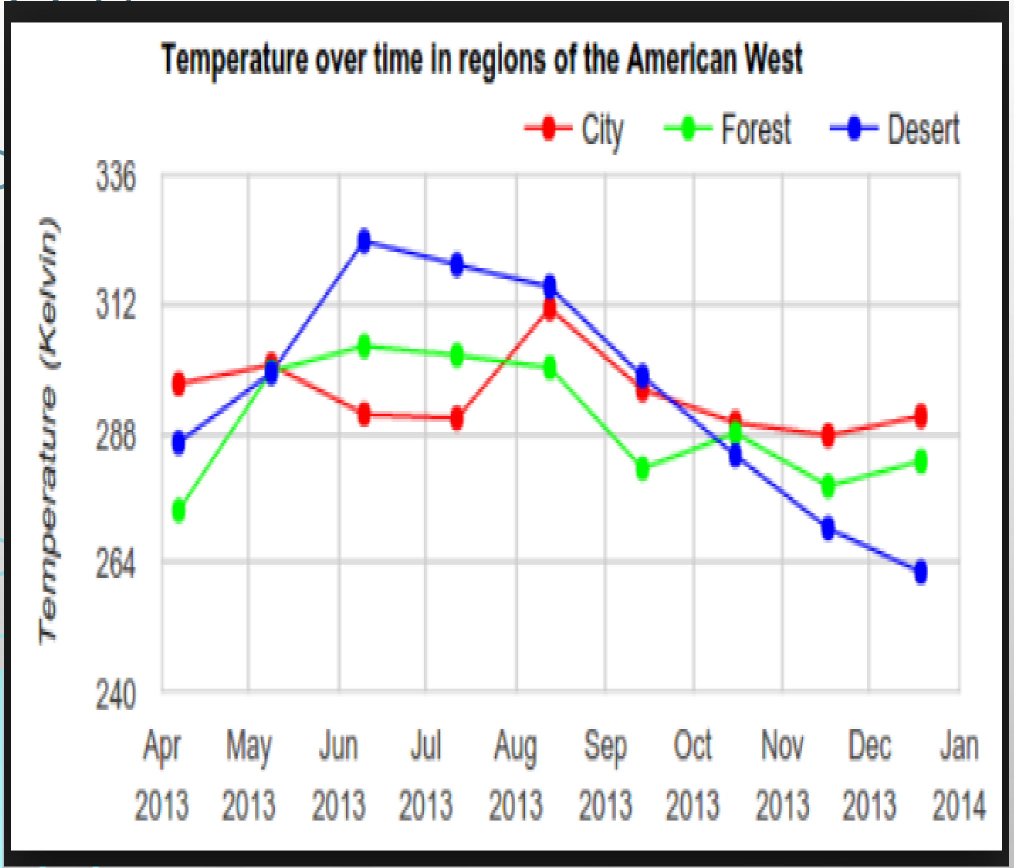


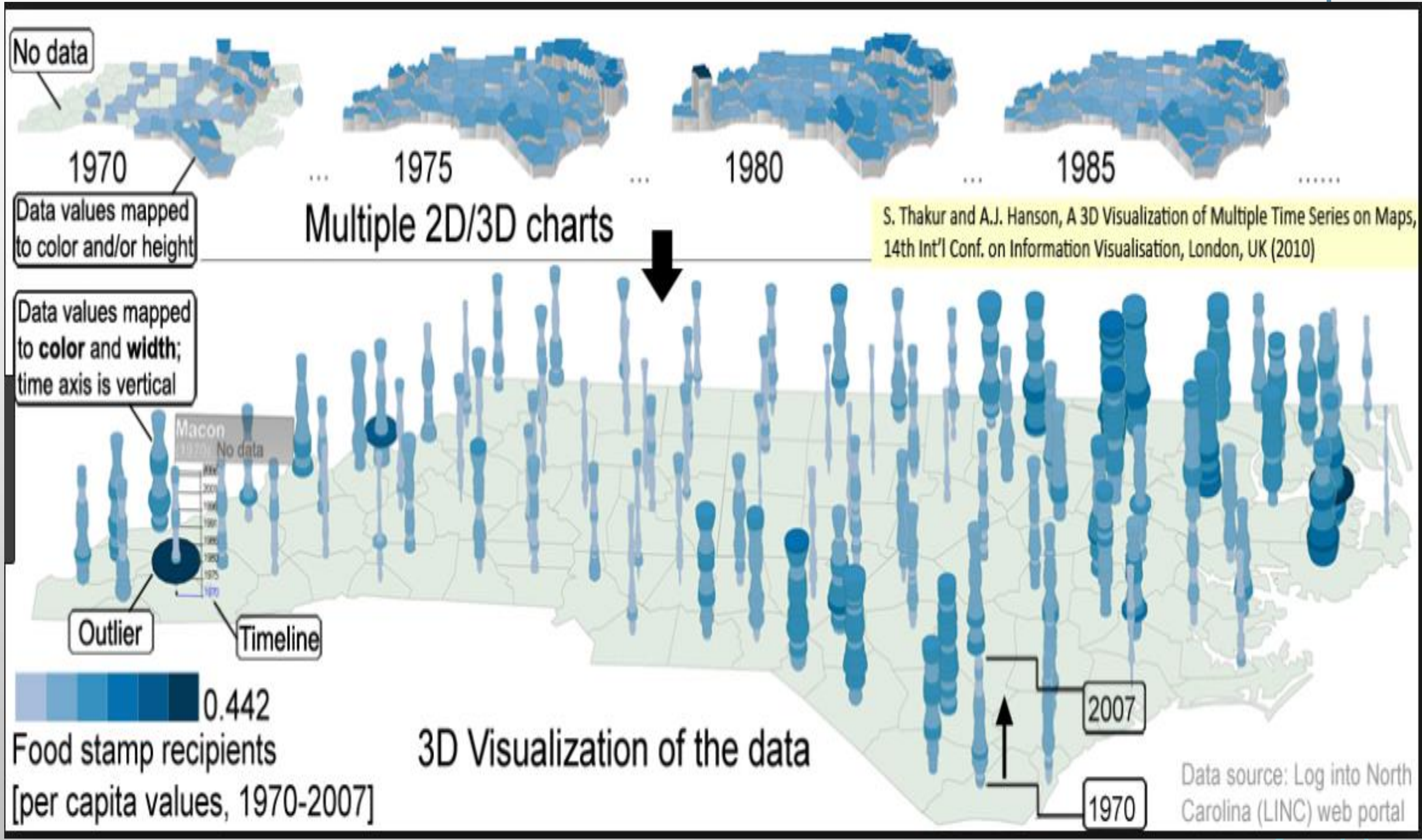
Stationary trend



Non-Stationary trend

MULTIPLE TIME SERIES





TIME SERIES FORECASTING

- Appropriate Time scale for time series and the level of noise in the data must be considered
- Time series forecasting – Using the information on time series to forecast future values of that series
- Dissect time series into 4 components
  - **Level** – Average value of series
  - **Trend** – Change in the series from one period to another
  - **Seasonality** – Short-term cyclical behavior of the series can be observed several times within the given series
  - **Noise** – Random variation that results from measurement error or other causes
- Examine a time plot – to identify the components
- XLminer (ASP) helps in evaluating the predictability of a series and improving forecast precision

Concepts based on 'Data Mining for Business Analytics' by Shmueli, Bruce, Patel

PARTITIONING IN TIME SERIES FORECASTING

Data Partitioning

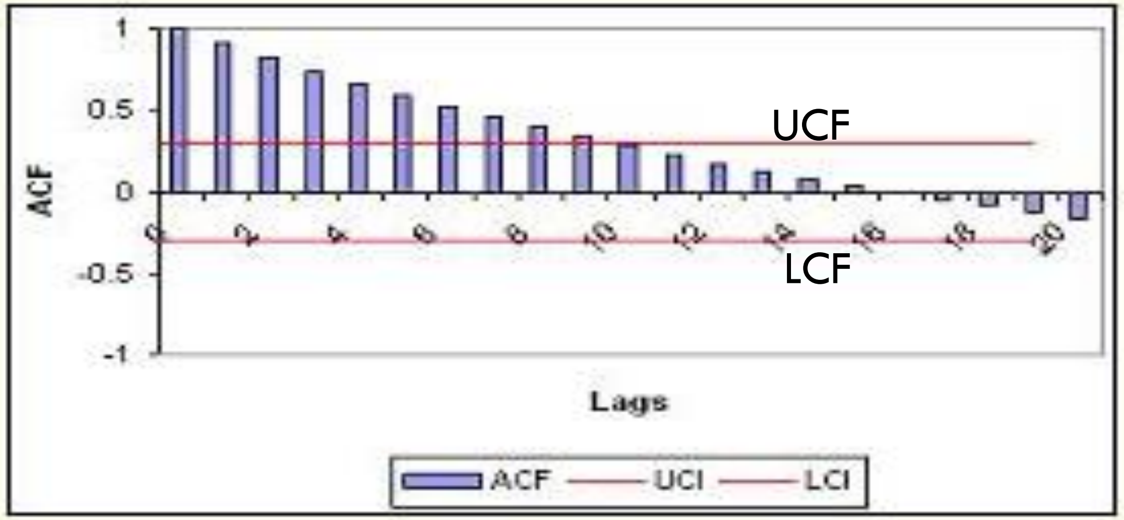
- Not done randomly as will create two time series with ‘holes’ (missing data)
- Training set (earlier period records)
- Validation set (later period records)
- Forecasting algorithm is trained on training data and their predictive performance accessed on validation data
- Validation set contains most recent period, closest in time to the forecast period
- If only the training set is used for forecasting it will require forecasting farther into future

LAG ANALYSIS

AUTOCORRELATION CONCEPT

- **Correlation between values within neighboring periods**
- **Times series observations in neighboring periods tend to be correlated**
- **Correlation info helps in improving forecasts -** If we know a high forecast value tends to be followed by high values then we can use that to adjust forecasts.
- Autocorrelation function computation
  - Compute correlation between series and a lagged version of series
  - Lag -1 –> original series moved forward by 1 time period, Lag – 2 – move forward by 2 time periods… etc
  - Xlminer's ACF (Autocorrelation function) – computes autocorrelation of a series at different lags
  - Upper confidence level (UCL) and the Lower confidence level (LCL). If the data is random and less correlated, then the plot should be within the UCL and LCL. You set the confidence % in Xlminer.
  - If the plot exceeds either of these two levels, as see in this plot - some correlation exists in the data.

Day	Observed Value	Lag-1	Lag-2
1	10		
2	20	10	
3	30	20	10
4	40	30	20
5	50	40	30



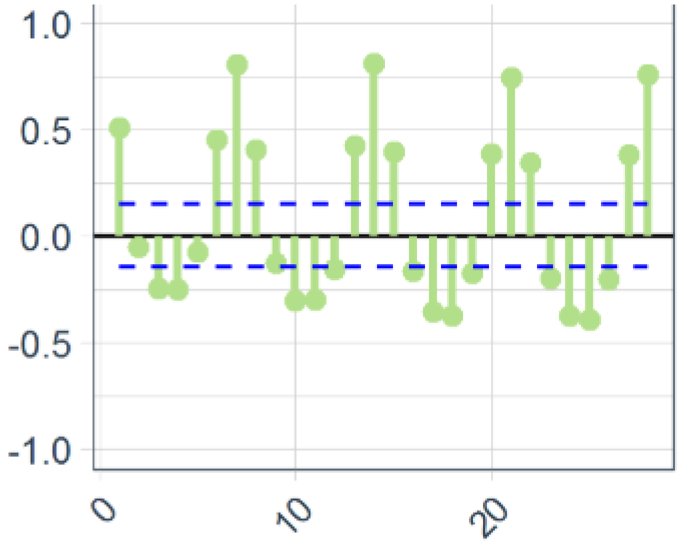
AUTOCORRELATION CONCEPT

- **Partial Autocorrelation Function (PACF)** - Computes and plots the partial autocorrelations between the original series and the lags. Eliminates all linear dependence in the time series beyond the specified lag.
- **ARIMA model (Autoregressive integrated moving-average)**
  - ❖ Regression-type model that includes autocorrelation.
  - ❖ The quality of the model evaluated by comparing the time plot of the actual values with the forecasted values - If both curves are close - model is a good fit.
  - ❖ The model should expose any trends and seasonality, if any exist. If the residuals are random then the model can be assumed a good fit.
  - ❖ However, if the residuals exhibit a trend, then the model should be refined.

AUTOCORRELATION CONCEPT

- Positive autocorrelation –> consecutive data values move generally in same direction
- Negative autocorrelation –> consecutive data values move generally in opposite direction (high values immediately followed by low values and vice versa)
- Strong autocorrelation at lag k larger than 1 and all its multiples typically reflects an annual seasonality

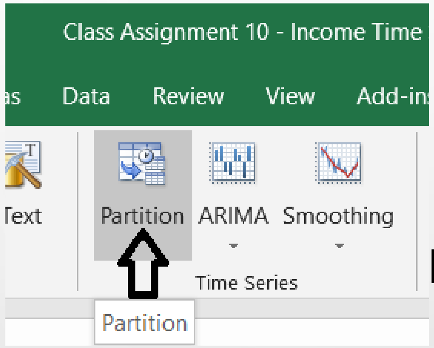
Positive and Negative autocorrelations

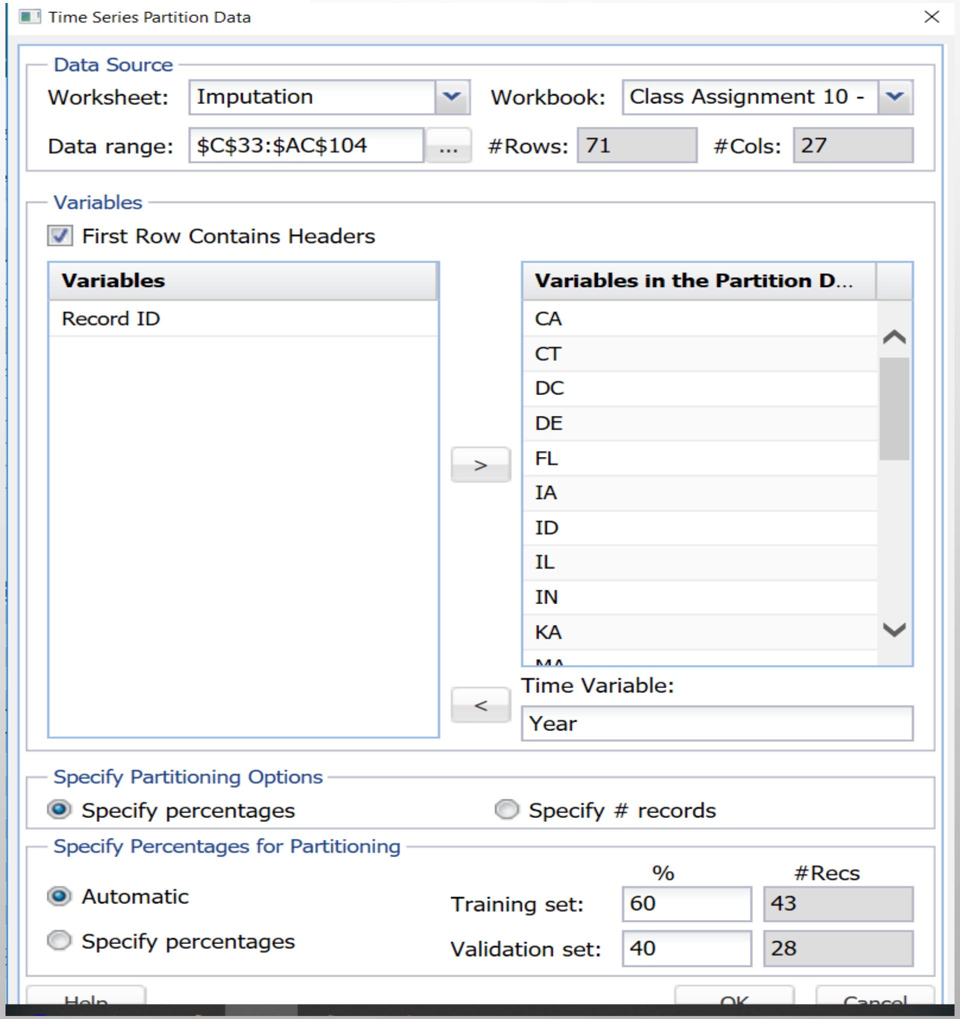


Positive autocorrelations

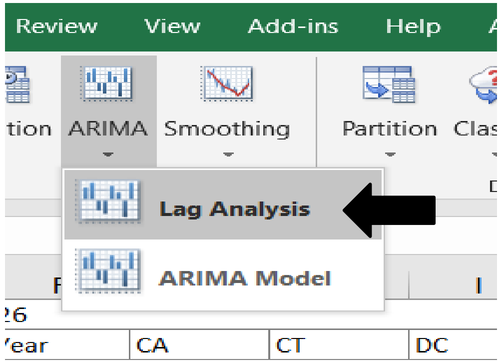
TIMESERIES FORECASTING STEPS

1. Obtain the dataset
2. Perform 'Missing Data Handling'
3. Perform Time Series Partitioning

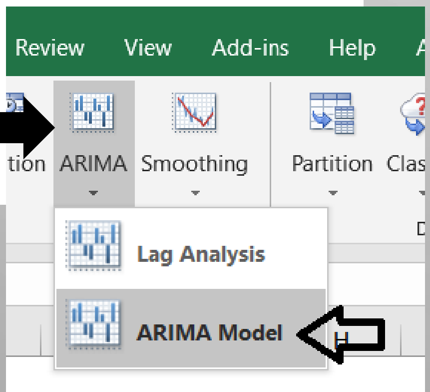




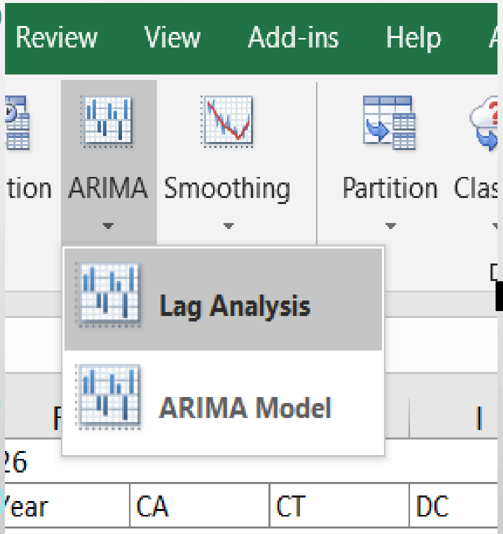
4. Perform Lag Analysis

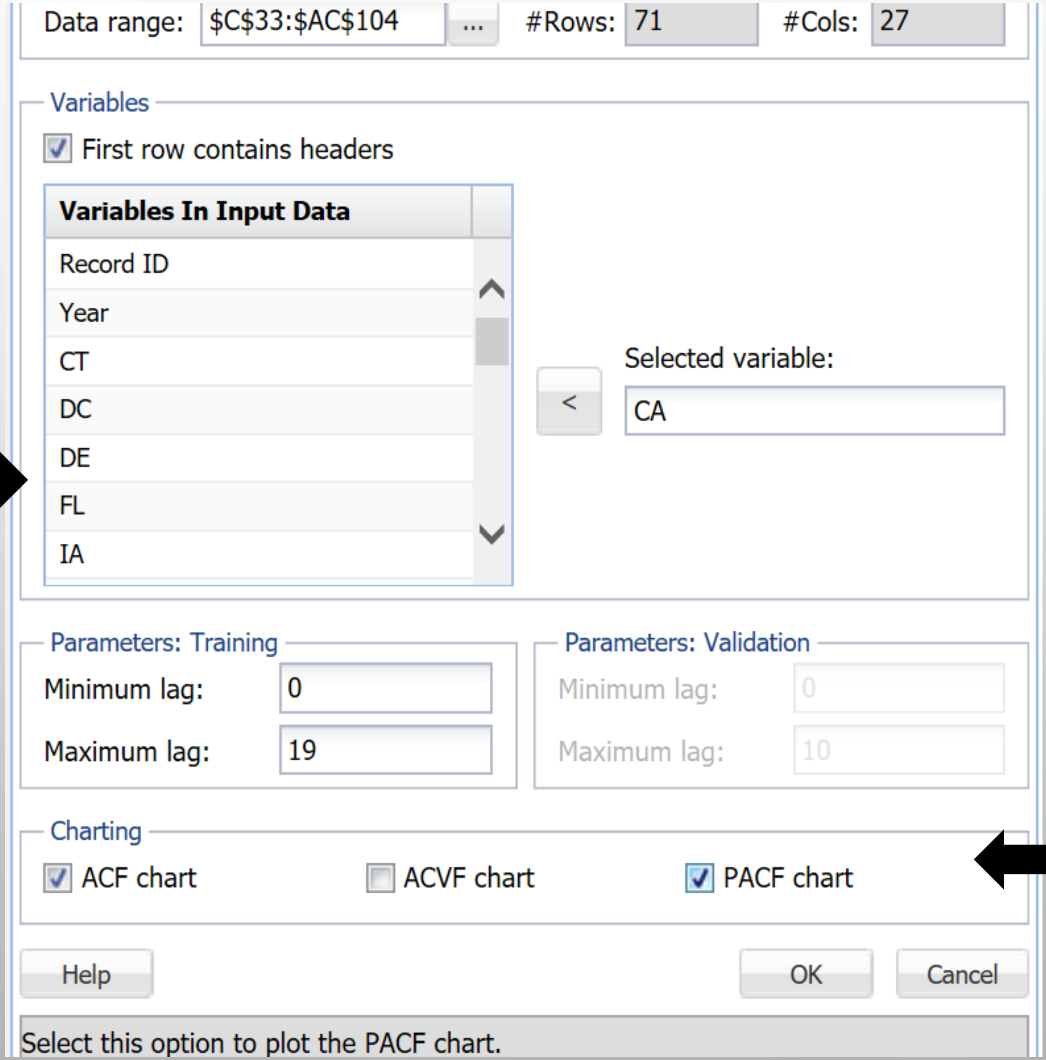


5. Back to Partitioned tab - Apply the ARIMA model



LAG ANALYSIS





Lag Analysis - Shows the autocorrelation plots

Check for:

- **Trend** – Change in the series from one period to another
- **Seasonality** – Short-term cyclical behavior of the series can be observed several times within the given series

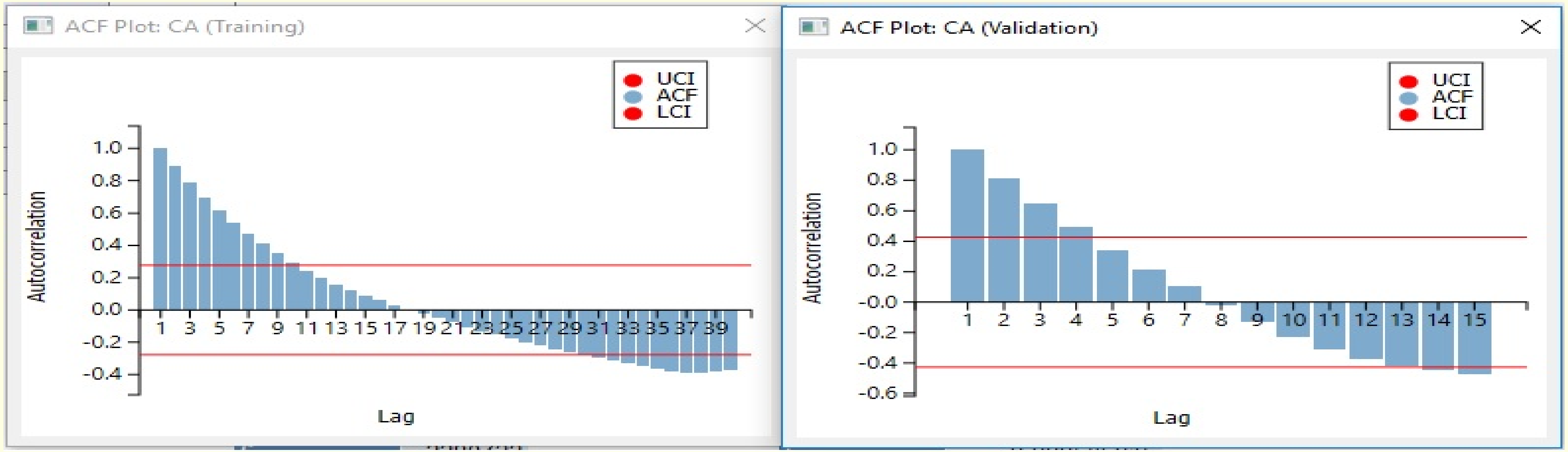
Check the ACF and PACF plots

ACVF – Autocovariance plot - optional

LAG ANALYSIS RESULTS

Analyze the Autocorrelation plots (ACF)

Click **OK**. *TS_Lags* is inserted into the task pane under Reports -- Autocorrelations.

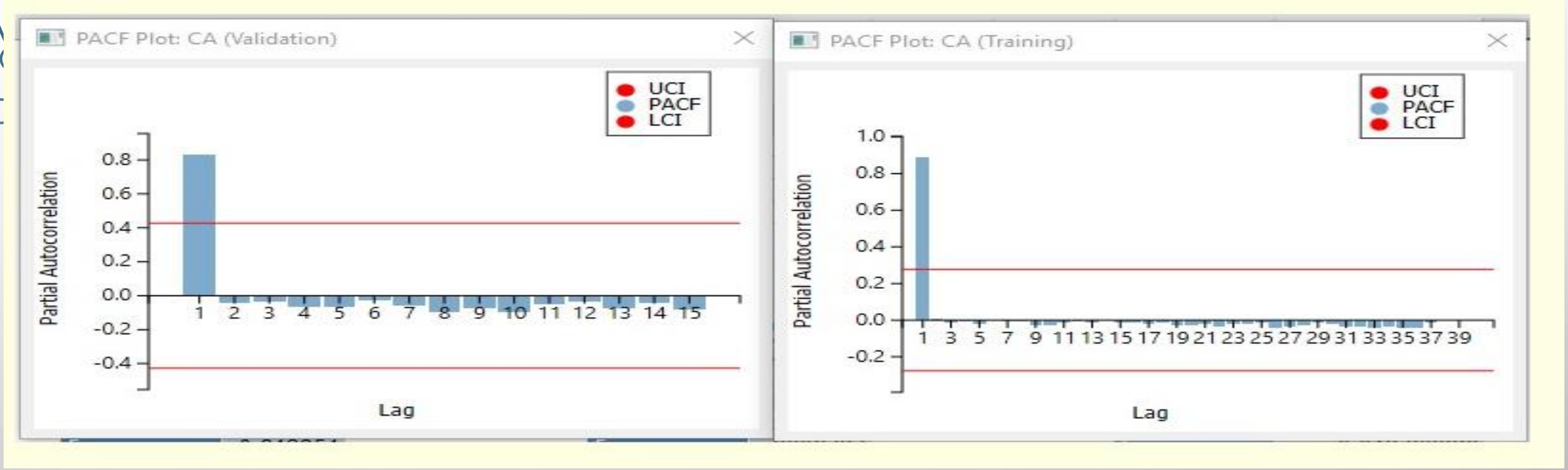


Note on each chart, the autocorrelation decreases as the number of lags increase. This suggests that **a definite pattern does exist in each partition.**

However, since the pattern does not repeat, it can be assumed that **no seasonality** is included in the data. In addition, **both charts appear to exhibit a similar pattern.**

LAG ANALYSIS RESULTS

Analyze the Autocorrelation plots (PACF)

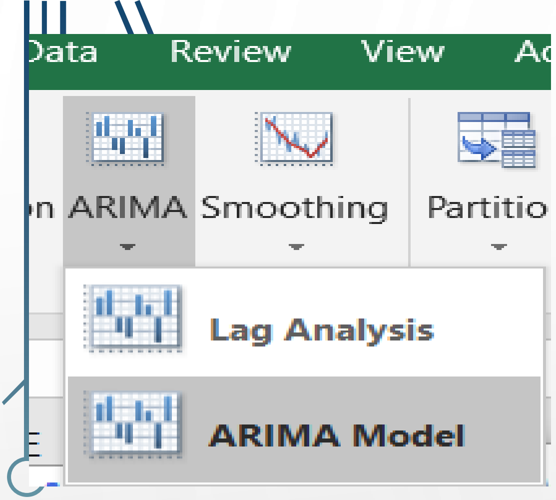


The PACF plots show a definite pattern which means there is a trend in the data.

However, since the pattern does not repeat, we can conclude that the data does not show any seasonality.

Both datasets exhibit the same behavior in both the training and validation sets which suggests that the same model could be appropriate for each.
Now we are ready to fit the model.

FITTING THE MODEL USING ARIMA

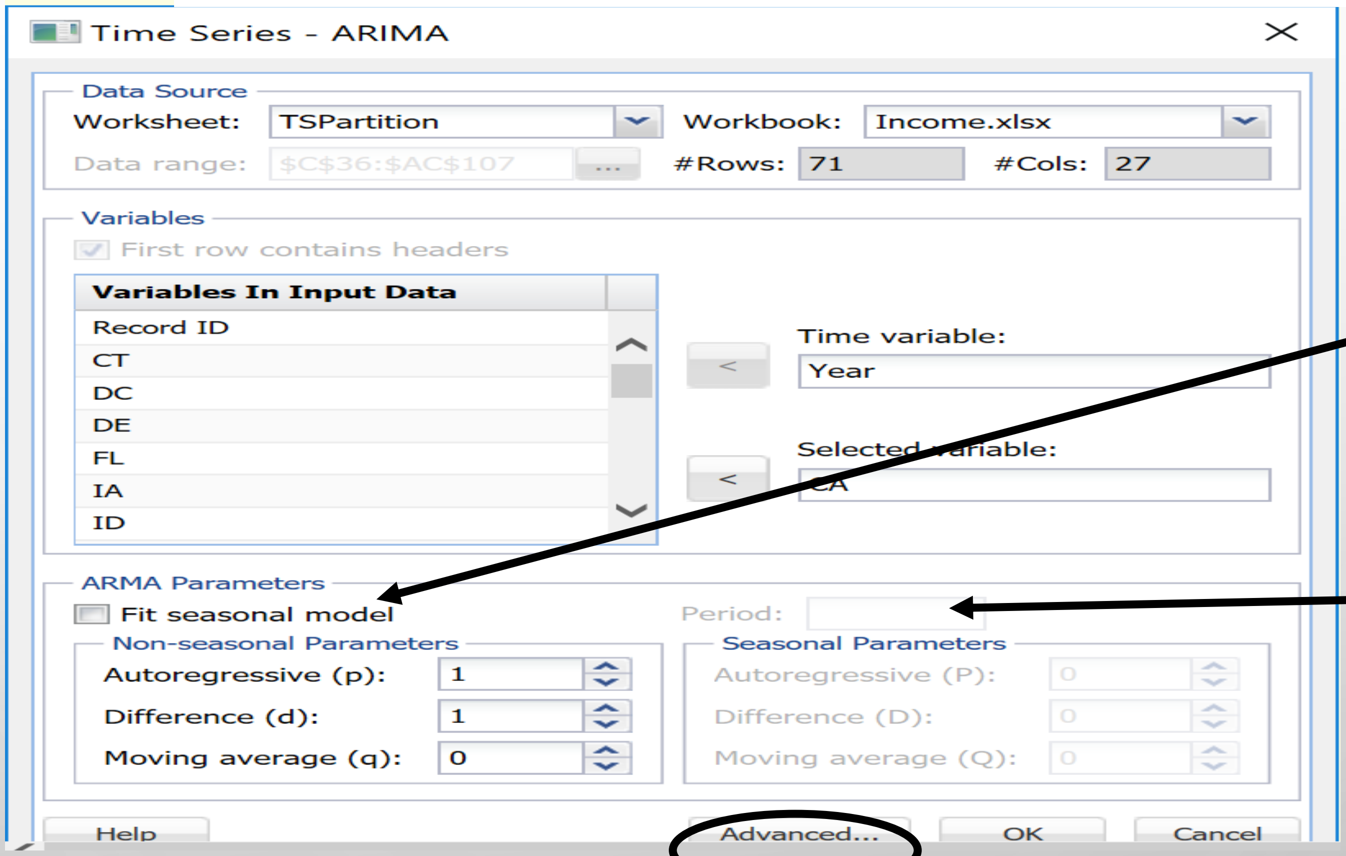


The ARIMA model accepts three parameters:

p - the number of autoregressive terms,

d - the number of non-seasonal differences,

q - the number of lagged errors (moving averages).



Fit seasonal model Select this option only for a seasonal model. The seasonal parameters are enabled when this option is selected.

Period If Fit seasonal model is selected, this option is enabled. Seasonality in a dataset appears as patterns at specific periods in the

From Lag Analysis –

- The PACF plot displayed a large value for the first lag but minimal plots for successive lags. This suggest setting p =1 since seems like there is one auto-regressive term.
- With most datasets, setting d =1 is sufficient or can at least be a starting point.
- ACF plot showed no seasonality in the data which means that autocorrelation is almost static - decreasing with the number of lags increasing. This suggests setting q = 0 since there appears to be no lagged errors.

ARIMA RESULTS

ARIMA - Advanced Options

Maximum number of iterations: 200

Output

- [x] Fitted values and residuals
- [x] Variance-covariance matrix

Forecast

- [x] Produce forecasts

Number of forecasts: 21

Confidence level for forecast confidence intervals: 95

Help | OK | Cancel

ARIMA Model

Record ID	Coeff	Std-Dev	p-value
Const	-16.20163	3.053324388	1.119E-07
AR 1	1.0920227	0.054394554	1.198E-89

Mean	176.06122
-2LogL	592.00842
Res. StdDev	103.97445
#Iterations	7

Good model should have low p – values close to zero

Error Measures: Training

Record ID	Value
SSE	284182.9
MSE	6608.905
MAPE	4.733355
MAD	66.12295
CFE	91.65556
MFE	2.131525
TSE	1.386138

Error Measures: Validation

Record ID	Value
SSE	4.03E+09
MSE	1.44E+08
MAPE	52.03059
MAD	10139.24
CFE	283898.6
MFE	10139.24
TSE	28

Good model should have low MSE values

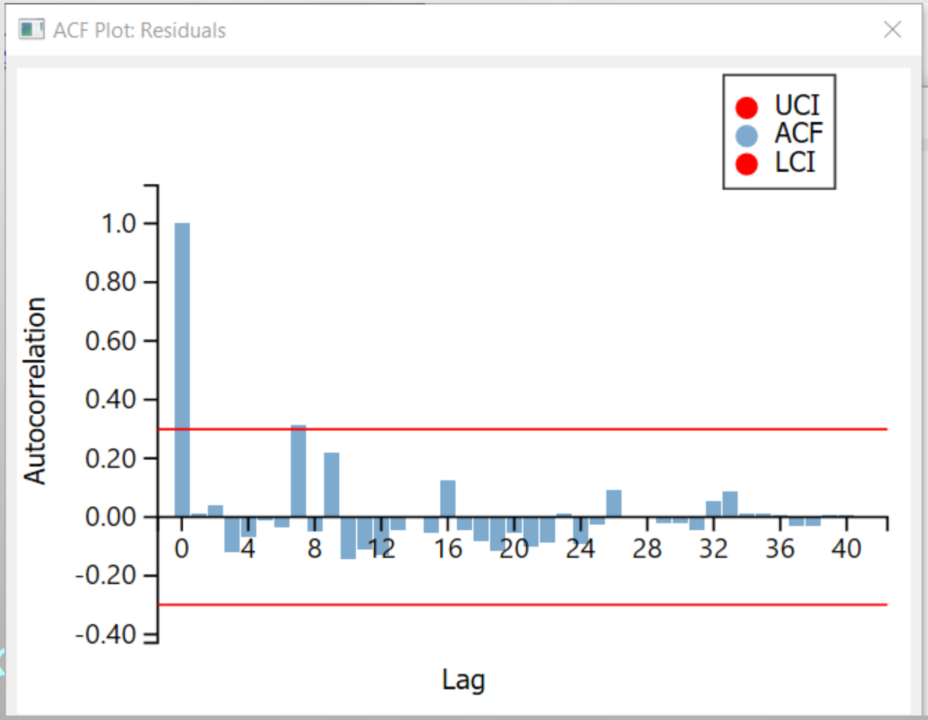
Model Selection Criteria

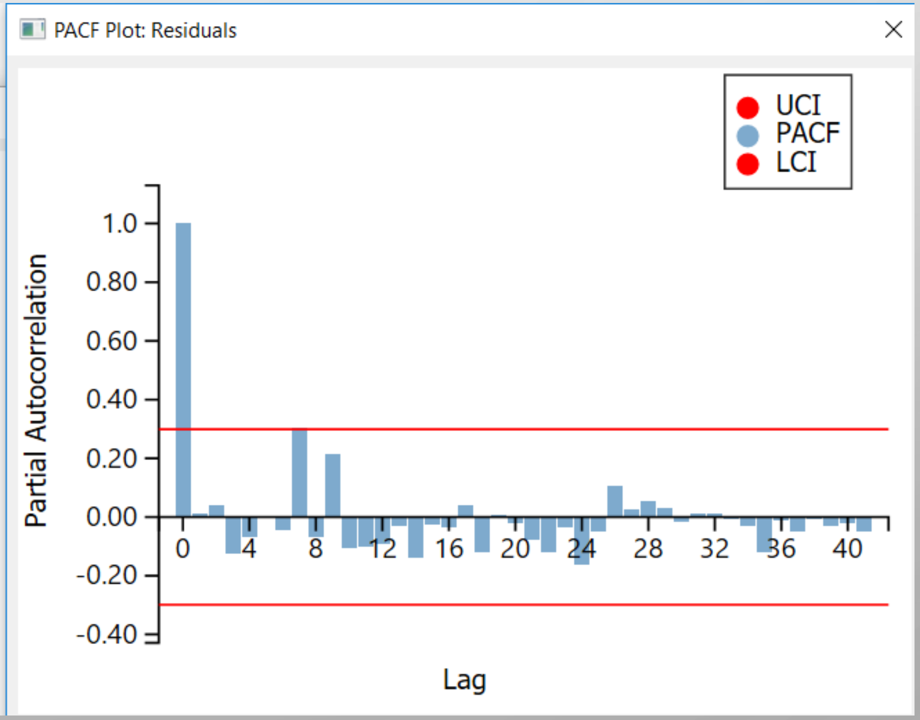
1. Low p values – close to zero
2. Low MSE values

ARIMA RESULTS

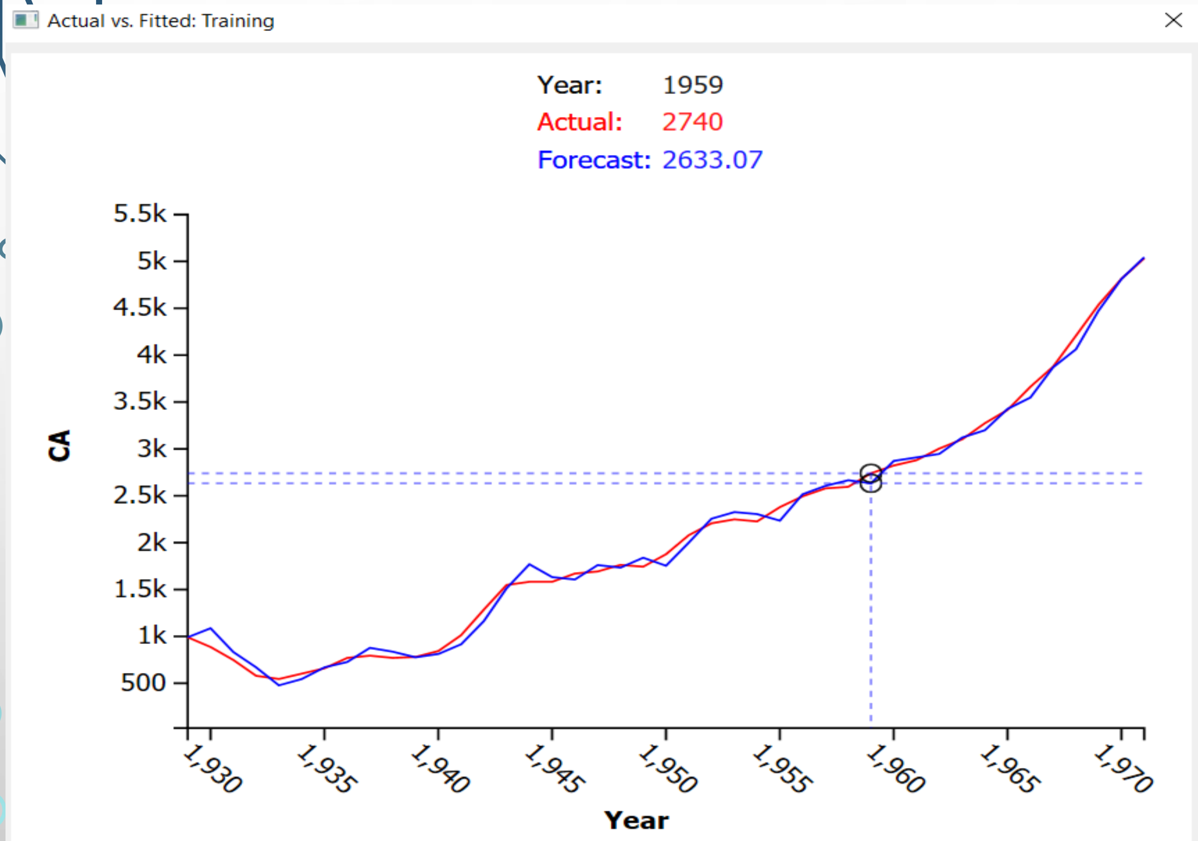
Examine the ACF and PACF residuals plots created by ARIMA

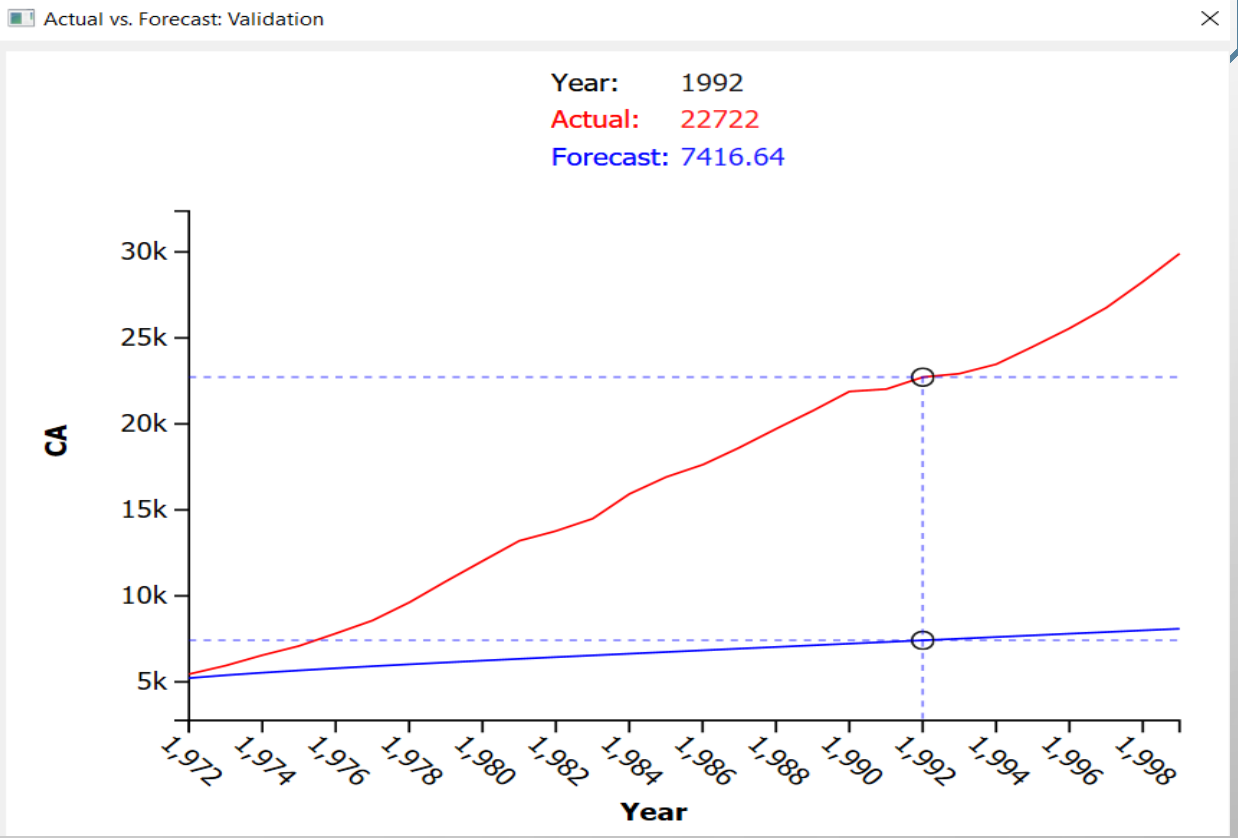
- All lags, except lag 1, are clearly within the UCL and LCL bands. **This indicates that the residuals are random and are not correlated, which is the first indication that the model parameters are adequate for this data.**





FINAL FORECAST RESULTS





There is a 95% chance (confidence level we set up in AIRMA model settings) that the forecasted value will fall into this range

Change the ARIMA model parameters p, q, d, number of iterations to get the most accurate model for a particular partitioning

So to forecast for future years ….you can use the best model and apply it to the future years being in the validation dataset with some 'estimated' actual values.

END