**Section A**

**1**. Seasonal Autoregressive Integrated Moving Average (SARIMA)

**2**. SARIMA is a different variation of the ARIMA time series forecasting method. The method deals with univariate autoregressive and moving average elements. SARIMA differs from ARIMA as it can handle time series data with repeating cycles, where the mean of observations within a dataset is not constant, changing on a cyclical basis. The SIRIMA model is used for modeling seasonal time series where the mean and variance of a given season is not fixed across the years. Before we go any further, let us talk about what is a time series? Times series are values that are recorded at a fixed interval which can be seconds, minutes, hourly, daily, weekly, monthly, and so (Analytics Vidhya) [i]. SARIMA model is used for predicting the sales of a products, it can be also used for estimating the electricity of households, to predict the number of disease incidence, or traffic predictionary model is very successful in analyzing and forecasting times series data with seasonal components (Antonanzas, Javier, et al) [ii].

**3.** There are multiple ways to express the formula for SARIMA, one being the short-handed notation: ARIMA(p,d,q) X (P,D,Q)

The lower case p,d,q variables represent non-seasonal autoregressive order, differencing, and moving average order, respectively. The upper case P, D,Q variables represent carry the same meaning as their lower case counterparts, with seasonal effects. The S represents the time span of each repeating seasonal pattern.

Φ(*BS*)φ(*B*)(*xt* - μ) = Θ(*BS*)θ(*B*)*wt*

This is the formal way of writing the SARIMA formula, where the autoregressive order is represented by AR:  φ(*B*) = 1 - φ*1B - ... -* φ*pBp.* Moving average order is represented by MA:  θ(*B*) = 1 + θ*1B + ... +* θ*qBq*. Seasonal autoregressive order is represented by Seasonal AR:  Φ(*BS*) = 1 - Φ*1BS - ... -* Φ*PBPS*. And Seasonal moving averages are represented by Seasonal MA:  Θ(*BS*) = 1 + Θ*1BS + ... +* Θ*QBQS*.

Kostas Hatalis states in datasciencecentral.com website that the SARIMA models are “denoted  as SARIMA (p, d, q) (P, D, Q) [S]”, “where S is the number of periods for each season, d is the degree of differencing (the number of times the data have had past values subtracted), and P, D, and Q refers to the autoregressive, differencing and moving average terms for the seasonal”[iii]. (p, d, q) and (P, D, Q) are non-seasonal and seasonal respectively.

**4a.** Numerical, and time series value

**4b.**  There is no minimum number of observations requires because the more we have data, the better the forecasting (Rob J. Hyndman and Andrey V. Kostenko)[iv]

**4c**. SARIMA does not work well if the series are not stationary.

SARIMA models apply to times series and numerical data

- Data must be seasonal in nature, discovered by plotting it over a time series.

- Differencing might need to be performed on the data set to make it stationary.

**5a** the algorithm will not work effectively if the data is non-stationary. Without stationary data, it becomes increasingly more difficult to predict accurate results. Seasonal differencing will help stabilize the mean and remove stochastic trends. There is also a chance of overfitting the model around the time periods for a seasonal cycle. There is also a chance that seasons (lags) within the data might be correlated with one another.

SARIMA model works well when the series are stationary. Otherwise, it will get biased coefficients. SARIMA model has two different seasonally: Additive and Multiplicative. The SARIMA model is based on the Box and Jenkins procedure and has three steps: Identification, estimation and diagnostic checking (Fredrik Nikolaisen Sävås)[v].

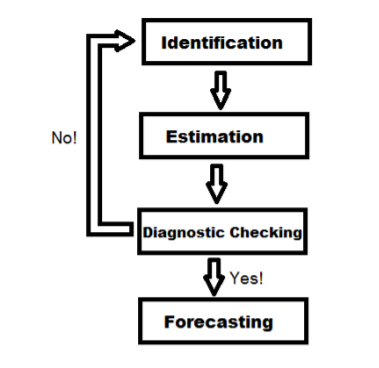


Figure 1: The Box and Jenkins Procedure

Identification refers to the finding of integration orders d and D, the autoregressive orders p and P, and the moving average orders q and Q. It can also be done by finding sample autocorrelation (SAC) and sample partial autocorrelation (SAPC)

Estimation helps to find the best model based on the identification.

Diagnostic Checking

Forecasting

**5b**. When working with time series, in most cases, we assume that the series are stationary which mean that the mean and the variance do not vary over time.

There is another case also where there is a trend and seasonal over time which we call non-stationarity. Meaning that there is a variation of mean at specific times frames.

**6a**

Pandas, numpy, statsmodels.tsa.statespace.sarimax import SARIMAX, statsmodels.tsa.stattools import acf,pcf, sklearn.model\_selection import train\_test\_split

import pandas as pd

import numpy as np

from scipy.stats import norm

import statsmodels.api as sm

import matplotlib.pylab as plt

%matplotlib inline

from matplotlib.pylab import rcParams

from statsmodels.tsa.statespace.sarimax import SARIMAX

**6b.**

model = SARIMAX()

model.fit()

model\_fit.forecast()

model\_fit.predict()

Find the trend and seasonality

Find auto correlation function (ACF) and partial auto correlation function (PACF)

Find coefficient of SARIMAX models based on ACF and PACF

Split the data into training and testing

Create the model

model=SARIMAX(training,order=(4,1,4),seasonal\_order=(1,0,0,12),

enforce\_stationarity=False,enforce\_invertibility=False)

Fit the model

model\_fit =model.fit(disp=False)

forecast and plot result.

R =len(testing)

forecast=model\_fit.forecast(Rforecast=np.exp(forecast)

The data must first be evaluated to determine if the data is seasonal and has a trend. Auto correlation and partial autocorrelation functions must be applied to determine moving averages and auto regressive terms. Once the moving averages, auto-regressive terms, and differencing, and seasonality have been discovered, values are put into SARIMA model. The fitted model is evaluated by …

**7a.**Plot the forecast result

Find the likelihood

Calculate the RMSE and less value prove best fitting

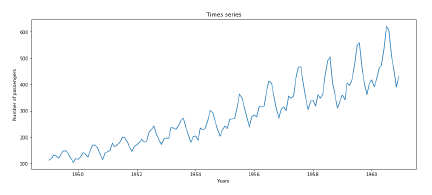
**7b**. Call the summary method on model\_fit.summary()

**SECTION B**

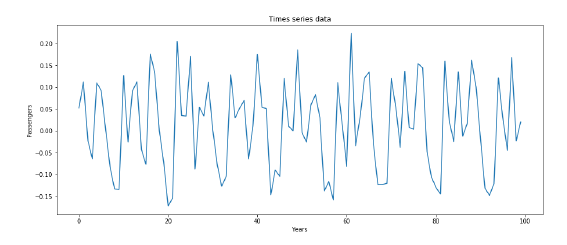
**8.** The dataset is of the number of air passengers that flew for each month over the course of a few years. There are only two variables within this data set, the month and #passengers variables. The variables for month consists of numerical continuous variables, and the number of passengers are discrete numerical variables. The data set was obtained from the following website:

<https://github.com/AileenNielsen/TimeSeriesAnalysisWithPython/blob/master/data/AirPassengers.csv>

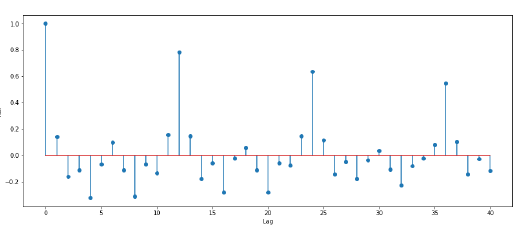
**9.** The algorithm could not be immediately applied to the dataset, as it did not contain stationary data. This was seen when the data was plotted using matplotlib, and where there was a clear pattern of multiplicity and a constant increasing trend. The data had to be differenced first. Using the ACF and PACF functions helped determine what values we had to input into the SARIMA model as well.



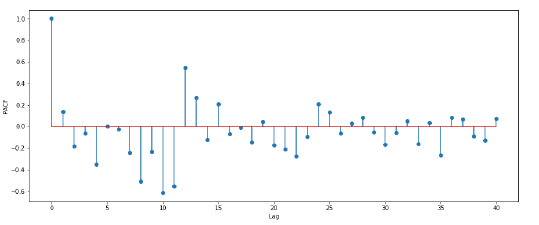
The first graph displays the number of passengers plotted over the period from 1949-1960, with time on the x axis and the number of passengers on the y axis.



The second graph displays the data after the trends and multiplicity has been removed.



Graph showing the amount of autocorrelation at across the number of lags in the time series.



Graph showing the level of partial auto correlation across the lags in the time series.

**References**

[i]<https://www.analyticsvidhya.com/blog/2018/08/auto-arima-time-series-modeling-python-r/>

[ii] Antonanzas, Javier, et al. "Review of photovoltaic power forecasting." *Solar Energy* 136 (2016): 78-111.

[iii]<https://www.datasciencecentral.com/profiles/blogs/tutorial-forecasting-with-seasonal-arima>

[iv]<http://www.bishophill.com/admin/sidebar_images/1741759940_test.pdf>

[v]<https://www.diva-portal.org/smash/get/diva2:631413/FULLTEXT01.pdf>