IOT Project 3: Time Series Forecating

Ojas Barve - ovbarve@ncsu.edu

Task 1:

This task involved the basic preliminary analysis of the data before applying any algorithmic techniques.

This involved the plotting of the time series data. The task was to inspect the data visually for any irregularities in the data which could lead us to believe that the time series is non stationary. Three things need to be checked while checking the stationarity of a time series namely;

1.) Trend 2.) Variability 3.) Seasonality

Figure 1. Below shows the visualisation of the time series data.

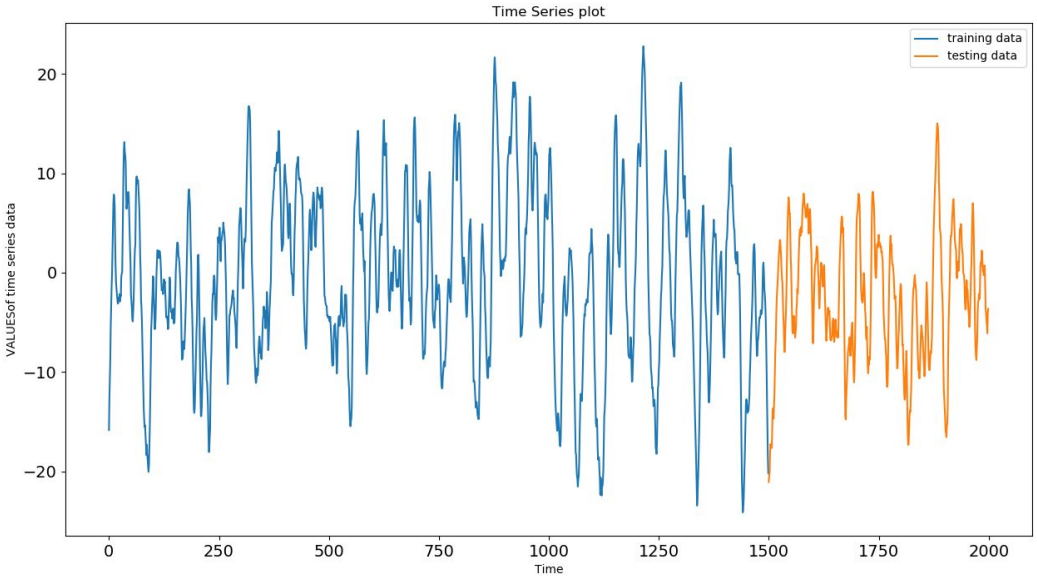

Figure 1. Plots of histogram of each independent variable followed the dependent variable on bottom right corner. Bin size is 30.

The data was divided into training and testing sets. 25/75 percentage split was carried

out on the dataset for testing and training.

Training = data[:1500]

Testing = data[1500:]

Comment On Task 1: As we can see from the plot of the time series we see no trends or any variability or any kind of seasonality in the time series. From preliminary visual inspection we see that the time series is stationary.

Also to be sure I carried out the Dickey Fuller test which puts forth a null hypothesis that the given time series is non - stationary. Below are the results of the Dickey Fuller test.

ADF Statistic:	-11.274280
p-value:	0.000000
Critical Values:	1%: -3.434 5%: -2.863 10%: -2.568

Here we see that the p - value obtaine from the Dickey Fuller test is 0.0. This means that we have to reject our null hypothesis in this case. Hence we accept the alternate hypothesis in this case which states that the time series is not non-stationary (Which means it is stationary).

Task 2 Simple Moving Average

In this case we are interested in fitting a simple moving average to our training data. The task is to calculate the value of window size for which we get the minimum Root Mean Squared Error. For this purpose I have taken a range of 2 to 50 values and then I plotted the values of RMSE v/s Window Size. Figure 2. Below shows the plot.

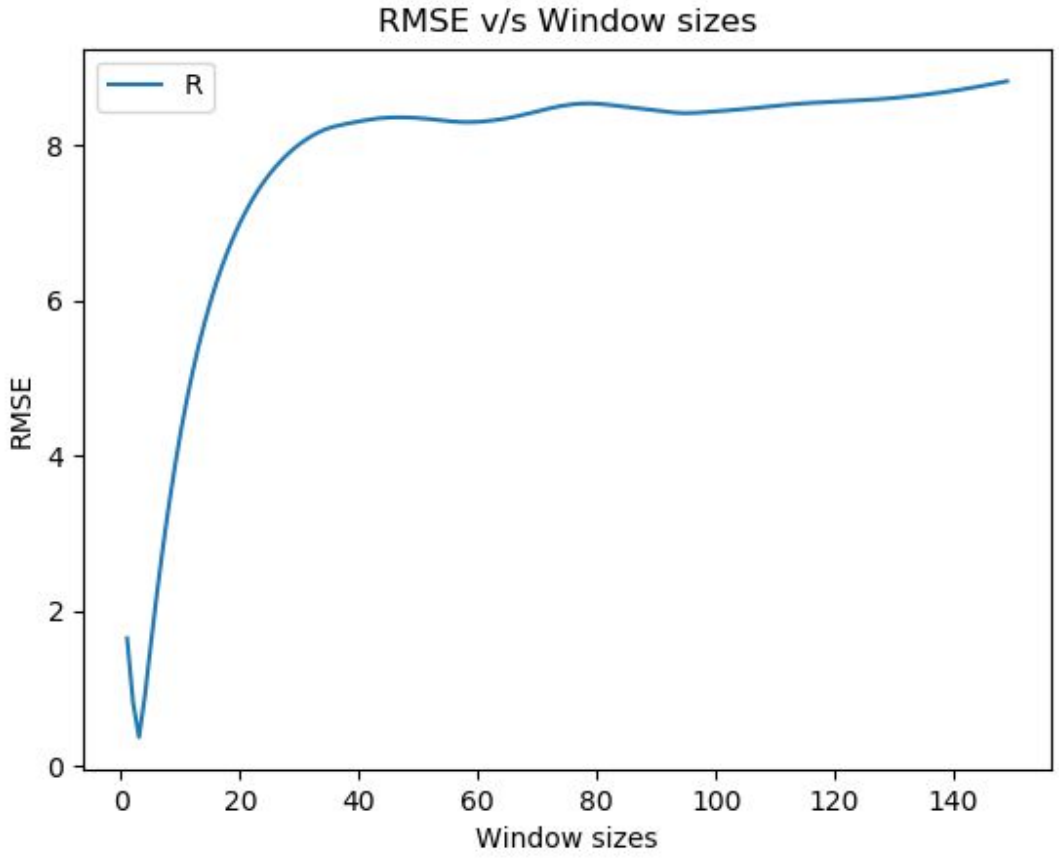

Figure 2: RMSE v/s Window Size

We see from the plot that the minimum value of RMSE occurs at a window size of 3.

Hence we plot the training data v/s Fit simple moving average model Below in Figure 3. You can see the plots of of training v/s Simple Moving average.

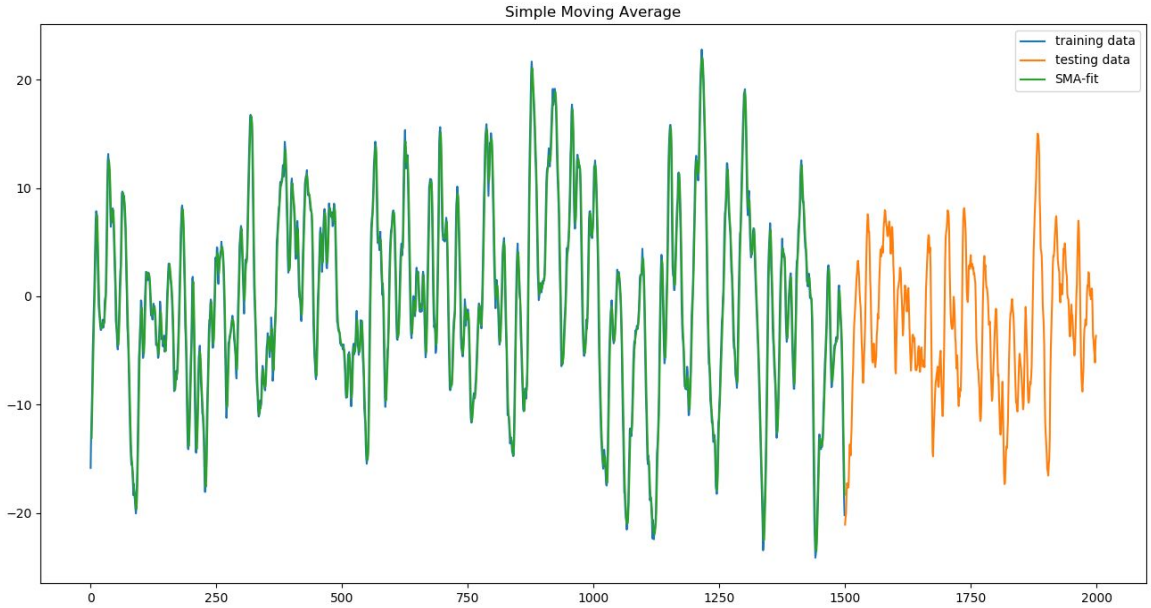

Figure 3: Training Data v/s Simple Moving average.

The minimum value of RMSE obtained in this case is : 0.37569271076696276

Task 3 Exponential Moving Average:

In this case we are interested in fitting a Exponential smoothing model to our training data. The task is to calculate the value of alpha (the smoothing coefficient) for which we get the minimum Root Mean Squared Error. For this purpose I have taken a range of 0.1 to 0.9 with an increment of 0.1 values and then I plotted the values of RMSE v/s the smoothing coefficient. Figure 4. Below shows the plot.

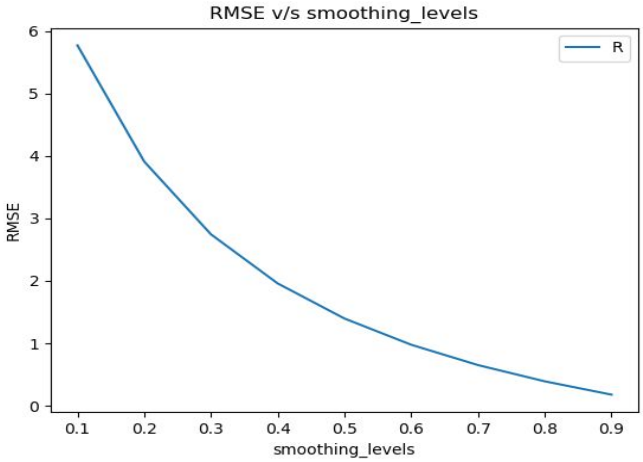

Figure 4: RMSE v/s Smoothing Coefficient Apha.

We see from the plot that the minimum value of RMSE occurs at a value of alpha equal to 0.9.

Hence we plot the training data v/s Fit exponential smoothing model Below in Figure 5. You can see the plots of of training v/s exponential smoothing.

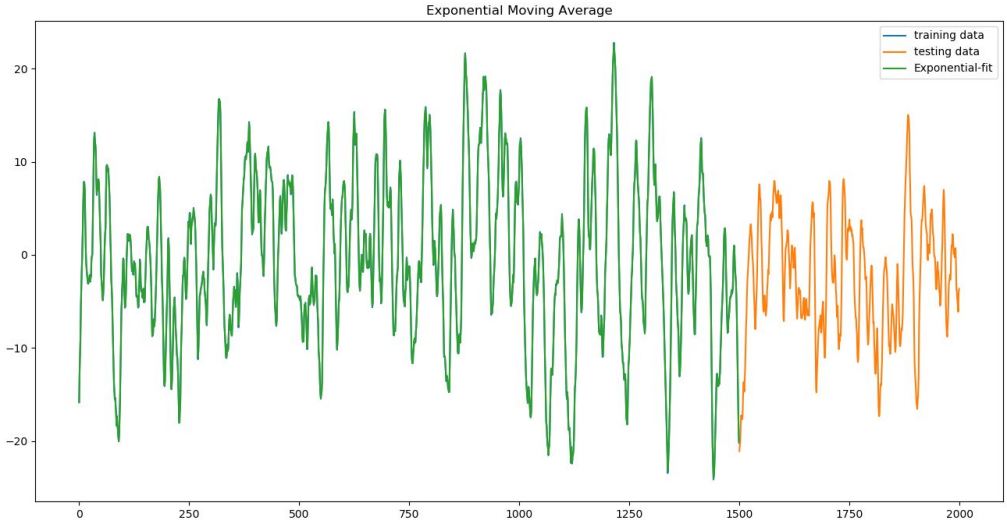

Figure 5: Training Data v/s exponential smoothing

The minimum value of RMSE obtained in this case is : 0.17888569848242378

Task 4 Auto-Regressive Models:

In this case we are interested in fitting an Auto-Regressive model to our training data. The task is to calculate the value of p (AR coefficient) for which we get the minimum Root Mean Squared Error. For this purpose the Partial Autocorrelation Function graphs are used. The order p of the AR model is obtained by plotting PACF in order to determine the lag k at which PACF cuts off the thresholds of +-0.15. Below is the PACF plot, Figure 6.

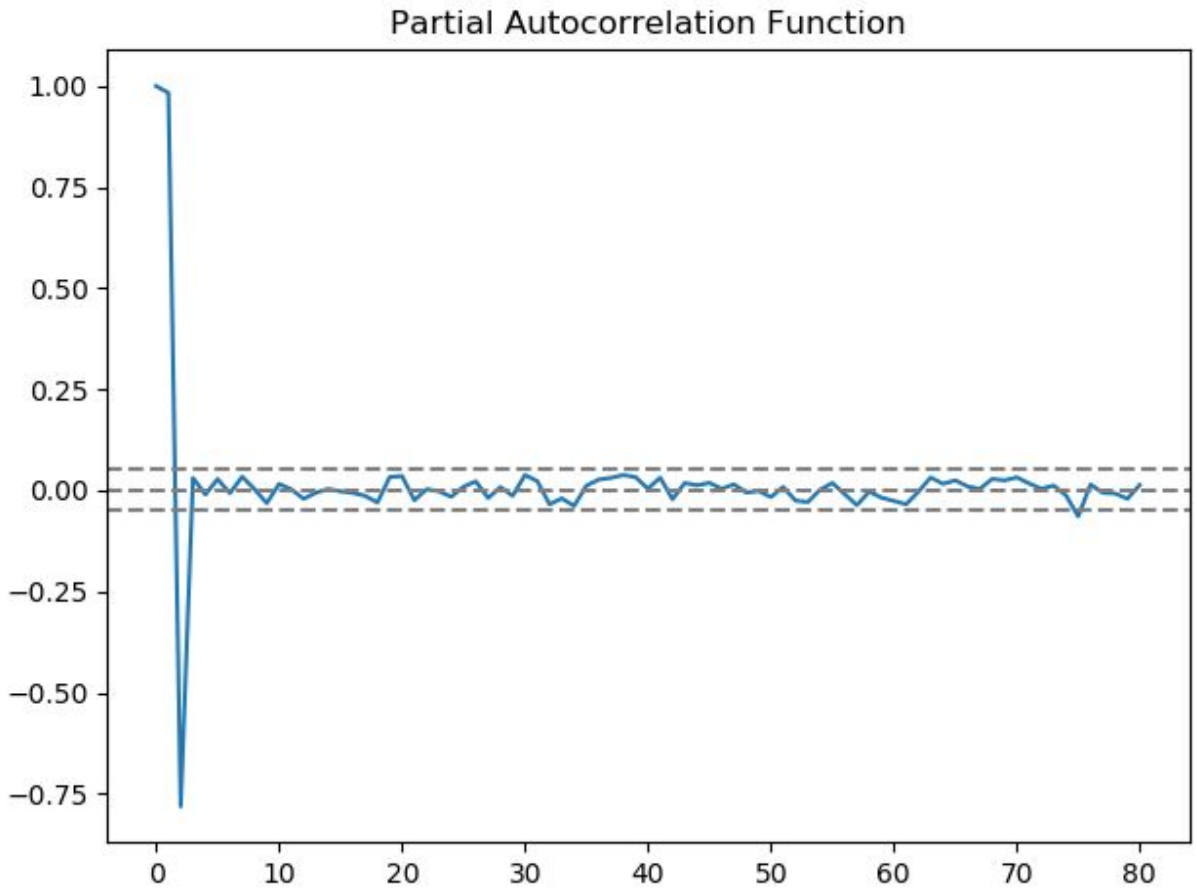

Figure 6 Training Data v/s Simple Moving average.

From the above graph we obtain the value of P for our AR model. In my case we can see

that the PACF plot cuts off the threshold limits at a value of 1. Here we use this value and check our fit of the AR model. Below is the fit of my AR model.

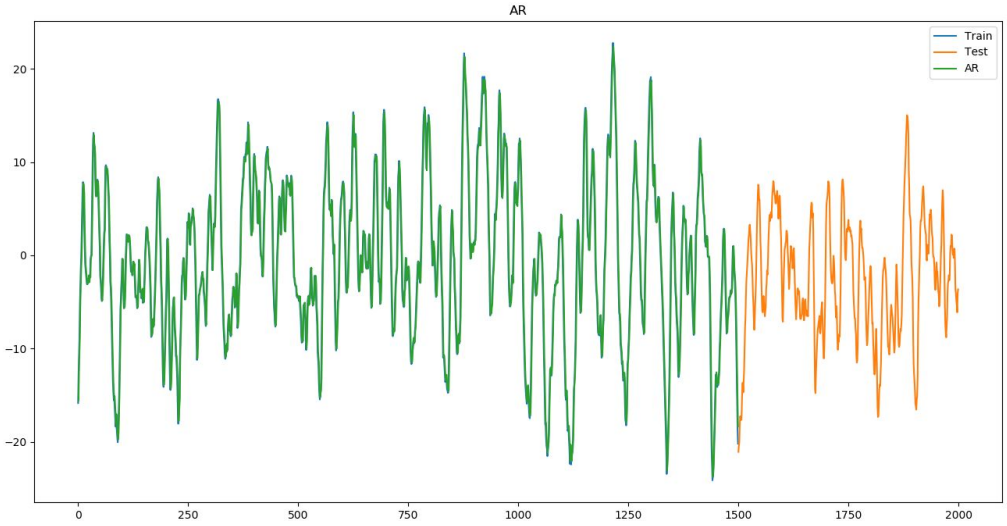

Figure 7. AR(1) fit on training data

We can see from the Figure 7 above above that the AR model fits the data quite well. If we try and increase our value of then the fit starts to worsen. Also we can see from the residual analysis below that the qq plot is proper when AR(1) is used.

Below I perform a residual analysis over the AR(1) model.

The ideal conditions for a AR fit to be perfect will be that all our tests on the residuals pass. That is the residuals are normally distributed and follow no trends and have a uniform variability. We use Residual scatter plots to visually inspect the data and check if it has a trend or any non uniform variability. Also we perform a Chi - Square test to check if my residuals follow a chi square distribution. Also we check if the qq plot of the residual is linear or not.

First we observe the scatter plot of the residuals. Here we can see from the Figure 8 below that we observe no trends and variability in the scatter plot

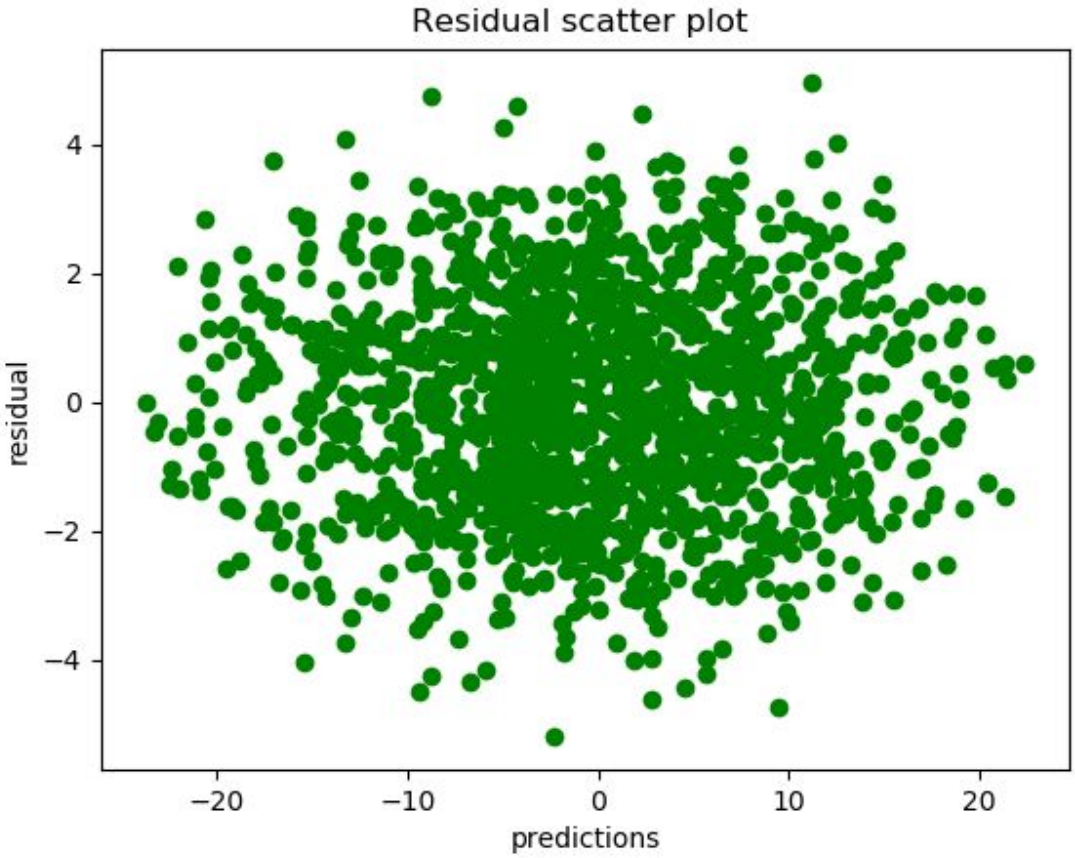

Figure 8: Residual Scatter plot.

Also below in figure 9 we plot the histogram of the residuals to check if if the residuals follow a normal distribution or not. We see that the histogram looks very similar to a normal distribution.

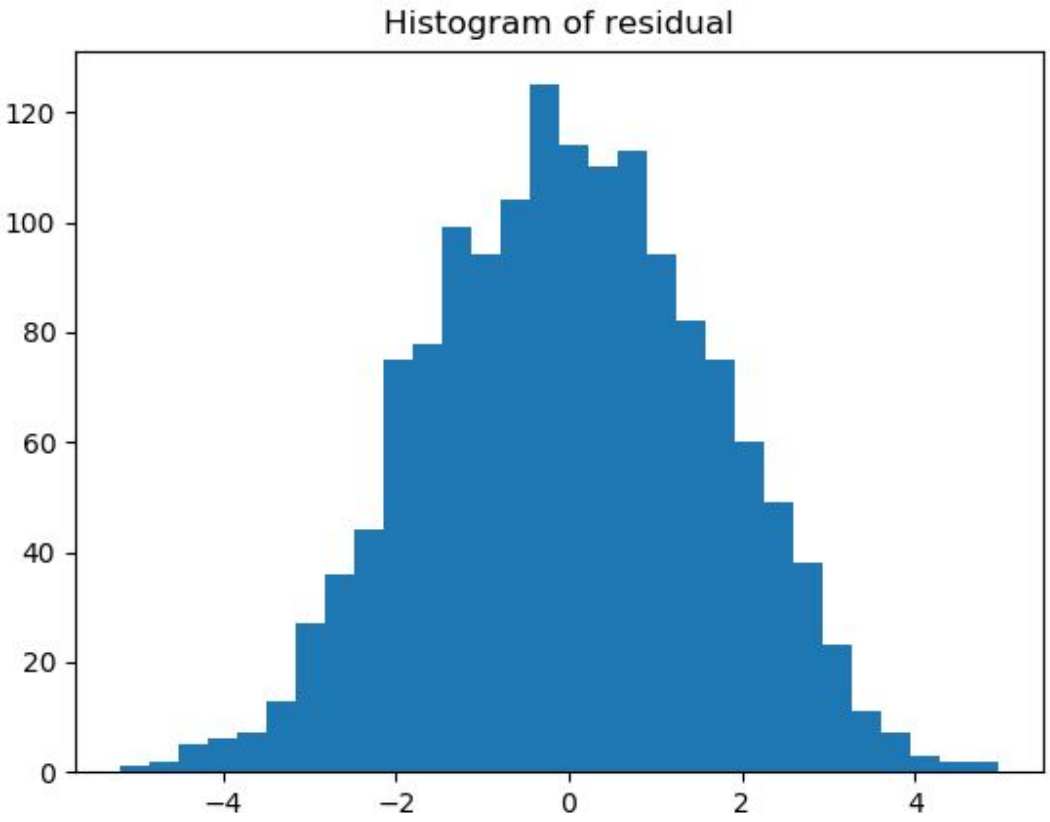

Figure 9 : Residual histogram

Now we plot the QQ- Plot of the Residuals in figure 10. We see that the qq-plot does follow a linear trend and we can conclude from this test also that the residuals are normally distributed.

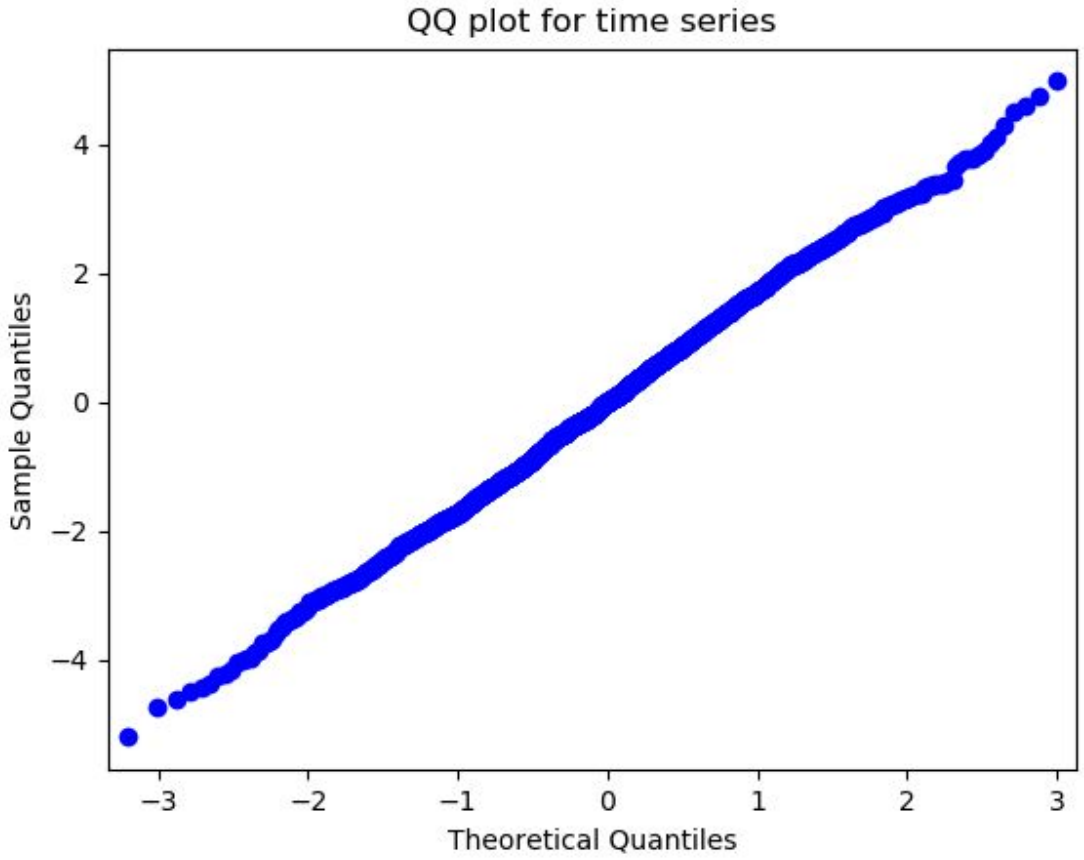

Figure 10 QQ-plot of residual

Now we perform the chi square test. For this purpose we use the test in python which is based on D'Agostino and Pearson's test for checking normality of a distribution.

The test puts forth a null hypothesis that the given distribution follows a normal distribution. We obtain the following p value : 0.019484163316611545. Here we are forced to reject the null hypothesis and carry on with the alternate that the given distribution does not follow a normal distribution. It can happen that one of the tests for checking normality does fails. But overall we can conclude that the best fit obtained in the AR model is for a p value equal to 1.

Task 5 Testing Data:

Here we use the best values of model parameters obtained from above tasks in case of all the three models and fit our models on the testing data to check how they perform on the testing data.

Below are the graphs which are obtained for each of the three models :

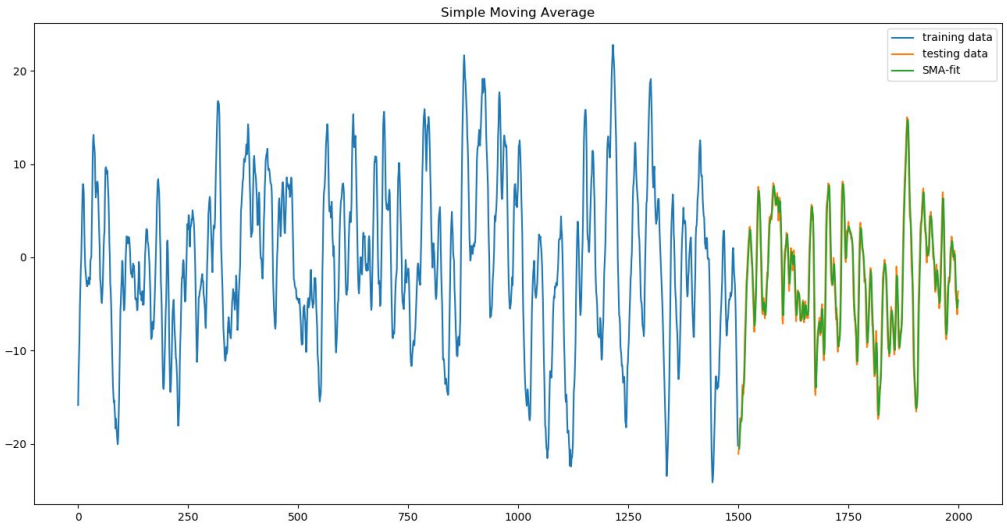

Figure 12: Simple moving average with window size 3 on Testing data

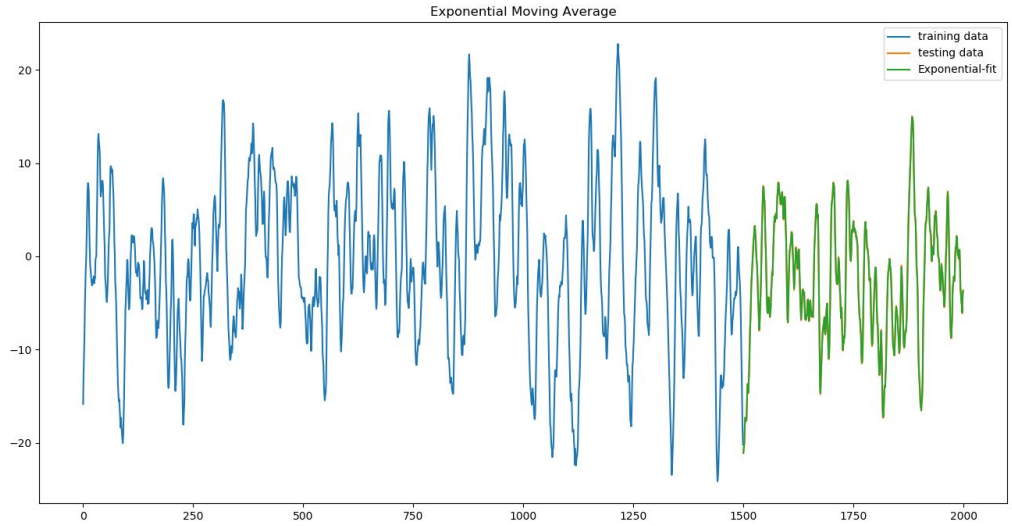

Figure 13: Exponential smoothing with alpha 0.9 on Testing data

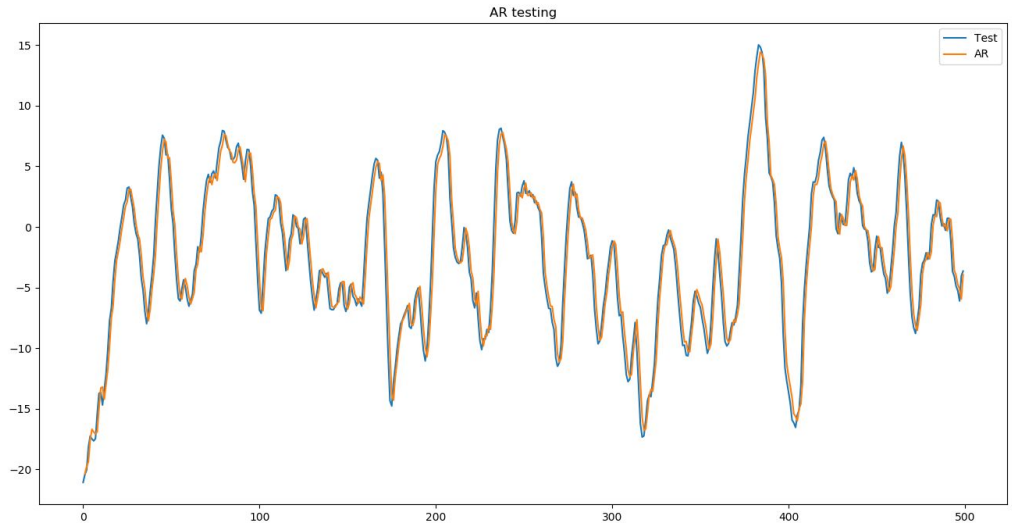

Figure 14: Autoregressive with p = 1 on Testing data

The following values of RMSE are obtained in the three cases:

RMSE SImple Moving Average 0.3787739172930101

RMSE Exponential 0.16229095254427903

RMSE Auto Regressive Model 1.4873352343833393

Comment Final:

We see from the RMSE values obtained that the best model which works on the testing data as well is the exponential smoothing model for the data given in my case. We can also do some more types of testing like we can use ARIMA models and check if the fits are good to better fit the given data. Also in order to pass the chi square test in AR model we can try including MA(q) parameters to see if we have any improvement in the RMSE values and if the residuals give us a positive chi square test.