ECE 592-005 IOT Analytics

Project 3 : Forecasting

Sreeraj Rajendran Email: srajend2@ncsu.edu ID: 200210462

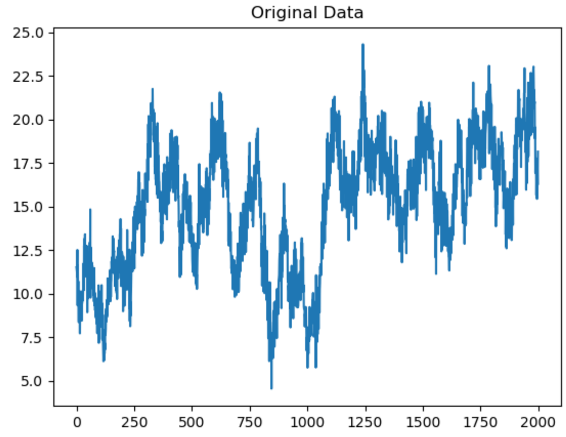
Objective:
To use various forecasting algorithms to determine the best model for a given time series data.

Dataset:
Dataset consists of 2000 observations which are partitioned into training set and test set with 1500 and 500 observations respectively.

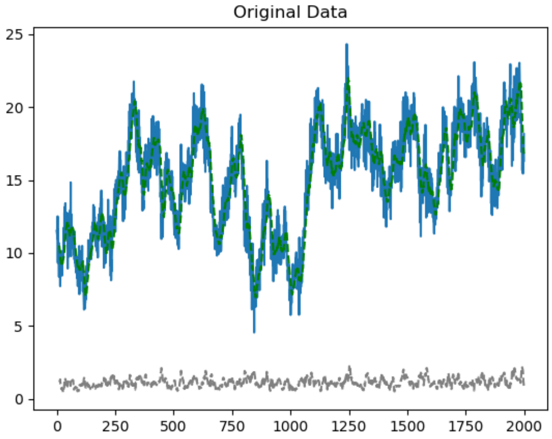
Task 1. Check for stationarity
Plot the entire time series (i.e. all 2000 observations) and check it visually for stationarity and make necessary appropriate transformations as discussed in section 6.1.2. Comment on your conclusions.

Plots:

The plot of the given data is as follows:



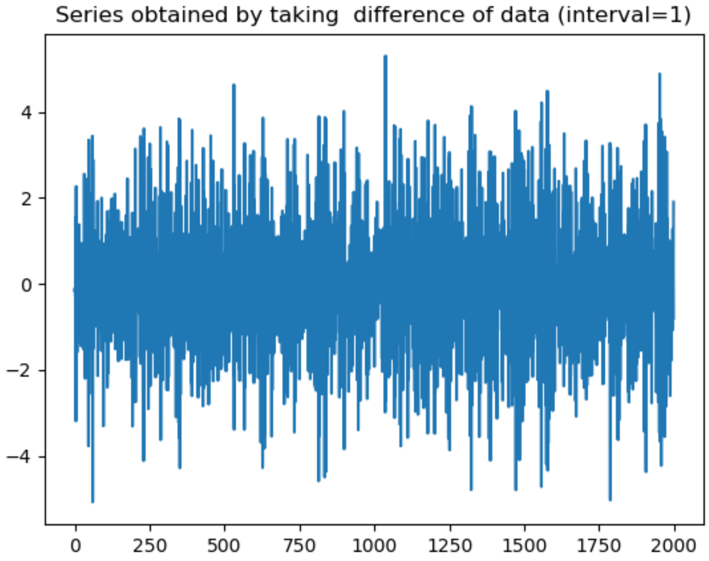
The plot clearly shows a trend with varying mean. The variation in variance is not significant can be seen in detail in the figure below with rolling mean(green) and standard deviation(gray) plotted with original data series(blue). Also, no seasonality is displayed by the data.



Stationarity is required to perform accurate predictions of time series data. Hence following methods were tried to make the data stationary.

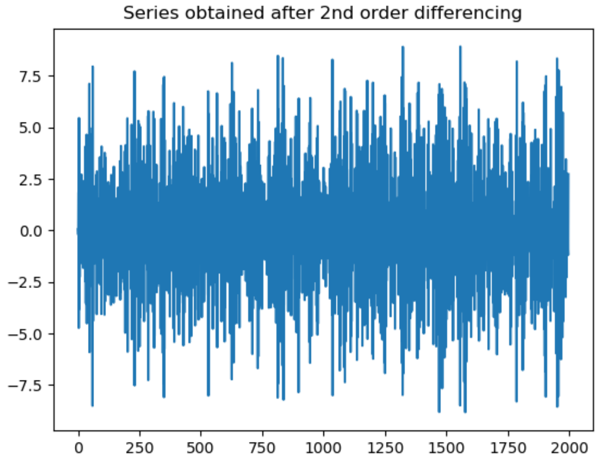
First order differencing:

The difference of the successive observations is taken and the resulting time series of the differences will be used to forecast the next value. As can be seen from the plot, the trend is removed by differencing. Mean is constant at about 0. The data will have to be reversed following prediction using a model to obtain real values.



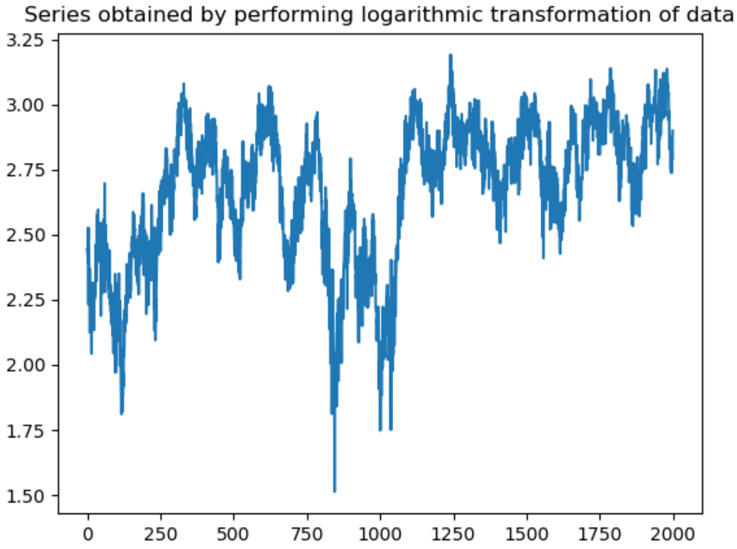
Second order differencing:

This takes the difference of differences obtained above to generate the series. This plot again shows that the trend is removed and the mean is constant at about 0.



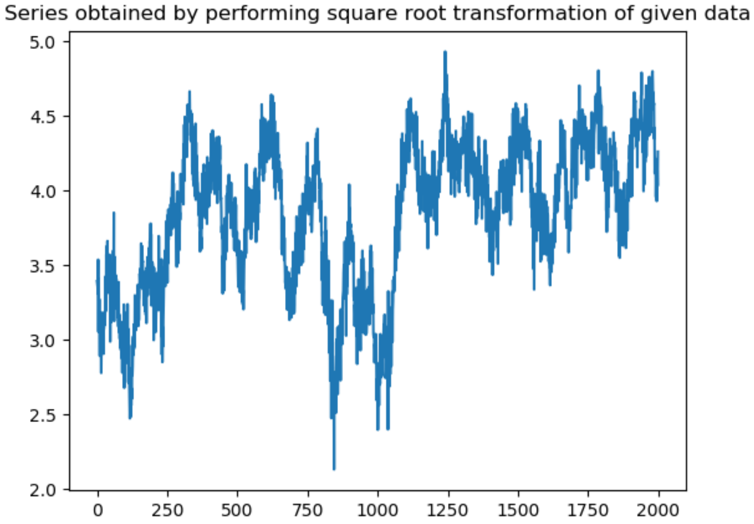
Logarithmic Transformation:

This plot shows a trend with varying mean and hence cannot be used.



Square Root Transformation:

This plot again shows a trend with varying mean and hence cannot be used.



Conclusion:

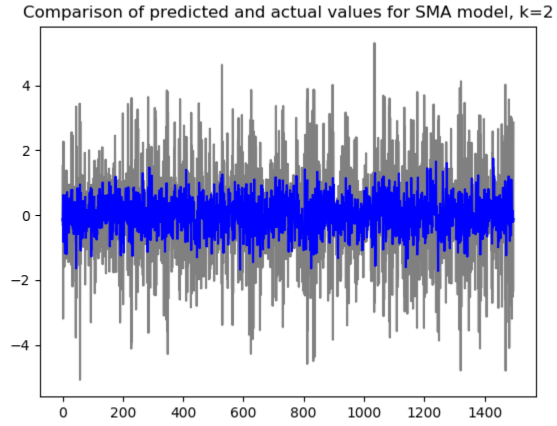
The series generated using first order differencing will be used for prediction as this successfully removed all trends. The data will be reversed after prediction.

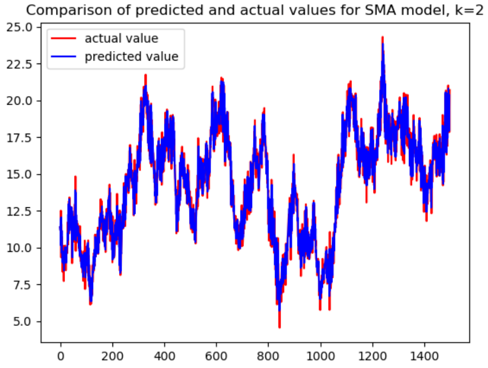
Task 2. Fit a simple moving average model for training set

$$s_t = \frac{1}{k}\sum_{i=t-k}^{t-1} x_i$$

2.1 Apply the simple moving average model to the training data set, for a given k.

For k=2, the predicted values are similar to the actual value as can be seen from the plot below. The first plot shows predictions using the difference values (grey=actual and blue = predicted). After prediction, the values have been inverted to obtain real values shown in second figure.



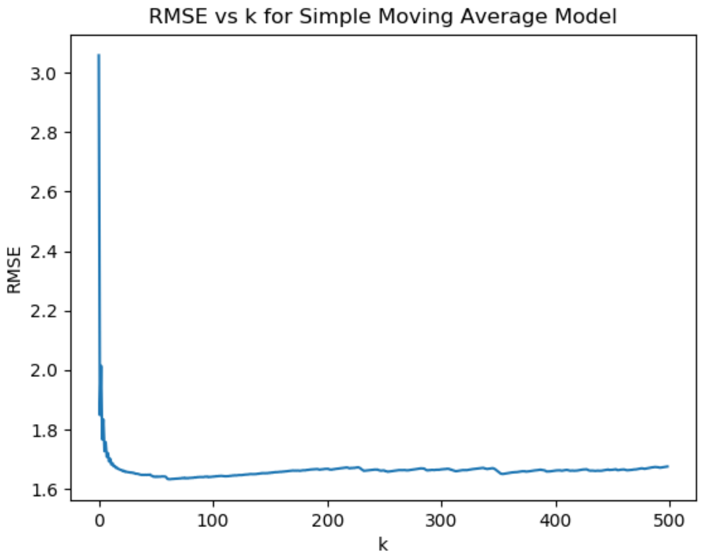


1.2 Calculate the error, i.e., the difference between the predicted and original value in the training data set, and compute the root mean squared error (RMSE).

RMSE for k=2 is: 1.8487457565074783

1.3. Repeat the above two steps by varying k and calculate the RMSE.
1.4 Plot RMSE vs k. Select k based on the lowest RMSE value. For the best value of k plot the predicted values against the original values.

Window size have been varied till 500 as test data is only 500.

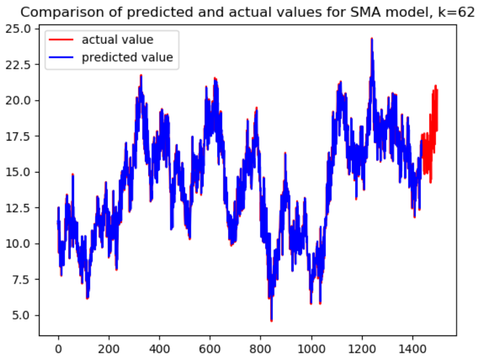


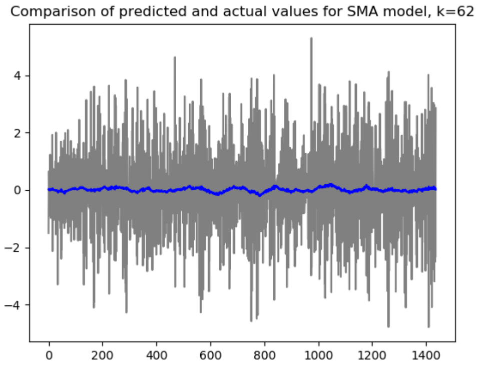
Minimum RMSE for Simple Moving Average Model: 1.6329213917267502

k value corresponding to min rmse simple moving average: 62
RMSE for k=62 is: 1.6331028201427915

The RMSE value decreases substantially for initial few values of k. The minimum observed is at k=63 and is coherent with the global minima in the plot above.

The plot for k= 63 is shown below. The predicted value is matching that of actual value and the model is a great fit.



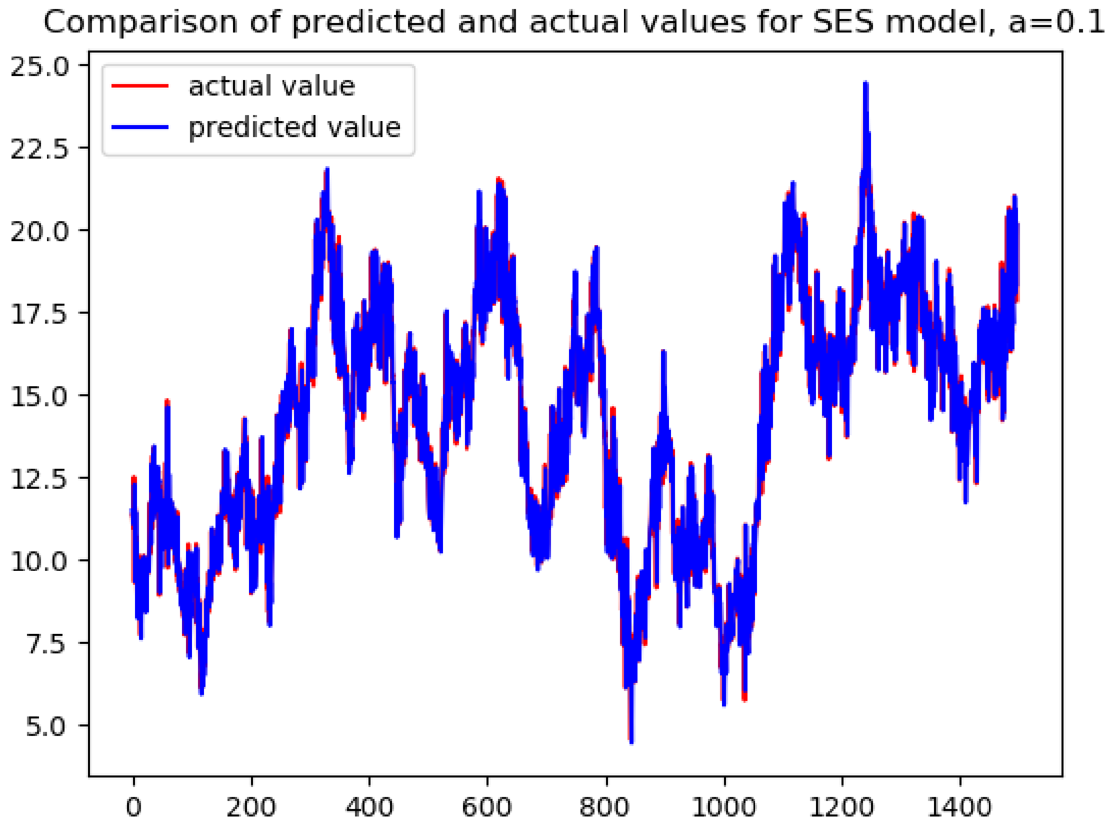


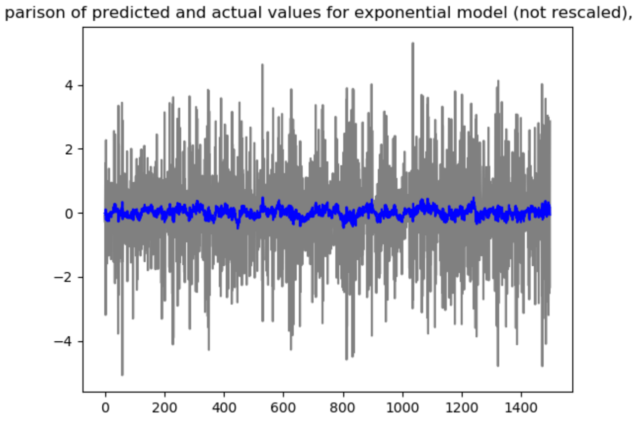
The figure to left is for data inverted back to original values and the data to right is for differenced actual and predicted data.

Task 3. Fit an exponential smoothing model (use the training set)

3.1 Apply the exponential smoothing model to training set for a=0.1

For a=0.1 , the predicted values are similar to the actual value as can be seen from the plot below. After prediction, the values have been inverted to obtain real values shown in second figure. The second plot shows predictions using the difference values (grey=actual and blue = predicted).





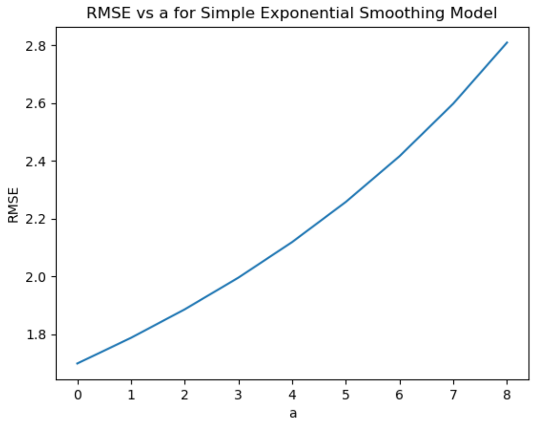
3.2 Calculate the error, i.e., the difference between the predicted and original value in the training data set, and compute the root mean squared error (RMSE).

RMSE for a=0.1 is: 1.699400945612357
The low value of RMSE indicates good fit between predicted and actual value.

3.3. Repeat steps 2.1 and 2.2 by increasing a each time by 0.1, until a = 0.9.

3.4. Plot RMSE vs a. Select a based on the lowest RMSE value.

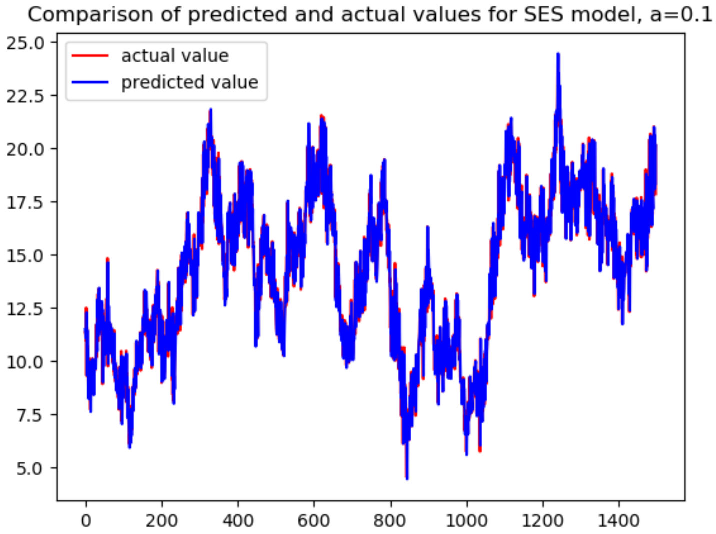
The plot while varying a between 0.1 to 0.9 is as shown below
X axis scale: 1 unit=0.1

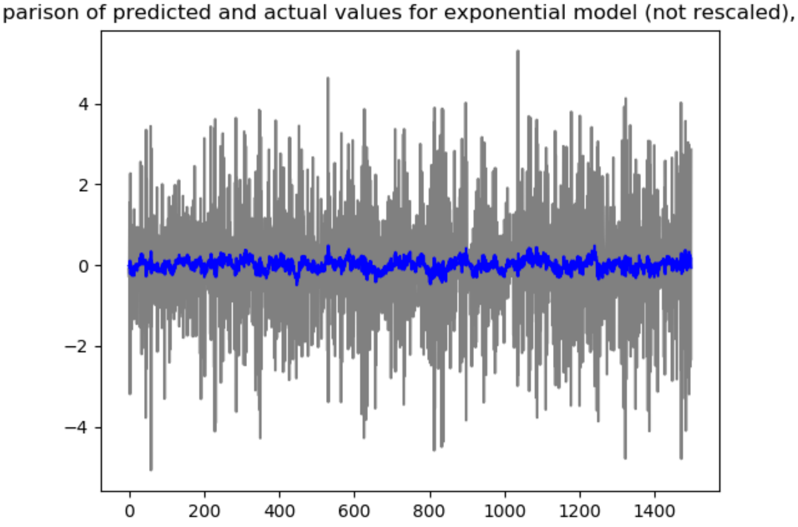


The exponentialy increasing curve indicates minimum in the beginning. The minimum RMSE is more than SMA model.

Minimum RMSE value for exponential smoothing model: 1.699400945612357
a value for simple exponential smoothing model: 0.1
RMSE for a=0.1 is: 1.699400945612357.

3.5 For the selected value of a plot the predicted values against the original values, and visually inspect the accuracy of the forecasting model.
Selected model: a=0.1

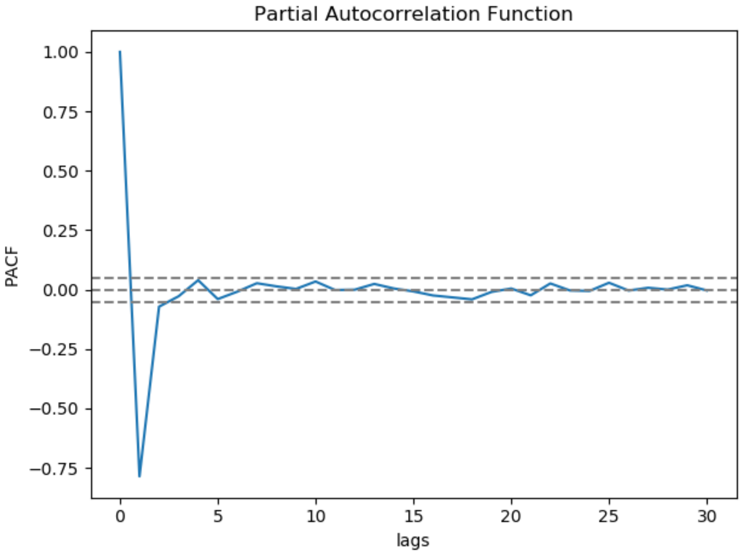


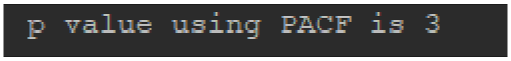


Task 4. Fit an AR(p) model (use the training set)

4.1: First select the order p of the AR model by plotting PACF in order to determine the lag k at which PACF cuts off, as discussed in section 6.4.4.

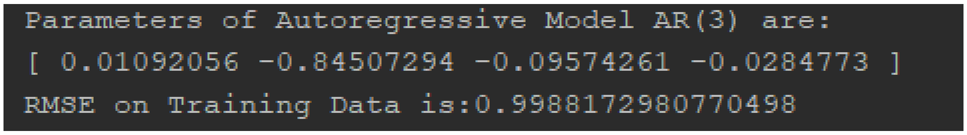
The order of a AR(p) model is determined by calculating the partial autocorrelation function (PACF) and then setting *p* to the lag *k* past which all lags are zero. That is, the value of *p* is set to the lag *k* past which all lags of the PACF are zero. It can be seen from the plot below that beyond lag value of 3, the PACF values are almost zero (<0.1).





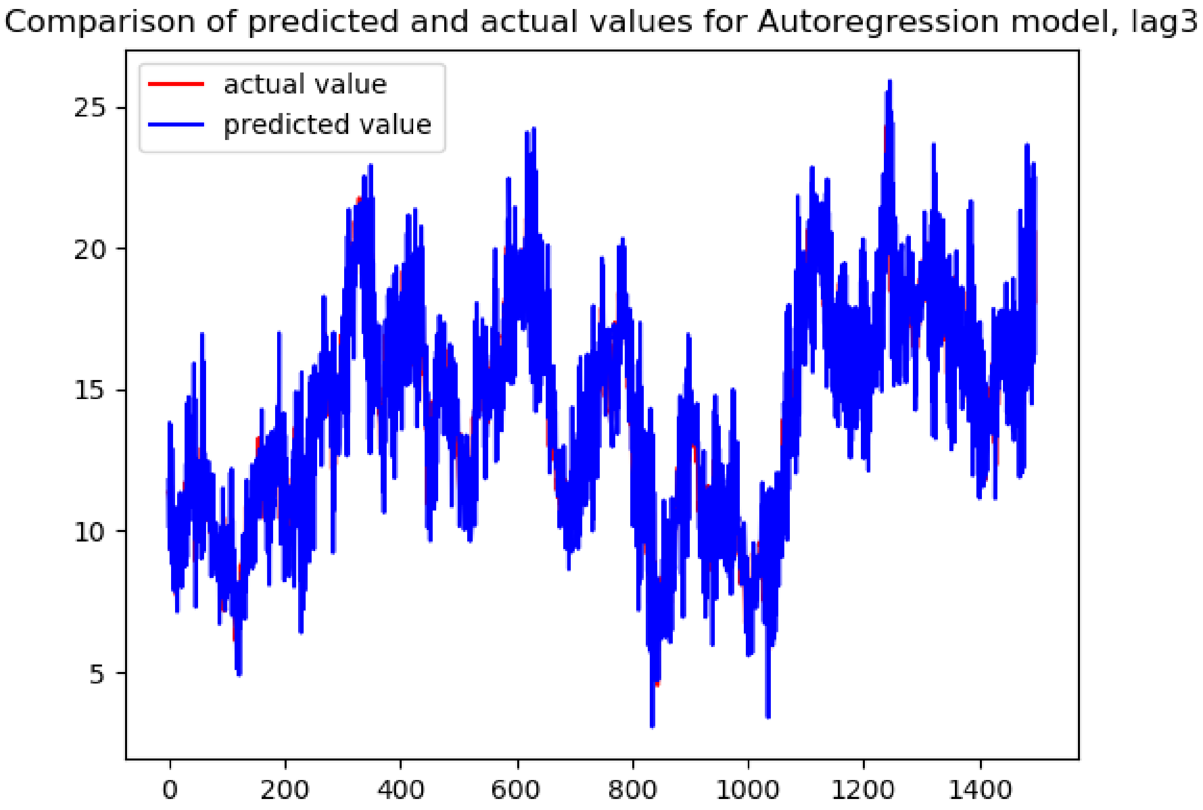
```
p value using PACF is 3
```

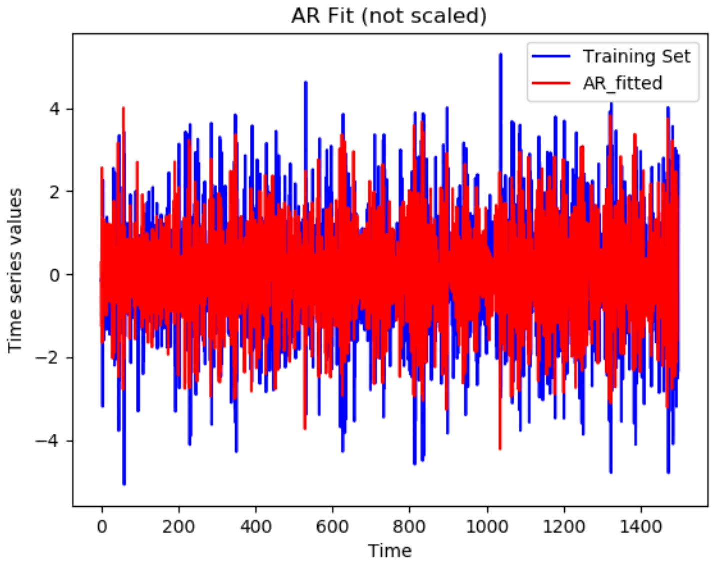
4.2 Estimate the parameters of the AR(p) model. Provide RMSE value and a plot the predicted values against the original values.



```
Parameters of Autoregressive Model AR(3) are:
[ 0.01092056 -0.84507294 -0.09574261 -0.0284773 ]
RMSE on Training Data is:0.9988172980770498
```

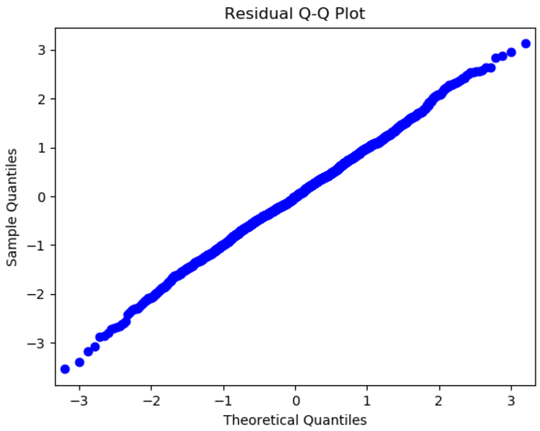
The RMSE value is much lower compared to previous models and hence a much better fit is obtained as can be seen in the plots below. First plot shows rescaled plot for actual and predicted values. The predicted plot almost overlaps the actual value. Second plot shows actual and predicted values for differenced values.


Comparison of predicted and actual values for Autoregression model, lag3
actual value
predicted value
25
20
15
10
5
0
200
400
600
800
1000
1200
1400


AR Fit (not scaled)
Training Set
AR_fitted
Time series values
4
2
0
−2
−4
0
200
400
600
800
1000
1200
1400
Time

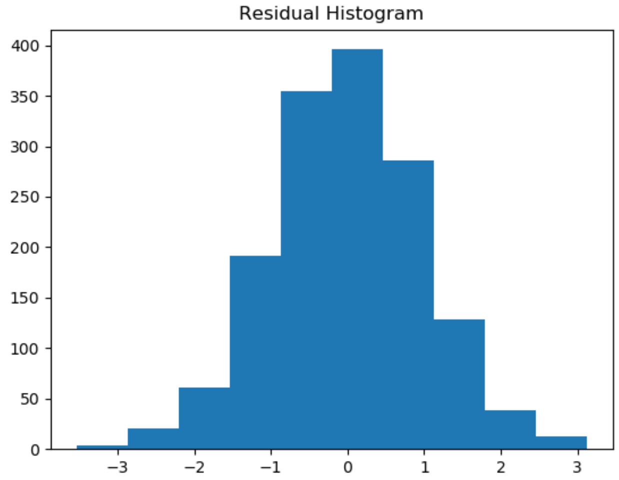
4.3 Residual Analysis:

Q-Q Plot:



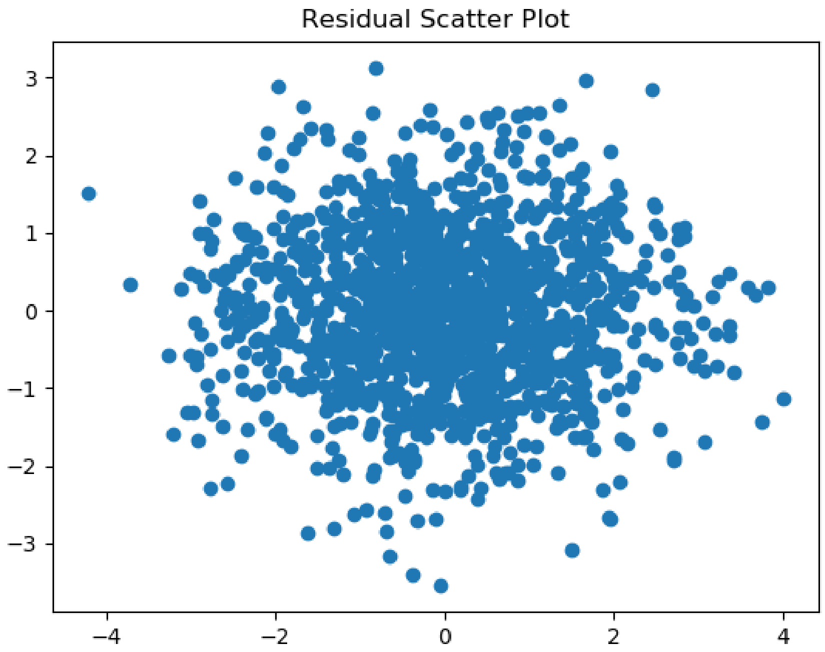
If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line $y = x$. Obtained Q-Q plot is not be diverging from reference line indicating a good fit.

Histogram:



The histogram shows that the data is concentrated over mean of 0 and doesn't show any skewness or presence of outliers. Hence indicating a good fit. The data is normally distributed with mean of 0 and standard deviation slightly above 3.

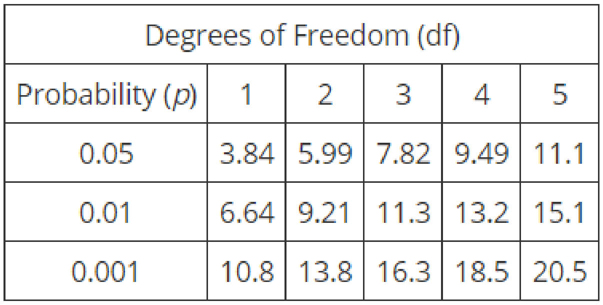
Residual Scatter Plot:



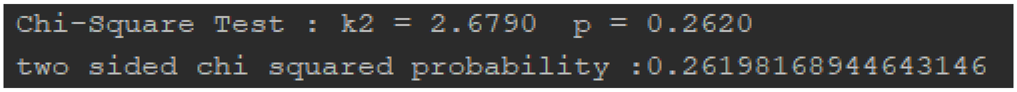
No trends of positive or negative correlation can be seen. Hence The model is good.

Chi Squared test:

Critical Chi squared values for different degrees of freedom are shown below. Obtained Chi squared value needs to be lower than critical value.



Degrees of Freedom (df)					
Probability (*p*)	1	2	3	4	5
0.05	3.84	5.99	7.82	9.49	11.1
0.01	6.64	9.21	11.3	13.2	15.1
0.001	10.8	13.8	16.3	18.5	20.5

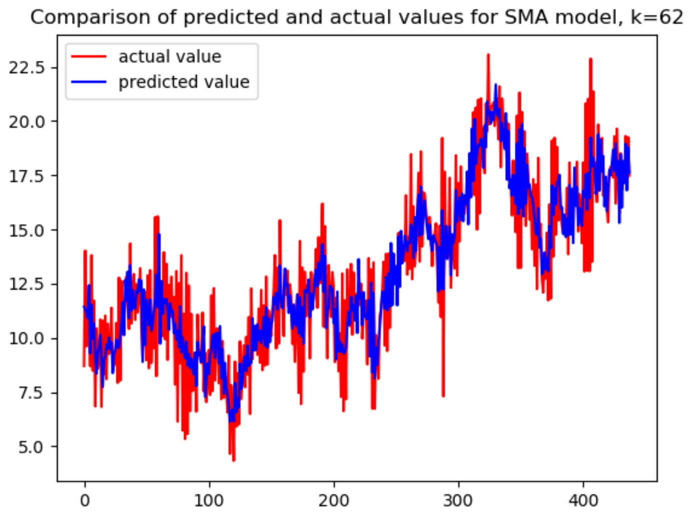


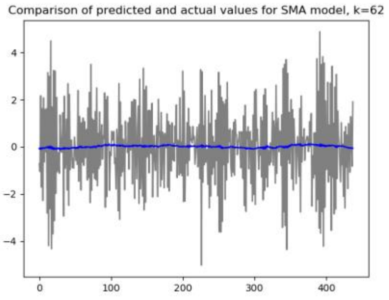
```
Chi-Square Test : k2 = 2.6790  p = 0.2620
two sided chi squared probability :0.26198168944643146
```

Since the two sided chi squared probability (0.262) is much lower than the critical value, the model is good.

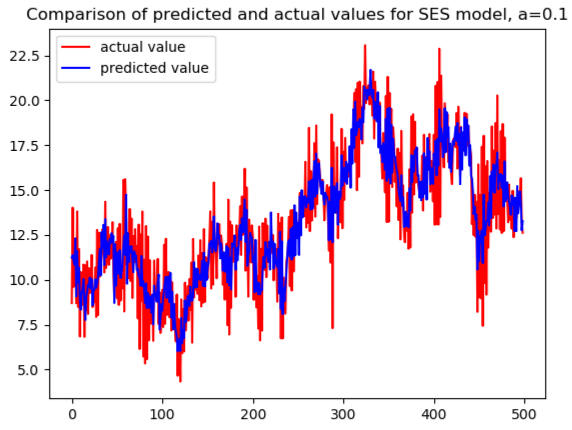
Task 5. Comparison of all the models (use the testing set)

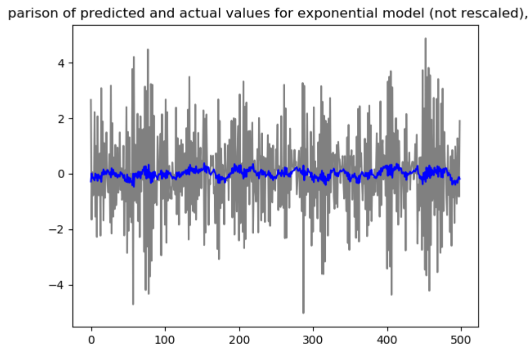
Model 1: Simple Moving Average model : k =62
RMSE for k=62 on testing data = 1.717





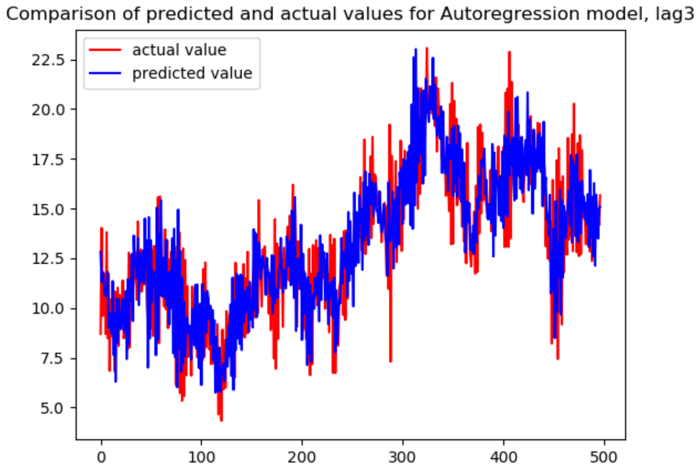
Model 2: Simple exponential Smoothing model : a=0.1
RMSE for a=0.1 on test data is 1.7926

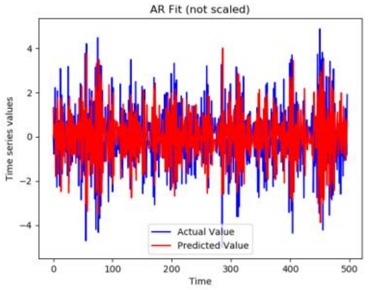


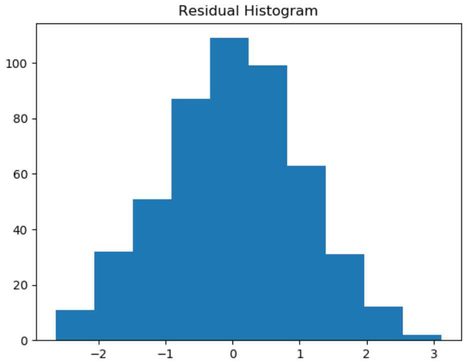


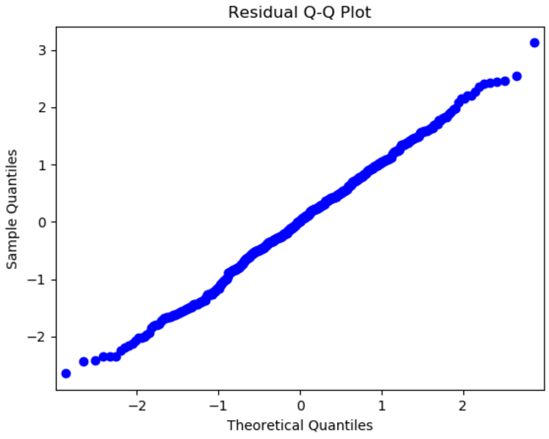
Model 3: Autoregression model of order 3 AR(3):

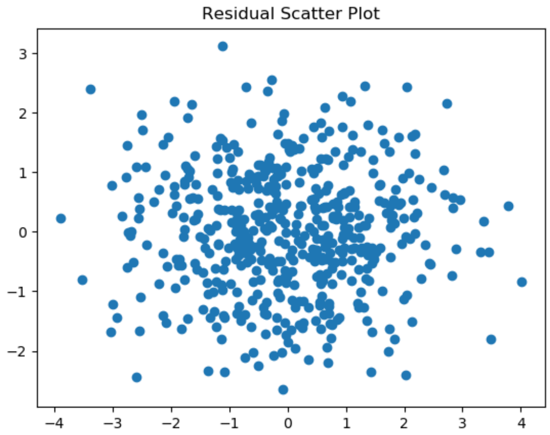
RMSE on test data using AR(3) model : 1.034


Comparison of predicted and actual values for Autoregression model, lag3

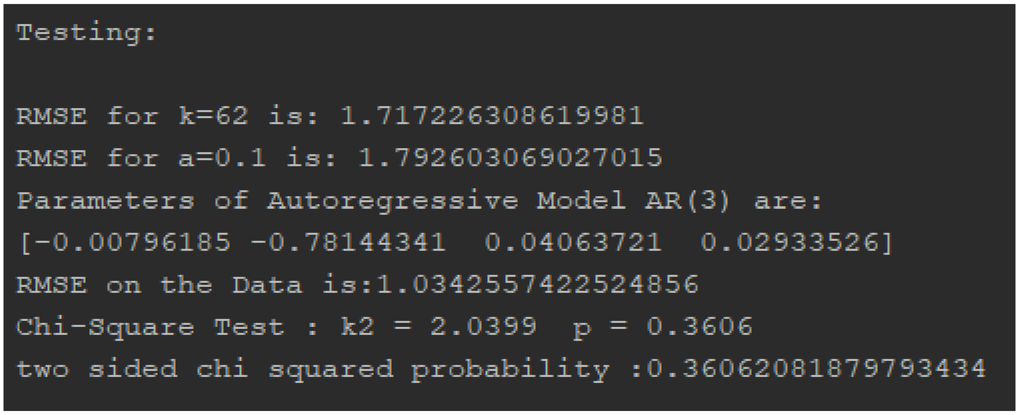

AR Fit (not scaled)


Residual Histogram


Residual Q-Q Plot



The results indicate a good fit and absence of any trends.



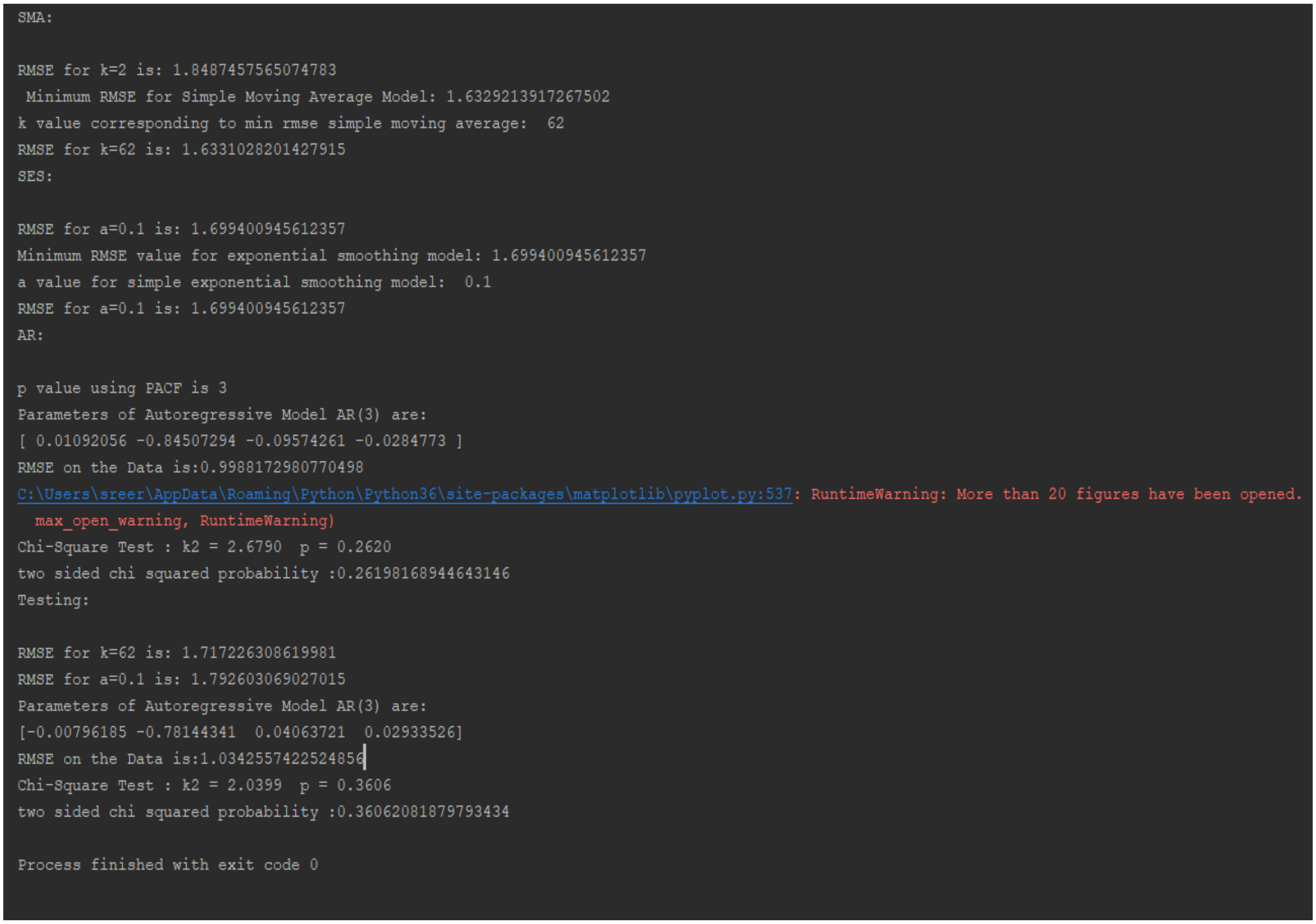
```
Testing:

RMSE for k=62 is: 1.717226308619981
RMSE for a=0.1 is: 1.792603069027015
Parameters of Autoregressive Model AR(3) are:
[-0.00796185 -0.78144341  0.04063721  0.02933526]
RMSE on the Data is:1.0342557422524856
Chi-Square Test : k2 = 2.0399  p = 0.3606
two sided chi squared probability :0.36062081879793434
```

Conclusion:

AR(3) better than SMA(k=62) better than SES (a=0.1)

The best results are obtained using autoregression model with order 3. The RMSE is minimum for AR(3) and the predicted best fits the actual values in case of AR(3). The performance is not as good as training set but the results are good and can predict a time series with good accuracy/ low error.

Output Log:



```
SMA:

RMSE for k=2 is: 1.8487457565074783
 Minimum RMSE for Simple Moving Average Model: 1.6329213917267502
k value corresponding to min rmse simple moving average:  62
RMSE for k=62 is: 1.6331028201427915
SES:

RMSE for a=0.1 is: 1.699400945612357
Minimum RMSE value for exponential smoothing model: 1.699400945612357
a value for simple exponential smoothing model:  0.1
RMSE for a=0.1 is: 1.699400945612357
AR:

p value using PACF is 3
Parameters of Autoregressive Model AR(3) are:
[ 0.01092056 -0.84507294 -0.09574261 -0.0284773 ]
RMSE on the Data is:0.9988172980770498
C:\Users\sreer\AppData\Roaming\Python\Python36\site-packages\matplotlib\pyplot.py:537: RuntimeWarning: More than 20 figures have been opened.
  max_open_warning, RuntimeWarning)
Chi-Square Test : k2 = 2.6790  p = 0.2620
two sided chi squared probability :0.26198168944643146
Testing:

RMSE for k=62 is: 1.717226308619981
RMSE for a=0.1 is: 1.792603069027015
Parameters of Autoregressive Model AR(3) are:
[-0.00796185 -0.78144341  0.04063721  0.02933526]
RMSE on the Data is:1.0342557422524856
Chi-Square Test : k2 = 2.0399  p = 0.3606
two sided chi squared probability :0.36062081879793434

Process finished with exit code 0
```