ECE 592: IoT Analytics

PROJECT 3: Time Series Forecasting

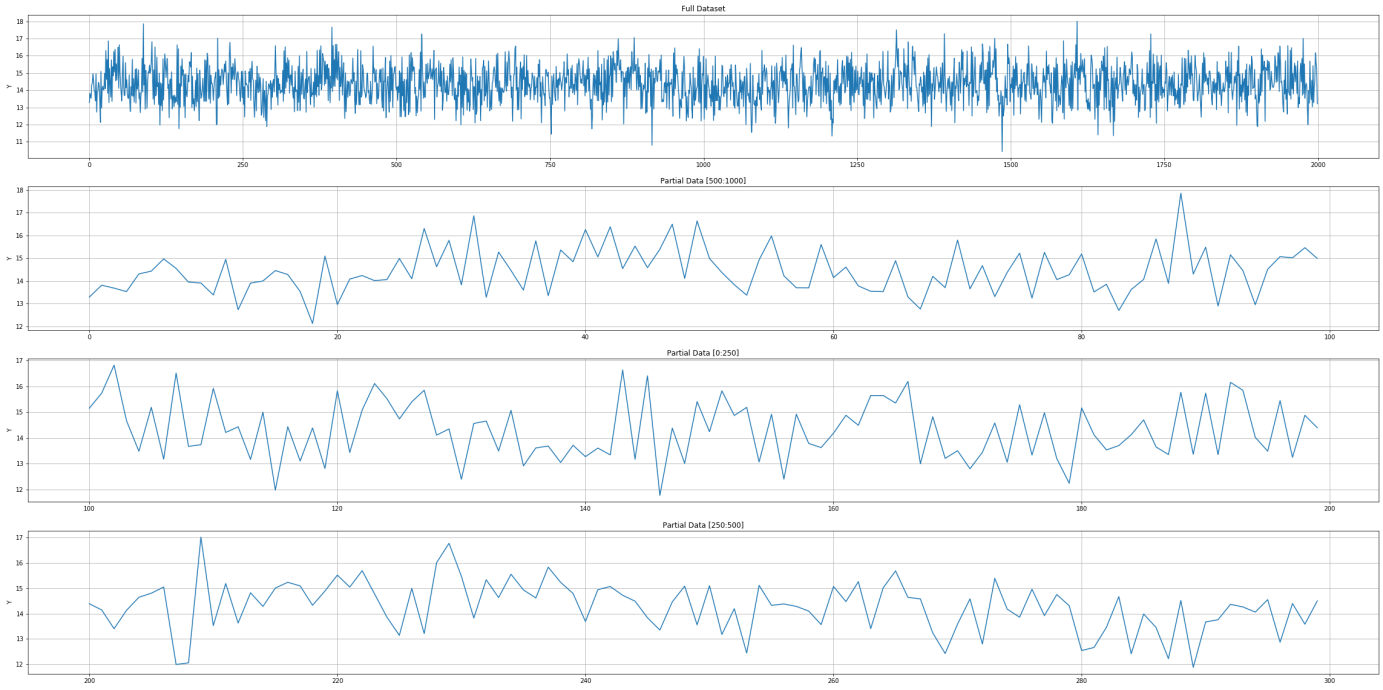
STUDENT ID: 200203773
STUDENT NAME: ESHAAN VENKAT KIRPAL

NOTE: There are five tasks in this report and each task is further divided into sub-sections.

TASK 1: BASIC STATISTICAL ANALYSIS

Summary of the Input Data:
•The input data had 2000 records, which was split into training and test data. To maintain the sequence, the first 1500 records were taken as training data and the rest as test data.

•The data seemed stationary with no variability, seasonality or trend in initial inspection of the timeseries data plots. Though to verify the stationarity, Augmented Dickey – Fuller test was carried out. In the below subplot, the top plot is the entire dataset while the bottom three graphs are zoomed in version of the dataset, to check for local trends.

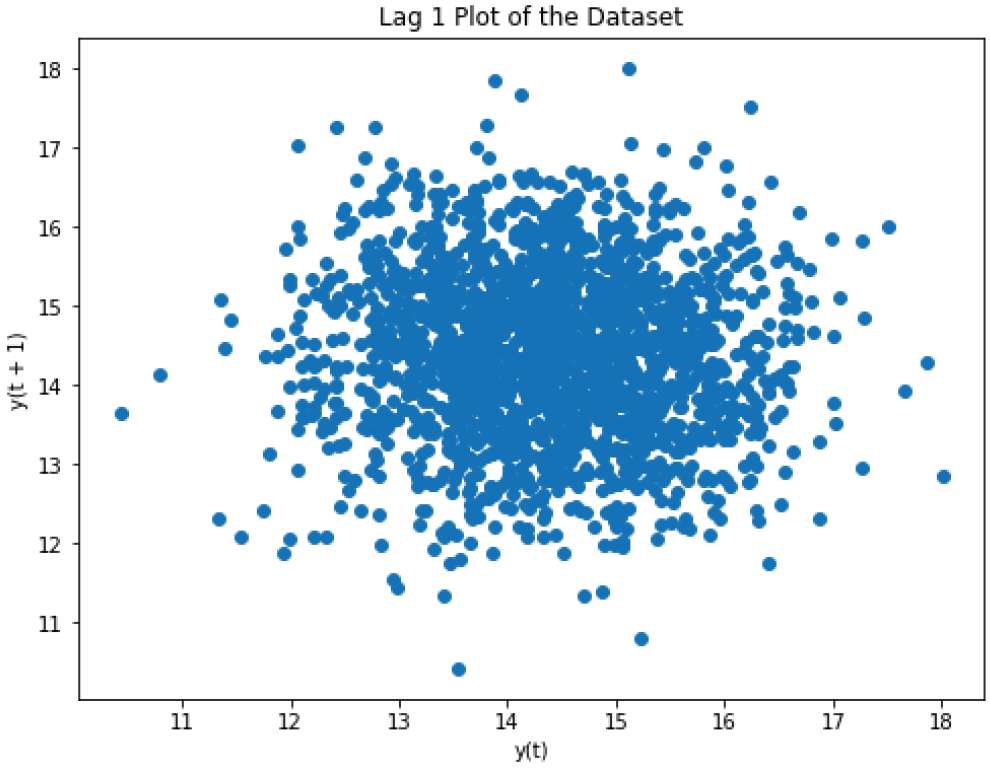


Based on the Augmented Dickey – Fuller test, we get the test statistic value of -22.93. The more negative this statistic, the more likely we are to reject the null hypothesis (i.e. we have a stationary dataset).

Also, we can see that our statistic value of -22.93 is less than the value of -3.434 at 1%. This suggests that we can reject the null hypothesis with a significance level of less than 1% (i.e. a low probability that the result is a statistical fluke).
Rejecting the null hypothesis means that the time series is stationary or does not have time-dependent structure.

```
ADF Statistic: -22.931547
p-value: 0.000000
Critical Values:
        1%: -3.434
        5%: -2.863
        10%: -2.568
```

Another way that we can check for stationarity is using a lag plot. Lag plots are used to check if a data set or time series is random. Random data should not exhibit any structure in the lag plot. Non-random structure implies that the underlying data are not random. In the figure below, we do not see any structure for our data, hence it is safe to conclude that our dataset has no non-stationary components and can be taken as such for forecasting without preprocessing steps.



Note: We use root mean square error (RMSE) as the performance metric for computing the optimal parameters for all the following models.

TASK 2: SIMPLE MOVING AVERAGE
SECTION 1.1 AND SECTION 1.2:

The simple moving average (SMA) model is implemented for first predicting values on the training data using which we find the optimum value of k. Thereafter using that k we predict values on the test data.

First, the SMA model was computed for value of window size of k=10.

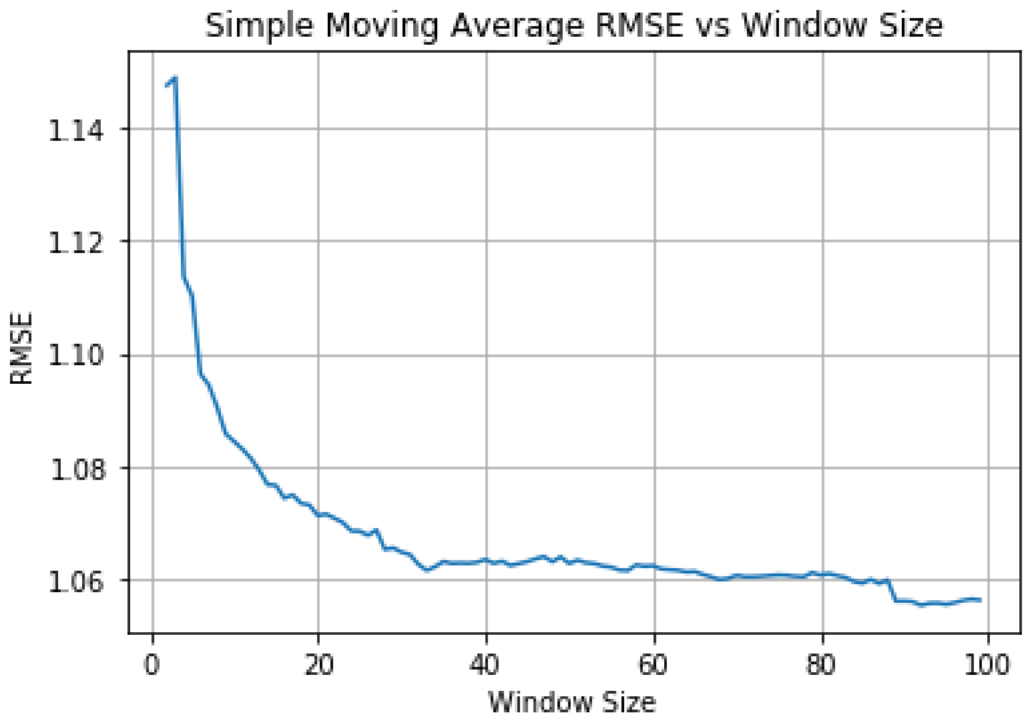
The Root mean squared (RMSE) was calculated and we got the result as below:

RMSE FOR WINDOW SIZE K=10 IS: 1.0844

SECTION 1.3, 1.4 AND 1.5:

Next, the simple moving average model is implemented on the training data for value of window size ‘k’ ranging from 1 to 100.

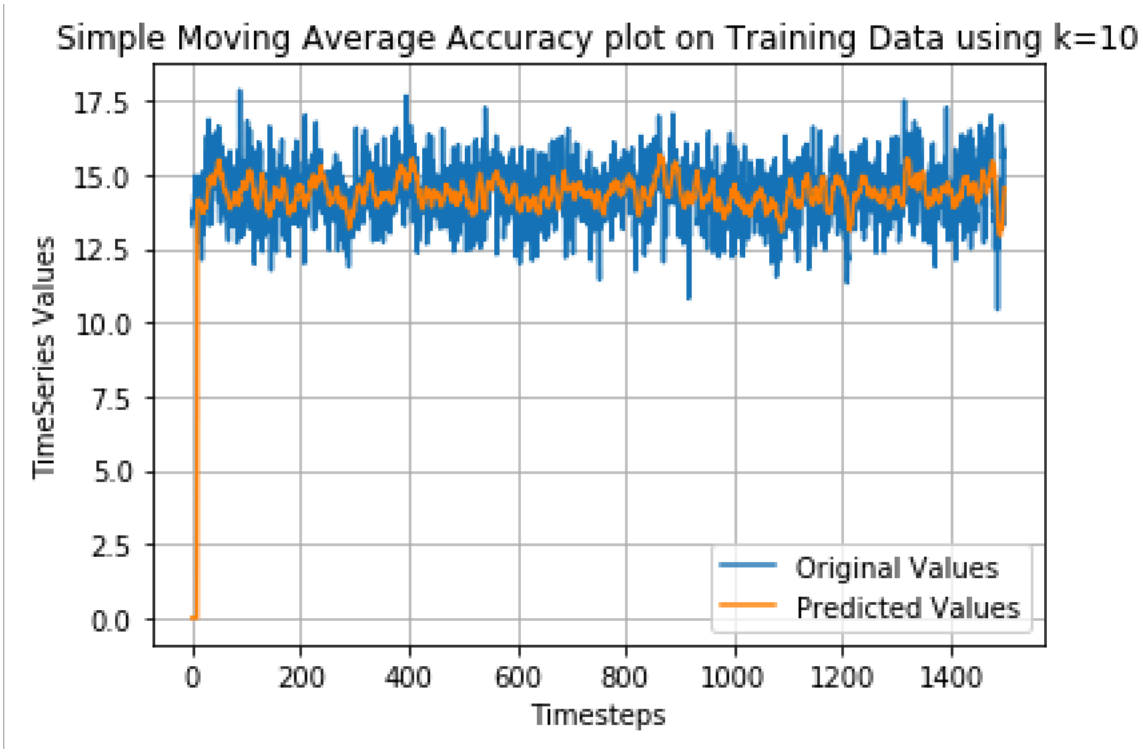
For each model the Root Mean Squared Error (RMSE) was calculated. The relationship between the value of ‘k’ and RMSE can be seen below.



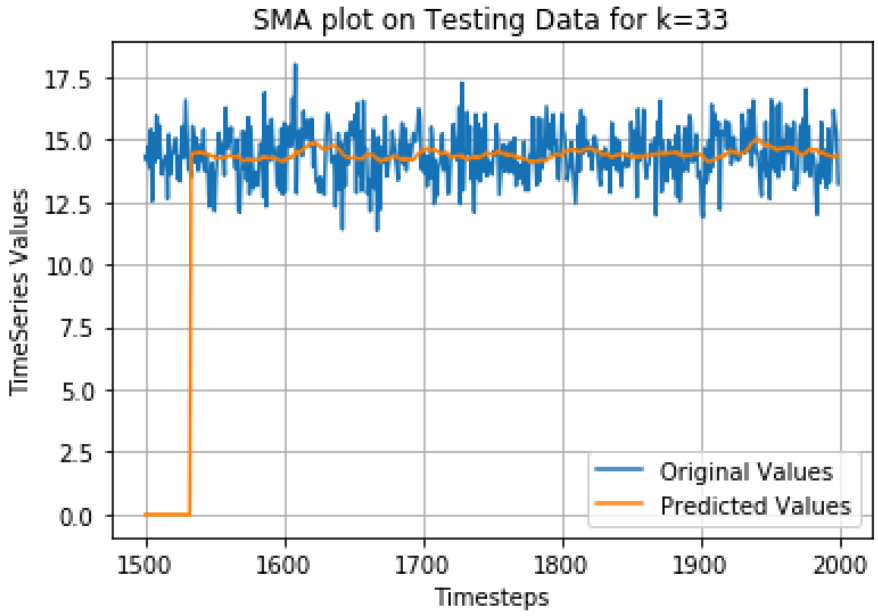
As it can be seen in the above picture, the RMSE seems to decrease with increase in the order size of the model. However, the rate of decrease in RMSE is not significantly less after window size 33.

Choosing a higher order SMA has the tradeoff of higher computational time and the model cannot forecast data until k readings are obtained. Also we bear in mind that we have a test data of only 500 points. Considering these factors, the optimum value of k was chosen to be 33.

The fitted and observed values of the SMA Model with k = 33 for the training data is shown in the following plot. Note that for the first 33 records of the training period, there is no forecast because there are less than 33 past values to average.



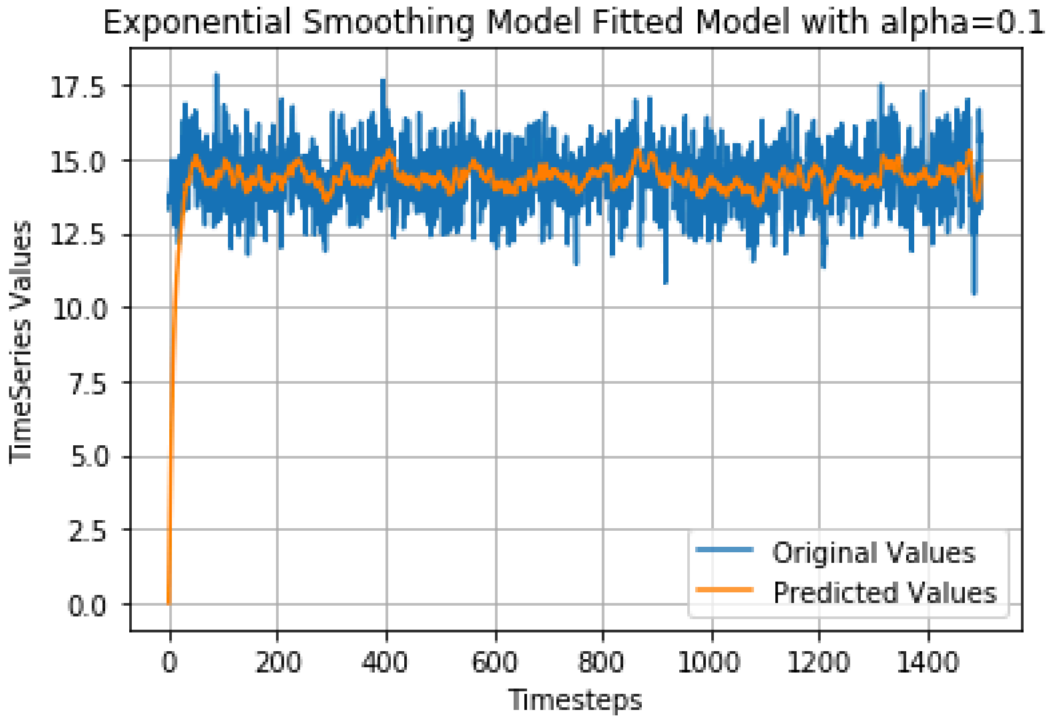
•The fitted values of the trained model follow the peaks and troughs of the observed data.
•For m = 33, the RMSE value of the model for training data was **1.062.**
•The Actual and the predicted time series on the test data are plotted below and we got an RMSE of 1.089.



TASK 2: EXPONENTIAL SMOOTHING MODEL
SECTION 2.1 AND SECTION 2.2:

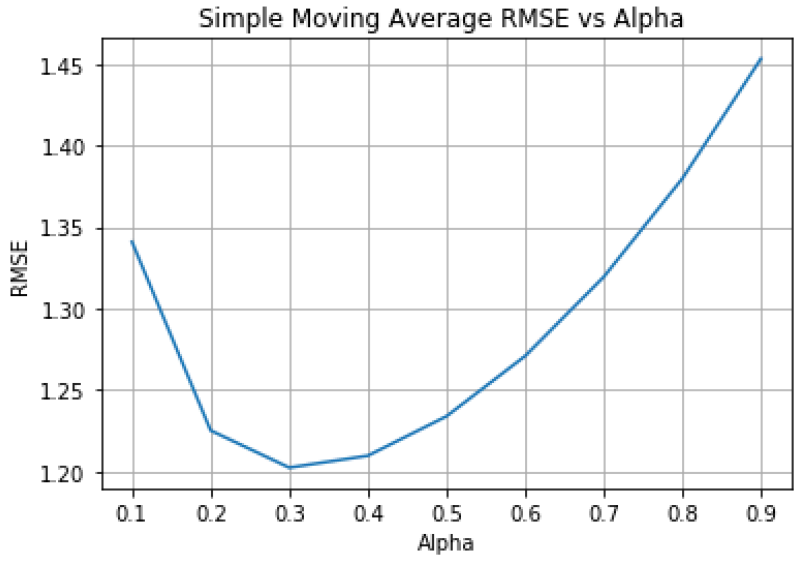
We first fit the exponential smoothing model (ESM) on the training data using an alpha of 0.1 and compute the error and the RMSE.

RMSE for window size alpha=0.1 is: 1.3408732517933069

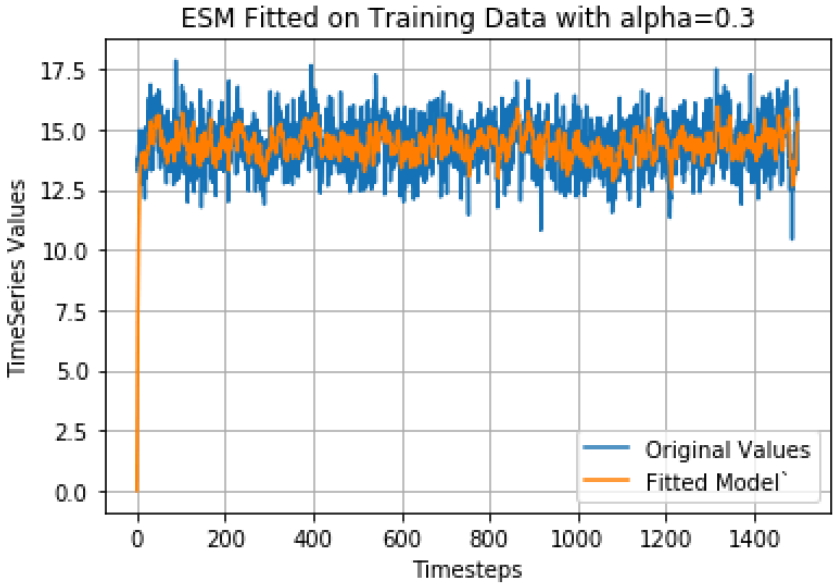


SECTION 1.3, 1.4 AND 1.5:

•Next, exponential Smoothening model was implanted on the training data with α ranging from 0.1 to 0.9 in steps of 0.1.
•For each model, Root Mean Square Error (RMSE) was calculated and the relationship between α and RMSE is plotted.
•As it can be seen in the relationship curve, the RMSE first decrease with increase in α but after α=0.3, RMSE increases. This behavior shows that there is almost equal importance to the historical value and the previous value.

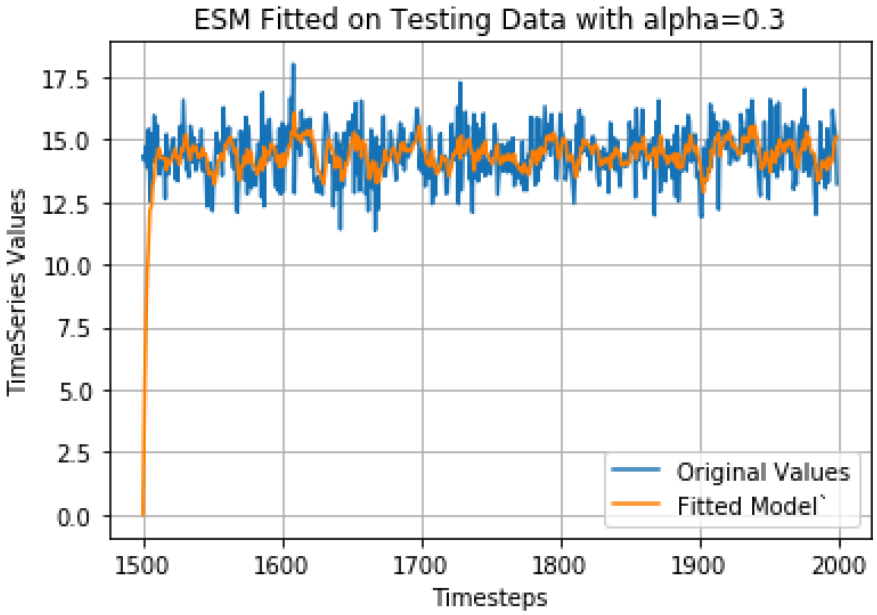


We select optimum alpha=0.3, as RMSE is minimum at this point, i.e. rmse=1.2025. Choosing alpha close to zero indicates slow learning, i.e. past observations have a large influence on the forecast.



Now we fit the exponential smoothing model with alpha=0.3 on the test data and obtain our predictions as shown in the image below.

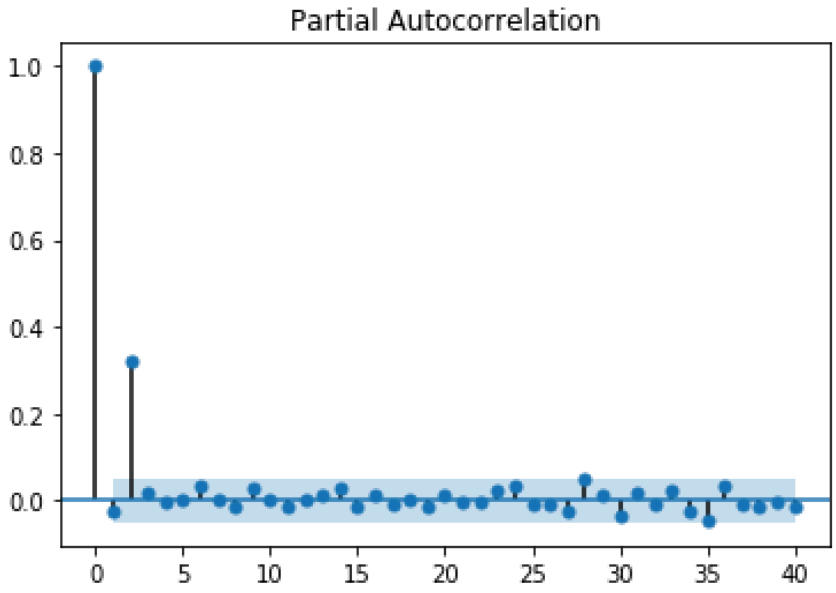
RMSE for window size k=0.3 on testing data was: 1.4358



Note that the first point in the test data is predicted as 0 as per the smoothing model(s_1=0).
The fitted data follows the observed data closely and the variances of the observed and fitted are also similar indicating a good fit.

TASK 4: AUTOREGRESSION MODEL
SECTION 3.1 AND 3.2:

The preliminary step in an autoregressive model is to determine the order P, which can be done in multiple ways such as determining the partial autocorrelation function or by Akaike Information Criterion.
The partial Autocorrelation function was determined and the corresponding correlogram is plotted below.

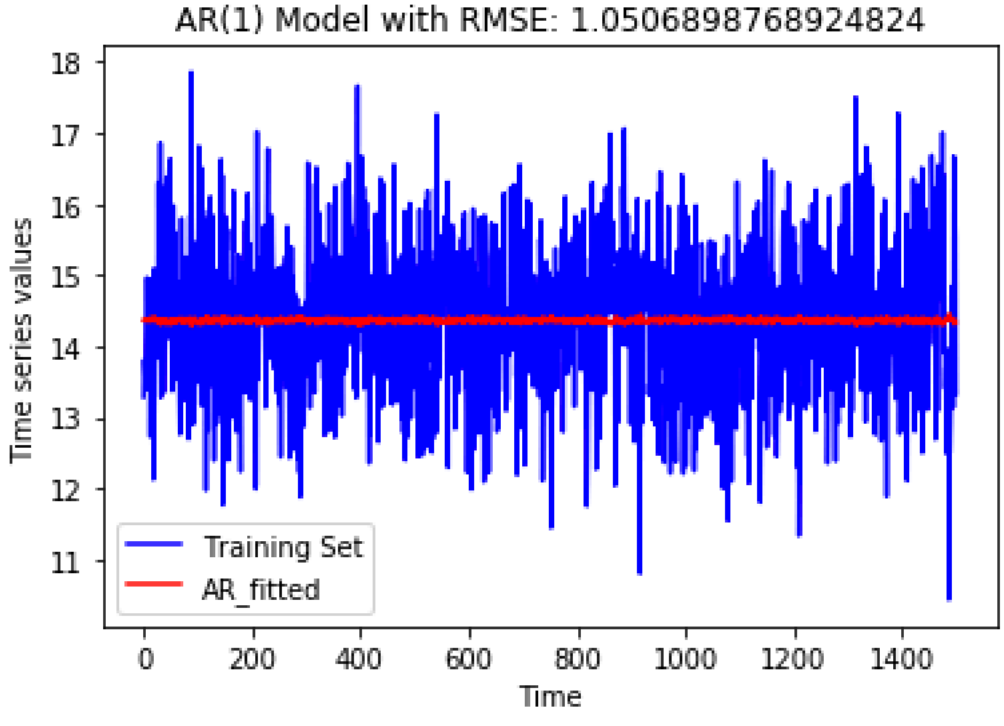


As can be seen in the above correlogram, lag 1 is the most significant as the PACF value of other lags are less than 0.15.
Thus the forecast model is a AR(1) model of form,

$$X_t = aX_{t-1} + \varepsilon_t$$

Where a is the intercept, and ε_t corresponds to a random variable of zero mean and finite variance (Random walk).

The model was made to fit the training data and the value of 'a' was found to be -0.023 while the constant was equal to 14.36.

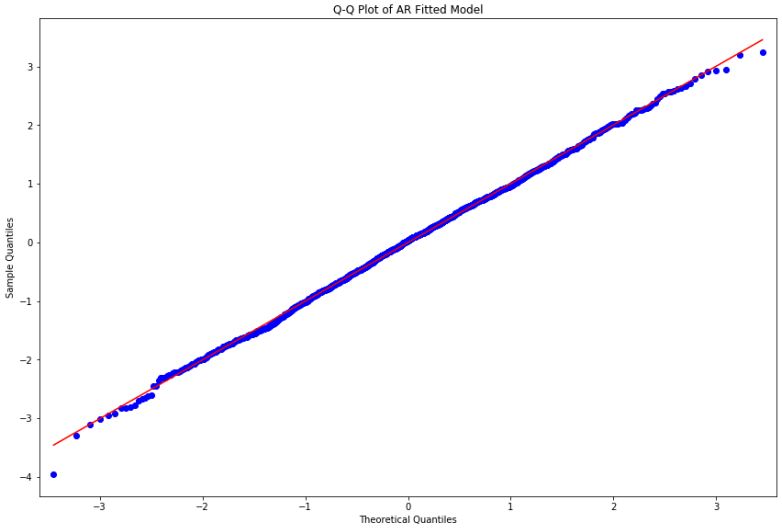


The RMSE value of AR(1) for the training data was 1.051.
The fitted and observed timeseries of the training data fitted to the AR(1) model can be seen in the above plot.

SECTION 3.3: RESIDUAL ANALYSIS FOR THE VALIDITY OF THE MODEL

We next perform several tests to test for the goodness of fit of the fitted AR model. The tests check whether the residuals from the model are normally distributed, have uniform variance and have no trend.

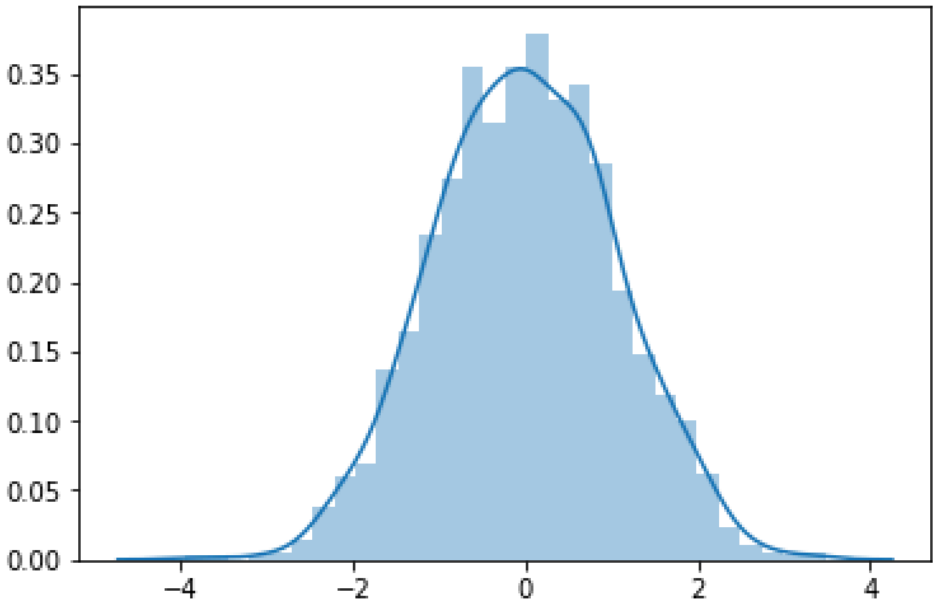
Q-Q plot:



We observe from the QQ-plot that the two distributions(i.e. the normal N(0,var) and the residual distribution) are very similar as almost all quantiles fall on the red line.

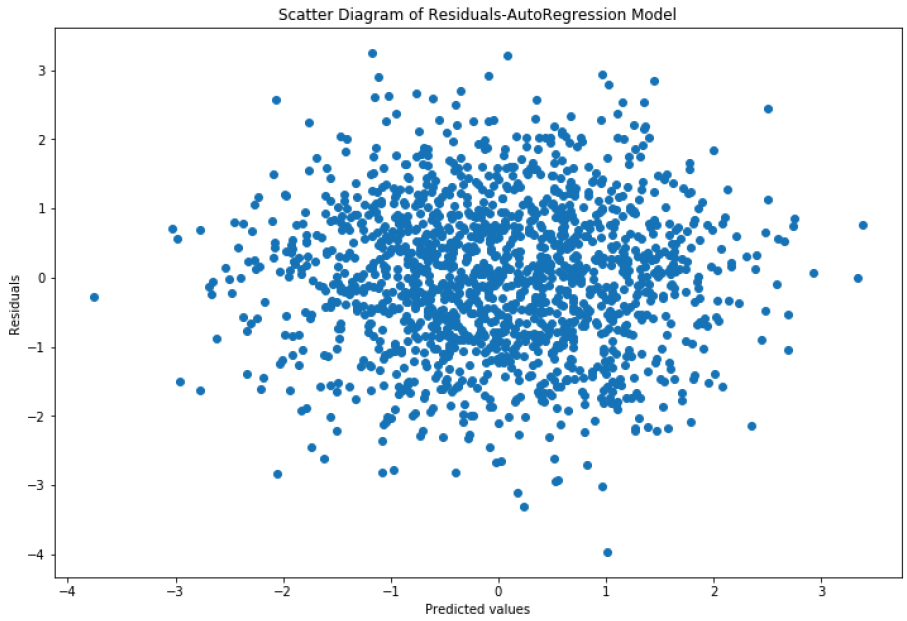
We can use the chi-squared test to verify if the residual distribution comes from the normal distribution. We get p-value of 0.777 which is greater than 0.05. Hence we can fail to reject the null hypothesis and again determine that the residual comes from the normal distribution.

Another way to check the residual distribution is by plotting the histogram of the residual. The below plot confirm that the residuals follows a gaussian distribution and has a standard deviation of 1.0510.

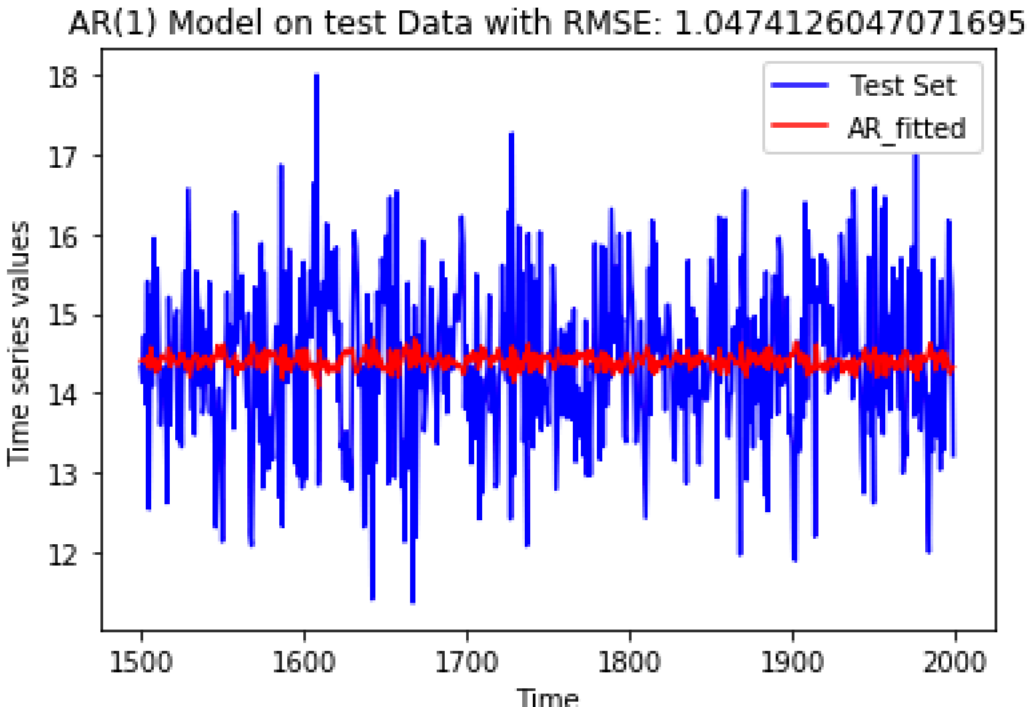


Scatter plot of residuals:

Finally, the scatter diagram of the residual as shown below does not show any trend. This confirms that the AR(1) model is a good fit for our dataset.



We can now use this model to predict on our test data and we get the graph as below where predicted values are show in red.



TASK 5:

CHOOSING THE BEST MODEL

The RMSE values of training and test data for all the three models are tabulated below:

	Simple Moving Average Forecasting	Exponential Smoothing Forecasting	AutoRegression(1) Forecasting
Training Data	1.062	1.2025	1.050
Testing Data	1.089	1.4358	1.047

•From the table above, auto-regressive forecasting has relatively the lowest value of RMSE both in training and test data. Also it can be noticed in the prediction plots of training and test data, that the variance of the fitted data is lesser than the observed value both in training and test data for the autoregressive model.
For obtaining a better average case performance as indicated by RMSE, the autoregressive model may be ideal.

•The Exponential Smoothening model on the other hand performs better in tracking the peak and troughs more effectively. Thus, if the forecasting model is required to perform better in the extremities Exponential Smoothening method can be chosen.