IC 272: DATA SCIENCE - III
LAB ASSIGNMENT – VI
Auto-regression

Student's Name: Aryan Ali **Mobile No: 9027209190**

Roll Number: B20279 **Branch:DSE**

1 a.

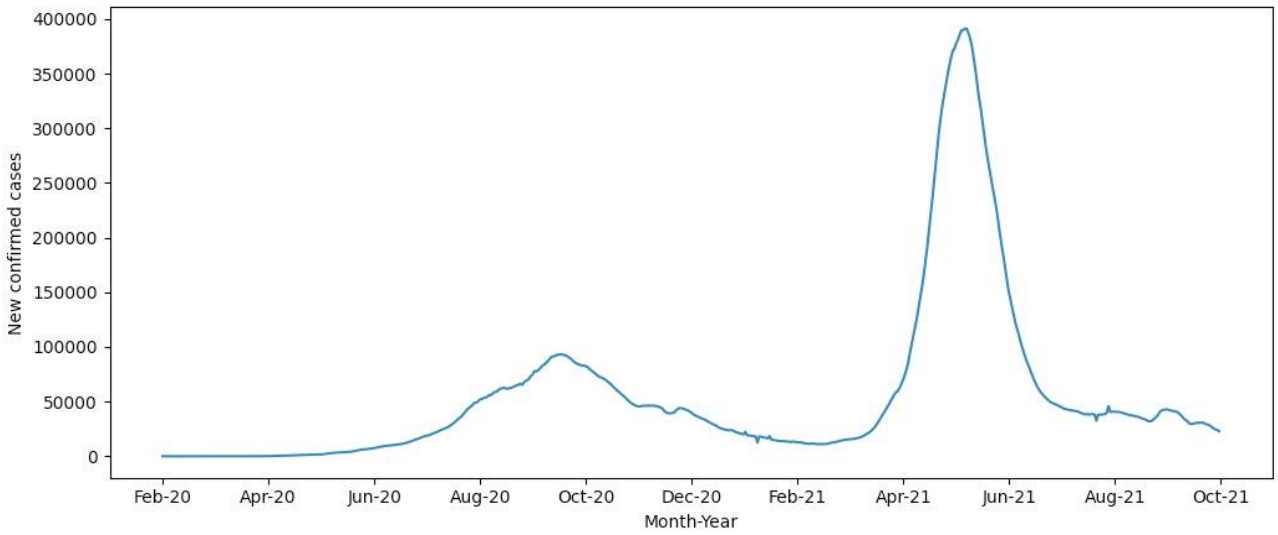

Figure 1 No. of COVID-19 cases vs. days

Inferences:

1. During the first and second wave the cases initially increases and then decreases rapidly but during initial period(Feb-20 to May-20) the cases are more or less the same on consecutive days.
2. From the above plot we can observe that the slope for second wave is highest followed by first wave and almost negligible during the initial period.
3. First wave : Jun-20 - Jan-21, Second wave : Apr-21 - Aug-21.

b. The value of the Pearson's correlation coefficient is 0.999.

Inferences:

1. One day time sequence has a very strong correlation with the given time sequence.

2. Observations on days one after the other are very similar since, the value of correlation coefficient is very high (0.999) it means future observation will be highly dependent on the past observations.
3. From the figure-1 we can observe that there is continuous plot between “number of cases” and “month-years”, so the given series will have high correlation with the lag series.

c.

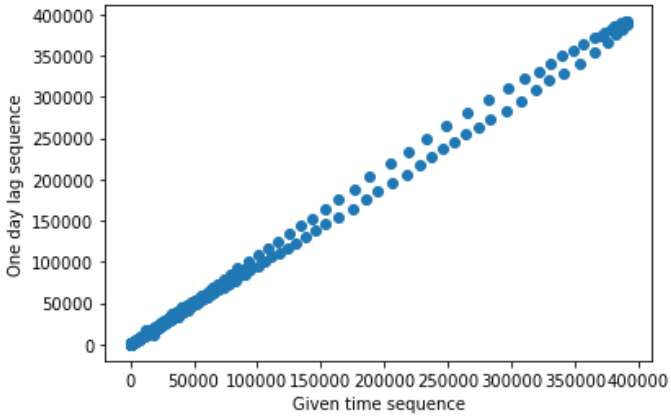

Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. Since the plot is same as a y = x plot, so we can state that there is a strong positive correlation between the given and one day lag time series.
2. Yes, from Pearson coefficient we computed the value 0.999 which validates the above statement.
3. From the figure-1 we can observe that there is continuous plot between “number of cases” and “month-years”, so the given series will have high correlation with the lag series.

d.

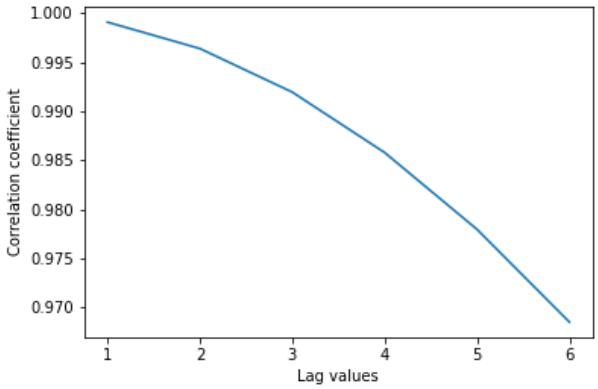

Figure 3: Correlation coefficient vs. lags in the given sequence

Inferences:

1. Correlation coefficients are 0.999, 0.996, 0.992, 0.986, 0.978, 0.969 respectively.
2. With increase in value of lag the correlation coefficient decrease.
3. The current day data is more closely related to the previous day data.

e.

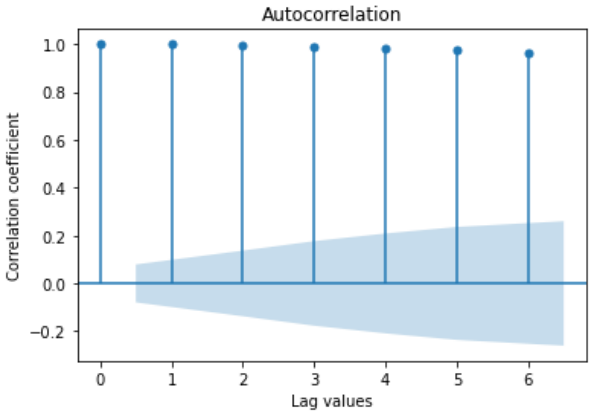

Figure 4 Correlation coefficient vs. lags in given sequencegenerated

using 'plot_acf' function

Inferences:

1. With increase in value of lag correlation coefficient decreases.
2. The current day data is more closely related to the previous day data if we increase the lag, the correlation coefficient will decrease further.

2.

a. The coefficients obtained from the AR model are : 59.955, 1.037, 0.262, 0.028, -0.175, -0.152.

b. i.

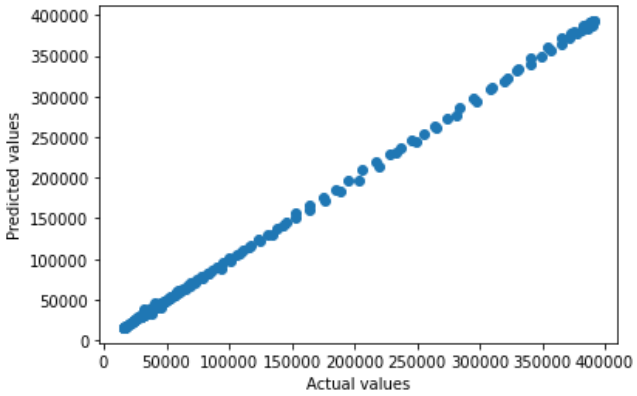

Figure 3 Scatter plot actual vs. predicted values

Inferences:

1. The plot is like a y = x plot, so we can state that the actual values and the predicted values are strongly related.
2. Yes, from Pearson coefficient we computed the value 0.999 which validates the above statement.
3. The given series and lag series were highly correlated therefore we got a much accurate prediction accuracy which we can observe in figure 5.

ii.

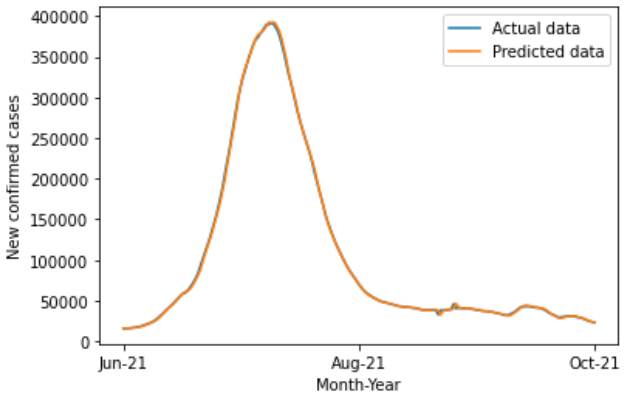

Figure 6 Predicted test data time sequence vs. original test data sequence

Inferences:

1. The model is very highly reliable as from the plot we can see predicted values are quite accurate. The curves for both actual and predicted data are almost overlapping each other.

iii.

The RMSE(%) and MAPE between predicted power consumed for test data and original values for test data are 1.825 and 1.575 respectively.

Inferences:

1. The model is highly accurate.
2. Low value of RMSE and MAPE represents that the difference between actual and predicted data is very small. Which means data predicted by model is almost same as actual data

3.

Table 1 RMSE (%) and MAPE between predicted and original data values wrt lags in time sequence

Lag value	RMSE (%)	MAPE
1	5.373	3.447
5	1.825	1.575
10	1.686	1.519
15	1.612	1.496
25	1.703	1.535

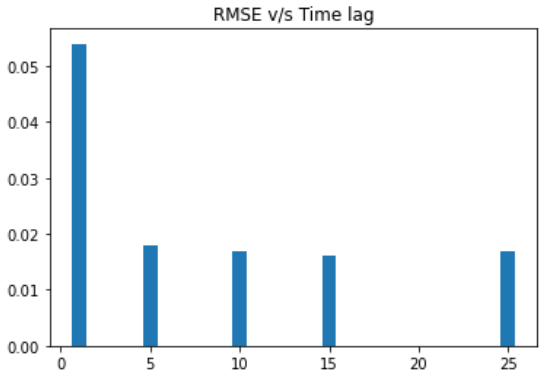

Figure 4 RMSE(%) vs. time lag

Inferences:

1. As the lag increases the RMSE(%) value decreases, therefore the prediction accuracy increases.
2. As the number of lags increases, number of independent variables also increases, giving us a more accurate prediction.

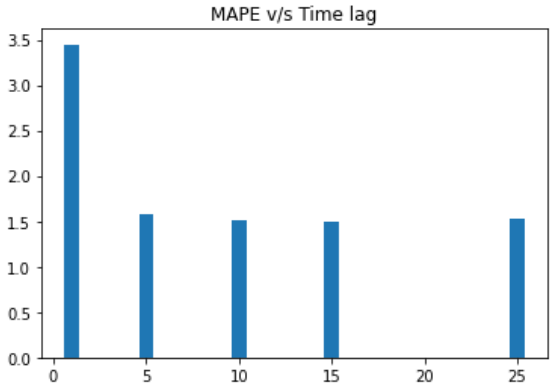

Figure 5 MAPE vs. time lag

Inferences:

1. As the lag increases the MAPE value decreases, therefore the prediction accuracy increases.
2. As the number of lags increases, number of independent variables also increases, giving us a more accurate prediction.

4.

The heuristic value for the optimal number of lags is 77.

The RMSE(%) and MAPE value between test data time sequence and original test data sequence are 1.759 and 2.026 respectively.

Inferences:

1. The RMSE(%) and MAPE values for lag = 77 (heuristic value) is higher when compared to other lag values thus the accuracy is lower.
2. Increasing the lag values increases our prediction accuracy but as we blindly increase the lag value, there might be a case where the tn series is low correlated with t series, so after a certain time lag the accuracy might decrease.
3. The predicted accuracy is low when compared to other lag values.