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

Student's Name: Khushi Ladha **Mobile No: 7665519043**

Roll Number: B20013 **Branch:CSE**

1 a.

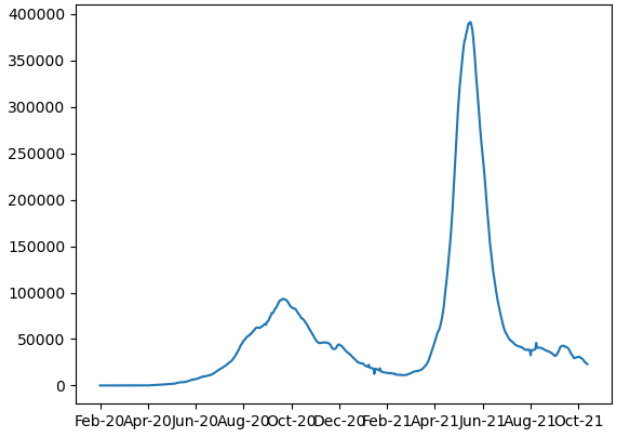


Figure 1 No. of COVID-19 cases vs. days

Inferences:

1. There is similar rise except at the wave points.
2. Cases in nearby future depends on present.
3. 1st wave: August to Oct 2020, 2nd wave: May to July 2021

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

Inferences:

1. Pearson's correlation coefficient, tells about the degree of correlation between the two time sequences that the 2 are associated.
2. We generally expect observations (here number of COVID-19 cases) on days one after the other to be similar. With respect to the value of Pearson's correlation coefficient, this holds very true as the parson correlation coefficient is very high.

3. The reason behind both these is that both have high Pearson correlation coefficient, hence the association is very high.

c.

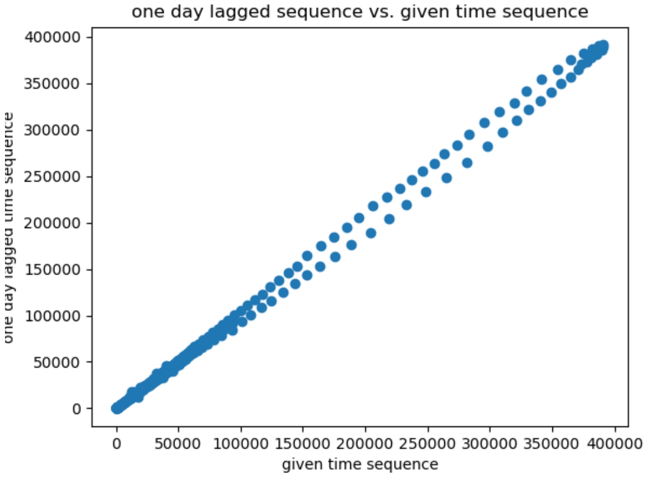


Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. From the nature of the spread of data points, we can infer about the nature of correlation between the two sequences that the points near the lower left are very densely packed and they become less as we move further. The higher density shows stronger correlation. Also, the points are spread around y= x line. This shows that One day lagged and given te sequence ear highly correlated. Also, a clear uphill trend is visible hence; the data are stringy positively related.
2. Yes, the scatter plot seems to obey the nature reflected by Pearson's correlation coefficient calculated in 1.b?
3. The 2 time series value is highly dependent and therefore they have high correlation value. The strog linear relationship can be observed in the scatter plot as well s by the numeric value of PCC.
4.

d.

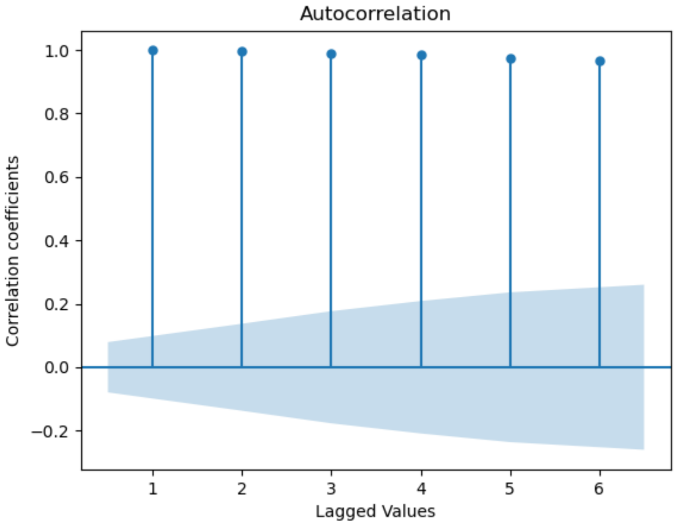


Figure 3 Correlation coefficient vs. lags in given sequence

Inferences:

1. The correlation value gradually decreases with increase in lag values.
2. t series is most dependent on t-1 series, lesser dependent on t-2 series and so on. Hence, as the number of t lag series values increases, the correlation values gradually decreases.

e.

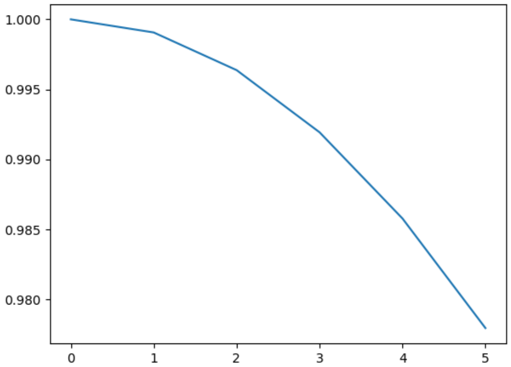


Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. The data isn’t stationary.
2. This is because the blue regions hold the significance threshold. IF lag values are consistently outside the threshold value, then the data is considered to be non stationary.

2

a. The coefficients obtained from the AR model are;

b. i.

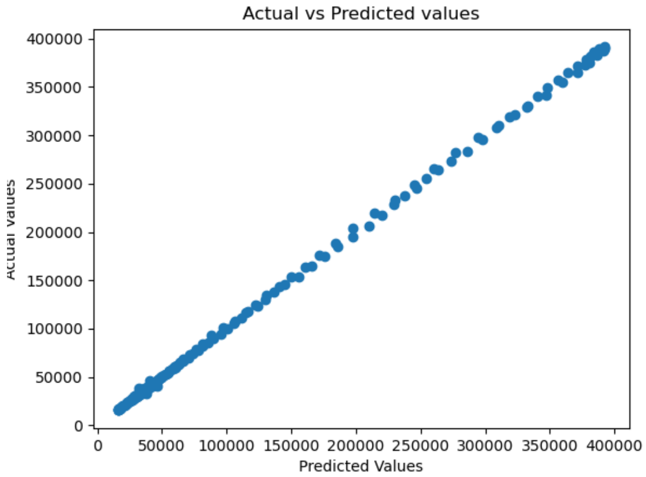


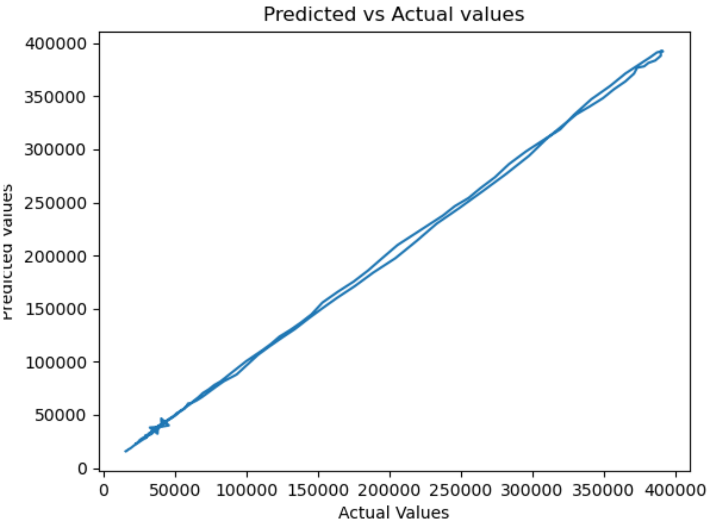
Figure 5 Scatter plot actual vs. predicted values

Inferences:

1. The correlation between the two sequences is highly positively correlated.
2. Yes, the scatter plot seems to obey the nature reflected by Pearson’s correlation coefficient calculated in 1.b?
3. The correlation value obtained was very high. This showed high dependency of previous lags on current time series. Hence, as we predicted the values by previous time lags, it came very close to the predicted values.

ii.

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



Inferences:

1. The model is reliable is for future predictions because actual values and predicted values are very close. Also, as calculated earlier, the correlation values were high and there was linear relationship between t and its lag p series.

iii.

The RMSE (\ %) and MAPE between predicted power consumed for test data and original values for test data are 1.82 and 1.57 respectively.

Inferences:

1. From the value of RMSE(\%) and MAPE value comment how accurate is the model for the given time series is. A MAPE score less than 10 percentages is considered ideal and the RMSE value is also very small, hence the model is suitable for future predictions.
2. The correlation value was high between time and the lag series, hence they are strongly depending. So, when lags series was used to make prediction, it came out to be very close to the actual values.

3

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

Lag value	RMSE (%)	MAPE
0	1	5.37
1	5	1.83
2	10	1.69
3	15	1.62
4	25	1.71

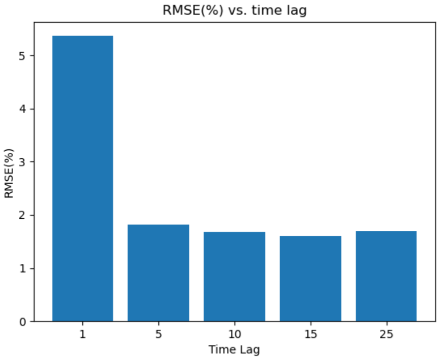


Figure 7 RMSE(%) vs. time lag

Inferences:

1. RMSE has a sharp decrease at p=5, after that decrease becomes gradual. And at p= 25 there is a slight decrease.
2. The reason is that, p =5 turns out to be the optimal value for time lag. After that less weighted time steps start to be a part of the model so if we still keep increasing the lags then accuracy starts to decrease.

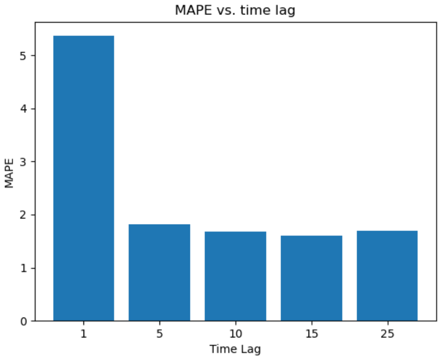


Figure 8 MAPE vs. time lag

Inferences:

1. Similar to RMSE, MAPE has a sharp decrease at p=5, after that decrease becomes gradual. And at p= 25 there is a slight decrease.
2. The reason is that, p =5 turns out to be the optimal value for time lag. After that less weighted time steps start to be a part of the model so if we still keep increasing the lags then accuracy starts to decrease.

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 RMSE(%): 1.76% , MAPE: 2.03%

Inferences:

1. Using heuristics, there wasn't any significant improvement. This can seen y the RMSE and MAPE values.
2. This is because after the significant lag value, RMSE v/s lag doesn't chnag significantly. Infact, the AR(77) is a s good a slag of one day only.
3. The prediction accuracies obtained with the heuristic for calculating optimal lag with respect to RMSE (%) and MAPE values is significantly better without the heuristic.