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

Student's Name: Vikas Dangi **Mobile No: 9406661661**

Roll Number: B20238 **Branch: EE**

1 a.

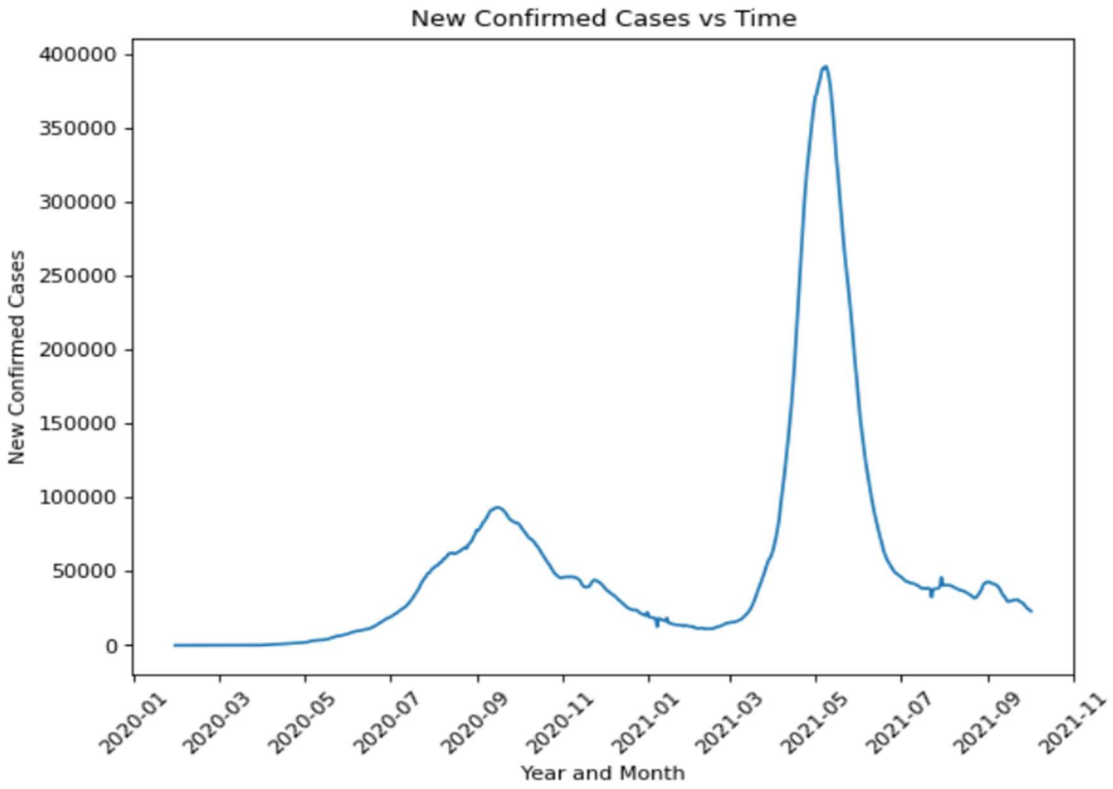

Figure 1 No. of COVID-19 cases vs. days

Inferences:

1. The plot is a continuous plot and we can see the whole trend of covid cases with the first and the second wave.
2. The duration of first is 6 months and second wave is 4 months.

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

Inferences:

1. From the value of Pearson's correlation coefficient, we infer that the degree of correlation between the two-time sequences is very high.
2. Yes, they are similar. It holds to a great extent. The value of correlation coefficient is almost 1.
3. The covid cases does not jump extremely in one day.

c.

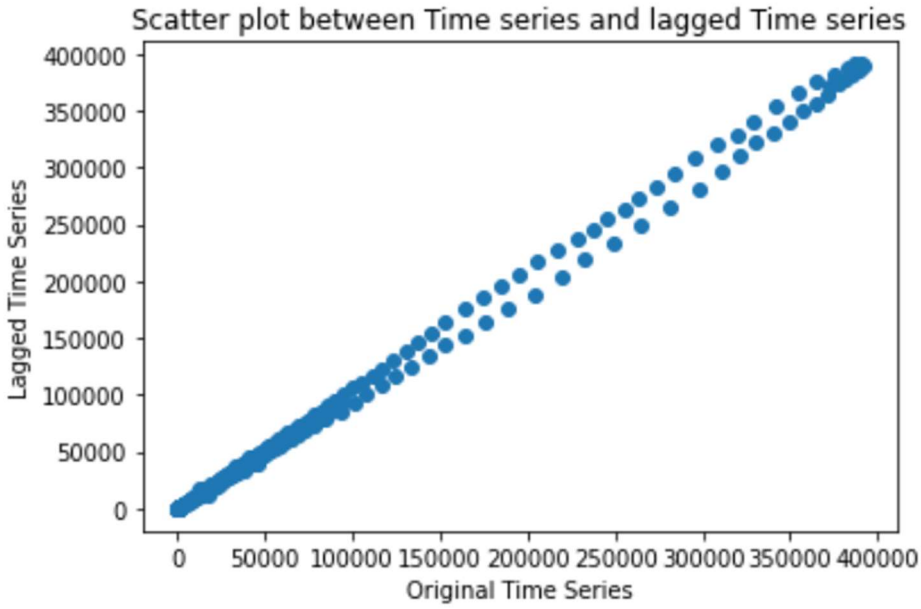

Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. The nature of the spread of data points shows us that the two sequences are highly correlated and increase of one is clearly seen with the other.
2. Yes, the scatter plot seems to obey the nature reflected by Pearson's correlation coefficient calculated in 1.b.
3. The new covid cases depends on the current active cases which is highly proportional to the previous new cases.

d.

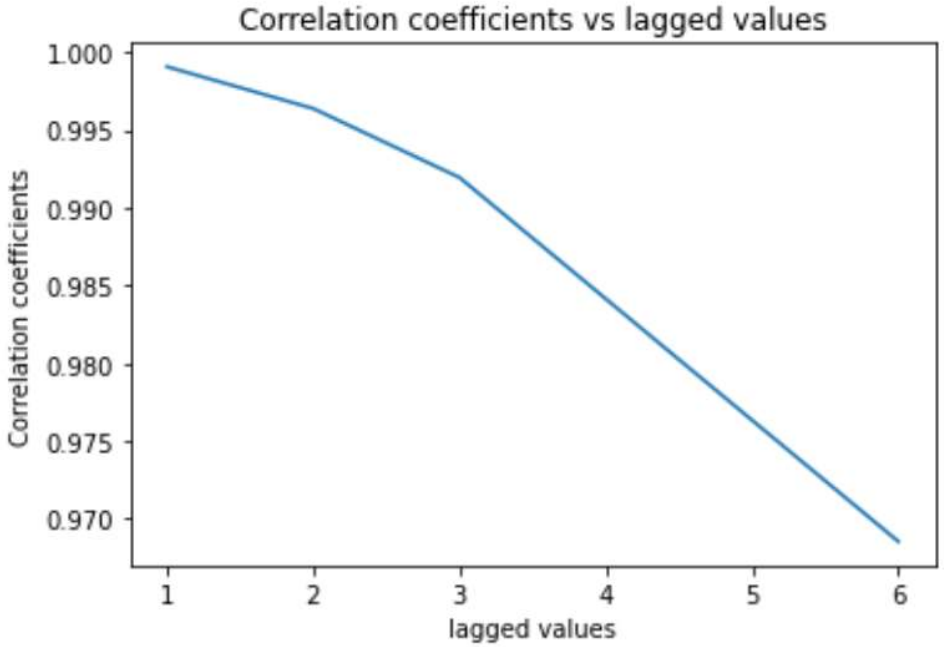

Figure 3 Correlation coefficient vs. lags in given sequence

Inferences:

1. Increase in lags in time sequence reduces the correlation coefficient.
2. The latest data gives more insight on how the things can proceed in case of covid-19.

e.

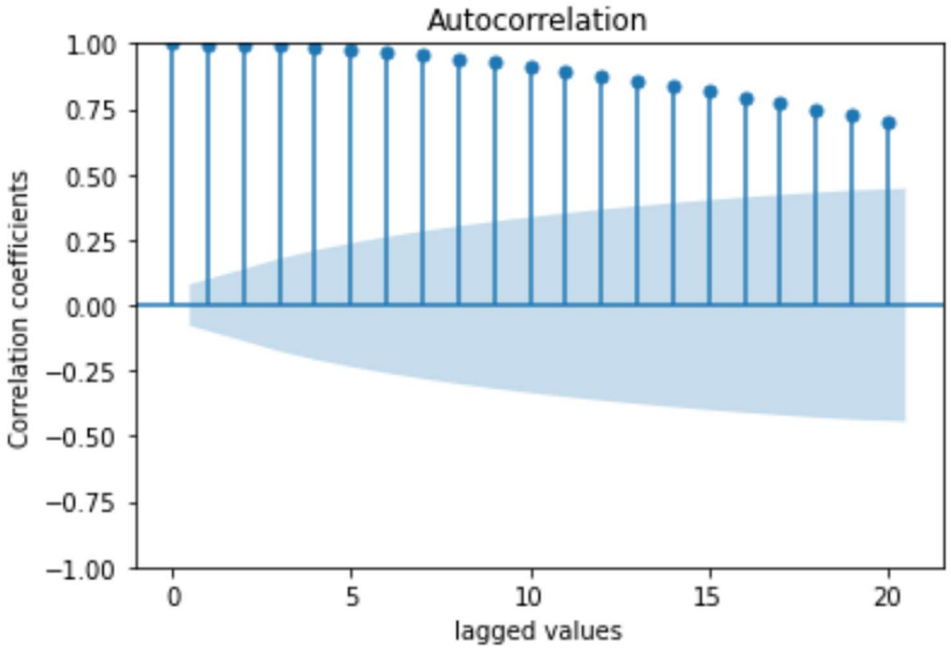

Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. Increase in lags in time sequence reduces the correlation coefficient.
2. The latest data gives more insight on how the things can proceed in case of covid-19.

2

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

b. i.

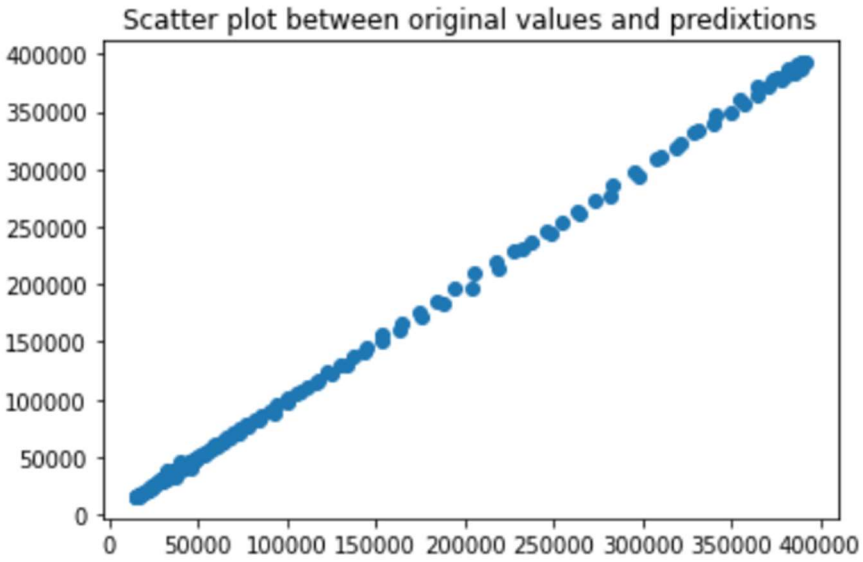

Figure 5 Scatter plot actual vs. predicted values

Inferences:

1. They are highly correlated.
2. Yes, it obeys.
3. The prediction works fine as there is a high dependency of the new cases over the previous cases.

ii.

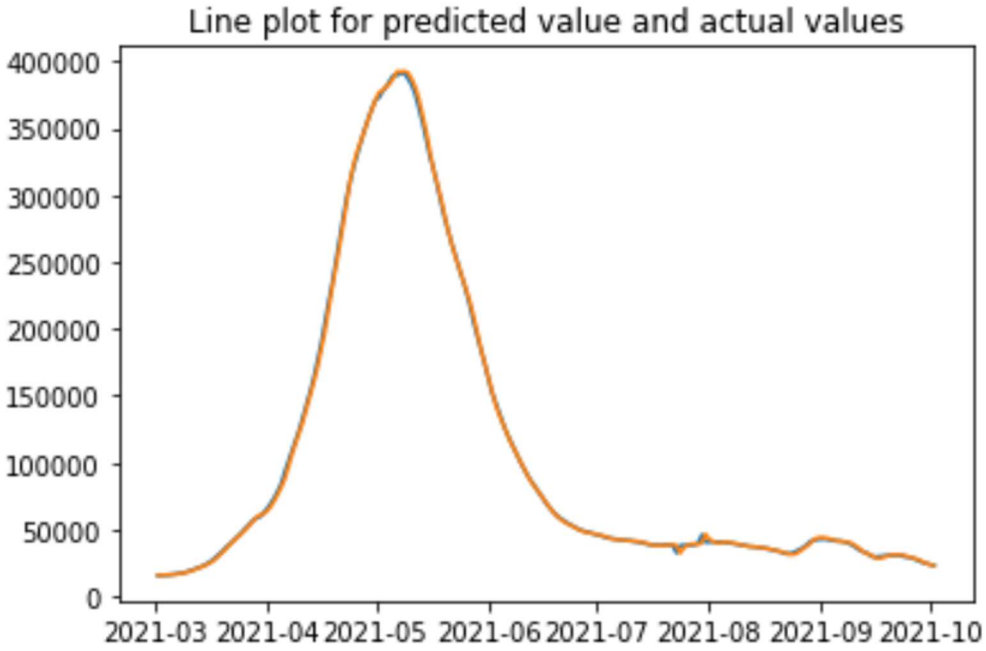

Figure 6 Predicted test data time sequence vs. original test data sequence

Inferences:

1. From the plot of predicted test data time sequence vs. original test data sequence we can say that the model is fairly reliable for future predictions as the predicted values are almost coinciding with the original values.

iii.

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

Inferences:

1. The value of RMSE(\%) and MAPE is quite low which means our model is accurate enough to make us rely on it.

2. The errors are less because there is a high dependency of the new cases over the previous cases. The current active cases decide the further new cases as the infection spreads hence the cases can be predicted.

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

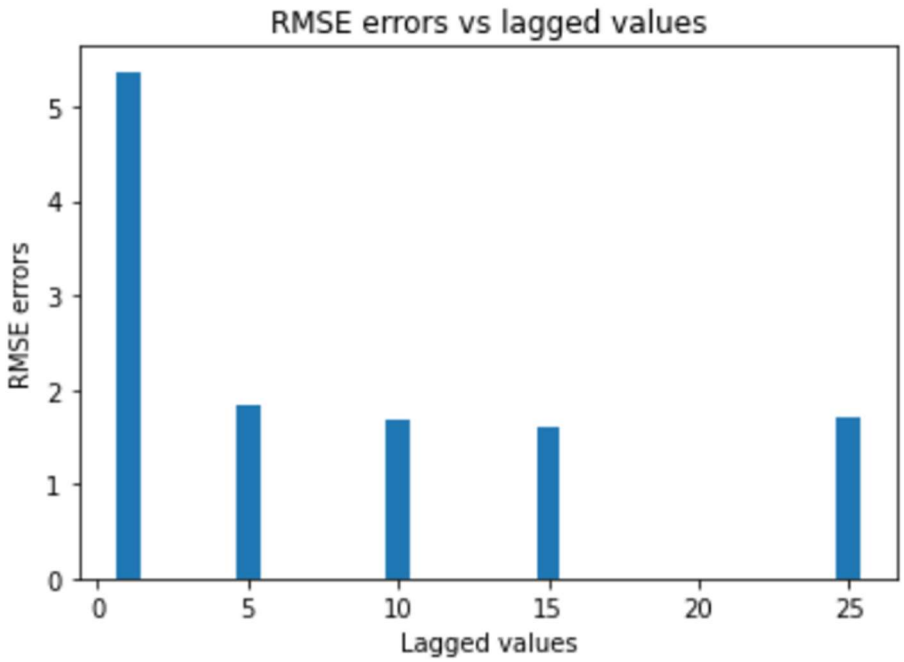

Figure 7 RMSE(%) vs. time lag

Inferences:

1. Increase in lags in time sequence reduces the RMSE error in prediction.

2. The fact that the cases not only depend on the previous day but also on some more of the previous days gives us a better prediction when we consider the cases of more than one day.

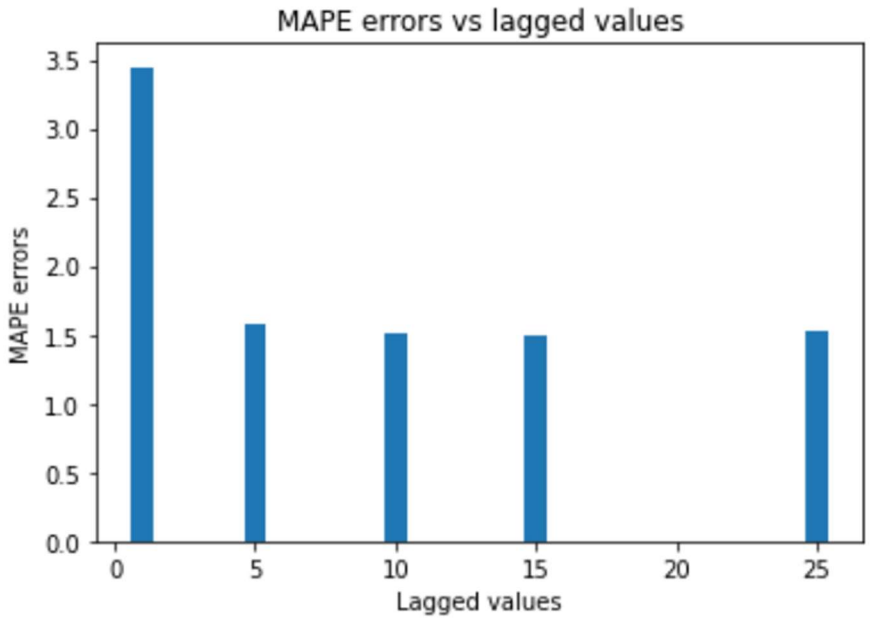

Figure 8 MAPE vs. time lag

Inferences:

1. Increase in lags in time sequence reduces the MAPE error in prediction.
2. The fact that the cases not only depend on the previous day but also on some more of the previous days gives us a better prediction when we consider the cases of more than one day.

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. Based upon the RMSE(%) and MAPE value, heuristics for calculating the optimal number of lags did not really improved the accuracy overall.
2. As we will keep on increasing the lag there has to be some point where it won't be so much depending on the previous 77^{th} value. Also it seems clear that the covid does not really depend much on the daily case of 77 days ago.
3. The prediction accuracies obtained is better if we compare it with lag 1 for MAPE and for RMSE (%) we can even go for lag 2 but later on it deteriorates.

THANK YOU