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

Student's Name: AKSHAR SINGH **Mobile No: 7428357700**

Roll Number: B20147 **Branch: COMPUTER SCIENCE**

1 a.

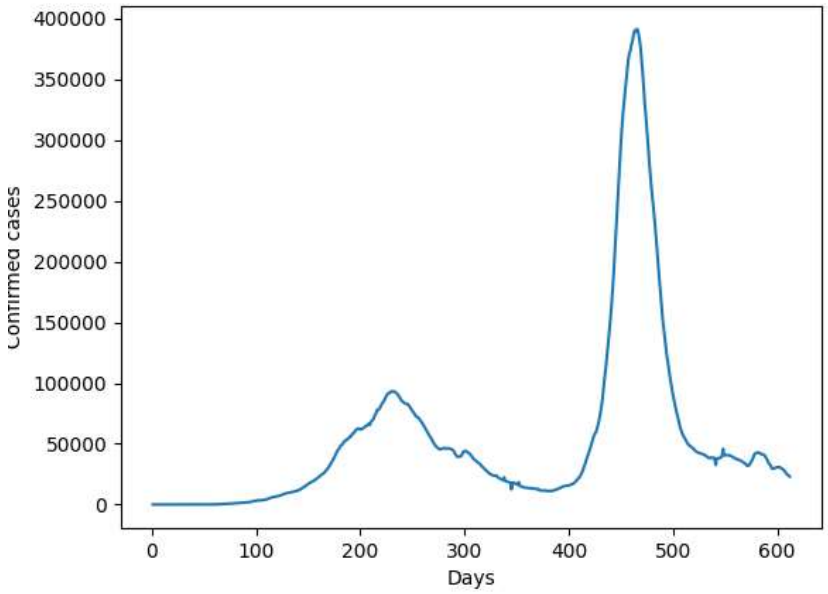


Figure 1 No. of COVID-19 cases vs. days

Inferences:

1. The second wave is more severe than first wave
2. The duration of covid cases in first wave is about 200 days and for second wave is about 100 days.

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

Inferences:

1. The two time sequences are highly correlated.

2. The value of Pearson's correlation coefficient is 0.999 which tells that the cases on days one after the other are similar.
3. The reason of the above two is that in real life changes happen very slowly and over a period of time.

c.

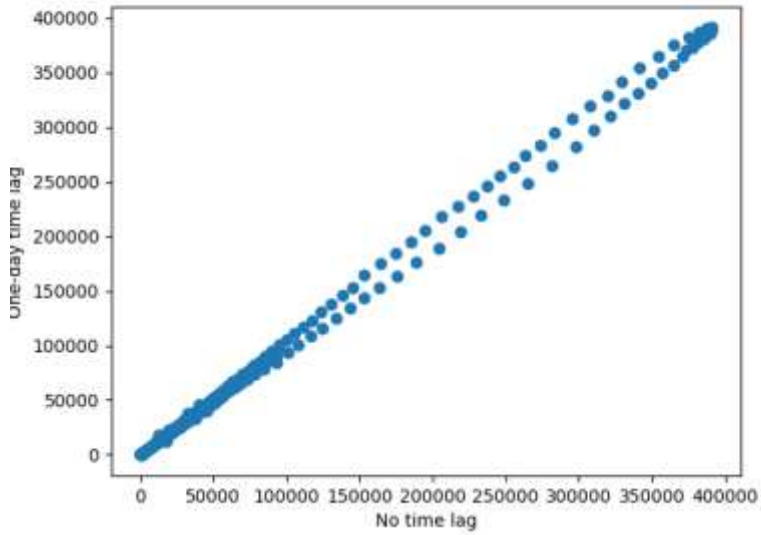
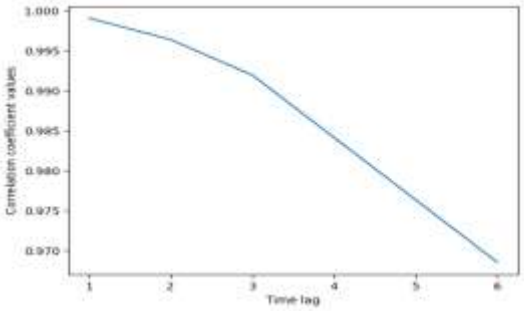


Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. The two time sequences are highly correlated.
2. The scatter plot obey the nature reflected by Pearson's correlation coefficient as the graph is pretty much linear.

d.


1.000
0.995
0.990
0.985
0.980
0.975
0.970
Correlation coefficient values
1
2
3
4
5
6
Time lag

Inferences:

1. The correlation coefficient value is decreasing with increase in lag.
2. This happens because the data becomes less dependent on more time lag data.

e.

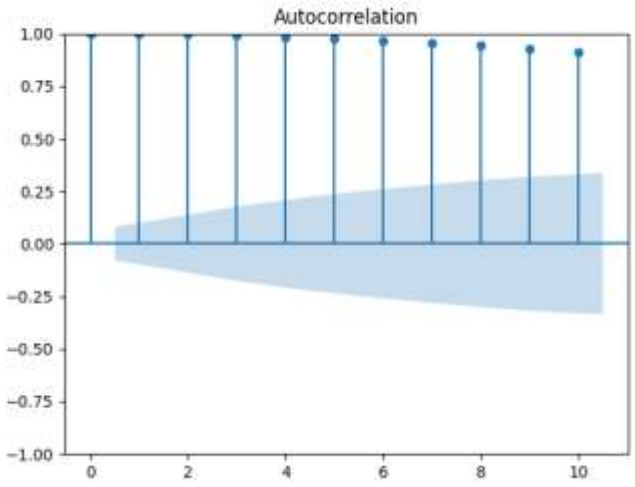


Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. The correlation coefficient value is decreasing with increase in lag.
2. This happens because the data becomes less dependent on more time lag data.

2

a. The coefficients obtained from the AR model are- 5.99548333e+01 1.03675933e+00 2.61712336e-01 2.75612628e-02

-1.75391955e-01 -1.52461366e-01

b. i.

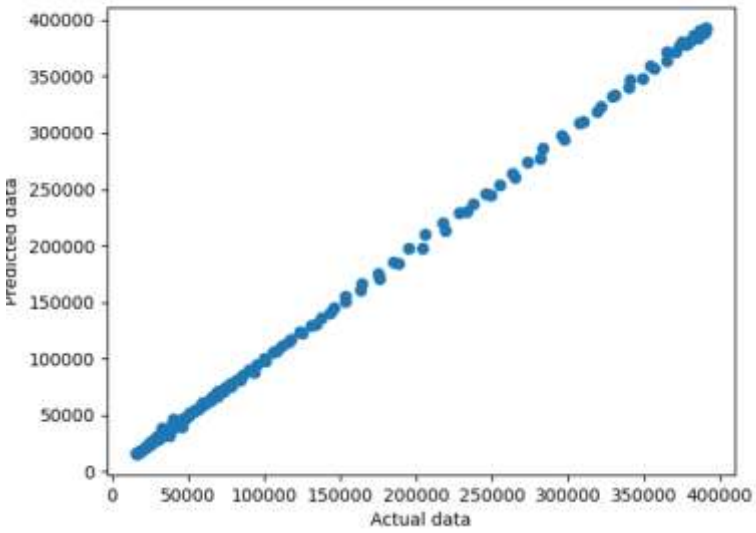


Figure 5 Scatter plot actual vs. predicted values

Inferences:

1. The two sequences are highly correlated.
2. The scatter plot seem to obey the nature reflected by Pearson's correlation coefficient calculated in 1.b because the two series are highly correlated and hence they are almost the same.

ii.

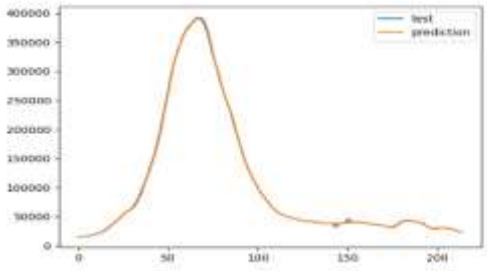


Figure 6 Predicted test data time sequence vs. original test data sequence

Inferences:

1. The predicted test data seems to overlap the original test data which shows that the model is very accurate for future predictions.

iii.

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

Inferences:

1. The value of RMSE(\%) and MAPE value is very low which tells that the model is very accurate .

3

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

Lag value	RMSE (%)	MAPE
1	5.373	3.446
5	1.825	1.575
10	1.689	1.519
15	1.612	1.496
25	1.703	1.535

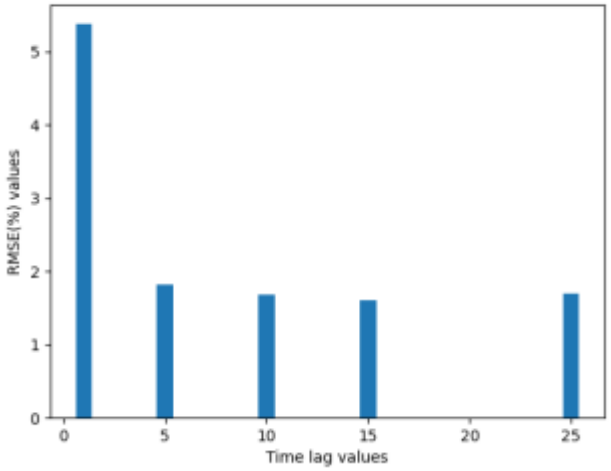


Figure 7 RMSE(%) vs. time lag

Inferences:

1. The RMSE(%) values decreases with increase in lags in time sequence due to more number of terms in regression though it increases from 15 to 25 which could be due to overfitting.

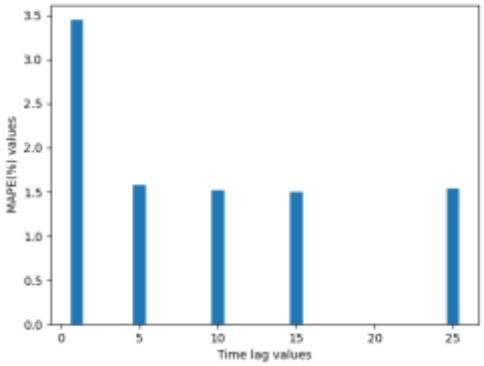


Figure 8 MAPE vs. time lag

Inferences:

1. The MAPE values decreases with increase in lags in time sequence due to more number of terms in regression though it increases from 15 to 25 which could be due to overfitting.

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 optimal value somewhat kept the accuracy more or less the same mainly because of negative effect of overfitting.
2. The prediction accuracies obtained without and with the heuristic for calculating optimal lag with respect to RMSE(%) and MAPE values are almost same .