IC 272: DATA SCIENCE - III

LAB ASSIGNMENT – VI

Auto-regression

Student's Name: PRIYANKA KUMARI **Mobile No: 8328354314**

Roll Number: B20307 **Branch: ELECTRICAL ENGINEERING**

1 a.

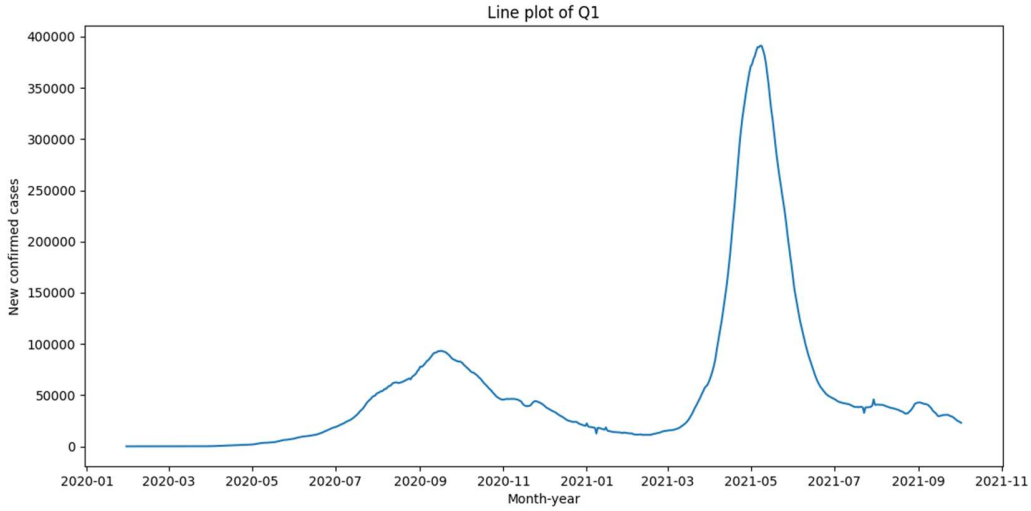


Figure 1 No. of COVID-19 cases vs. days

Inferences:

1. From the plot we can see that the number of cases rise is same during month gap except at the time of first and second covid.

2. It can be seen from the graph that the number cases in coming days depend on the present number of cases.

3. The first wave starts around August 2020 and lasts till October 2020 while the second wave is at its peak around May 2021.

b. The value of the Pearson's correlation coefficient is 0.9990

Inferences:

1. From the value of correlation we can see that the degree of correlation between lag and actual is very strong

2. We generally expect observations on days one after the other to be similar. It holds to a very good extent here as we can see that the correlation coefficient is almost 1.

3. as we have taken assumption that the observations at lag time steps are useful to predict the value at the next time step is true in this case

c.

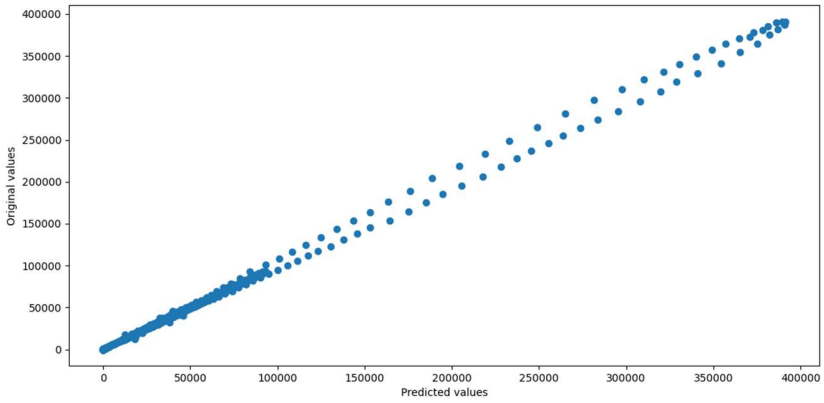


Figure 2 Scatter plot one day lagged sequence vs. given time sequence

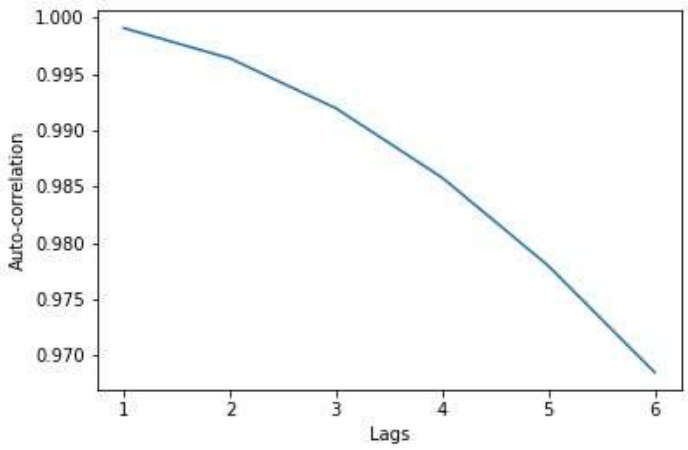
Inferences:

1. From the nature of the spread of data points, the nature of correlation between the two sequences is very high but not perfect 1.

2. yes, the scatter plot obeying the nature reflected by correlation coefficient calculated in 1.b in the start and towards the end of the graph.

3. It is because of when the first and second wave came the number of observations rises so much within less days weakening the correlation between previous and current observations.

d.



Inferences:

1. The correlation coefficient value decreases gradually with respect to increase in lags in time sequence.

2. as we keep increasing the lag, number of possible matches decreases .

e.

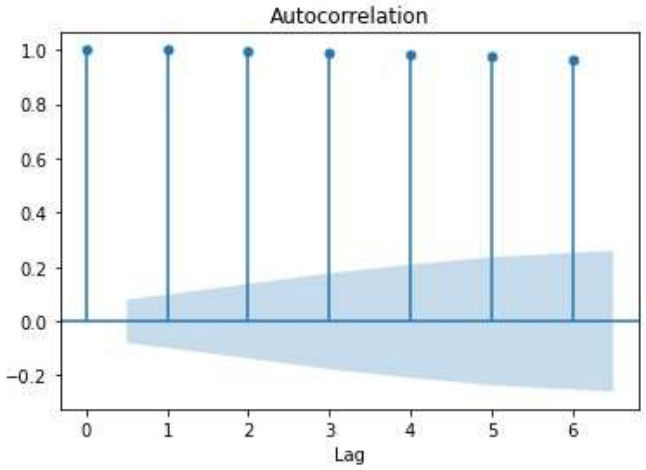


Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. The correlation coefficient value decreases gradually with respect to increase in lags in time sequence.

2. as we keep increasing the lag the number of possible matches decreases.

2

a. The coefficients obtained from the AR model are; [5.9955, 1.037, 2.62, 2.758, -1.75, -1.52] **b.**

i.

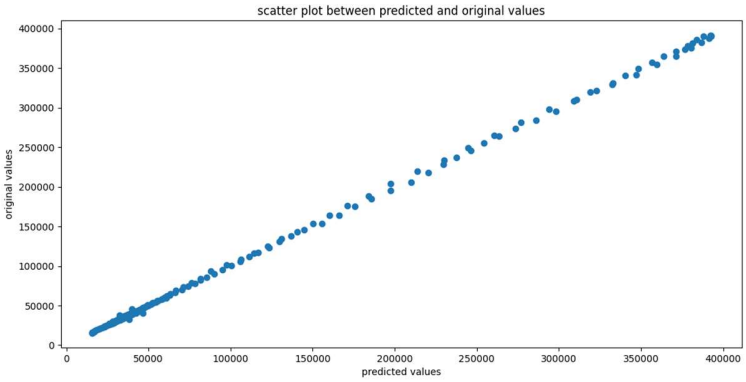


Figure 5 Scatter plot actual vs. predicted values

Inferences:

1. From the nature of the spread of data points, the nature of the correlation between the two sequences is very strong.

2. Yes the scatter plot seem to obey the nature reflected by Pearson's correlation coefficient calculated in 1.b .

3. As the lag is increased, more variables are added to our regression model and it inherently improves the fit.

ii.

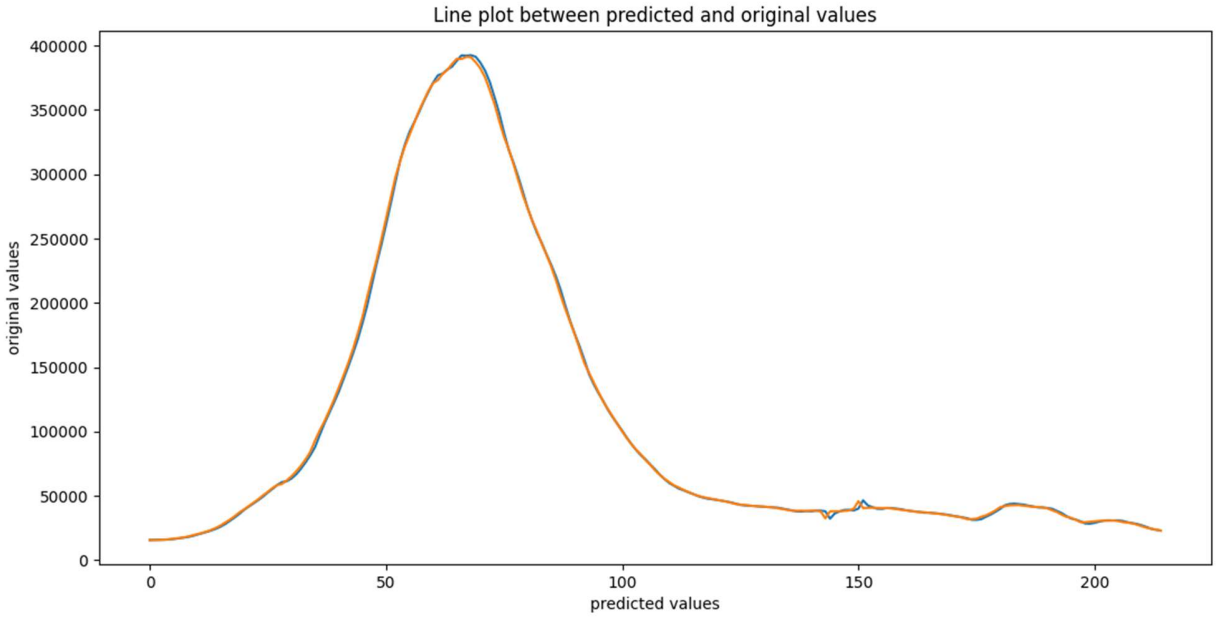


Figure 6 Predicted test data time sequence vs. original test data sequence

Inferences:

1. From the plot of predicted test data time sequence vs. original test data sequence our model is not that reliable for future predictions because even if it is giving quite a good accuracy but there's still a scope of improvement further.

iii.

The RMSE(\%) and MAPE between predicted power consumed for test data and original values for test data are 1.824,1.574 respectively.

Inferences:

1. From the value of rmse and mape value our model is so much that accurate for the given time series.
2. because if we further increase the p, the RMSE will decrease which means that there exists a more accurate model.

3

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

Lag value	RMSE (%)	MAPE
1	5.372	3.446
5	1.824	1.574
10	1.685	1.519
15	1.611	1.496
25	1.703	1.535

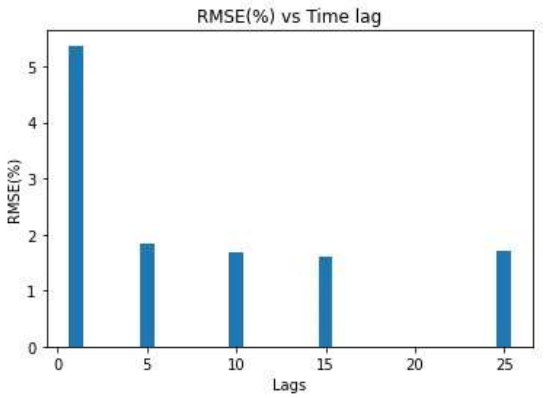


Figure 7 RMSE(%) vs. time lag Inferences:

1. The rmse percentage decreases quickly from 1 to 5 but then decreases from 5 to 15 and from 15 to 25 increases with respect to increase in lags in time sequence.

2. It is because a complex model is needed to fit our data more accurately so when the lag is increased from 1 to 5 the accuracy improves significantly but then the increase in accuracy in gradual

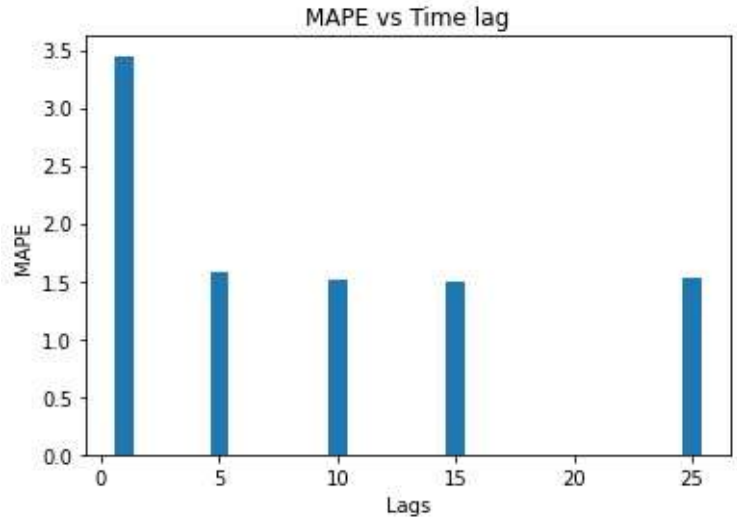


Figure 8 MAPE vs. time lag **Inferences:**

1. The MAPE decreases quickly from 1 to 5 but then decreases gradually with respect to increase in lags in time sequence.

1. It is because a complex model is needed to fit our data more accurately so when the lag is increased from 1 to 5 the accuracy improves significantly but then the increase in accuracy in gradual.

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 1.759,2.026 respectively.

Inferences:

1. Based upon the rmse(%) and mape value, the heuristics for calculating the optimal number of lags didn't improve the prediction accuracy of the model as much we expected as we can see the RMSE(%) for lag=15 was less than that for optimal lag.

2. Because as we keep increasing the lag, after certain time the pattern of rmse vs lag will become random and we can also see that as the observations are made for every day AR(77) doesn't make sense than that of a lag of around one day.

3. The prediction accuracies obtained without heuristic is less as compared to with heuristic for calculating optimal lag with respect to rmse(%) and mape values.