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

Student's Name: Aditi Singh

Mobile No: 9005943744

Roll Number: B20272

Branch: CSE

1 a.

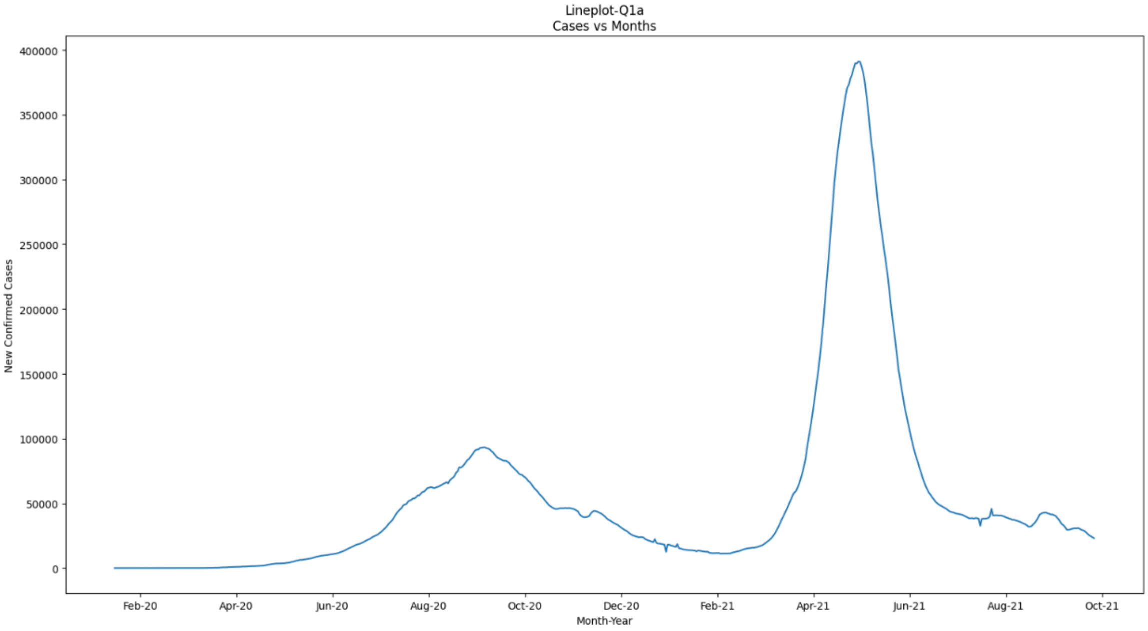

Figure 1 No. of COVID-19 cases vs. days

Inferences:

1. From the plot, we can infer that the days one after the other have similar power consumption.
2. It is clearly evident 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 0.999

Inferences:

1. From the value of Pearson’s correlation coefficient, we can infer a strong degree of correlation between the two time sequences.
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 that is 1, we can infer that the statement holds well.
3. Because our assumption that the observations at previous time steps are useful to predict the value at the next time step is true in this case.

c.

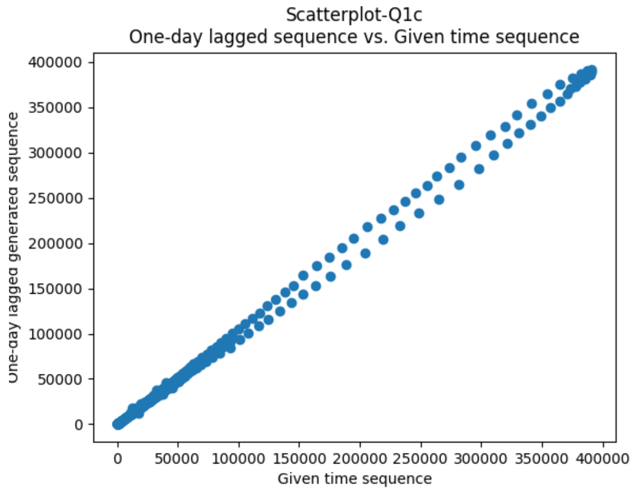

Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. From the nature of the spread of data points, nature of correlation between the two sequences is very high but not perfect zero.
2. Yes, the scatter plot seems to obey the nature reflected by Pearson’s 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 rose abruptly weakening the correlation between previous and current observations.

d.

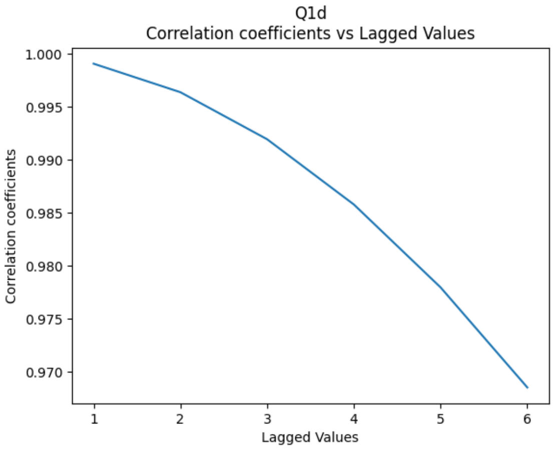

Figure 3 Correlation coefficient vs. lags in given sequence

Inferences:

1. The correlation coefficient value decreases gradually with respect to increase in lags in time sequence.
2. Because as we keep increasing the lag the number of possible matches decreases because the series “hang out” at the ends and do not overlap.

e.

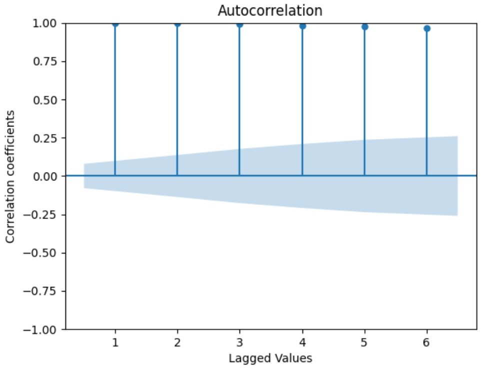

Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. The correlation coefficient value decreases gradually with respect to increase in lags in time sequence.
2. Because as we keep increasing the lag the number of possible matches decreases because the series “hang out” at the ends and do not overlap.

2

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

b. i.

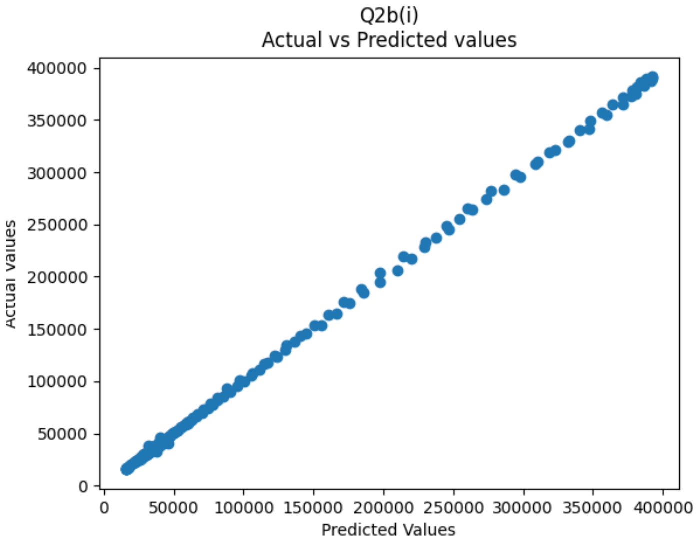

Figure 5 Scatter plot actual vs. predicted values

Inferences:

1. 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 seems 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.

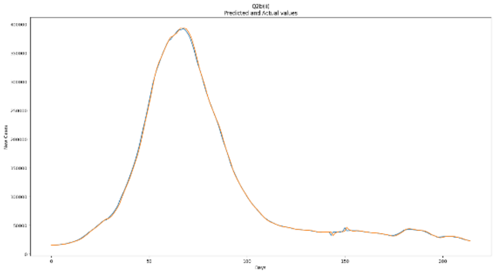

Figure 6 Predicted test data time sequence vs. original test data sequence

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 1.825 and 1.575 respectively.

Inferences:

1. From the value of RMSE(\%) and MAPE value our model is not that accurate for the given Cme 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	**5.373**	**3.447**
5	**1.825**	**1.575**
10	**1.686**	**1.519**
15	**1.612**	**1.496**
25	**1.703**	**1.535**

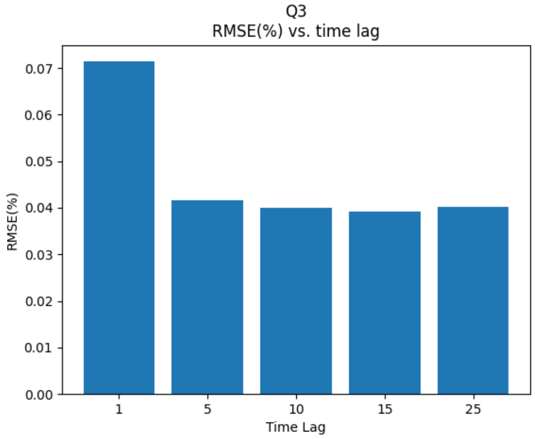

Figure 7 RMSE(%) vs. time lag

Inferences:

1. The RMSE(%) decreases quickly from 1 to 5 but then decreases gradually with respect to increase in lags in time sequence.
2. 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.

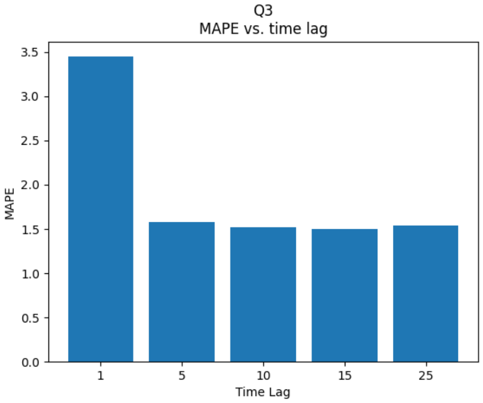

Figure 8 MAPE vs. time lag

Inferences:

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

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.

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, the heuristics for calculating the optimal number of lags didn’t improve the prediction accuracy of the model significantly as we can see the RMSE(%) for lag=10 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.
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.