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

Student's Name: Shubham Shukla **Mobile No: 8317012277**

Roll Number: B20168 **Branch: CSE**

1 a.

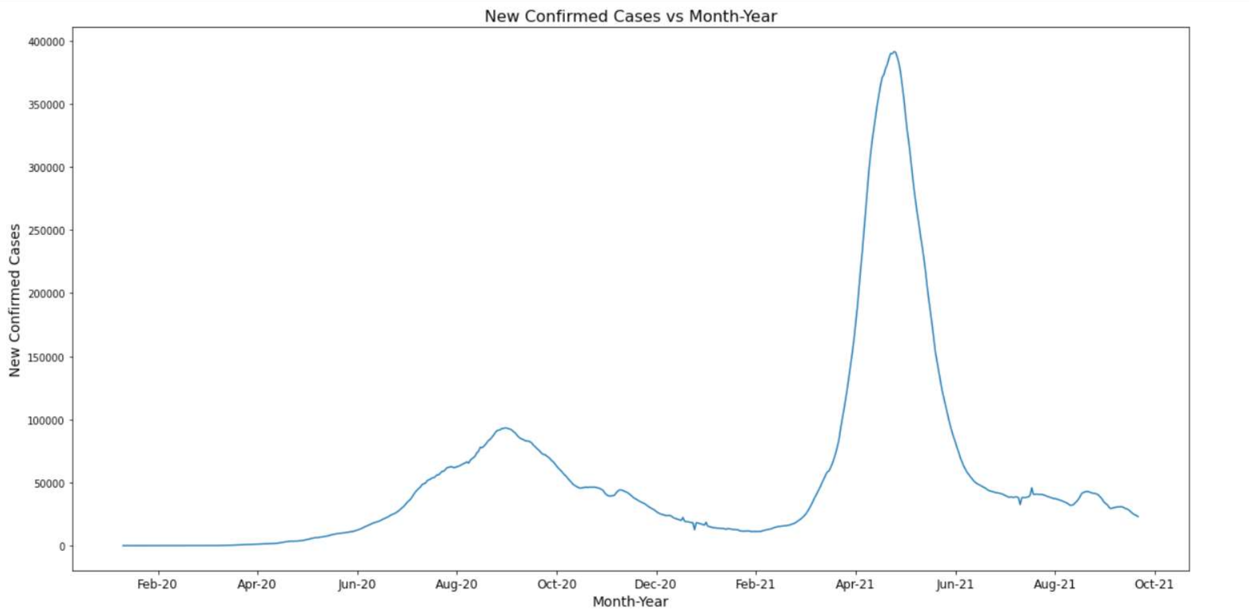

Figure 1 No. of COVID-19 cases vs. days

Inferences:

1. Yes, they have almost equal power consumption.
2. The confirmed cases are having less change number for consecutive days i.e. there is not sudden increase wrt consecutive days.
3. Duration of first covid wave :July 2020 to Dec 2020.
   Duration of second covid wave : April 2021 to June 2021.

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

Inferences:

1. They are having highly correlated.
2. Our expectation that the cases on days one after the other to be similar is highly correct.
3. Since we are studying on a large dataset hence an abrupt change is going to occur very rarely.

c.

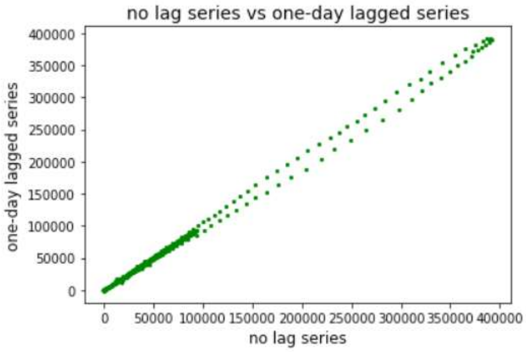

Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. There must be strong correlation.
2. Yes, it correlation coefficient is matching with graph.
3. Since if the value of the correlation coefficient is very high or low then we will directly analyze the correlation on seeing the graph.

d.

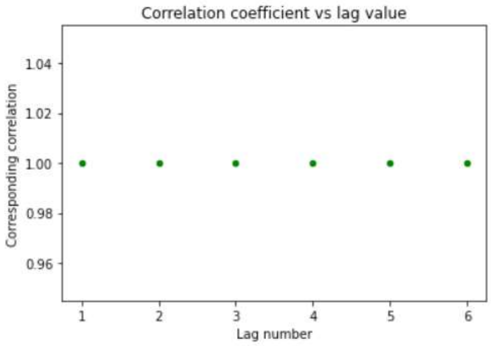

Figure 3 Correlation coefficient vs. lags in given sequence

Inferences:

1. Not much changes for low lags.
2. Since the data is large so abrupt change can’t be seen in a short span of time.

e.

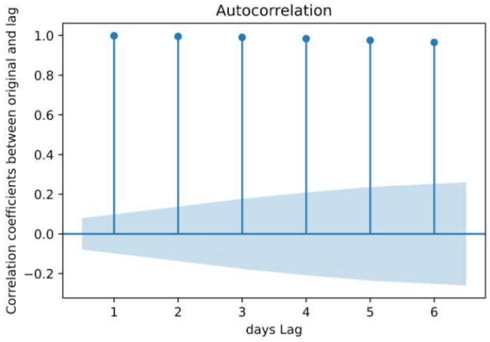

Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. Very slightly decrease in correlation coefficient value with respect to lags in time sequence.
2. As the lag time increases then we the situation maybe changed to some large extent.

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.

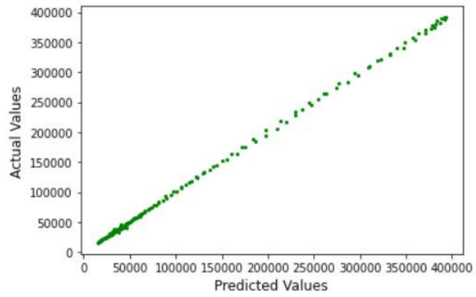

Figure 5 Scatter plot actual vs. predicted values

Inferences:

1. There must be strong positive correlation
2. Yes, it is obeying the calculated correlation coefficient.
3. Since there is strong positive autocorrelation for lag 5 series and that's what we are using here so we are getting a strong positive correlation here.

ii.

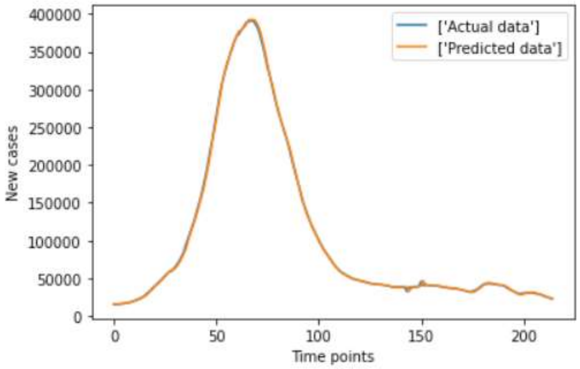

Figure 6 Predicted test data time sequence vs. original test data sequence

Inferences:

1. We can see that both are just overlapped so we can say that we have a very good model to predict the data so we can very much rely on the model.

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. Seeing very small value of RMSE(\%) and MAPE value comment, our model is very much accurate.
2. This is because there was very high correlation in between the lag series for the corresponding 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	1.685	1.519
15	1.612	1.496
25	1.703	1.535

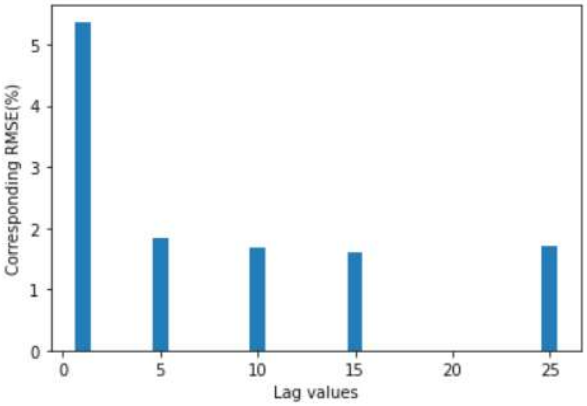

Figure 7 RMSE(%) vs. time lag

Inferences:

1. The RMSE(%) decreases with respect to increase in lags in time sequence.
2. Since it shows that our better lag series has high value means the higher lag data has better autocorrelation than the smaller one.

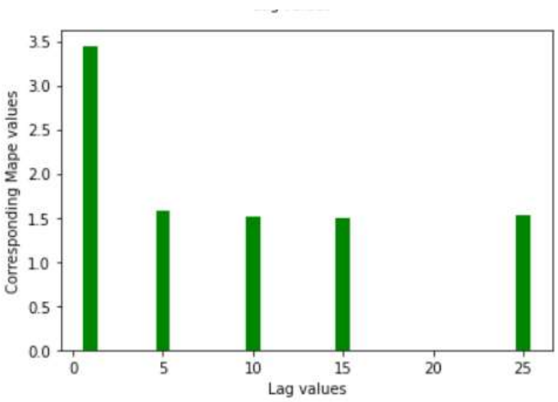

Figure 8 MAPE vs. time lag

Inferences:

1. The MAPE decreases with respect to increase in lags in time sequence.
2. It infers there is better autocorrelation for the higher series.

4

The heuristic value for the optimal number of lags is 79

The RMSE(%) and MAPE value between test data time sequence and original test data sequence are

1.765 and 2.061.

Inferences:

1. The RMSE decreased while MAPE increased.
2. Since the 79th lag series has optimum correlation values so it infers that the optimum value calculation uses RMSE at the backend but the same is not true for MAPE calculation.