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

Student's Name: Aryan Tiwari **Mobile No: 8982562898**

Roll Number: B20187 **Branch:** Electrical Engineering

1 a.

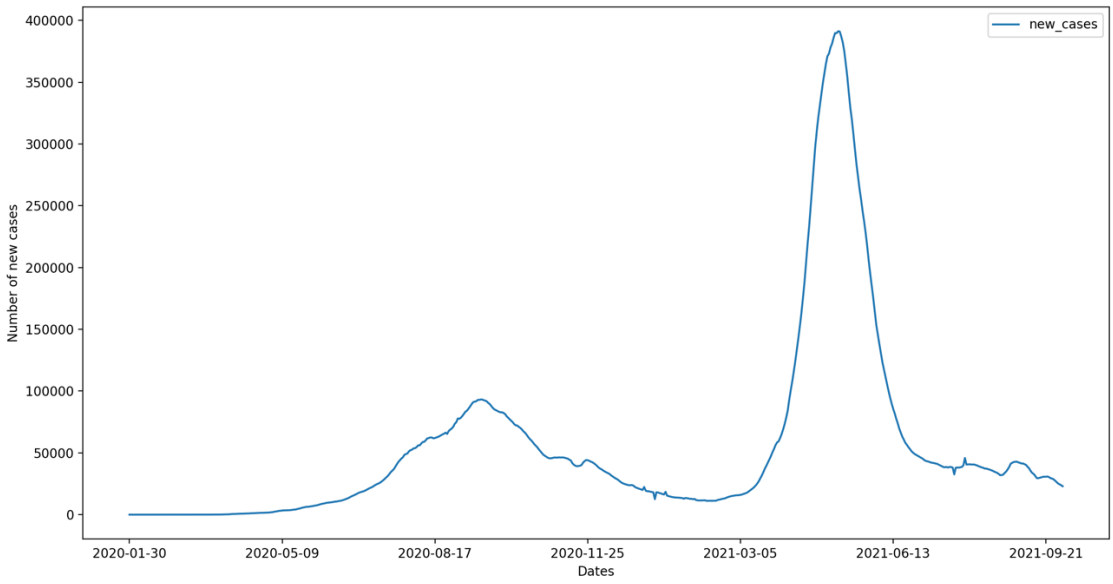

Figure 1 No. of COVID-19 cases vs. days

Inferences:

1. The time-series depicts bimodal data with 2 distinct peaks.
2. First wave - around August-2020
   Second wave - May-2021

b. The value of the Pearson's correlation coefficient is 0.99906

Inferences:

1. The two series are highly correlated (positive correlation).
2. Higher the pearson coefficient, higher the extent of similarity.

c.

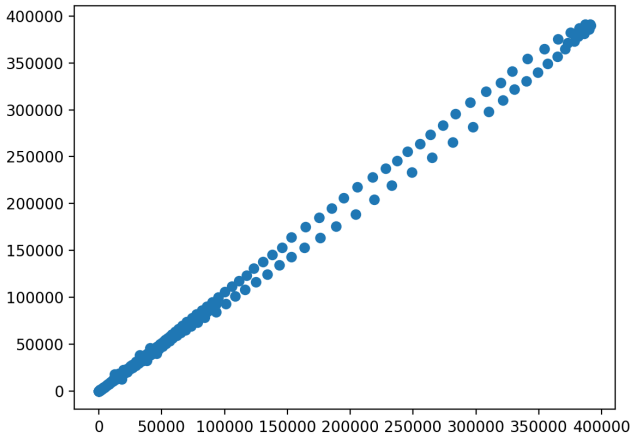

Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. The two series are highly correlated (positive correlation).
2. The scatter plot deviates very little from a straight line with slope 1 and hence obeys the pearson coefficient.

d.

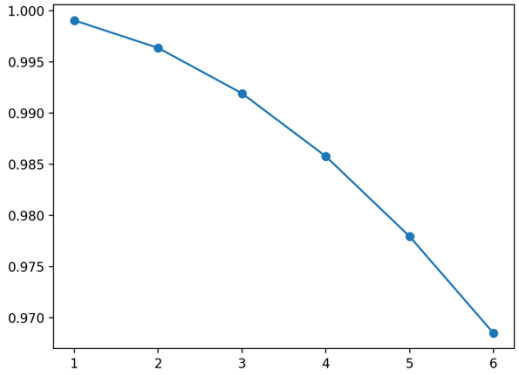

Figure 3 Correlation coefficient vs. lags in given sequence

Inferences:

1. The correlation decreases with increase in lags.
2. The number of new cases in 2 consecutive days are similar with subtler changes.

e.

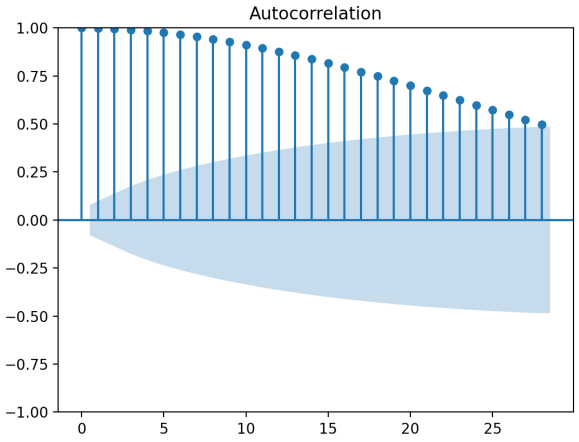

Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. The correlation decreases with increase inn lags.
2. The new cases per day depends on the existing number of active cases and hence is more related to lesser lagged series.

2

a. The coefficients obtained from the AR model are;

59.95483328406361 1.0367593349641009 0.2617123358706088
0.027561262816078624 -0.17539195532509488 -0.15246136637643914

b. i.

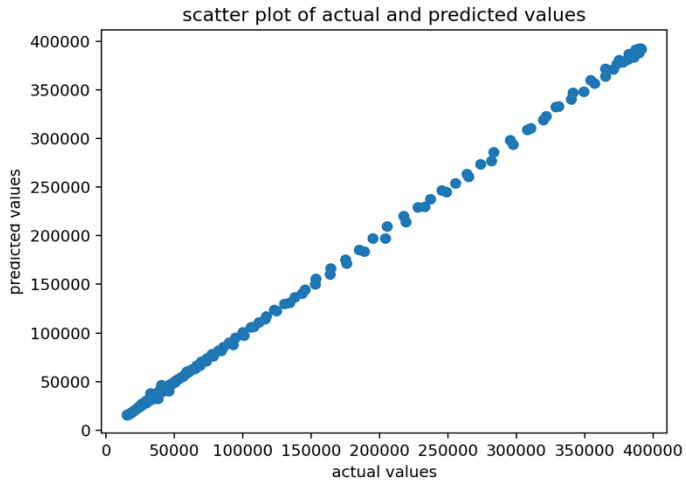

Figure 5 Scatter plot actual vs. predicted values

Inferences:

1. The two series are highly correlated.
2. Since the auto-correlation was very high in q1, the graph obeys the same.

ii.

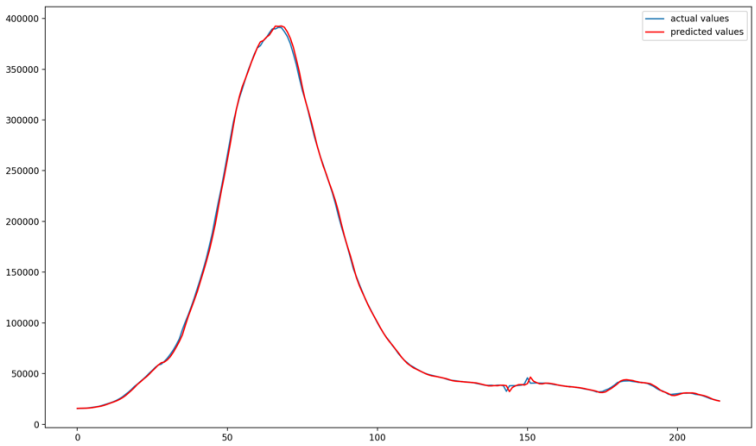

Figure 6: Predicted test data time sequence vs. original test data sequence

Inferences:

1. Since the predicted and actual plots overlap for most part of the timeline, the prediction model is quite effective for future predictions

iii.

RMSE(%) in predication: 1.8247684769390877

Mean absolute percentage error 1.5748363824058313

Inferences:

1. Both RMSE and MAPE are well below 5% which make the prediction quite accurate.
2. The high value of auto-correlation makes the prediction so accurate.

3

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

Lag value	RMSE (%)	MAPE
1	5.37	3.44
5	1.82	1.57
10	1.68	1.52
15	1.61	1.49
25	1.70	1.53

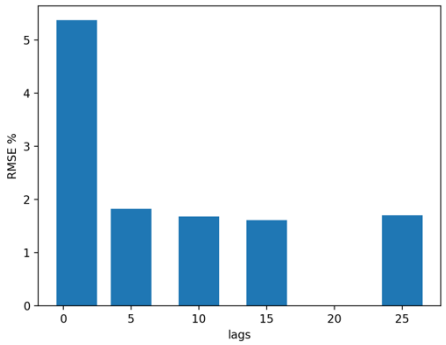

Figure 7 RMSE(%) vs. time lag

Inferences:

1. RMSE decreases up to lag = 15, after which the RMSE starts increasing.
2. The lag = 15 is best suited for prediction

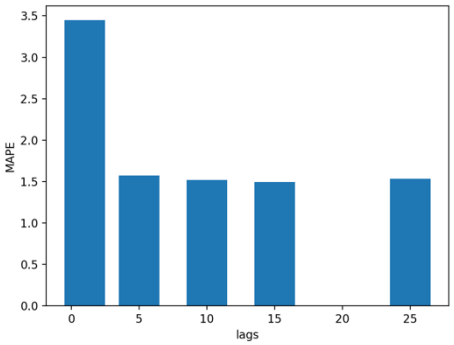

Figure 8 MAPE vs. time lag

Inferences:

1. MAPE decreases up to lag = 15, after which MAPE starts increasing.
2. Lag =15, is optimal for timeseries future prediction

4

The heuristic value for the optimal number of lags is 77.

RMSE(%) : 1.7593780528866607

MAPE : 2.0264439052850114

Inferences:

1. Both rmse and MAPE value for heuristic 77 are more than lag=15, but lesser than lag =1 which shows that the optimal solution from heuristic is better than single lagged series but not as good as statistical analysis of lagged data.
2. The heuristic method of parameter optimization is an intelligent guess, but statistical method of analysing different lag series is more exhaustive and hence more accurate.