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

**Student's Name: Akash Anand** **Mobile No: 7677858773**

**Roll Number: B20243** **Branch:EP**

**1 a.**

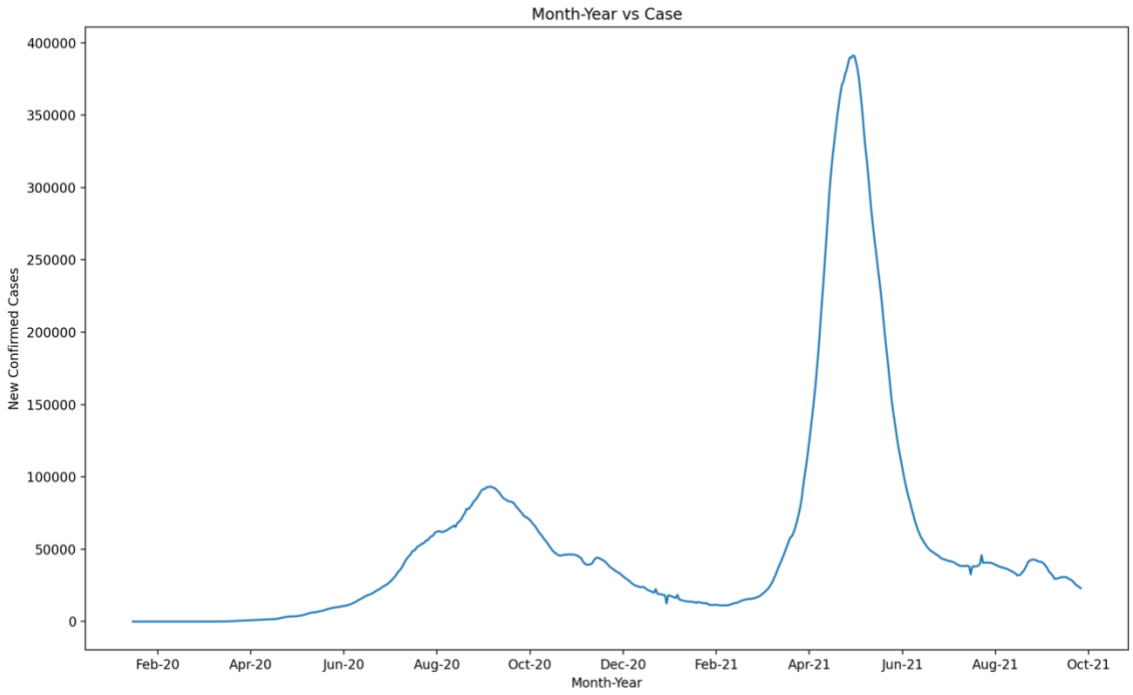


**Figure 1 No. of COVID-19 cases vs. days**

**Inferences:**

1. The days one after the other have similar power consumption.
2. As we can see that the curve, we have plotted is continuous. Therefore, the consequent days have similar no. of confirmed cases.
3. The duration of first wave is around August 2020 and second wave is 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 say that both are highly correlated.
2. The number of confirmed cases on consequent days are almost similar as the correlation coefficient is approximately equals to 1.
3. The reason behind 1 and 2 is that they are highly correlated and consequent days have similar covid cases because coefficient of correlation is approximately 1.

**c.**

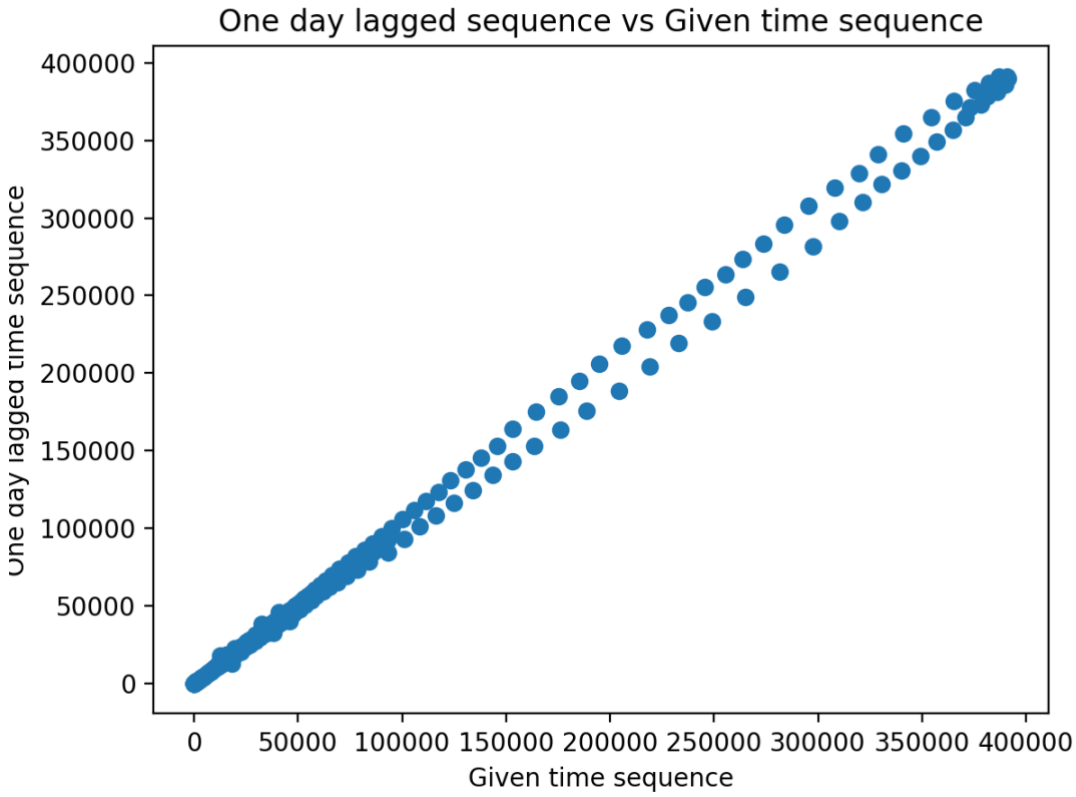


**Figure 2 Scatter plot one day lagged sequence vs. given time sequence**

**Inferences:**

1. From the nature of the spread of data points the correlation is nearly 1.
2. Yes, the scatter plot seems to obey the nature reflected by Pearson's correlation coefficient calculated in 1 b.
3. As we can see that the curve is around the diagonal of the plot, so its correlation must be approximately 1.

**d.**

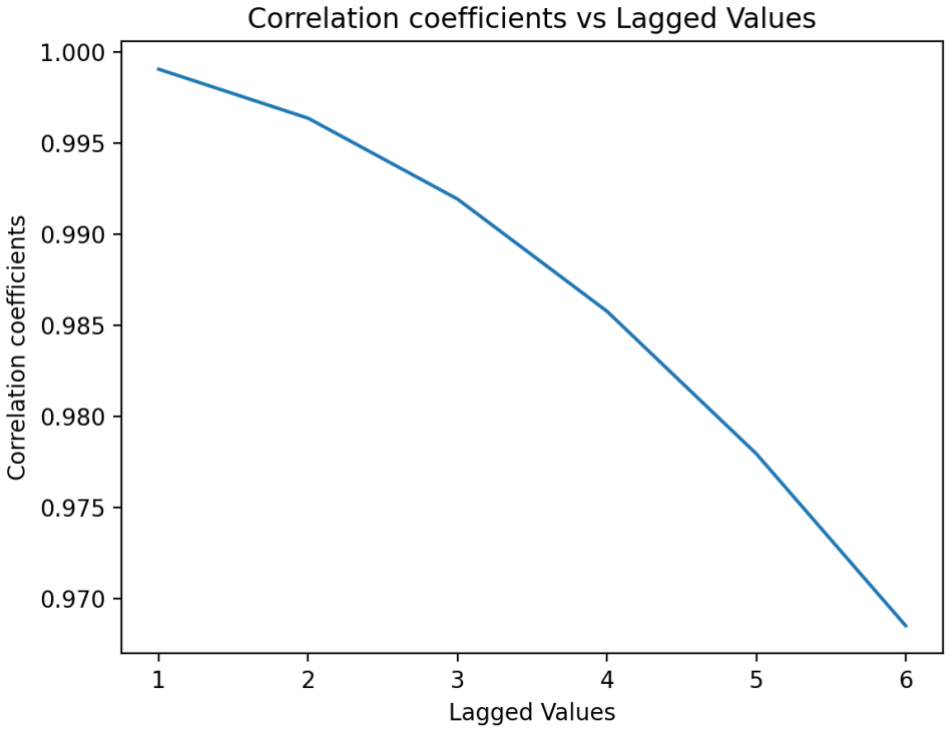


**Figure 3 Correlation coefficient vs. lags in given sequence**

**Inferences:**

1. The trend of correlation coefficient value with respect to increase in lags in time sequence is that the value of correlation coefficient decreases as the number of lags increase.

2. The reason behind the observed trend is that as the covid cases on consequent days are nearly same, the correlation coefficient between two consequent days will be more than that of correlation between two days with more gap in between.

**e.**

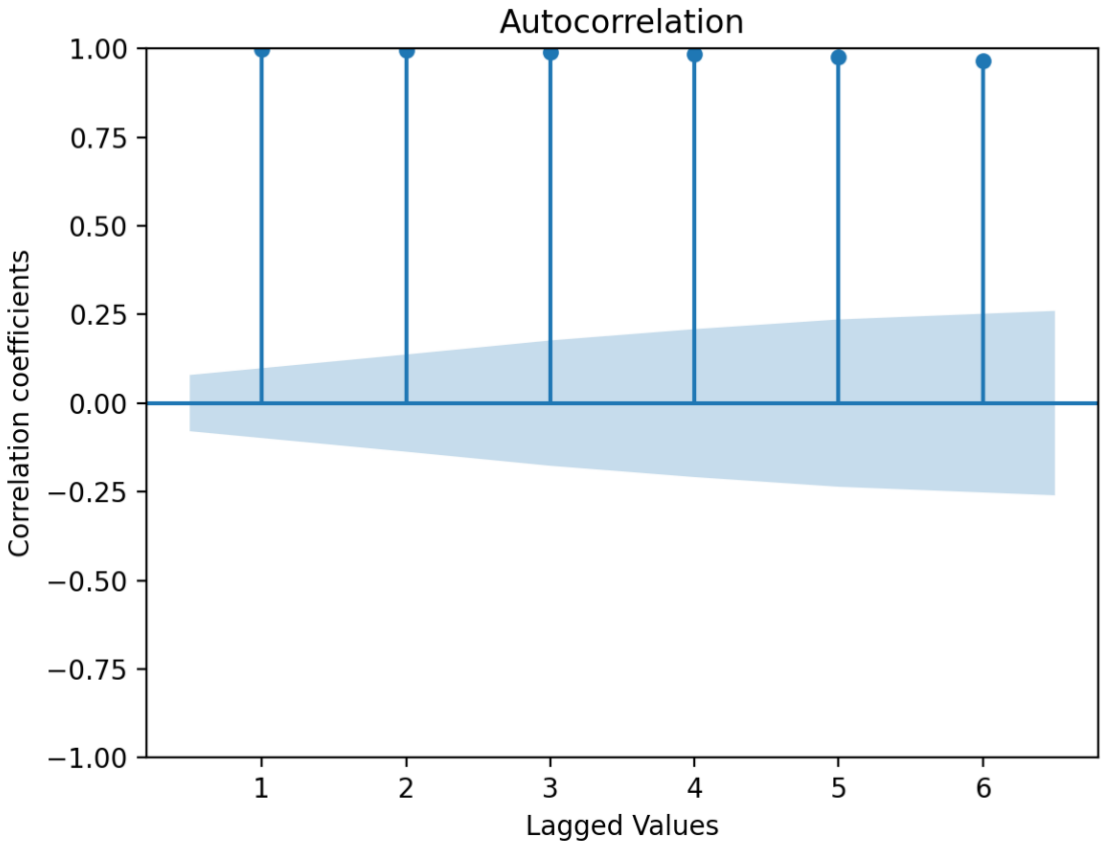


**Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function**

**Inferences:**

1. The trend of correlation coefficient is almost 1 but it is decreasing continuously.

2. As the covid cases on consequent days are nearly same, the correlation coefficient between two consequent days will be more than that of correlation between two days with more gap in between.

**2**

**a.** The coefficients obtained from the AR model are [ 59.954, 1.036, 0.261,  0.027, -0.175, -0.152]

**b. i.**

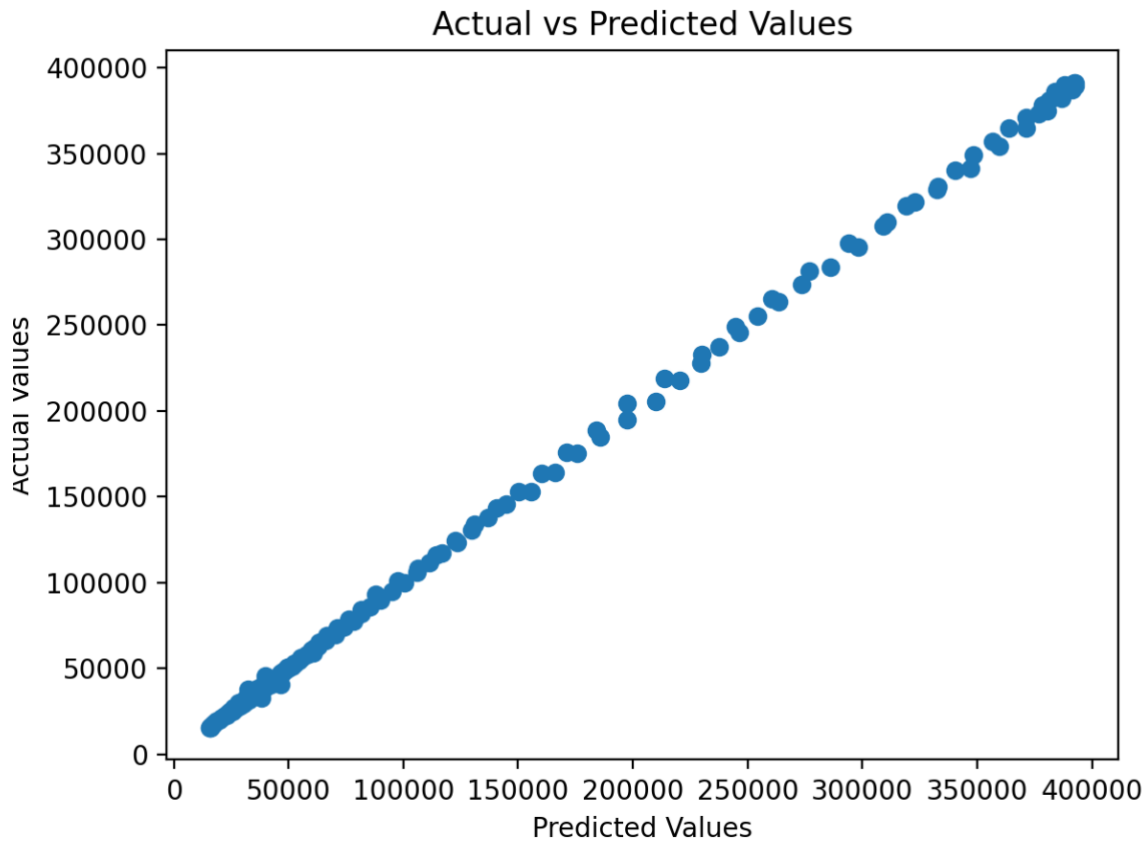


**Figure 5 Scatter plot actual vs. predicted values**

**Inferences:**

1. From the nature of the spread of data points, the nature of the correlation between the two sequences is high which is approximately 1.
2. Yes, the scatter plot seem to obey the nature reflected by Pearson's correlation coefficient calculated in 1 b.
3. The model predicts correctly, as we can see that correlation is almost 1. So, the model had predicted good.

**ii.**

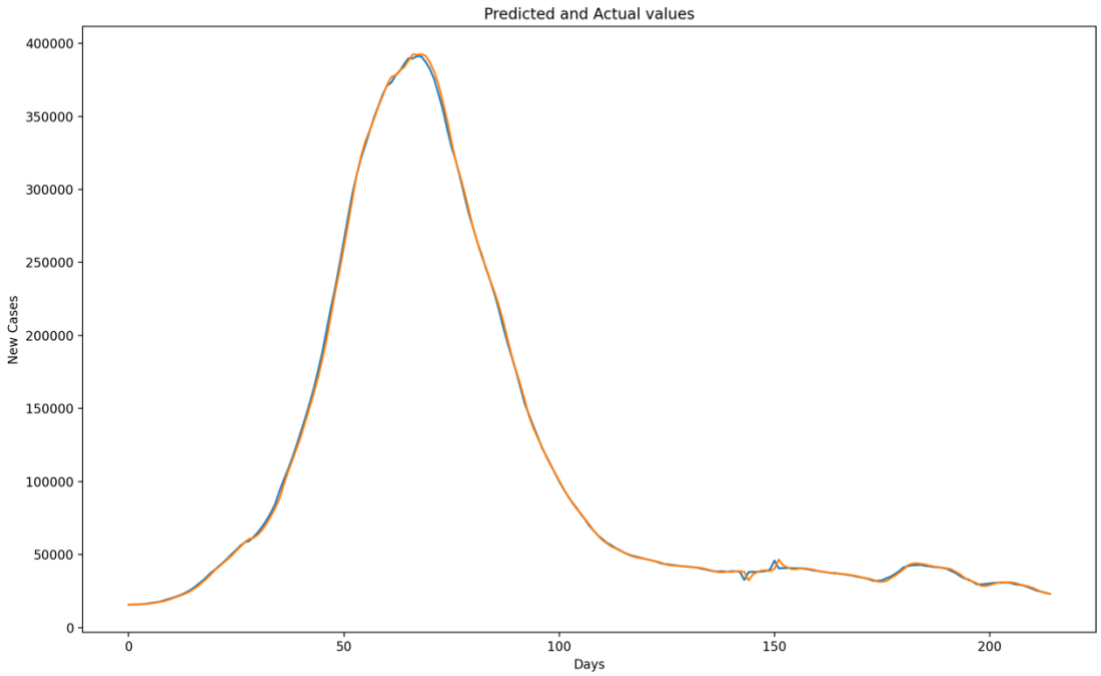


**Figure 6 Predicted test data time sequence vs. original test data sequence**

**Inferences:**

1. The model is reliable for future predictions as it has given the same value that we had from the test data.

**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 we can say that our model is good and reliable.
2. Both RMSE and MAPE have value less than 2. So, we can say that the model predicts correctly.

**3**

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

| Lag value | RMSE (%) | MAPE |
|---|---|---|
| 1 | 5.372948 | 3.446540 |
| 5 | 1.824768 | 1.574836 |
| 10 | 1.685532 | 1.519370 |
| 15 | 1.611935 | 1.496236 |
| 25 | 1.703391 | 1.535421 |

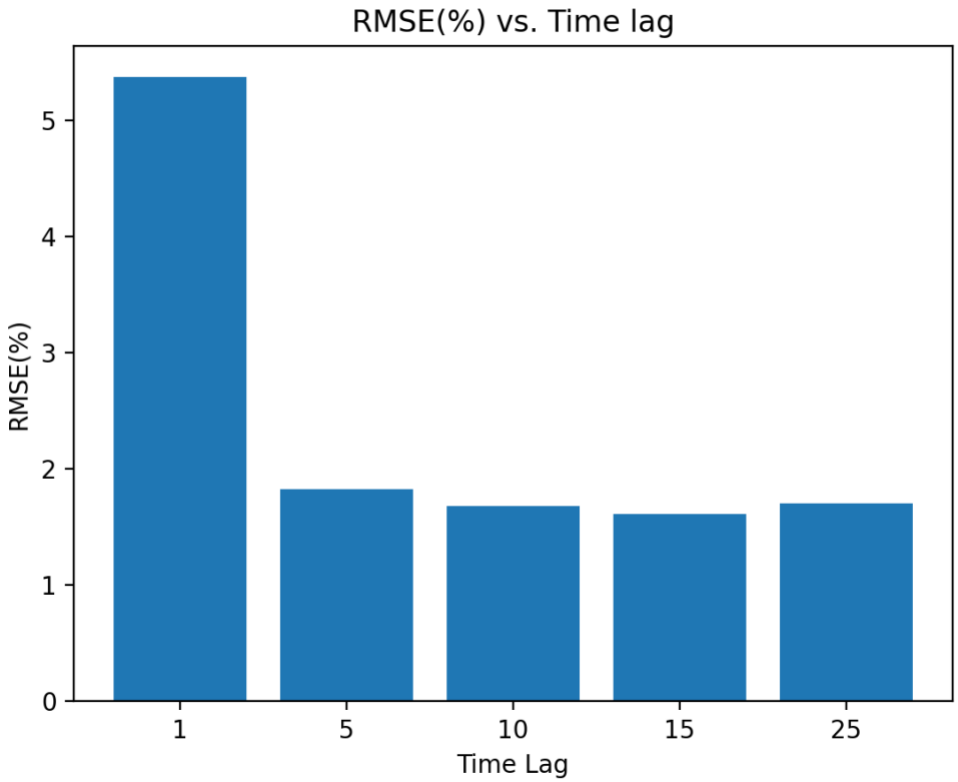


**Figure 7 RMSE(%) vs. time lag**

**Inferences:**

1. RMSE(%) decreases quickly from value 1 to 5 but after that it decreases gradually with increase in lag value.
2. It is because the complex model needed to fit our data more accurately. Therefore, when the lag is increased from 1 to 5 the accuracy improves significantly but after that the increase is gradual.

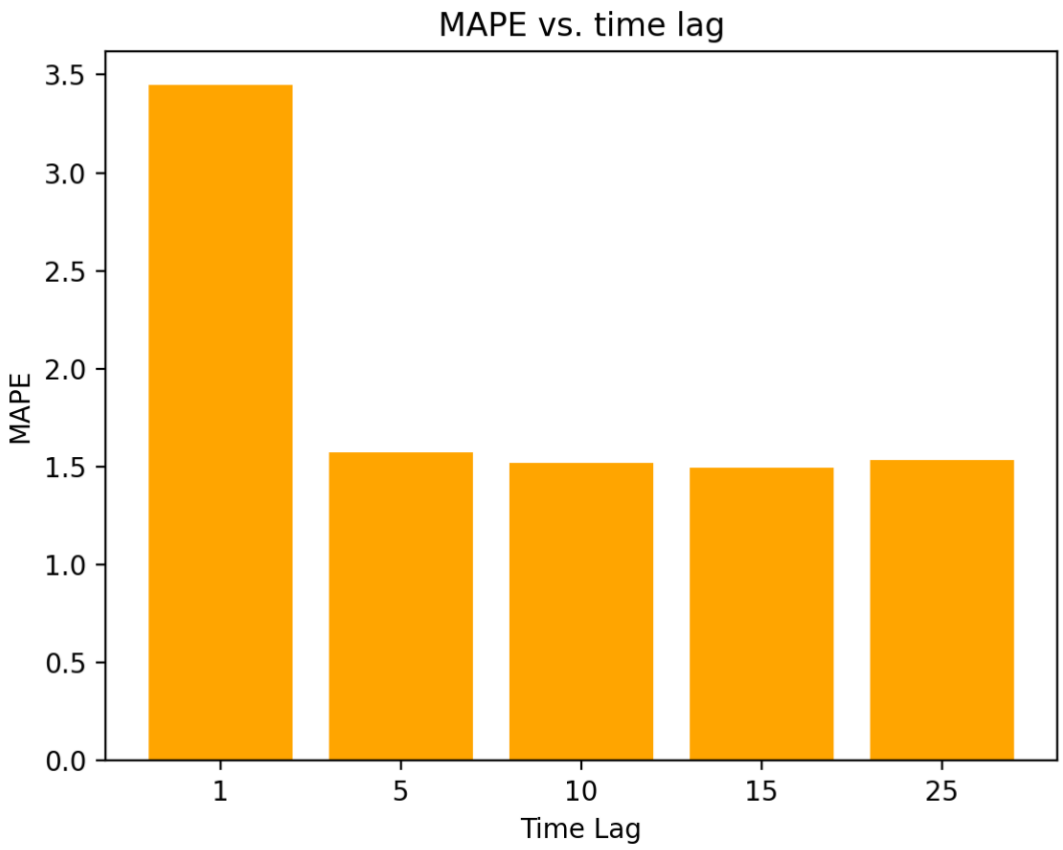


**Figure 8 MAPE vs. time lag**

**Inferences:**

1. MAPE first deceases quickly from 1 to 5 but after that it decreases gradually with increase in lags.
2. This is because a complex model is needed to fit our data more accurately. Therefore when the lag is increased from 1 to 5 the accuracy improves significantly but after that the increase is 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 optimal lag value didn't improve the prediction accuracy. As we can see that the RMSE for lag value 10 is less than that of the optimal lag value.
2. Because as we keep increasing the lag, after certain time the pattern RMSE vs Lag will become random and we can also see that as the observation 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.