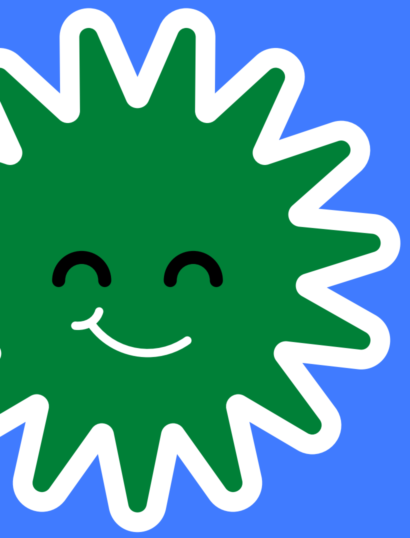
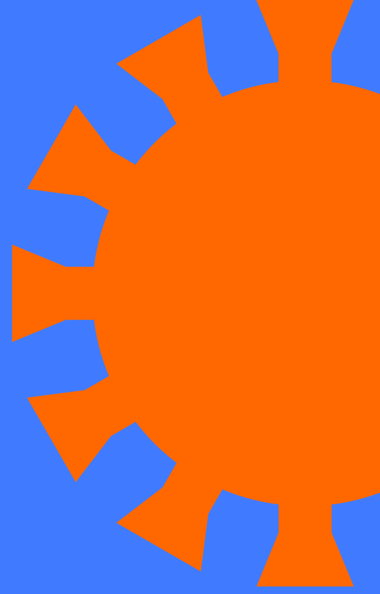
SPATIAL-TEMPORAL DISTRIBUTION OF COVID-19 IN A STATE AND ITS PREDICTION: A DATA-DRIVEN MODELING ANALYSIS

Group Members

Medha Balani	IIT2019021
Vidushi Pathak	IIT2019027
Aarushi	IIT2019032
Jyotika Bhatti	IIT2019036





CONTENT

- Introduction
- Problem Statement
- Objective
- Literature Review
- Workflow
- Initial Work
- Future Work/ Activity Diagram
- References

Introduction

- Covid 19 is a disease that has caused great havoc in the world, since it originated in China and began to spread over to every corner of the world.

- The World Health Organisation (WHO) has declared the coronavirus disease 2019 (COVID-19) a pandemic.

- A novel coronavirus (nCoV) is a new strain that has not been previously identified in humans.

Purpose of Project

- We will analyze the spread of this disease from one province to another.
- The probability to find the expected number of cases in the surrounding cities using spatial-temporal relations.
- We would use entropy, correlation matrix and a strong statistical model, to show spatial-temporal relation of COVID-19 infection spread from a certain particular province to neighboring areas.

Problem Statement

- We have observed, there is spatial temporal relationship in the covid 19 data.
- The problem here is the spread of this disease is causing a lockdown situation everywhere it is spreading in large numbers.
- Thus having a fair idea of how the present situation of a place may vary according to the number of cases of the surrounding areas would give a fair opportunity to predict and prevent extreme fatality due to covid.
- In order to get these stats of the surrounding areas we need to study and train a model accordingly for the prediction of fatality, the number of cases, recovery, and their ratios.

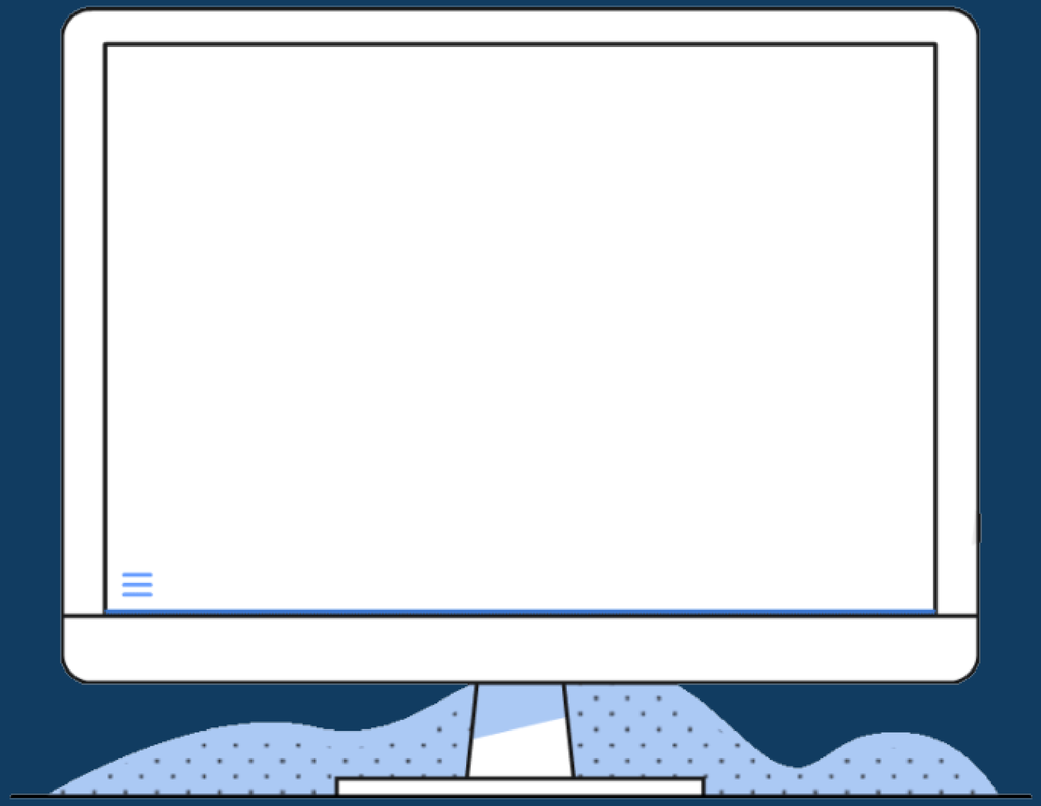
OBJECTIVE

- The objective here is the spatial-temporal distribution of COVID 19 data and its prediction in the nearby areas.

- Our main objective is to use a strong statistical model to show that COVID 19 infection is spatially dependent and is mainly spread via to the neighboring areas.

- Our task is to analyze how the increase in covid cases in one state/city(hotspot) will affect the neighboring states/cities.

LITERATURE REVIEW



Spatial-temporal distribution of COVID-19 in China[1]

- The attribute values of adjacent region units are characterized by Moran Index.
- SIER MODEL which has been used to predict the rate of spread of viruses. In this model, if an individual is in an infected state
- Inorder to test the spatial autocorrelation of COVID 19 , the destination between the cities and the provinces is required, which is estimated through Moran Index.
- This paper adopted the Moran index to deduce the COVID-19 has had a spatial correlation in China and paid more attention to analyze the confirmed cases of this outbreak from the perspective of spatial measurement and statistical modeling.

Spatial-temporal generalized additive model for modeling COVID-19 mortality risk in Toronto, Canada summary [3]

- The author has done a comparison analysis of three model's performance with three different functions namely, logistic link function, cloglog link function, and probit link function, out of which the model with probit link function out-performed the rest of the models.

- It yielded the lowest AIC and deviance, highest percentage of deviance explained, highest AUC and lowest Brier's score.

Adaptive Multi-Kernel SVM With Spatial-Temporal Correlation for Short-Term Traffic Flow Prediction[4]

- The paper mainly deals with estimation of the traffic state and can help to address the issue of urban traffic congestion, providing guiding advice for people's travel and traffic regulation, using adaptive multi-kernel support vector machines (AMSVM).

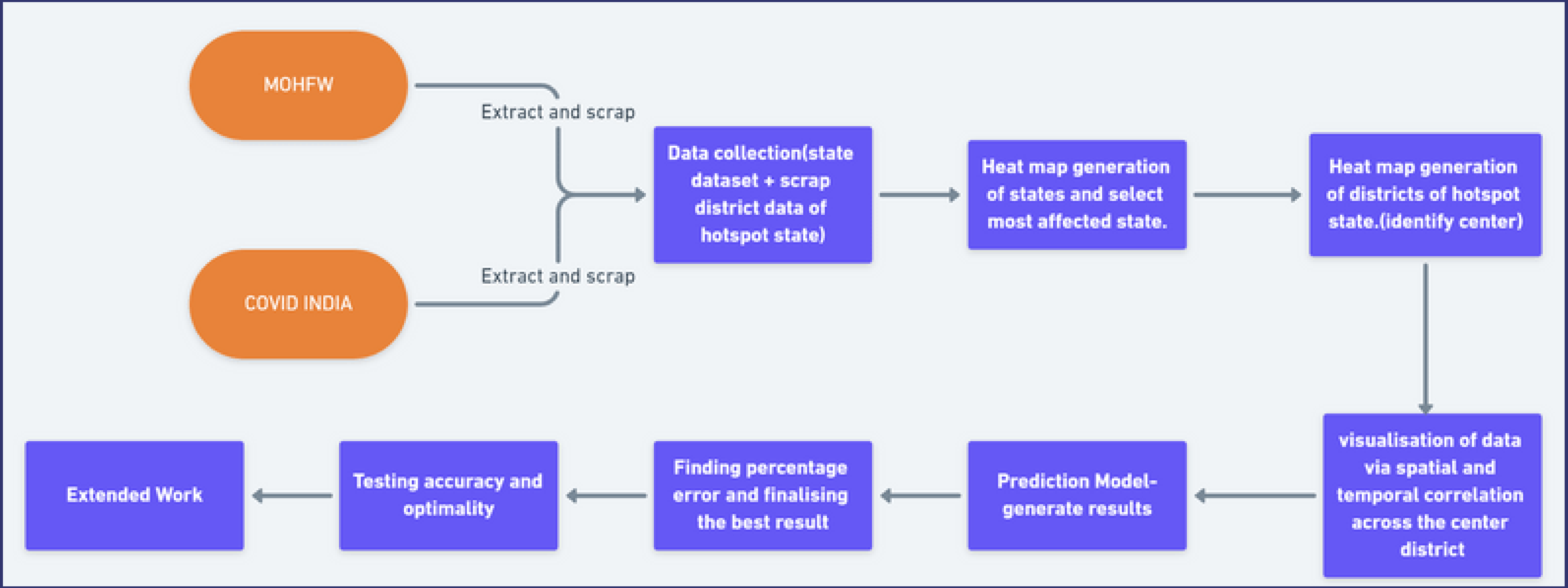
- The spatial temporal correlation is incorporated with AMSVM to predict the short term traffic flow, which can fuse spatial-temporal correlation predicted values with different weights. Selection of kernel function depends on the distribution of sample data and the relationship between sample data and predicted variables. Since different feature spaces have different data distribution, the performance of SVM depends largely on the choice of kernel function.

Spatial and temporal differentiation of COVID-19 epidemic spread in mainland China and its influencing factors [5]

- The local spatial correlation characteristics were mainly composed of the 'high-high' and ‘low-low’ clustering types, and the situation of the contiguous layout was very significant.

- It mainly focuses on the spatial and temporal evolution characteristics of the epidemic. In this paper, the number of confirmed COVID-19 cases in mainland China was taken as the measurement index, and the spatial and temporal differentiation of the epidemic spread were described by the exploratory spatial data analysis method.

Workflow



MOHFW

COVID INDIA

Extract and scrap

Extract and scrap

Data collection(state dataset + scrap district data of hotspot state)

Heat map generation of states and select most affected state.

Heat map generation of districts of hotspot state.(identify center)

visualisation of data via spatial and temporal correlation across the center district

Prediction Model- generate results

Finding percentage error and finalising the best result

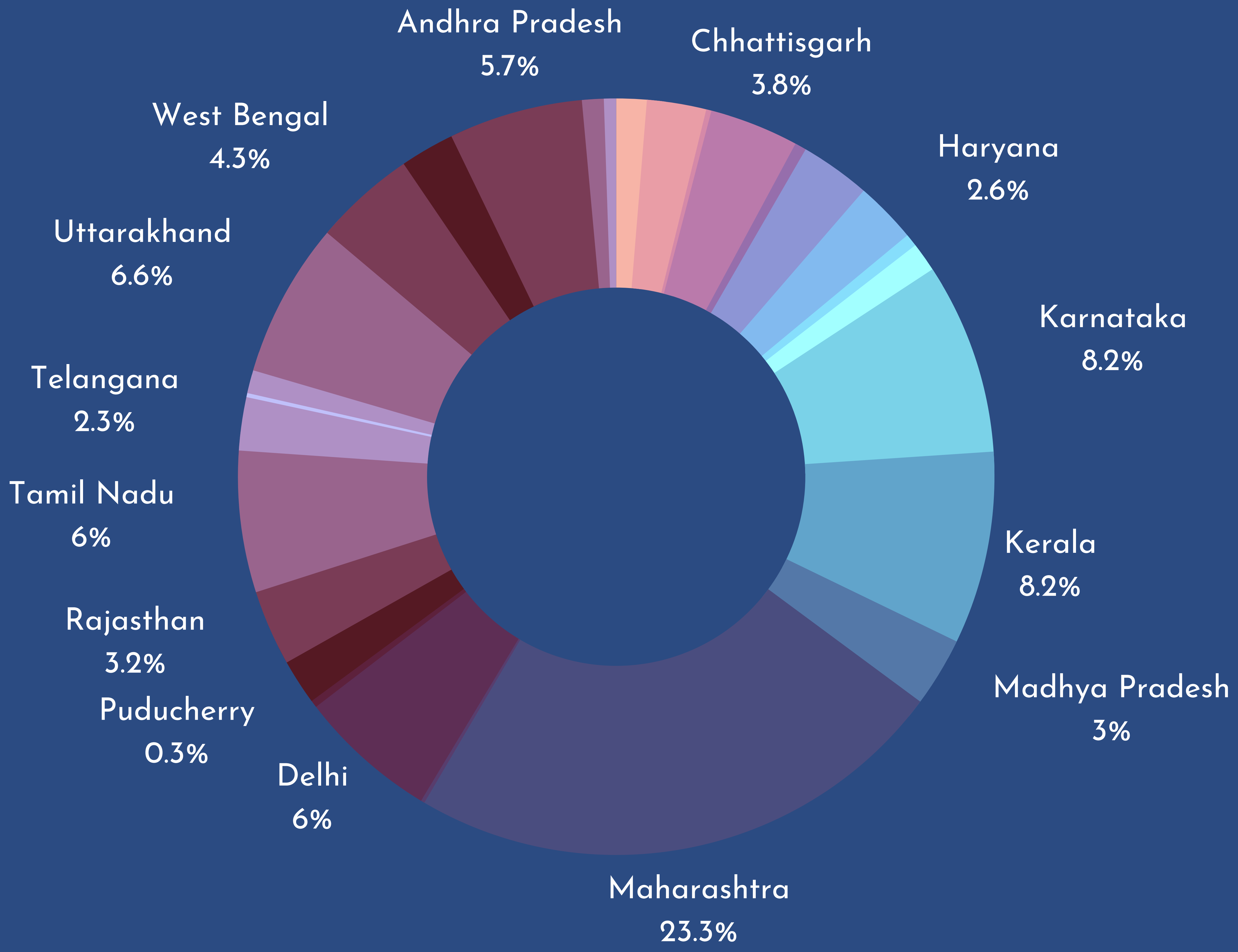
Testing accuracy and optimality

Extended Work

Selection of hotspot

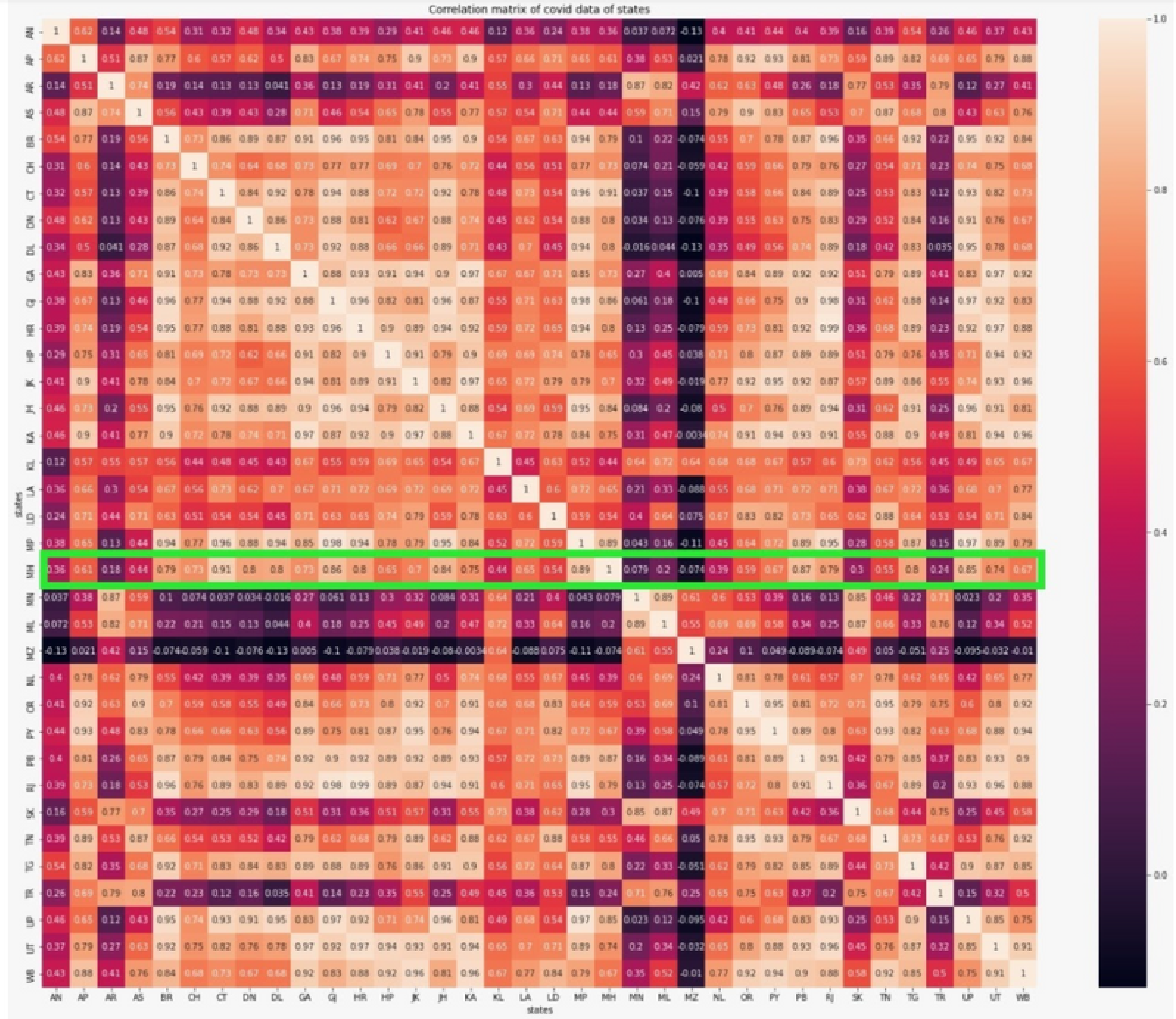
COVID CASE DISTRIBUTION

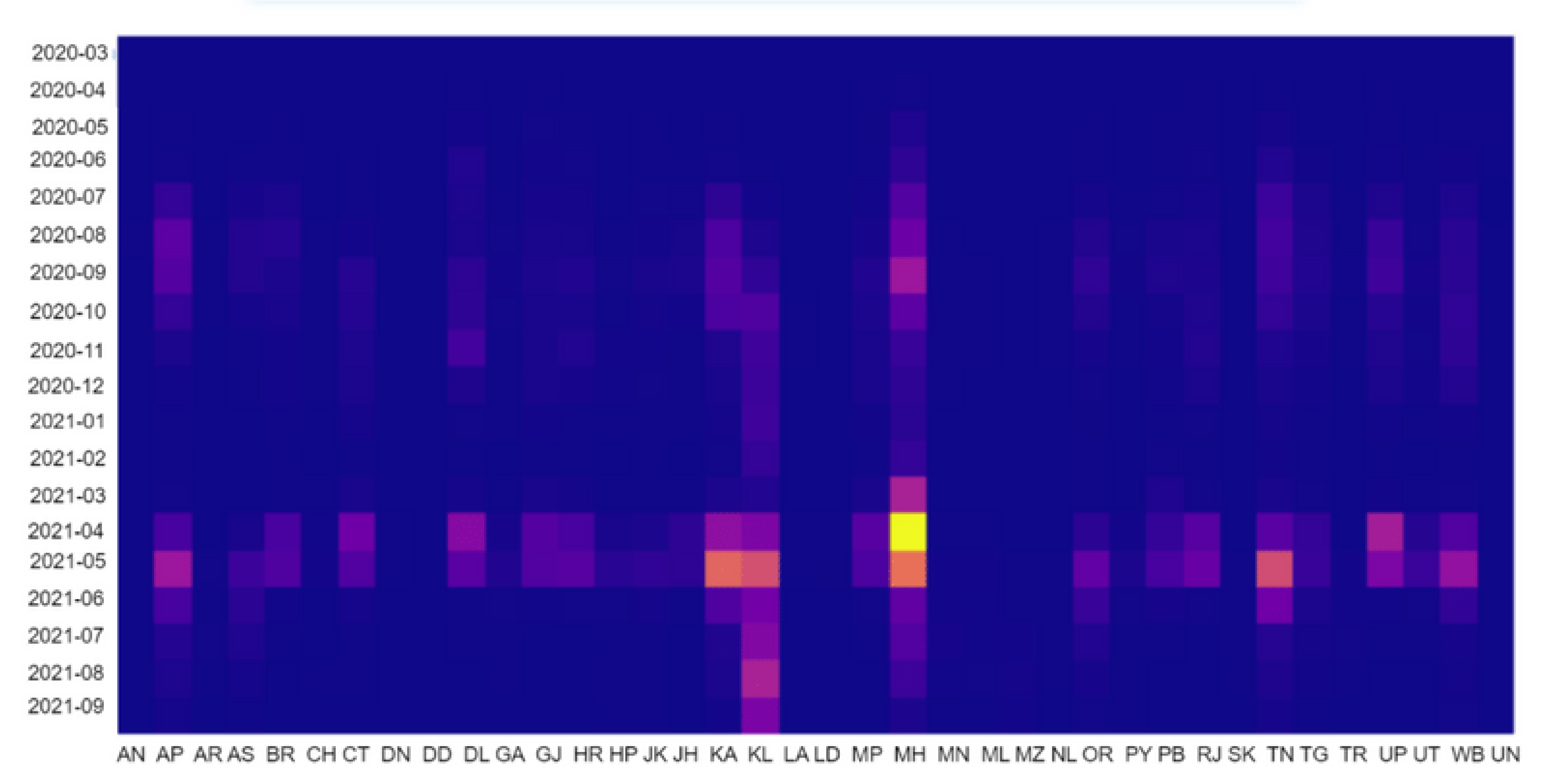
Presenting the inference by studying the data available of the distribution of cases most of the confirmed cases lies in the Maharashtra region.



Correlation Matrix

According to the correlations generated, we can see that the correlations of Maharashtra are pretty higher with the neighbouring states. Suggesting that the increase in cases here, makes the cases of other neighbouring states higher. Thus, an overall rise in the country.



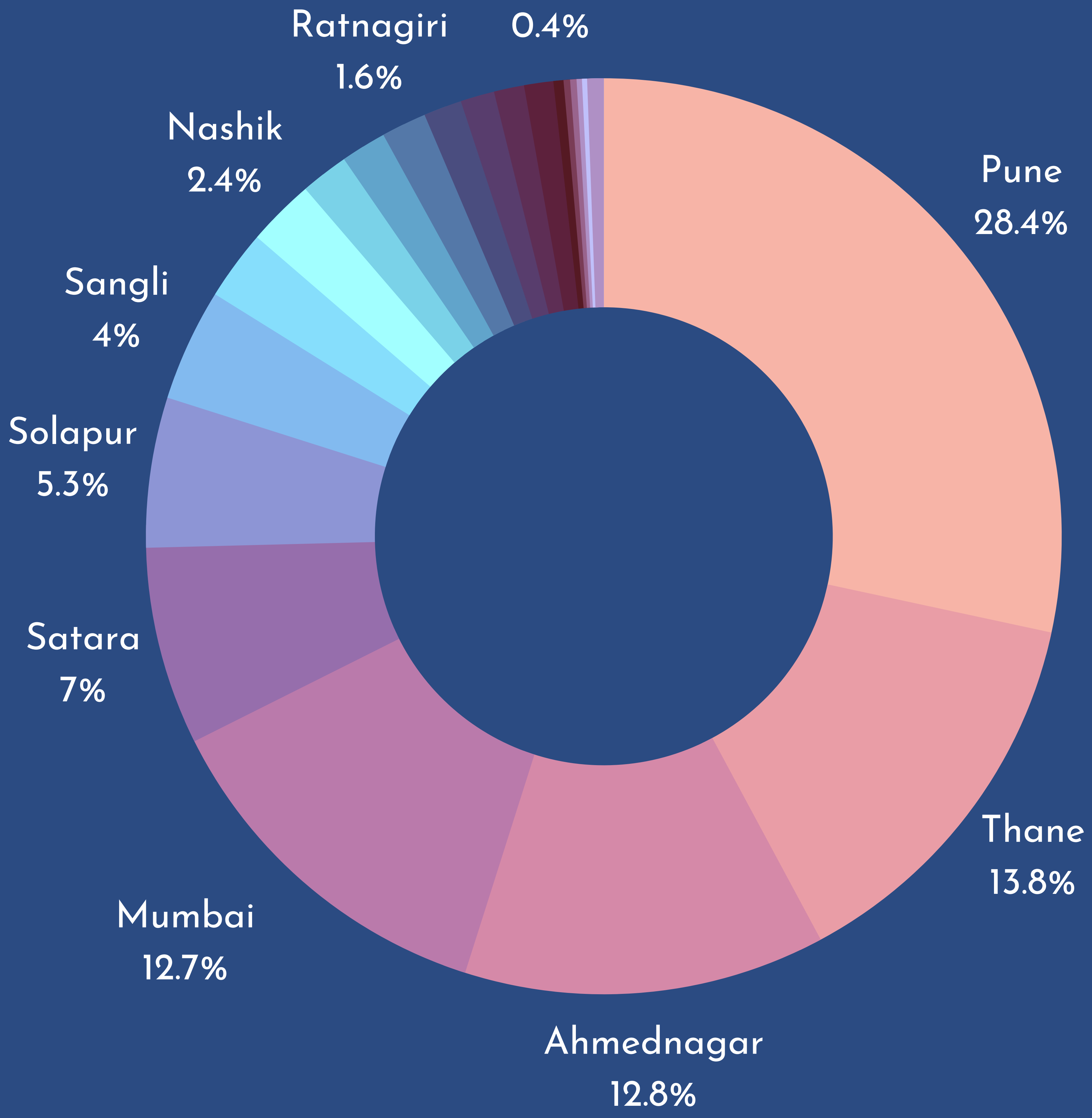


From this plot it is quite visible that most cases lie during the phase of 2021-03 to 2021-06. And Maharashtra being the brightest indicates that it has the largest cases of that time. With Kerala being on the 2nd number.

Maharastra

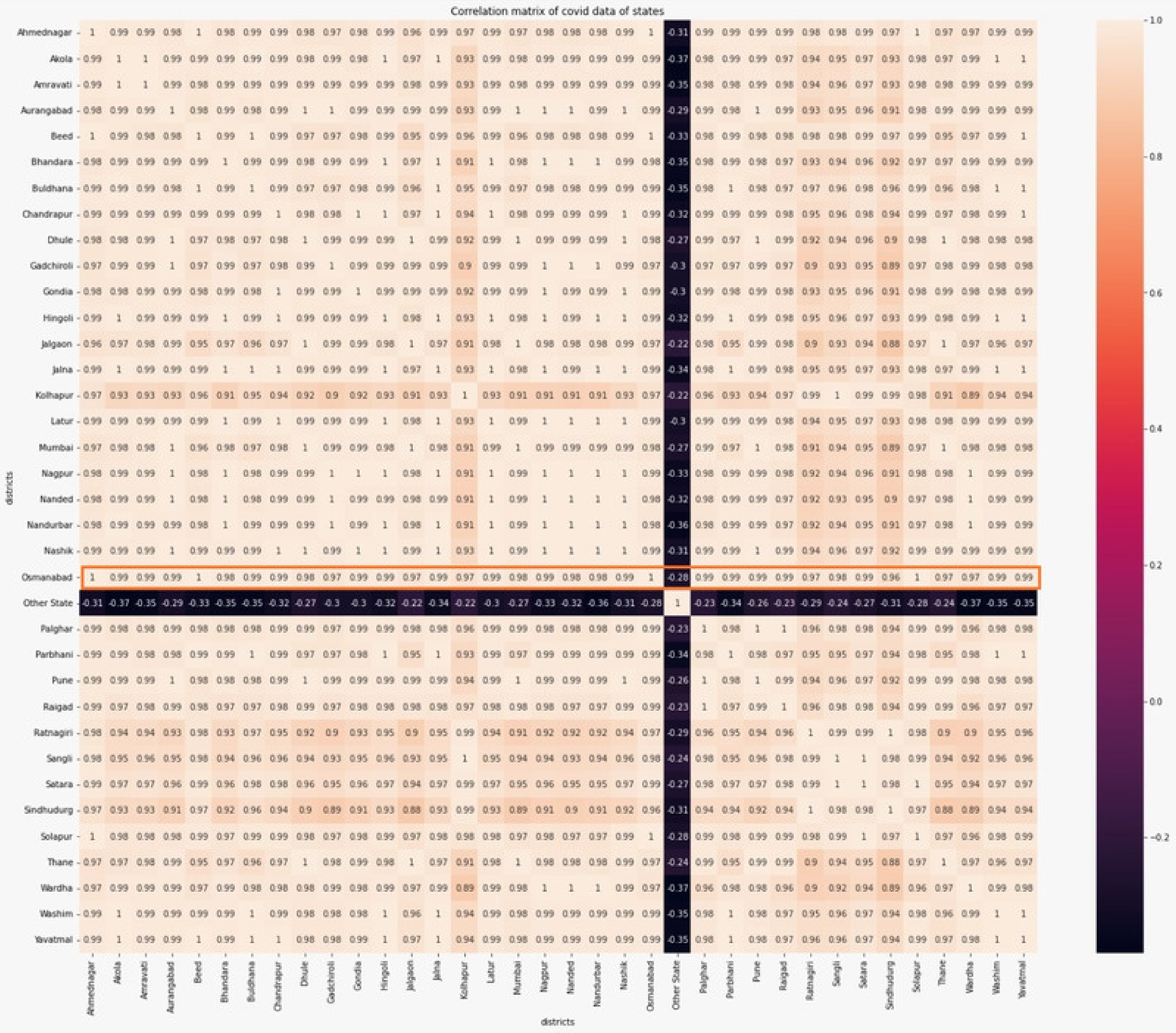
DISTRICT WISE

Taking a more closer look on district wise cases distribution of the state.

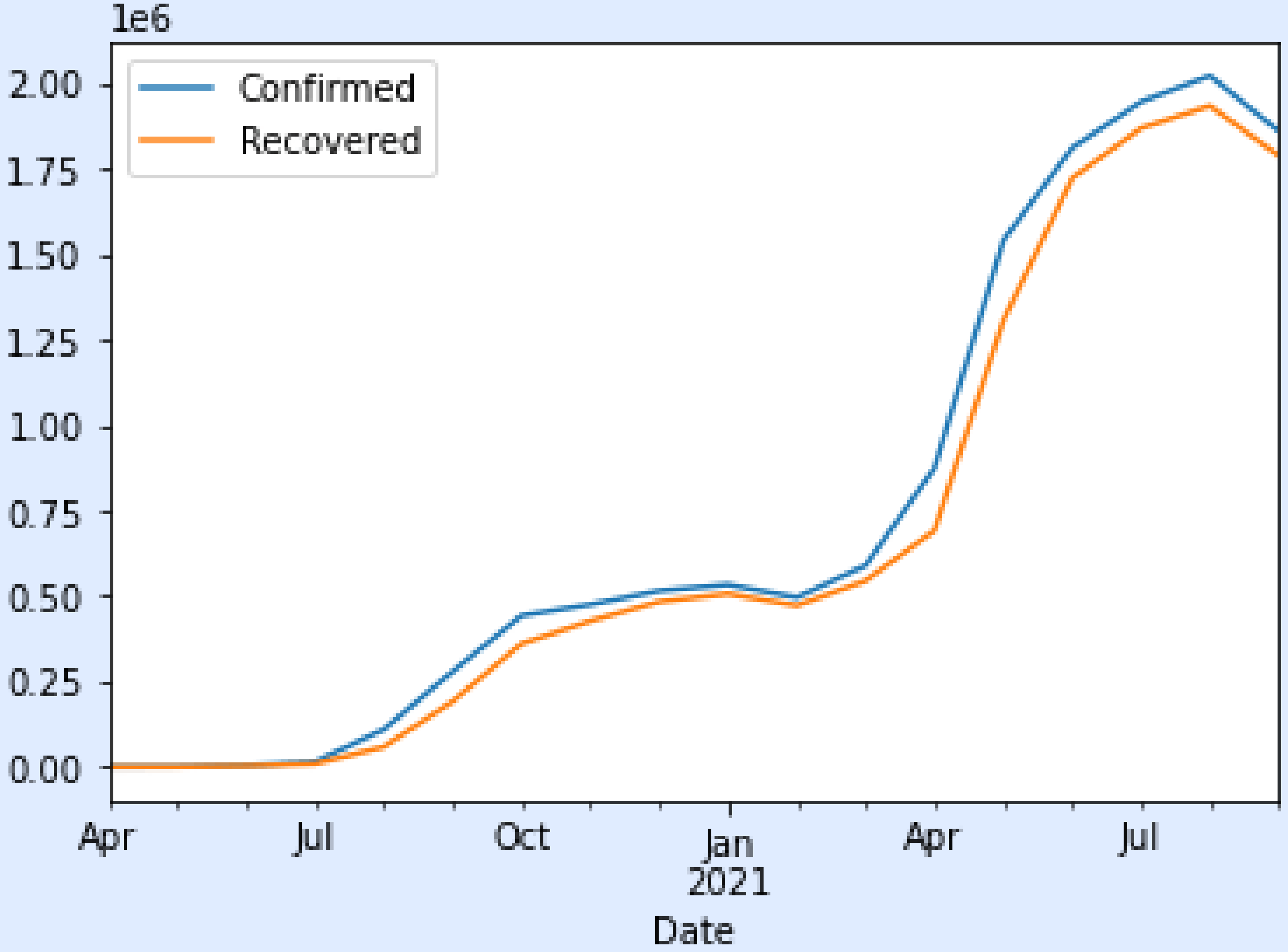


Maharashtra districts Spatial correlation

We can see that the maximum correlation is shown in the district Osmanabad. Hence we will consider Osmananbad as the district hotspot for our prediction..

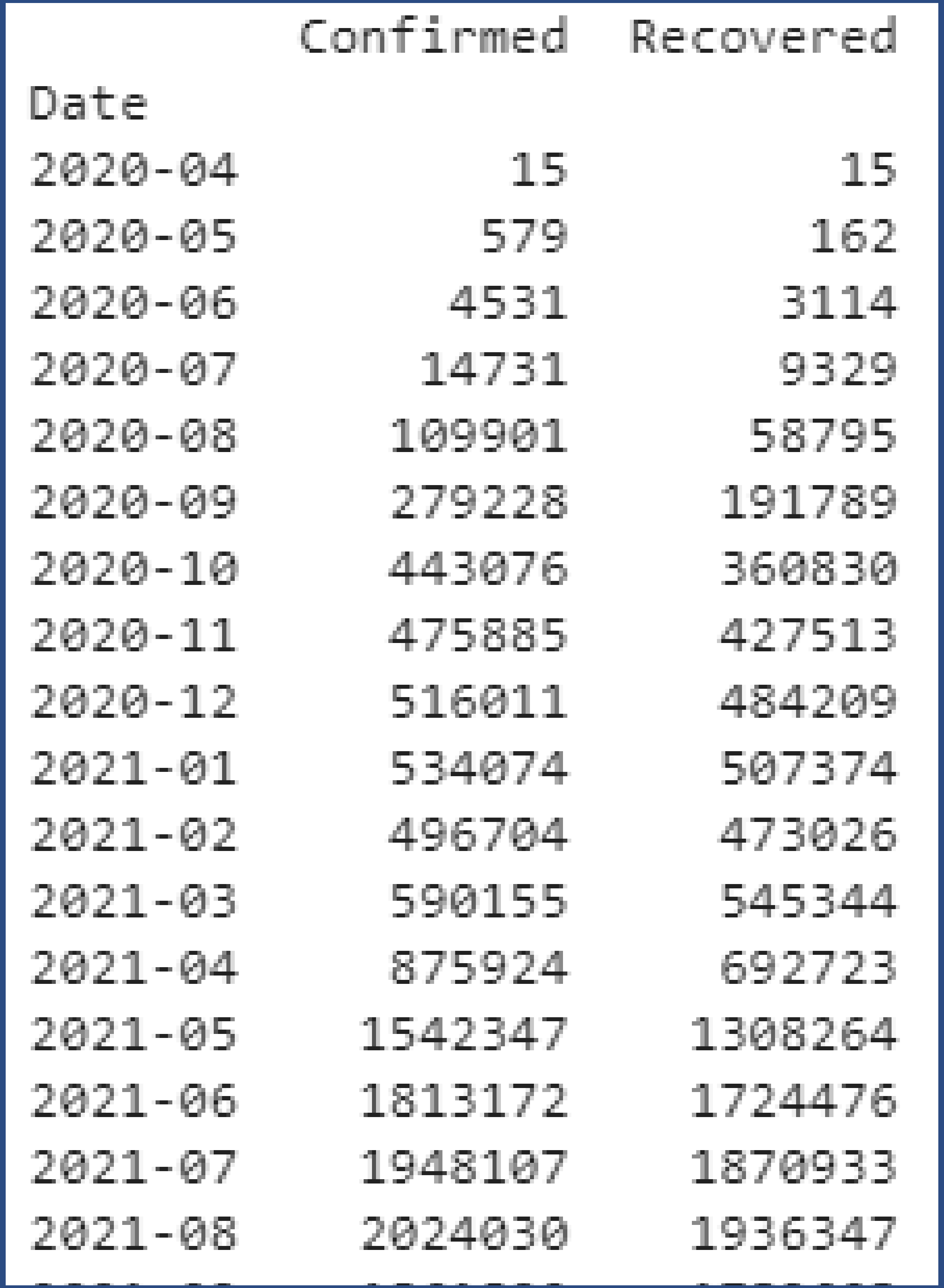


The Covid cases in Osmanabad can be seen increasing exponentially,



A general pattern of rising cases of Osmanabad

The monthly summation of the COVID confirmed cases and the recovered cases, of Osmanabad is,



```
         Confirmed  Recovered
Date
2020-04         15         15
2020-05        579        162
2020-06       4531       3114
2020-07      14731       9329
2020-08     109901      58795
2020-09     279228     191789
2020-10     443076     360830
2020-11     475885     427513
2020-12     516011     484209
2021-01     534074     507374
2021-02     496704     473026
2021-03     590155     545344
2021-04     875924     692723
2021-05    1542347    1308264
2021-06    1813172    1724476
2021-07    1948107    1870933
2021-08    2024030    1936347
```

ADF Testing

The Augmented Dickey Fuller Test (ADF) is unit root test for stationarity. Unit roots can cause unpredictable results in time series analysis.

Upon doing the following test we got the following values:

ADF Statistic: -0.27925597101134114

n_lags: 17

p-value: 0.928391356920488

Critial Values: 1%: -3.4434175660489905

Critial Values: 5%: -2.8673031724657454

Critial Values: 10%: -2.5698395516760275

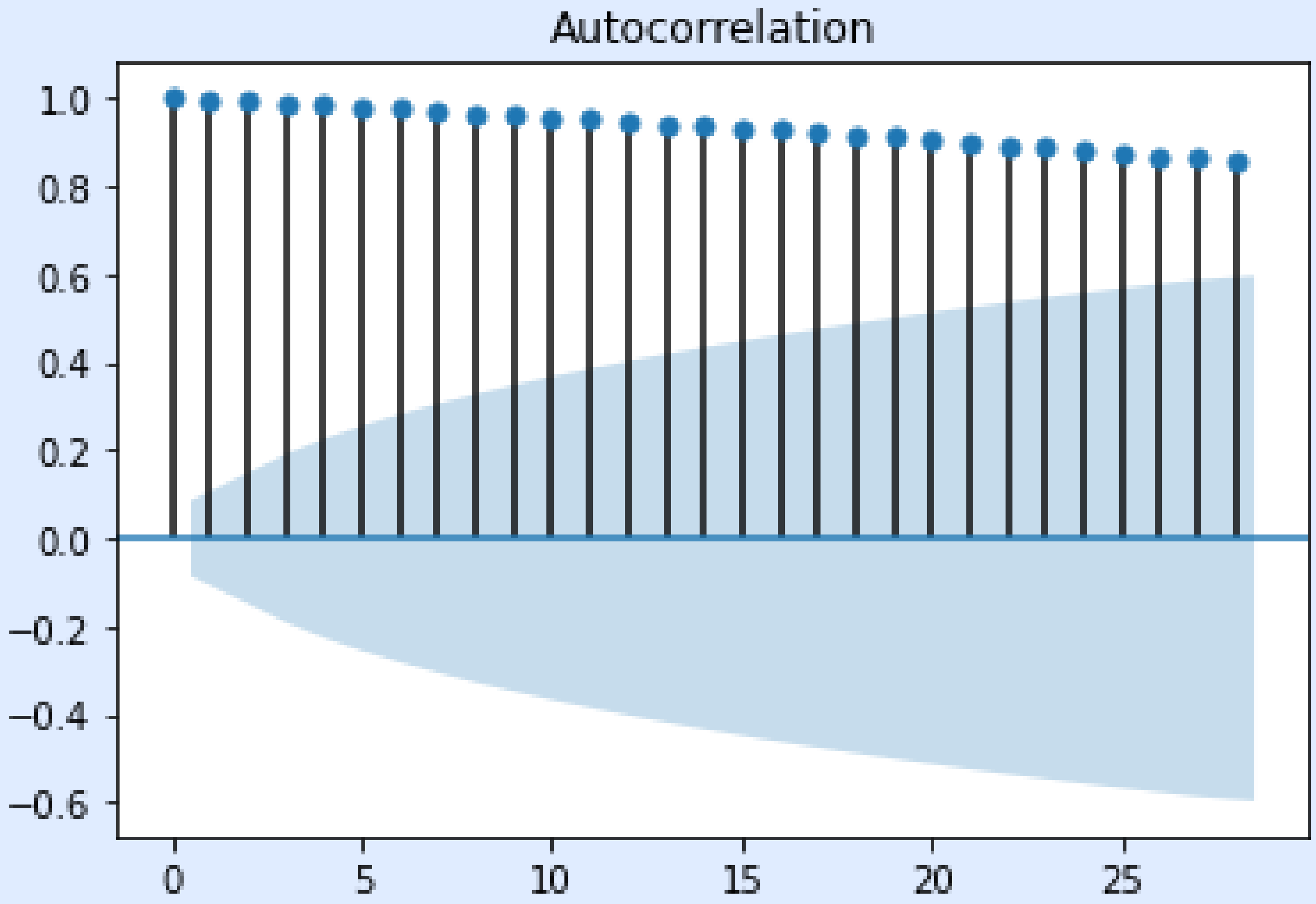
The p-value obtained is greater than the significance level of 0.05 and the ADF statistic is higher than any of the critical values. providing strong evidence that the data is non-stationary.

Temporal Correlation

- For the hotspot district, here we have calculated temporal correlation to find out how much the cases are correlated to the subsequent days.

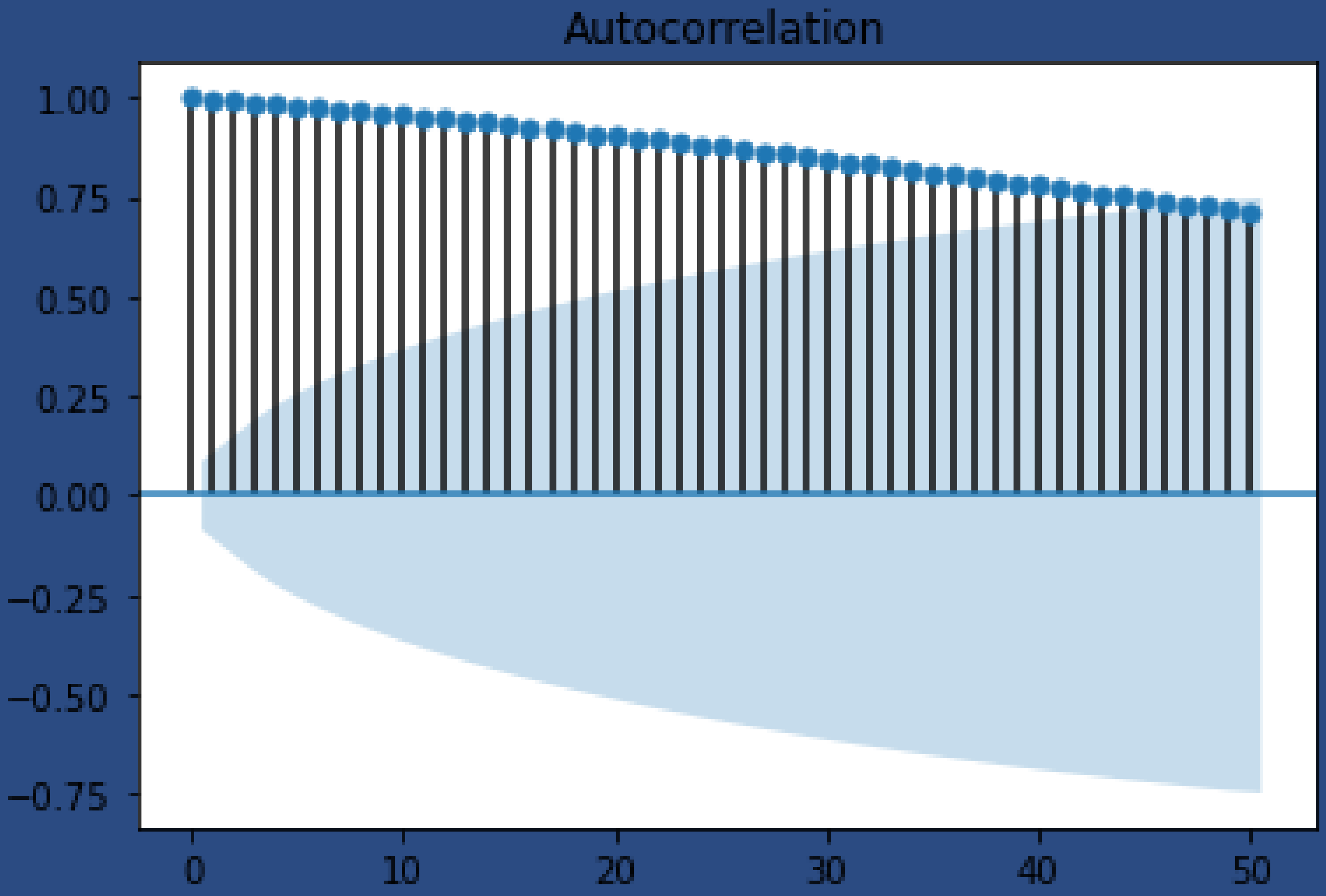
- We have taken the lag to be around 25 and 50 days for the calculation of autocorrelation because we cannot observe any significant changes among the cases if the difference is very low, i.e. consecutive days would show a nominal rise in cases, which would not be sufficient to draw any conclusions.

For lag=25, we can conclude that the temporal data is extremely correlated to each other on the basis of the previous case, as they lie in the range 0.8<x<1.0. Highly correlated.



For lag=50, the further we get the lower is the autocorrelation among the temporal data.

We can conclude that the temporal data is extremely correlated to each other on the basis of the previous case, as they lie in the range of 0.7<x<1.0.



Data for training model

In order to fully optimize and use the information available in the spatial and the temporal data of the hotspot district, and to feed the model this info.

We are calculating the entropies for the corresponding columns.

Permutation Entropy (PE) is a robust time-series tool that provides a quantification measure of the complexity of a dynamic system by capturing the order relations between values of a time series and extracting a probability distribution of the ordinal patterns

Equation used:

derived(x) = current_data*entropy_current_data + spatial_data*entropy_spatial_data + temporal_data* entropy_temporal_data

Why LSTM?

We have chosen the LSTM model because

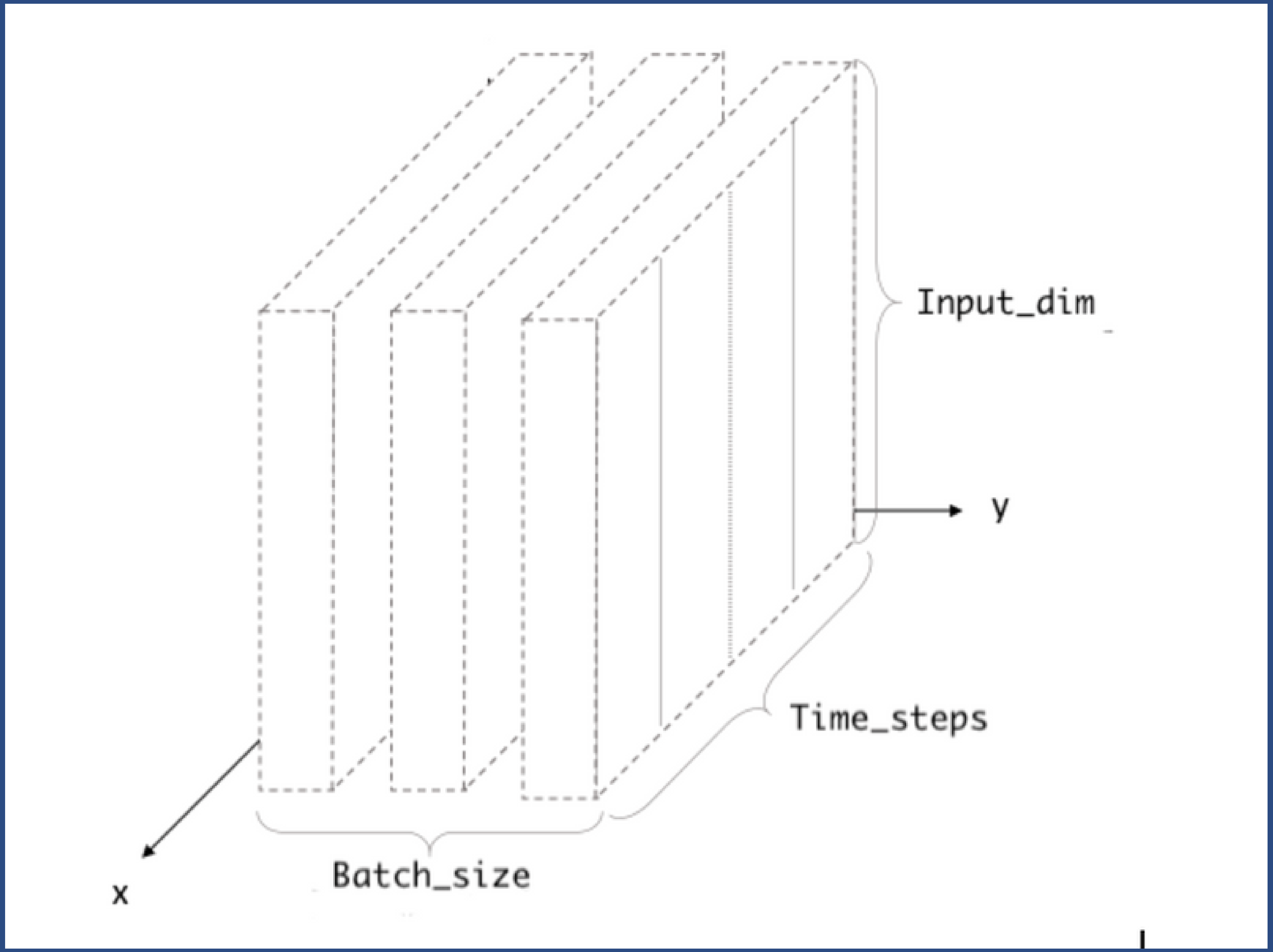
- It is an extension of the Recurrent Neural Network.
- And as for the covid data we know that the data is not completely independent of itself, it depends on past data as well as the cases of areas around the hotspot, i.e. the ones highly correlated.
- LSTM uses the previous data to predict the future value. Thus it does not leave any important aspect or external factor that may drastically affect the prediction of future cases.

About the Prediction Model

- The modified data which was shown above is feature scaled . After feature scaling, it had a total of (521 rows of data). In order to use LSTM, our input and output data should have a specific shape.

- The input data in an LSTM model is a 3D array where the first dimension represents the number of samples (or batch size) as the number of rows of data in a two-dimensional setting, the second dimension stands for time steps which indicate the amount of time that we want to go back through time, and the third dimension shows the number of features that we want to include in the model for every element in our batch.

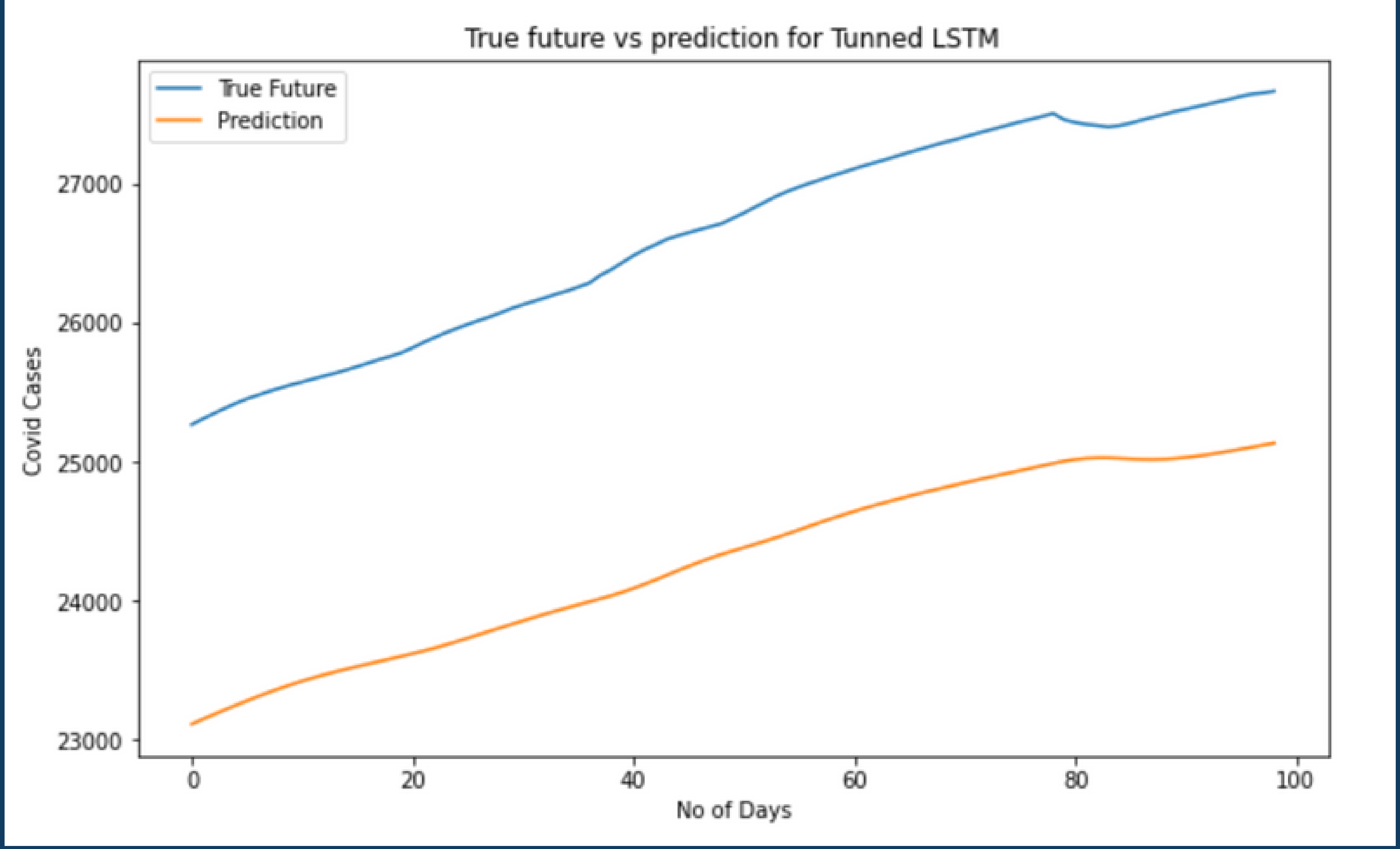
- So, it is like [number_of_samples, time_steps, input_dim].



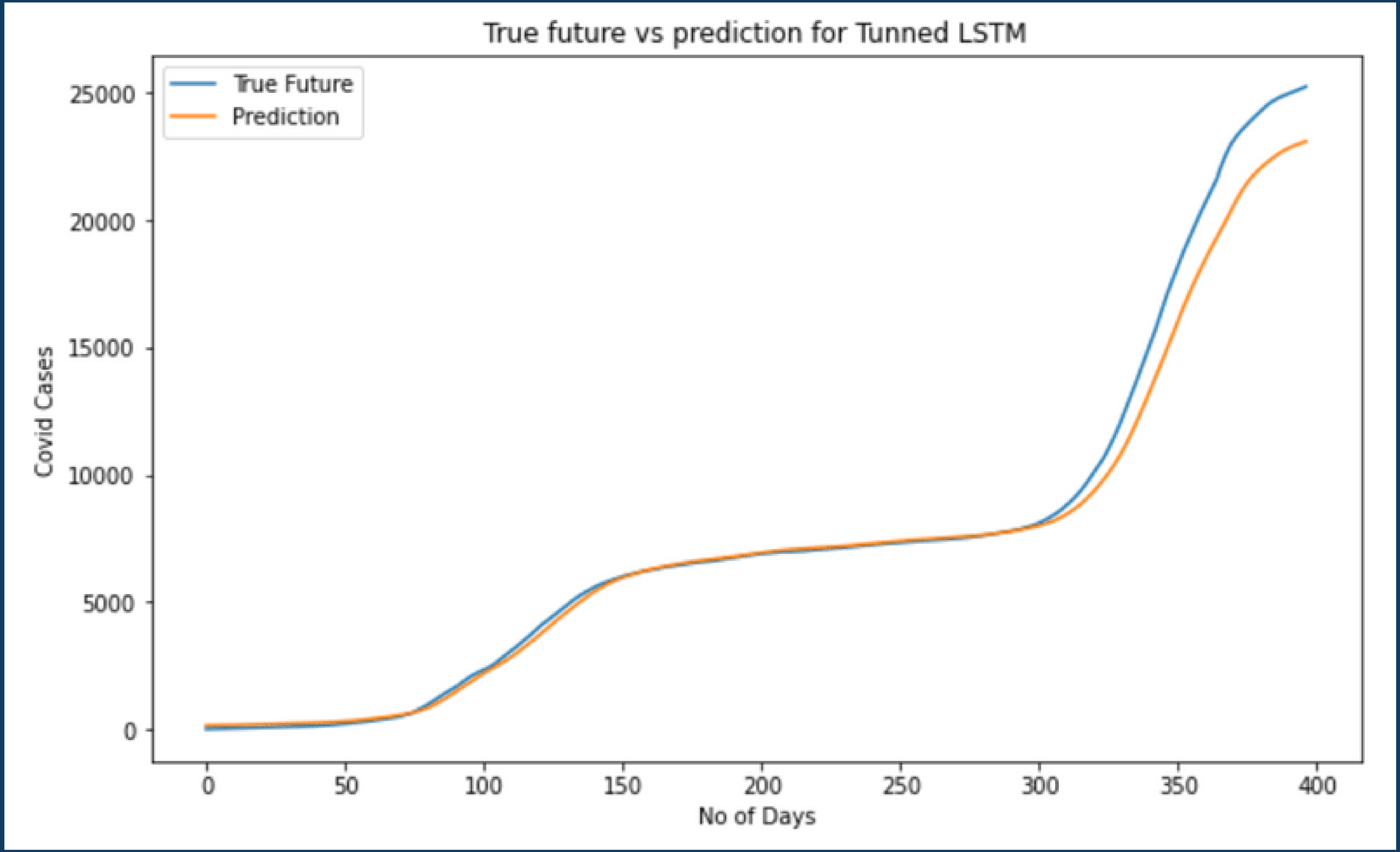
For the training and testing data we have used 80-20 split, with a batch or n value for lags as 64. Around 2 months for the prediction of the next day's cases.

Illustration of LSTM input and output data shape.

Results:



x= Testing data
y= predicted scaled values

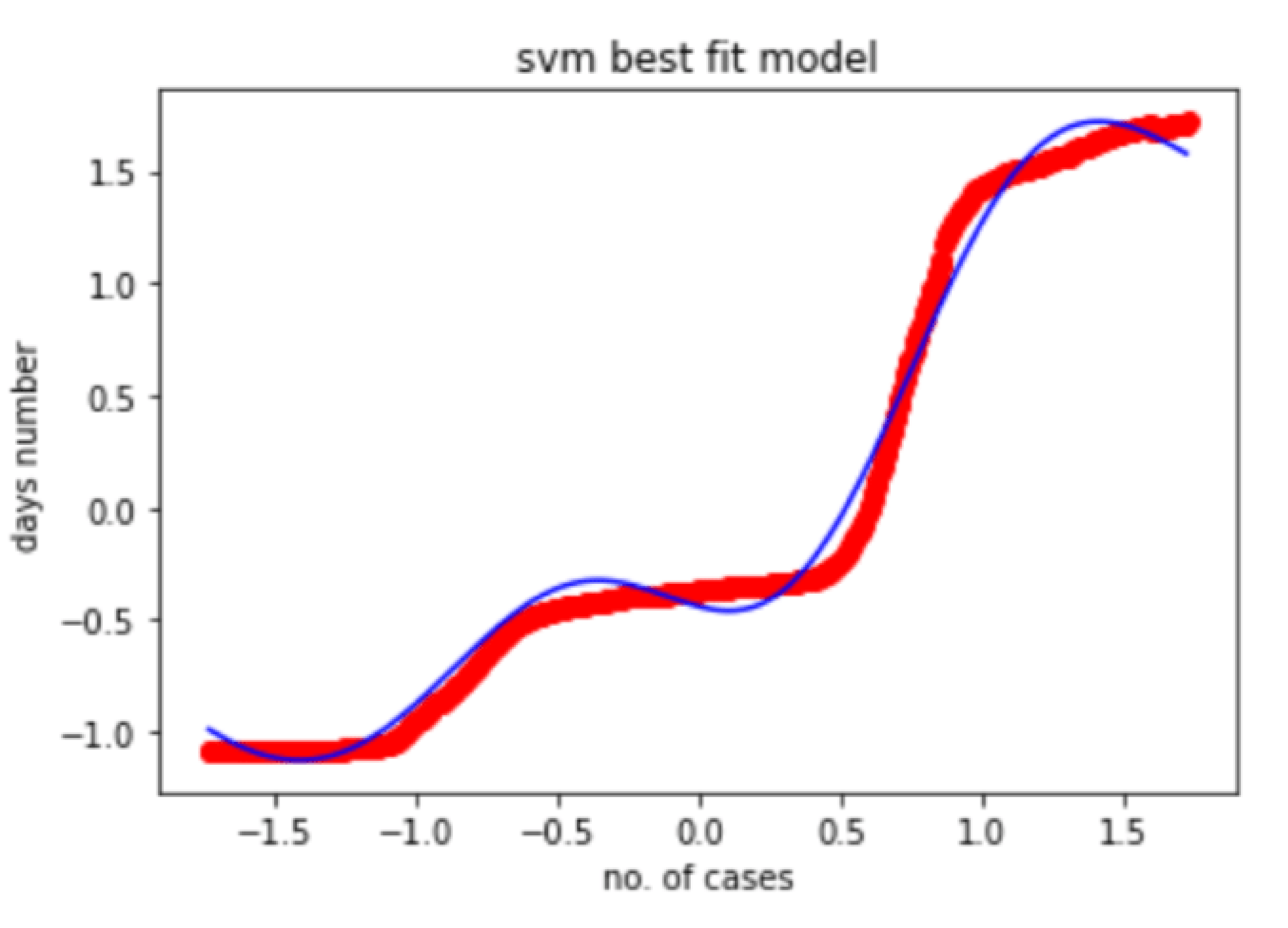


For training data

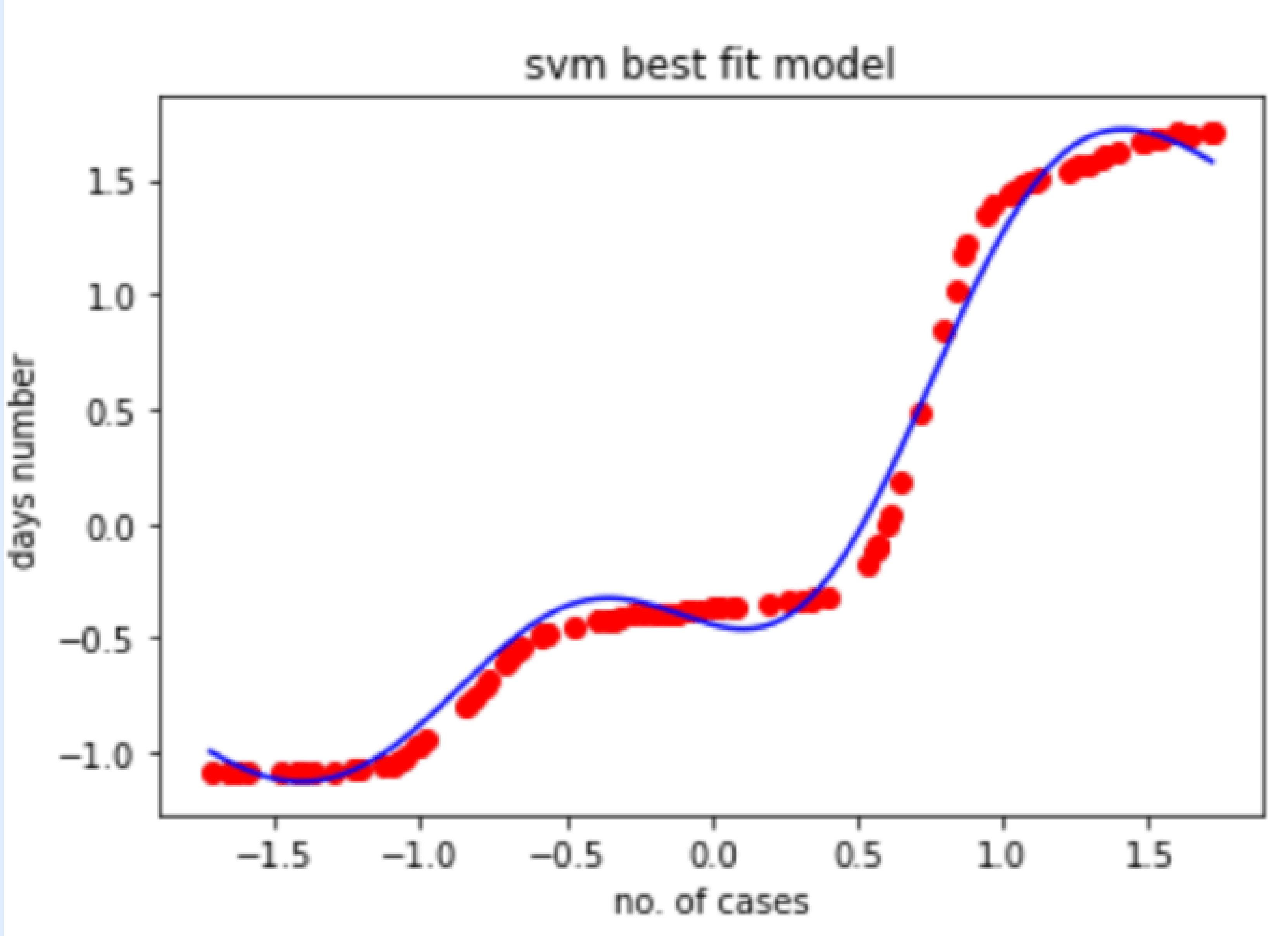
SVM RESULTS

If we use the svm model for the same purpose, the best fit line obtained looks like this.

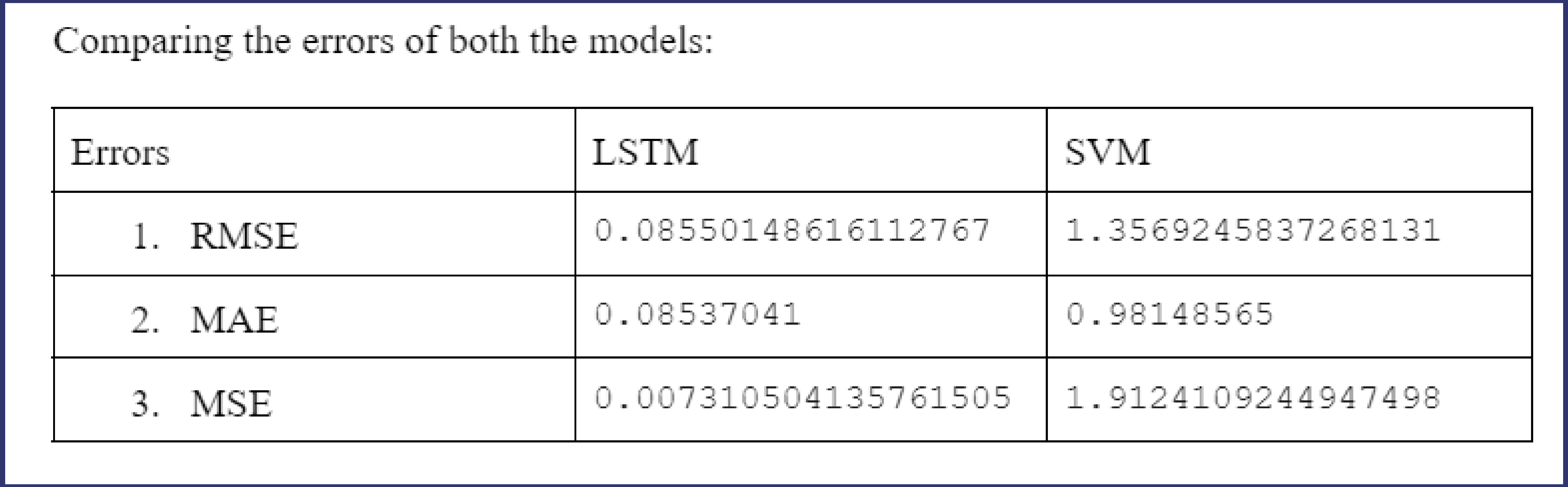
Testing data



Training data



Comparing of the errors in both the models:



Comparing the errors of both the models:

Errors	LSTM	SVM
1. RMSE	0.08550148616112767	1.3569245837268131
2. MAE	0.08537041	0.98148565
3. MSE	0.007310504135761505	1.9124109244947498

RMSE - root mean squared error
MAE - mean absolute error
MSE - mean squarred error

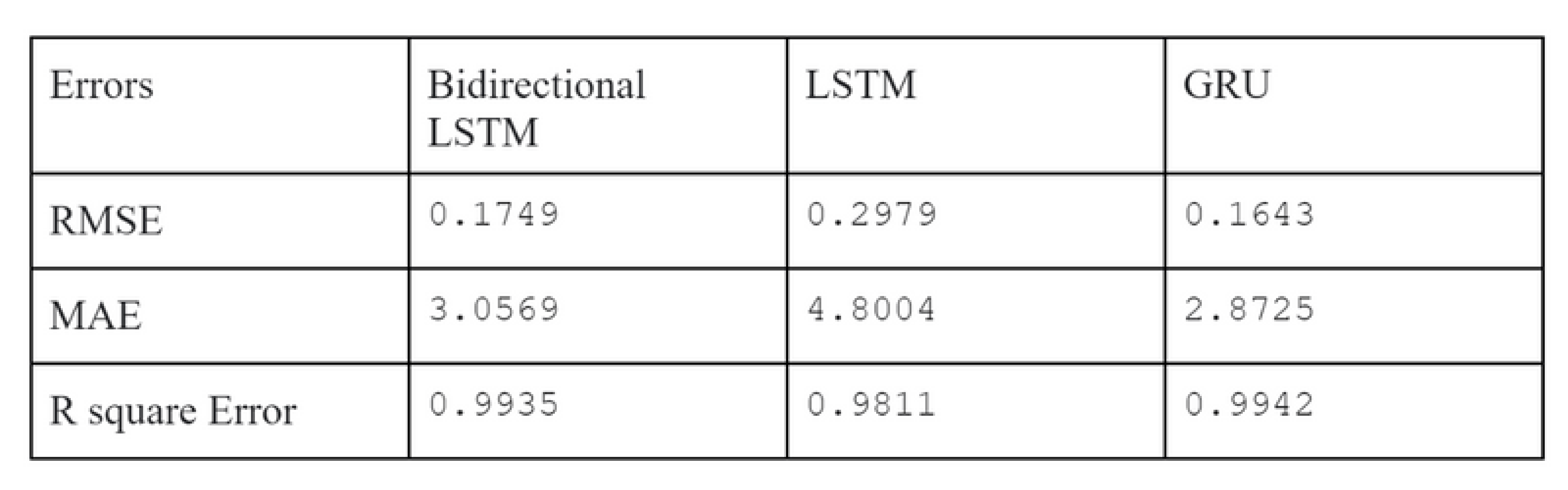
Clustering Model using DBSCAN

- This is also part of our future work.
- We consider Pune as hotspot for our clustering model.
- From observation, taking Pune as the hotspot for our clustering model is most appropriate because it is the reason for a covid spread in the entire state so rapidly due to a large number of new cases each day, and the presence of international airport. It may not be the most spatially correlated district but still has the most significant effect on the covid spread in Maharashtra.

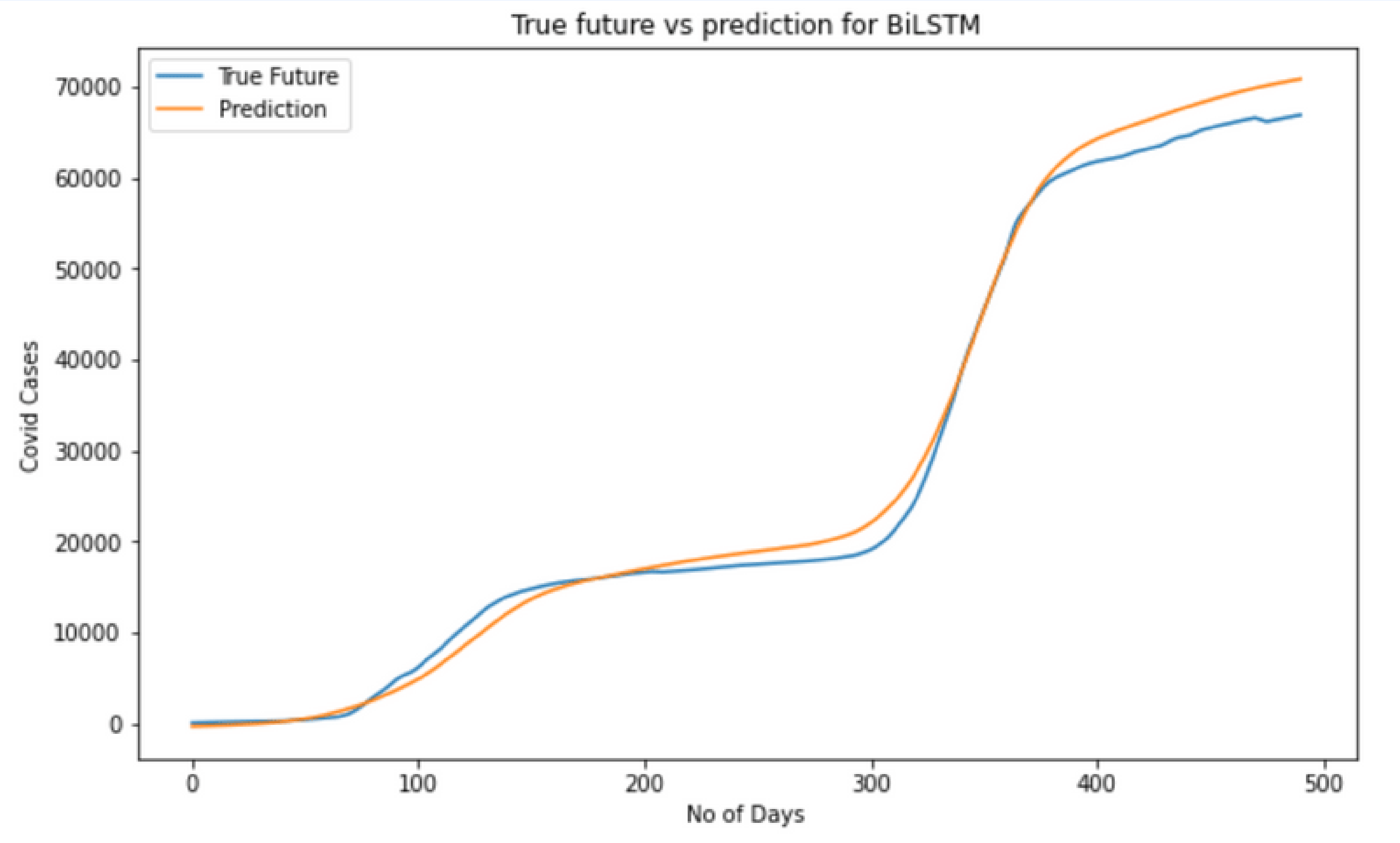
Predicting current values using Spatial and Temporal Values:

We have used three models, namely Bidirectional LSTM, LSTM, and GRU for the prediction of the current Covid cases from 2020 to 2021, from spatial and the temporal data.

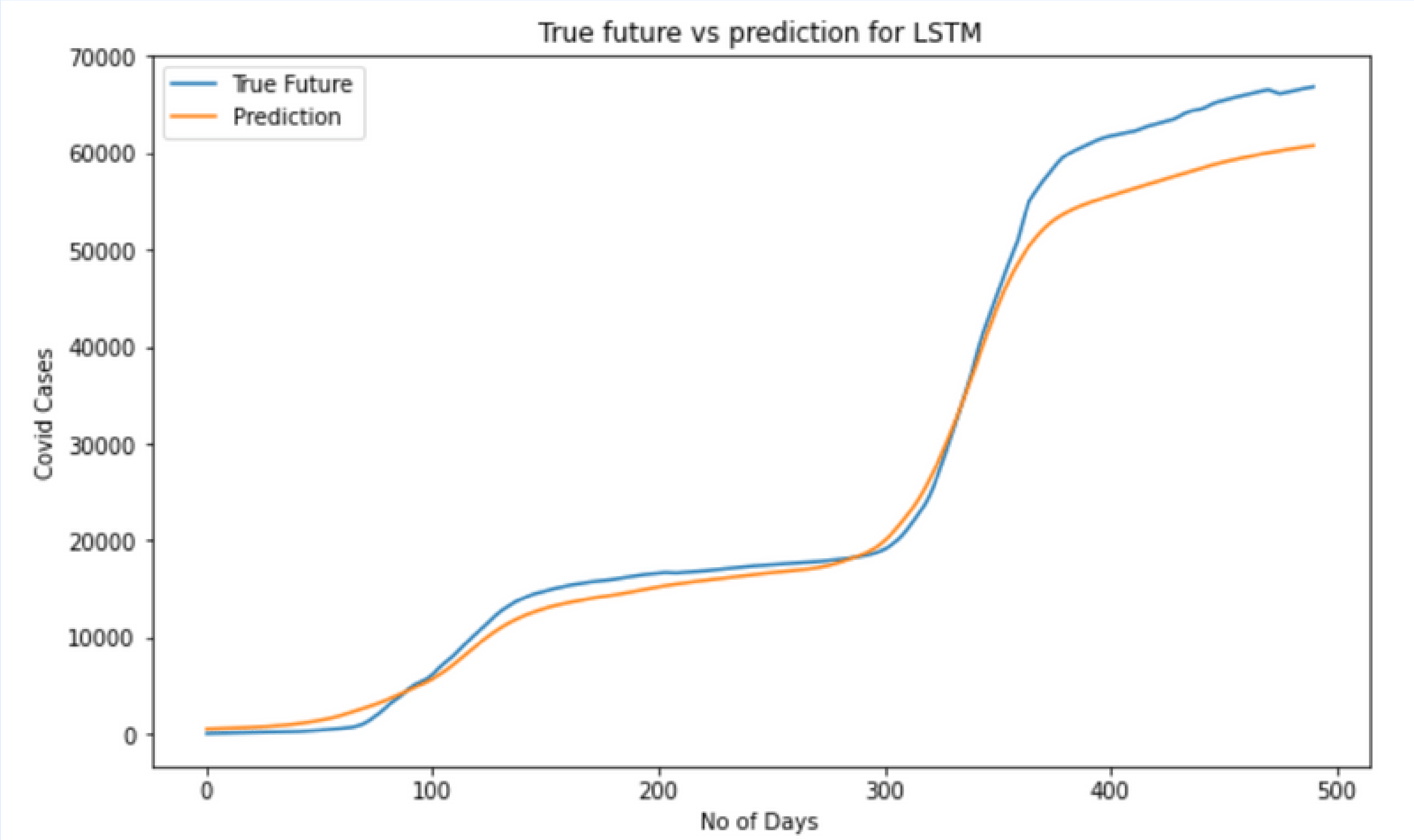
Error comparison from results.



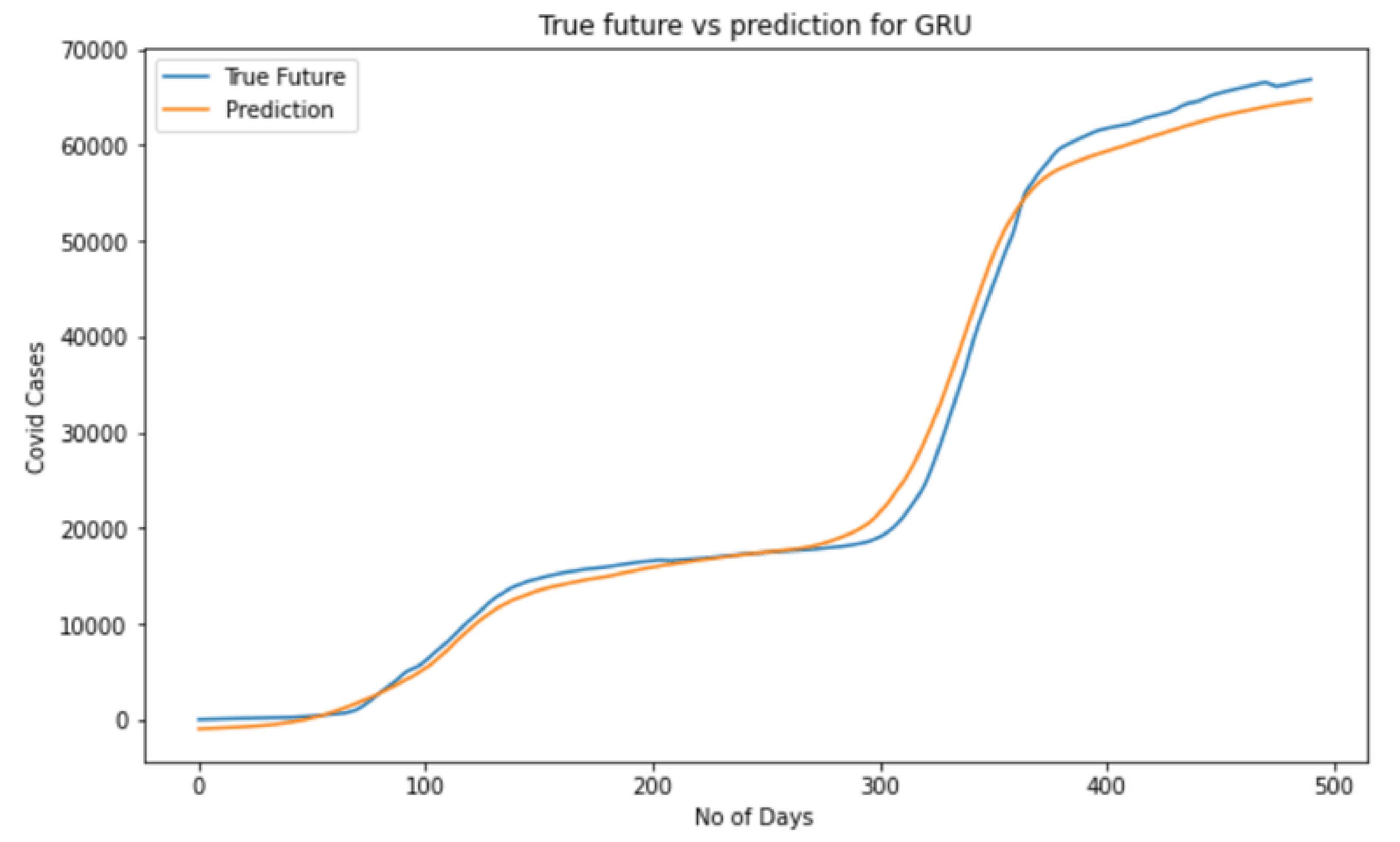
Errors	Bidirectional LSTM	LSTM	GRU
RMSE	0.1749	0.2979	0.1643
MAE	3.0569	4.8004	2.8725
R square Error	0.9935	0.9811	0.9942



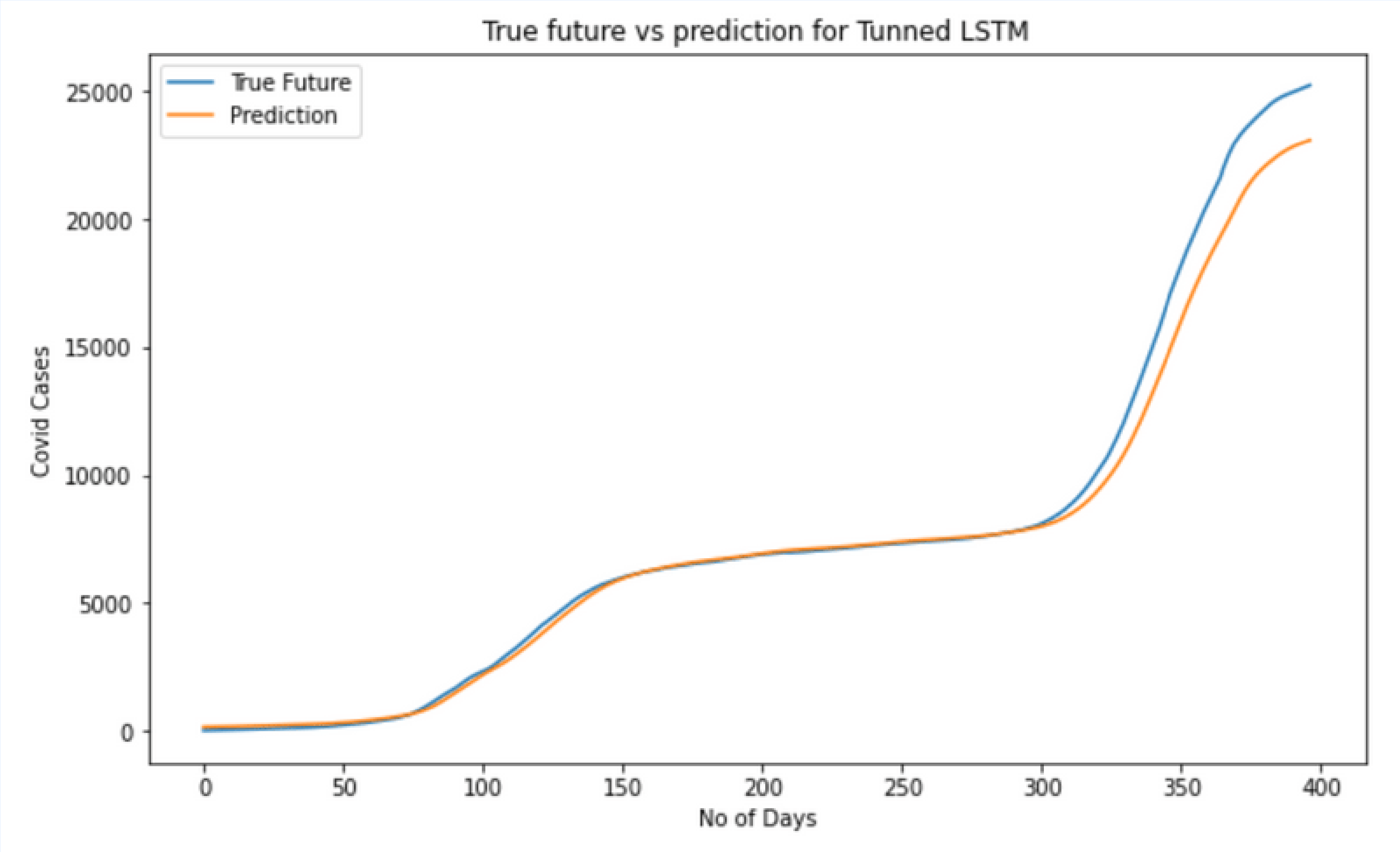
True future vs prediction for BiLSTM



True future vs prediction for LSTM

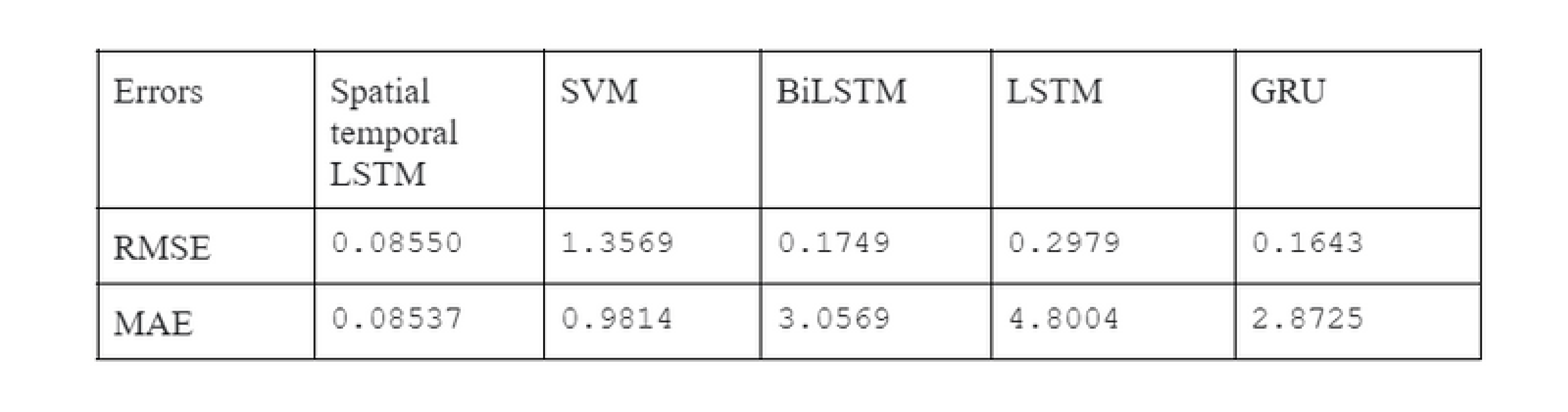


True future vs prediction for GRU



True future vs prediction for Tunned LSTM

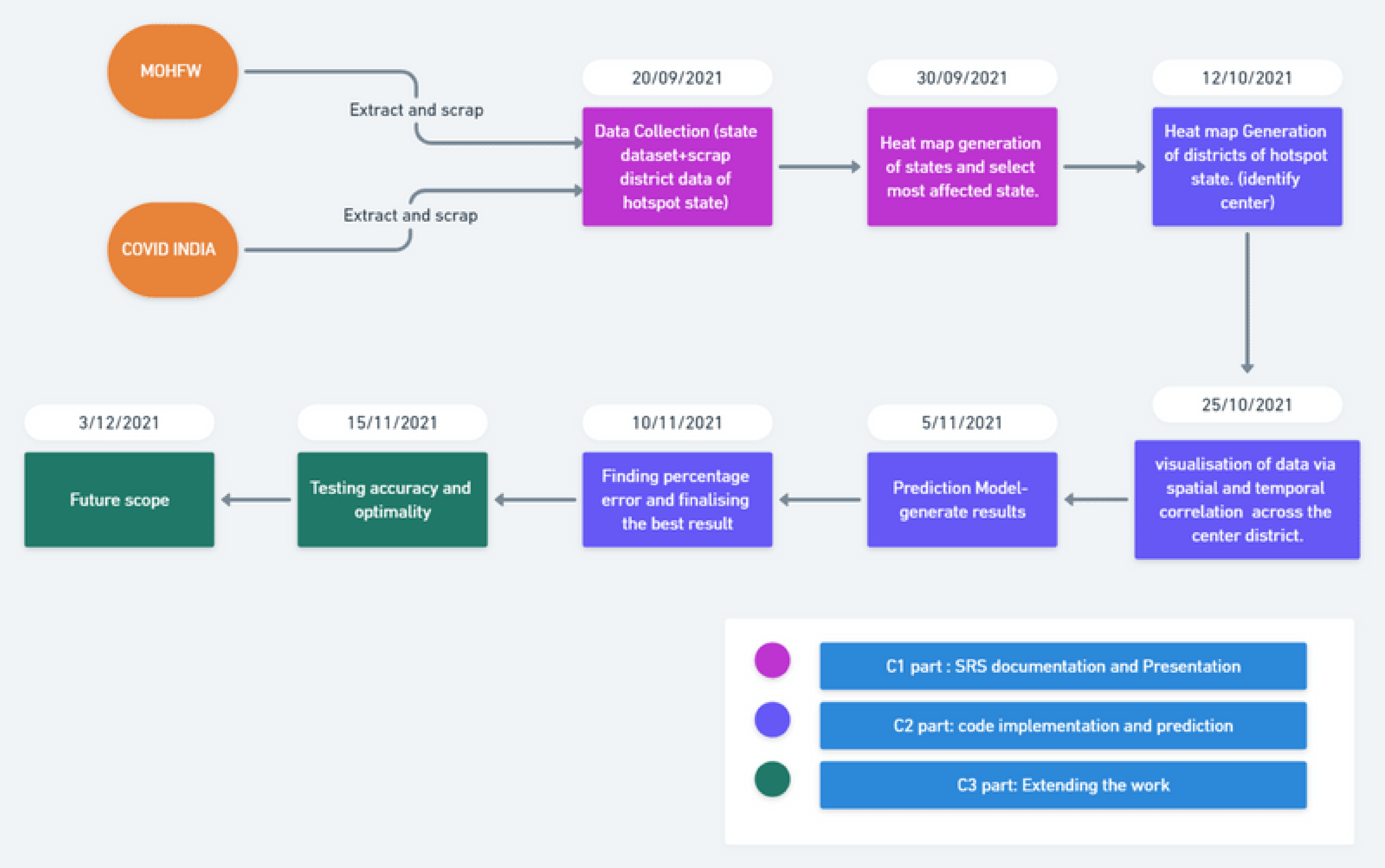
Final results



Errors	Spatial temporal LSTM	SVM	BiLSTM	LSTM	GRU
RMSE	0.08550	1.3569	0.1749	0.2979	0.1643
MAE	0.08537	0.9814	3.0569	4.8004	2.8725

Thus concluding, our spatial temporal LSTM has the best results with least values for errors.

Work division/Activity Diagram



MOHFW

Extract and scrap

COVID INDIA

Extract and scrap

20/09/2021

Data Collection (state dataset+scrap district data of hotspot state)

30/09/2021

Heat map generation of states and select most affected state.

12/10/2021

Heat map Generation of districts of hotspot state. (identify center)

25/10/2021

visualisation of data via spatial and temporal correlation across the center district.

5/11/2021

Prediction Model- generate results

10/11/2021

Finding percentage error and finalising the best result

15/11/2021

Testing accuracy and optimality

3/12/2021

Future scope

● C1 part : SRS documentation and Presentation

● C2 part: code implementation and prediction

● C3 part: Extending the work

References:

[1] Huang R, Liu M, Ding Y. Spatial-temporal distribution of COVID-19 in China and its prediction: A data-driven modeling analysis. J Infect Dev Ctries. 2020 Mar 31;14(3):246-253. doi: 10.3855/jidc.12585. PMID: 32235084.

[2] Devi, Rani, Smrutishree Lenka, Kiran M. Hungud, and S. Himesh. "Analyzing Spatio-Temporal Spread of Covid19 in India."

[3]. Feng, Cindy. "Spatial-temporal generalized additive model for modeling COVID-19 mortality risk in Toronto, Canada." *Spatial statistics* (2021): 100526.

References Cont.

[4] Feng, Xinxin, Xianyao Ling, Haifeng Zheng, Zhonghui Chen, and Yiwen Xu. "Adaptive multi-kernel SVM with spatial–temporal correlation for short-term traffic flow prediction." *IEEE Transactions on Intelligent Transportation Systems* 20, no. 6 (2018): 2001-2013.

[5] Xie, Zhixiang, Yaochen Qin, Yang Li, Wei Shen, Zhicheng Zheng, and Shirui Liu. "Spatial and temporal differentiation of COVID-19 epidemic spread in mainland China and its influencing factors." *Science of The Total Environment* 744 (2020): 140929

[6]shiva-verma.medium.com/understanding-input-and-output-shape-in-lstm-keras-c501ee95c65e

THANK YOU