Indian Institute of Information Technology Allahabad

Mini Project- Semester 5

Spatial-temporal distribution of COVID-19 in India and its prediction: A data-driven modeling analysis

Supervisor - Dr. Vijay K. Chaurasiya

Submitted by :

IIT2019021 Medha Balani
IIT2019027 Vidushi Pathak
IIT2019032 Aarushi
IIT2019036 Jyotika Bhatti

Contents:

1. 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 global coordinated effort is needed to stop the spread of this disease.
Coronavirus disease (COVID-19), caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), presents an unprecedented threat to global health worldwide. Despite public health responses aimed at slowing the spread of the epidemic, confirmed case counts and hospitalizations continue to surge[3]. A novel coronavirus (nCoV) is a new strain that has not been previously identified in humans. It is very much fatal to people with low immunity and its ability to spread from person to person has made people be more serious about this disease.

In this paper, we are about to discuss the spread of this disease from one province to another and their probability to find the expected number of cases in the surrounding states using spatial-temporal. We would use the application of the Moran index, a strong statistical tool, to the spatial panel to show that COVID-19 infection is spatially dependent and mainly spread from a certain particular Province in Central India to neighboring areas. The logistic model would be employed according to the trend of available data, which shows the difference between the particular Province and outside of it. This would be made a kind of generalized model for any of the particular provinces where the number of cases is extremely high.

2. Problem Statement:

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, by giving the government a fair amount of time to take appropriate measures. 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.

3. Objective:

The objective here is the **spatial-temporal distribution of COVID 19 data** and its prediction in the nearby areas. The outbreak of COVID 19 in India claimed the lives of thousands of people. Our main objective is to use such a strong statistical model in order to show that COVID 19 infection is spatially dependent and is mainly spread via to the neighboring areas.

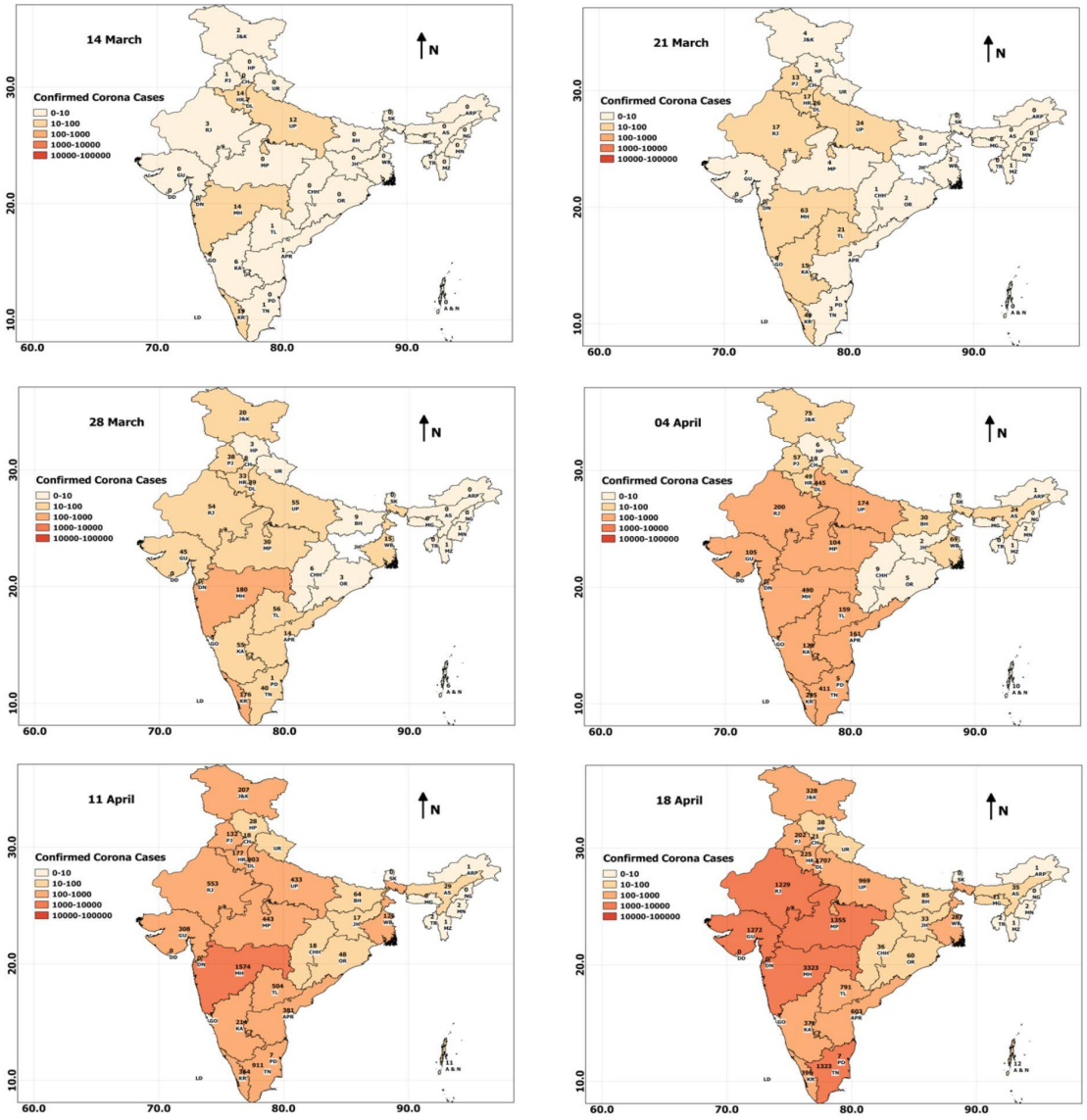


Figure 1: Spatial distribution of COVID-19 cases in India for weeks ending on 14th March to 18th April 2020.
(K C Gouda, Nikhila Suman P, Kumari R, Singh P, Benke M, et al. (2020) Analyzing Spatio-Temporal Spread of Covid19 in India.)

In the above figure, you can see that from 14 March 2020 cases started spreading from one state to its neighboring states. So our problem statement is to analyze how the increase in covid cases in one state/city(hotspot) will affect the neighboring states/cities.

4. Literature Review:

Talking about the methods/theories that have been read/discussed till now over this problem.

4.1 Spatial-temporal distribution of COVID-19 in China and its prediction [1]

The attribute values of adjacent region units are characterized by Moran Index, There are two general descriptions for spatial weight matrix, one of which is based on geographical distance and another on binary adjacency matrix, whose row i and column j can be expressed as
W(i, j) = { 1 if region i and j are connected , else 0}.
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(I), contacts an individual in a susceptible state, the probability of getting infected is β. Individuals in the incubation period (E) become infected with probability $\gamma\gamma 1$ in unit time (day) (I); An individual in an infected state (I) is converted to a cured state (R) with probability $\gamma\gamma 2$ per unit time (day). [1]

Real Data analysis :
Spatial feature: In order to test the spatial autocorrelation of COVID 19, the destination between the cities and the provinces is required, which is estimated through Moran Index.[1]
The values of the moran coefficients will vary according to the confirmed cases in accordance with its geographical structure as well as the spatial geographical distance. [1]

Epidemic prediction based on logistic model:
The Transmission of COVID 19 takes place in a limited population and this transmission rate will decline after reaching a certain threshold with the implementation of various other contaminants and factors. This whole phenomenon is fully consistent with the logistic curve. The paper [1] adopted the Moran index to deduce the COVID-19 has had a spatial correlation in China and paid more attention to analyzing the confirmed cases of this outbreak from the perspective of spatial measurement and statistical modeling. Because nucleic acid testing is time-consuming and labor-intensive and requires

professional equipment and technical personnel, data bias was inevitable, especially in Wuhan, the city with the most severe epidemic. [1]

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

The city contains 140 neighborhoods that were aggregated from census
tracts and created by the Social Policy Analysis and Research Unit in the City's Social Development and the Administration Division. Data on n = 49,216 COVID-19 confirmed cases from March 1, 2020, until December 10, 2020, in Toronto, Ontario, Canada, were retrieved from the Ontario Ministry of Health. Of those, 1938 (3.94%) died from COVID-19.[3]
Here, Yi denotes the mortality status due to COVID-19 for the ith individual, from i = 1, . . . , n, which follows Yi ~ Bernoulli(πi) where πi denotes the probability of mortality. A spatial-temporal GAM for modeling the mortality risk of COVID-19 is formulated as follows,

$$g(\Pi_i) = \alpha_0 + X_i\beta + f_t(day_i) + f_s(lat_i + long_i) + f_{ngb}(pop\ density_i, income_i) + f_{st}(lat_i, long_i, day_i) \quad (1)$$

where g(·) denotes the link function.
The author's study demonstrates a model developed in a generalized additive modeling (GAM) framework (Hastie and Tibshirani, 1990; Wood, 2004, 2017, 2011) is sufficiently flexible to model the spatial–temporal dynamics of COVID-19 mortality risk, which is computationally fast with good discrimination and calibration.

Here, the author has done a comparative analysis of three models' performance with three different functions namely, logit 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.

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

The paper mainly deals with the 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).
Here, the author discusses the randomness characteristic of traffic flow and hybridizes Gaussian kernel and polynomial kernel with different weights to constitute the AMSVM. Then they propose the APSO algorithm to optimize the parameters of AMSVM [4]. 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. The 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 distributions, the performance of SVM depends largely on the choice of the kernel function.[4].

For temporal correlation obtain the traffic data of current point I, and then calculate the correlation of traffic flow between current point and its distant history in the same day of previous h weeks according to Pearson correlation coefficient, which is

$$R_{mT} = \frac{Cov(XI, XmT),}{\sigma I\, \sigma mT} \qquad mT = 1, 2, \cdots, h,$$

where σI and σmT are standard deviations of XI and XmT separately, and they can be calculated by the corresponding sample data. Finally, the predicted value of current point in the next period is calculated by

$$P_{iT}(t+1) = \frac{1}{h} \sum_{m^T=1}^{h} [R_{m^T} . P_{m^T}(t+1)] \qquad \text{----(i)}$$

Talking about spatial correlation, the spatial correlation is obtained by analyzing the traffic flow at the current point and its adjacent areas on the same day of the previous week. The effects of the selected correlative point to the current point are evaluated by calculating the Pearson correlation at certain times, which is

$$R_{n^s} = \frac{cov(X_1 . X_{n^s})}{\sigma_1 . \sigma_{n^s}} \qquad n^s = 1, 2, \cdots, r,$$

where σ_{n^s} means the standard deviation of X_{n^s} , r is the number of these correlative points.

By doing some optimizations using spatial-temporal correlation on the previous AMSVM, the new proposed model showed better timely and adaptive prediction even in the rush hour when the traffic conditions change rapidly[4].

4.4 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 was described by the exploratory spatial data analysis method. Then, the key factors affecting the COVID-19 epidemic spread were identified by using the detector method, so as to provide references for clarifying the epidemic spread rule, formulating some protection policies, and promoting the resumption of work and production.

Firstly, the epidemic spread rate is calculated for different regions and then the exploratory spatial data analysis method is applied to verify whether the observed value of a unit has a spatial correlation with the observed values of its neighboring units. The global Moran's I index is used to measure the global spatial correlation, while the local Moran's I index in LISA (local indicators of spatial association) was used to measure the local spatial correlation.

In this paper, the cumulative number of confirmed COVID-19 cases and the epidemic spread rate were taken as variables, the spatial weight matrix based on geographical adjacency was selected, and the global Moran's I index, the P test value and the Z statistic score of the cumulative number of confirmed COVID-19 cases and the epidemic spread rate were calculated by using the GeoDa software, so as to clarify the global spatial correlation characteristics.

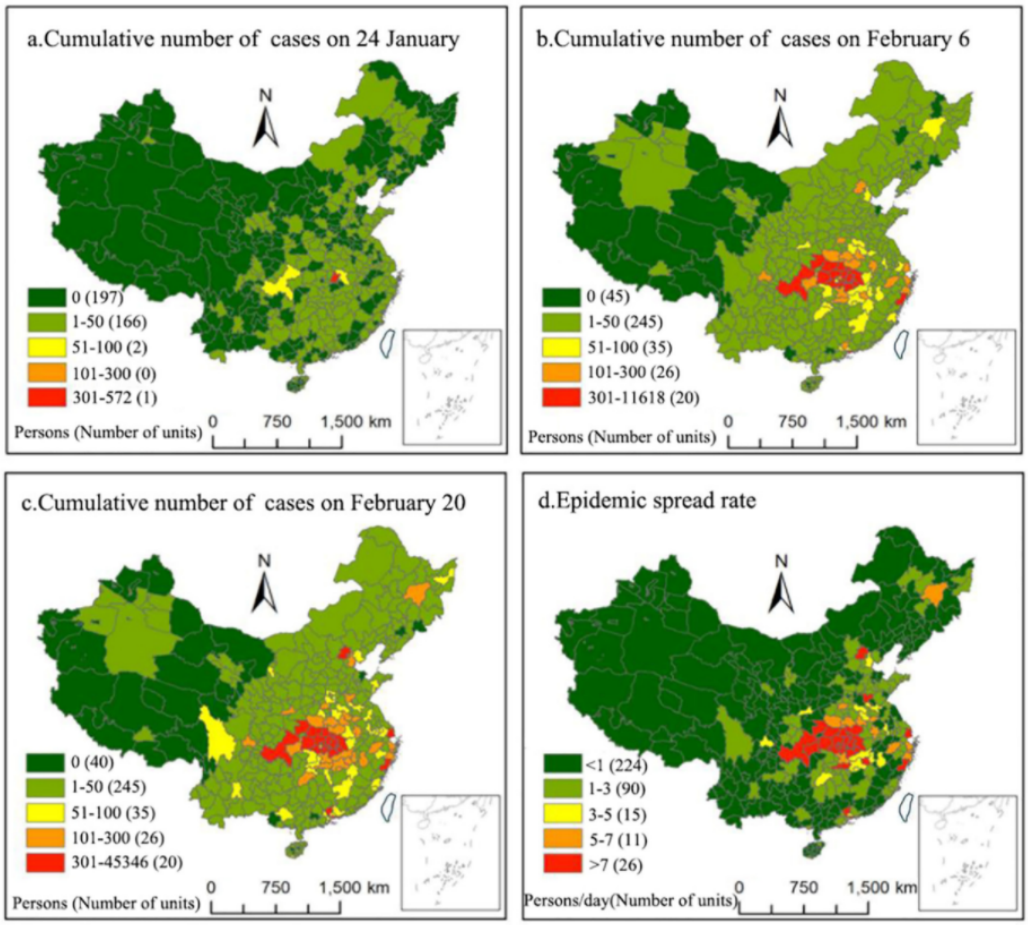


Fig. 1. Spatial distribution of the cumulative number of COVID-19 cases and epidemic spread rate.

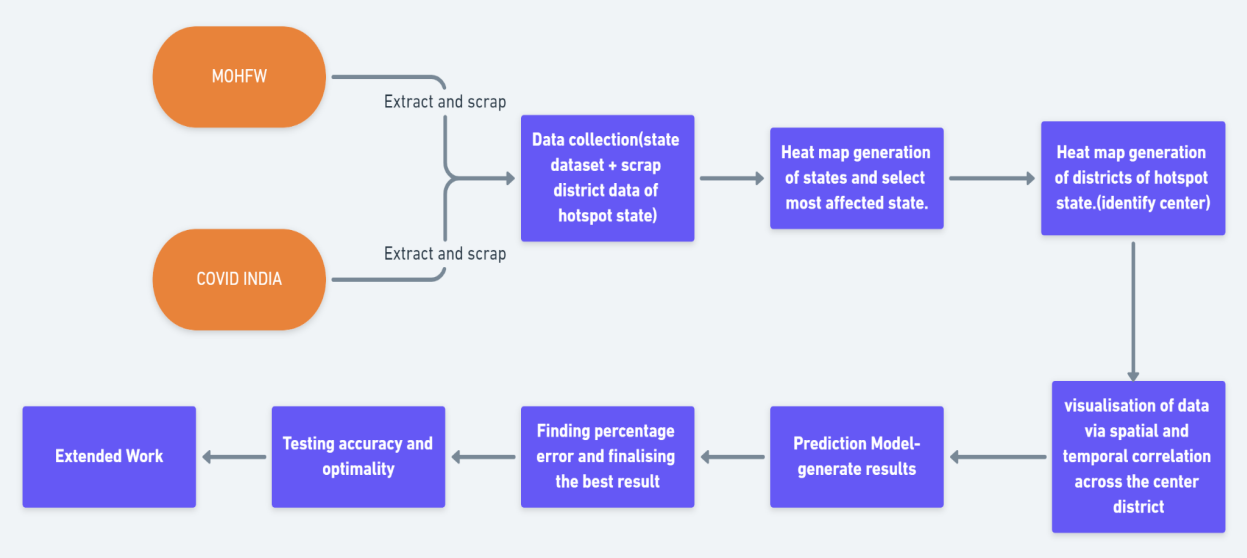
[5]

Overall, the 'high-high cluster areas showed a layout trend from centralization to decentralization. Accordingly medical facilities can be taken care of and all other necessary actions can be taken. So, this paper focused on the relevance of spatial-temporal data and trends verification, influencing factors.

5. Proposed Flow/Methodology:

Firstly, we would take a dataset of states of India, and then we would identify the hotspot state among them. Then we will scrap the data of cities/districts of that state.

After scraping and cleaning the data we would apply the above-described strategy to analyze and visualize data of neighboring cities. Then we will select a center among those cities where the covid cases are the highest, then we would further use the data in order to train our model for prediction, and test its accuracy, and optimality.



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.

6. About the dataset

6.1 Initial Dataset (for hotspot state):

The dataset initially taken was a time series dataset consisting of all the statewise COVID 19 data of India from 14th March 2020 till 18th September 2021. This dataset was downloaded from Kaggle which has a total of 1662 rows and 42 columns.

Its description is given below.

	Date	Date_YMD	Status	TT	AN	AP	AR	AS	BR	CH	CT	DN	DD	DL	GA	GJ	HR	HP	JK	JH	KA	KL	LA	LD	MP	MH	MN	ML	MZ	NL	OR	PY	PB	RJ	SK	TN	TG	TR	UP	UT	WB	UN
0	14-Mar-20	2020-03-14	Confirmed	81	0	1	0	0	0	0	0	0	0	7	0	0	14	0	2	0	6	19	0	0	0	14	0	0	0	0	0	0	1	3	0	1	1	0	12	0	0	0
1	14-Mar-20	2020-03-14	Recovered	9	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	4	0	0	0
2	14-Mar-20	2020-03-14	Deceased	2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	15-Mar-20	2020-03-15	Confirmed	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	18	0	0	0	0	0	0	1	0	0	2	0	1	0	0	0	0
4	15-Mar-20	2020-03-15	Recovered	4	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0	0	0	0	0	0

	Date	Date_YMD	Status	TT	AN	AP	AR	AS	BR	CH	CT	DN	DD	DL	GA	GJ	HR	HP	JK	JH	KA	KL	LA	LD	MP	MH	MN	ML	MZ	NL	OR	PY	PB	RJ	SK	TN	TG	TR	UP	UT	WB	UN
1657	17-Sep-21	2021-09-17	Recovered	33833	2	1296	69	477	14	1	31	1	0	56	75	20	12	198	135	17	1199	20388	3	0	16	4410	545	99	1069	46	719	129	31	7	56	1565	298	83	16	23	727	0
1658	17-Sep-21	2021-09-17	Deceased	285	0	8	0	9	0	0	1	0	0	1	1	0	0	0	1	0	18	131	0	0	0	67	5	1	3	4	4	1	0	0	0	17	2	1	1	0	9	0
1659	18-Sep-21	2021-09-18	Confirmed	31130	2	1174	32	365	10	4	28	0	0	41	123	13	6	174	152	3	889	19325	0	3	6	3391	140	119	1476	60	695	128	30	8	54	1653	255	22	0	21	728	0
1660	18-Sep-21	2021-09-18	Recovered	39652	2	1309	76	465	10	3	42	3	0	44	81	24	6	142	113	23	1080	27266	0	1	18	3841	228	192	928	37	756	122	23	10	81	1581	329	45	0	14	757	0
1661	18-Sep-21	2021-09-18	Deceased	306	0	9	0	2	0	0	0	0	0	0	2	0	0	2	0	0	14	143	0	0	0	80	4	3	3	1	6	1	0	0	0	22	1	0	0	1	12	0

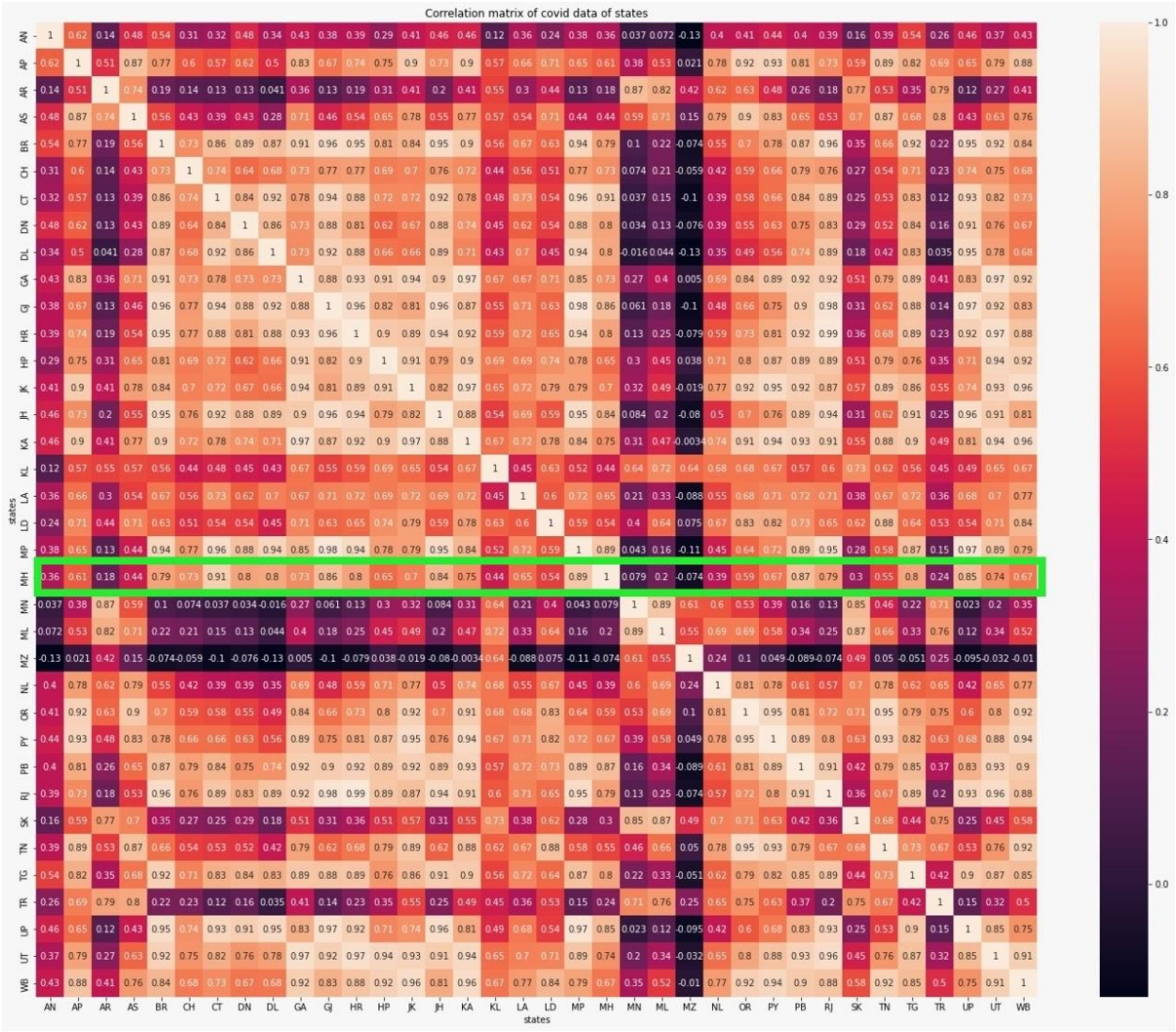
On preprocessing the above dataset, by only considering the confirmed cases, and dropping the null values, we get the final dataset to visualize the hotspot state of COVID spread in India.

The final dataset is:

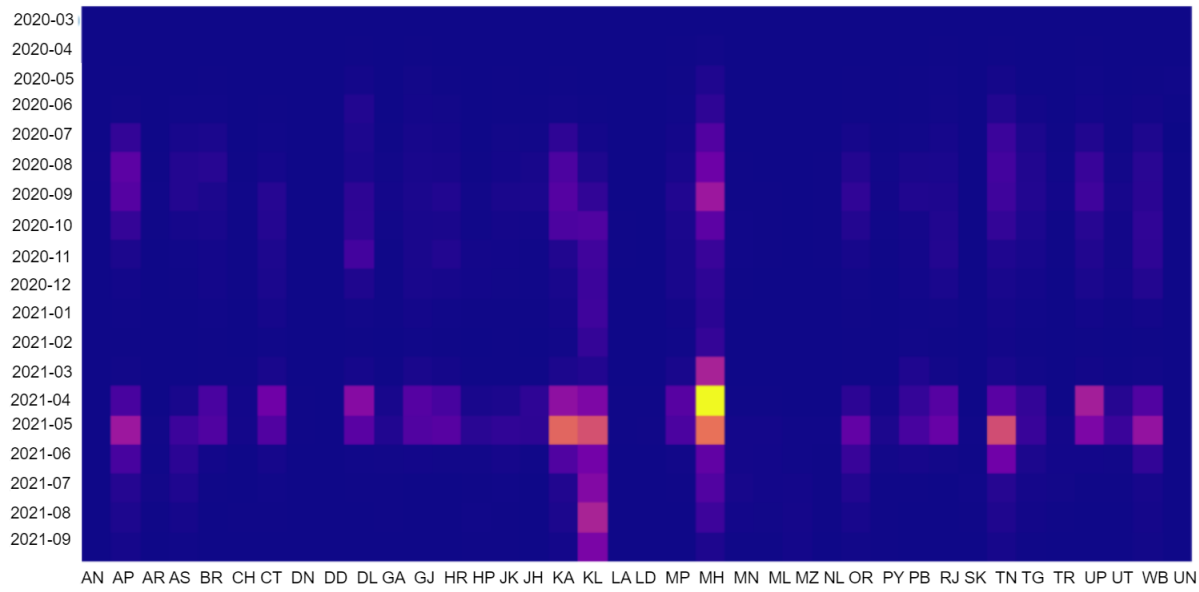
	Date	Date_YMD	AN	AP	AR	AS	BR	CH	CT	DN	DD	DL	GA	GJ	HR	HP	JK	JH	KA	KL	LA	LD	MP	MH	MN	ML	MZ	NL	OR	PY	PB	RJ	SK	TN	TG	TR	UP	UT	WB	UN
1647	14-Sep-21	2021-09-14	8	1125	65	493	12	2	35	0	0	38	109	11	17	195	150	7	559	15876	8	0	7	3530	231	127	1502	37	428	103	37	11	46	1591	336	52	32	19	703	0
1650	15-Sep-21	2021-09-15	0	1445	71	444	7	2	20	0	0	57	88	15	12	215	156	10	1116	17681	9	0	7	3783	236	243	1185	37	457	124	41	17	42	1658	324	43	18	49	743	0
1653	16-Sep-21	2021-09-16	3	1367	47	468	12	4	31	0	0	28	95	22	9	127	170	6	1108	22182	6	0	7	3595	216	229	1402	32	580	107	30	4	64	1693	259	0	23	20	707	0
1656	17-Sep-21	2021-09-17	1	1393	38	270	7	4	26	0	0	55	108	25	8	209	155	9	1003	23260	71	0	6	3586	208	248	1121	17	628	86	26	8	37	1669	241	72	15	25	719	0
1659	18-Sep-21	2021-09-18	2	1174	32	365	10	4	28	0	0	41	123	13	6	174	152	3	889	19325	0	3	6	3391	140	119	1476	60	695	128	30	8	54	1653	255	22	0	21	728	0

Further on visualizing the correlation between them using the correlation matrix, which is **the measure that is best used in variables** that demonstrate a linear relationship between each other.

The correlation Matrix of COVID DATA OF STATES for India is shown below.

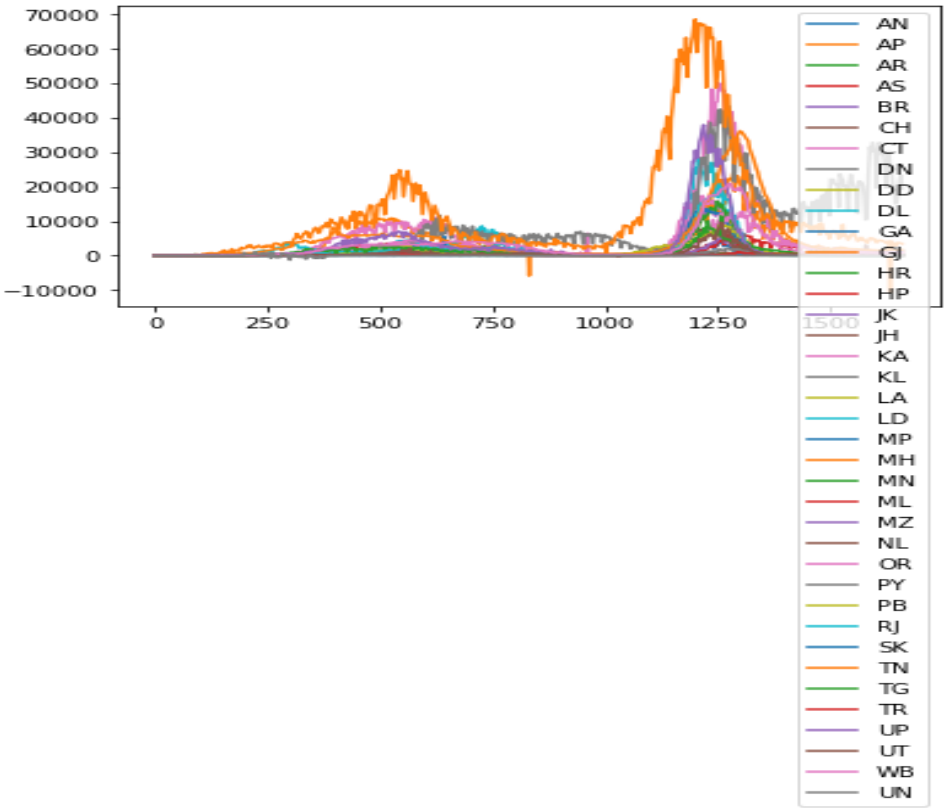


It was observed from here, that the maximum number of covid cases were reported in Maharashtra with the maximum correlation from most of the states. [highlighted with green border].

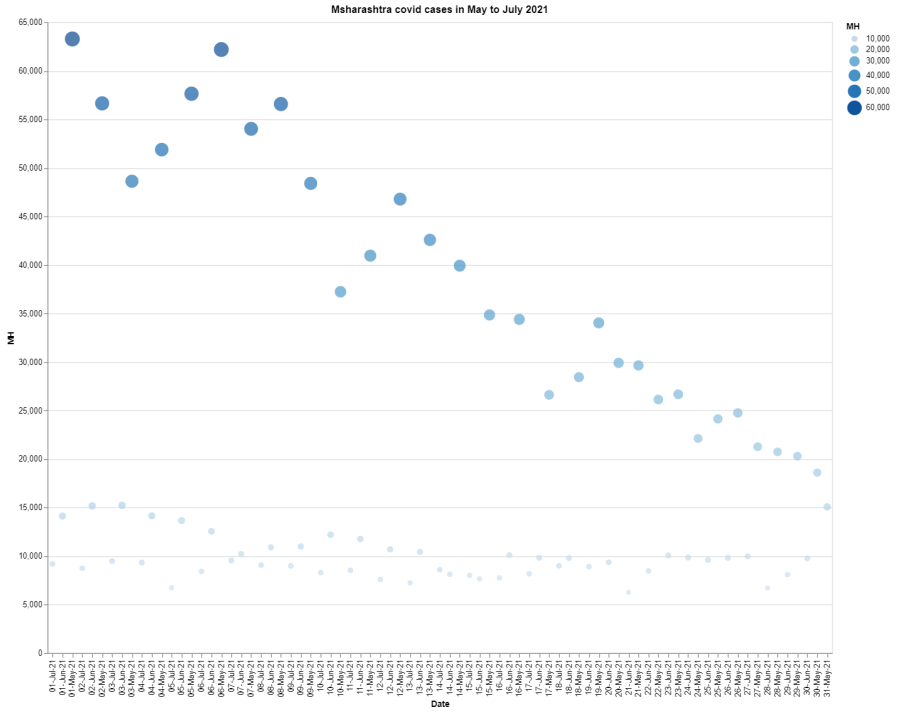


From this plot, it is quite visible that most cases lie during the phase of 2021-03 to 2021-06. And Maharashtra being the brightest indicates that it has the largest cases of that time. With Kerala being on the 2nd number.

Therefore, for choosing the hotspot we have 2 options Kerala and Maharashtra, of which we are choosing Maharashtra.



Below is the visualization of the covid cases of Maharashtra.



So, from here we have visualized that Maharashtra is the hotspot to be considered for spatial-temporal distribution of COVID Cases in India. Now, the further task was to visualize the hotspot district within Maharashtra. For that, we have further taken the time series data of all the districts of Maharashtra for further consideration of the District Hotspot of COVID Cases.

Date	State	District	Confirmed	Recovered
2020-04-26	Maharashtra	Ahmednagar	36	22
2020-04-26	Maharashtra	Akola	29	7
2020-04-26	Maharashtra	Amravati	20	4
2020-04-26	Maharashtra	Aurangabad	50	22
2020-04-26	Maharashtra	Beed	1	1

6.2 Actual dataset (Districts of hotspot state):

The dataset for districts was taken from https://data.covid19india.org/ which consists of districts of all states. We have dropped the null values, and the data of the other states, hence after preprocessing the above-mentioned dataset, it is :

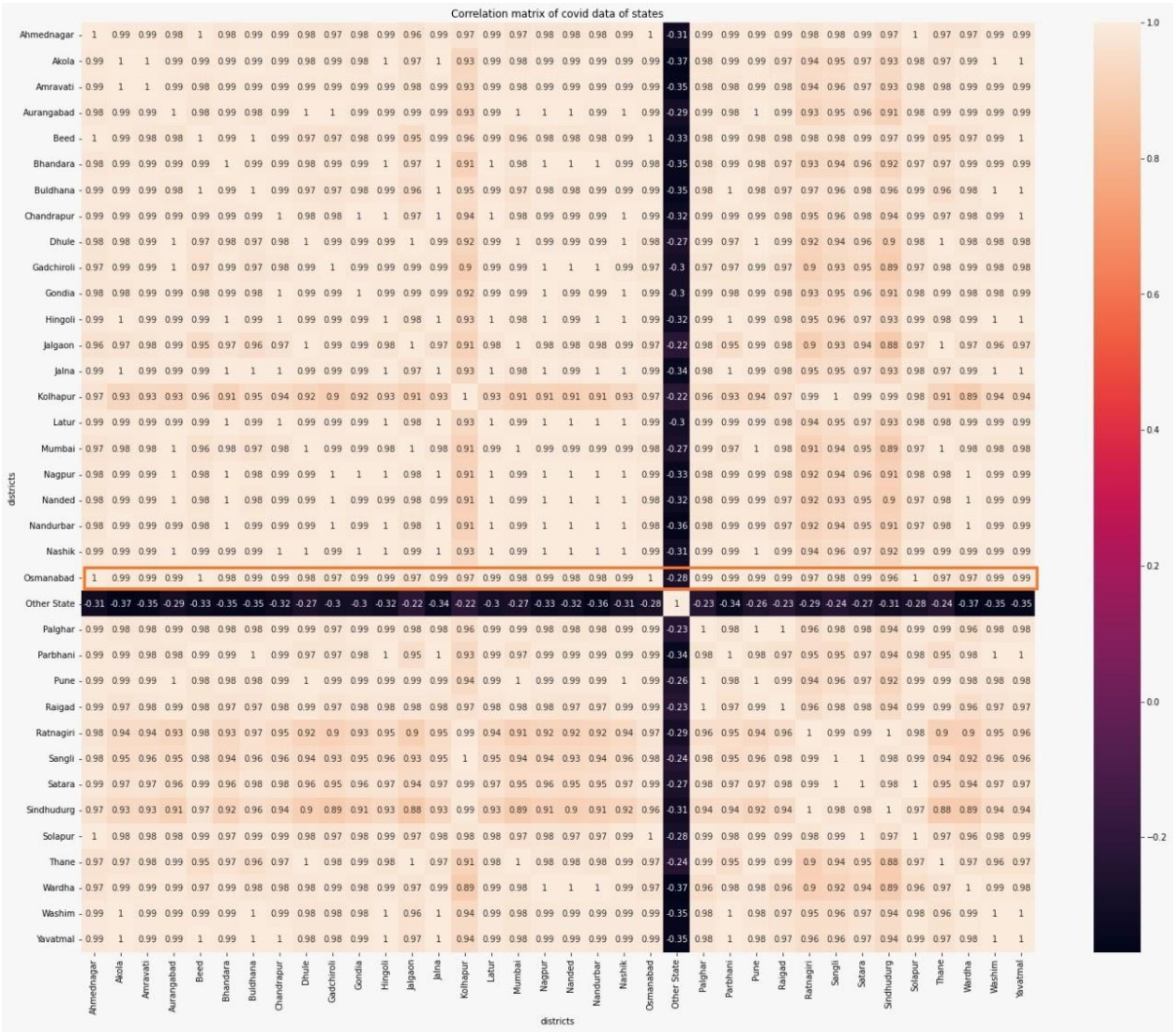
```
Int64Index: 18719 entries, 165 to 336531
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   Date       18719 non-null  object
 1   District   18719 non-null  object
 2   Confirmed  18719 non-null  int64 
 3   Recovered  18719 non-null  int64 
dtypes: int64(2), object(2)
memory usage: 731.2+ KB
```

The following is the view of the cleaned, and preprocessed dataset:

Date	District	Confirmed	Recovered
2020-04-26	Ahmednagar	36	22
2020-04-26	Akola	29	7
2020-04-26	Amravati	20	4
2020-04-26	Aurangabad	50	22
2020-04-26	Beed	1	1
...	...	...	...
2021-09-28	Solapur	207746	200635
2021-09-28	Thane	604241	586837
2021-09-28	Wardha	57321	55933
2021-09-28	Washim	41633	40985
2021-09-28	Yavatmal	75934	74126

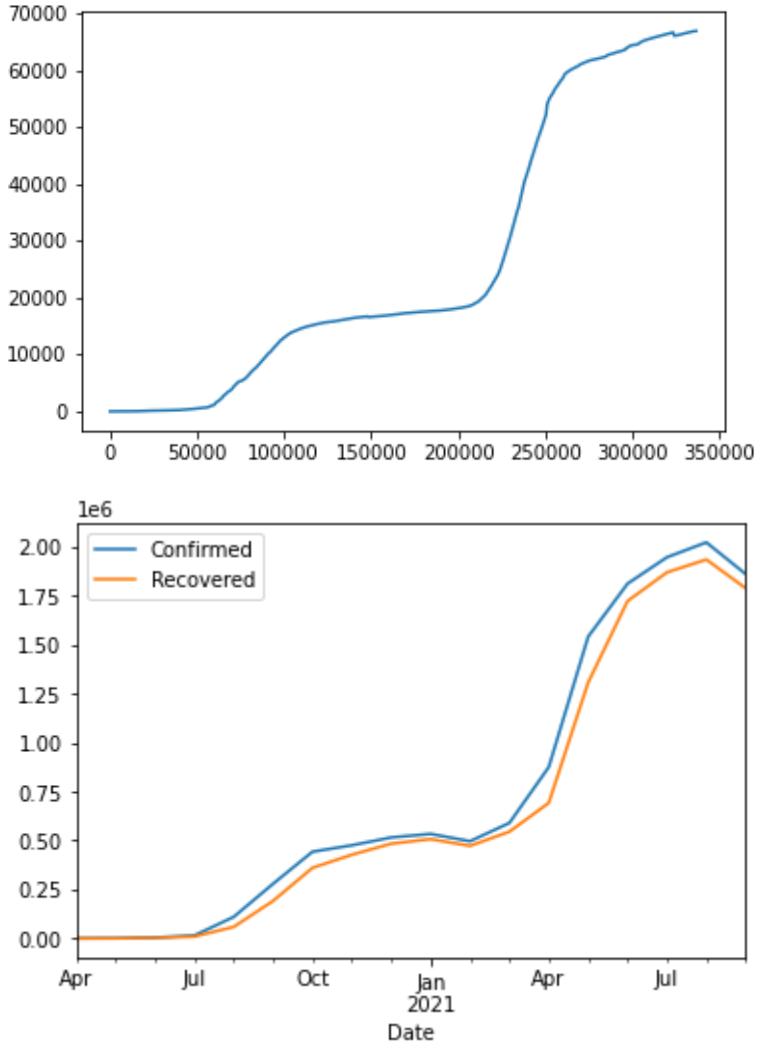
6.2.1 Spatial Correlation

The spatial correlation matrix with values, amongst the districts, is



From the above correlation, we see that the maximum correlation is shown in the district **Osmanabad**. Hence we will consider **Osmananbad** as the district hotspot of Maharashtra. It shows maximum correlation with every district of the state, thus upon predicting its cases we can get a closer idea of all the other districts.

The Covid cases in this district can be seen increasing exponentially,

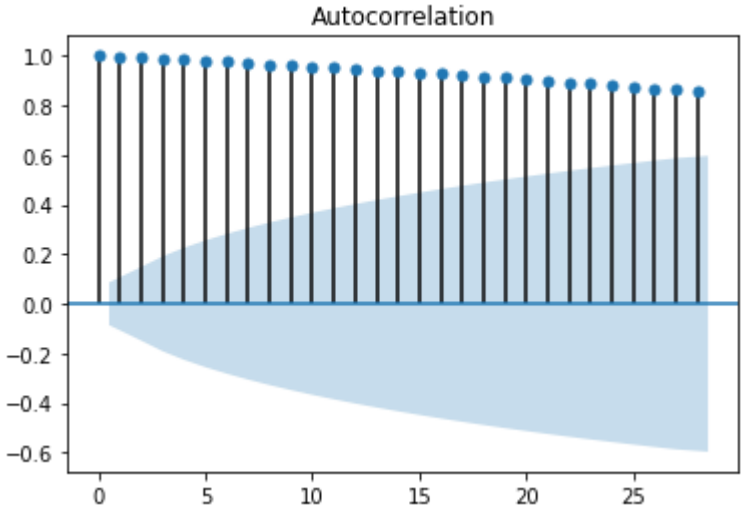


The monthly summation of the COVID confirmed cases and the recovered cases, of Osmanabad, is,

Date	Confirmed	Recovered
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
2021-09	1861822	1789603

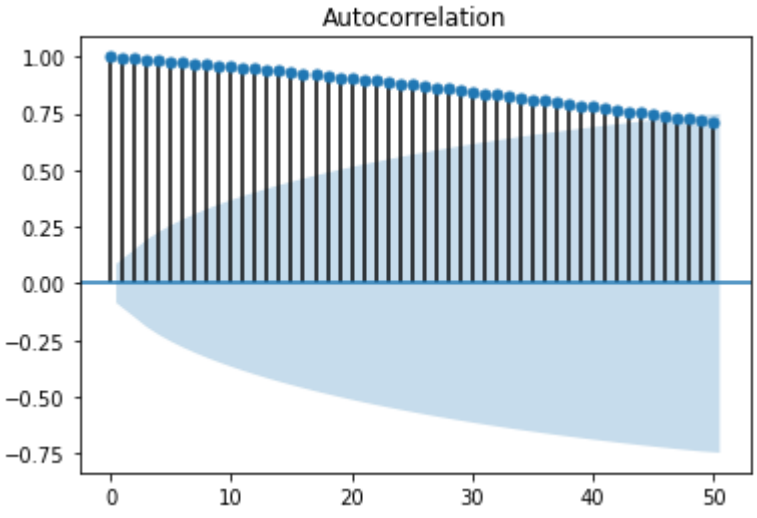
6.2.2 TEMPORAL CORRELATION of the hotspot district is shown as :

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

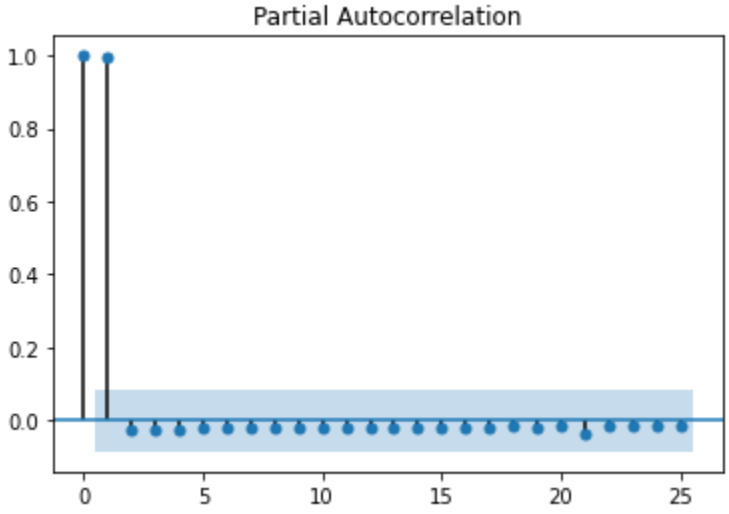
We can conclude that the temporal data is extremely correlated to each other on the basis of the previous case, as they lie in the range 0.8<x<1.0. Highly correlated.



for lag =50

Here 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.

We have also calculated partial autocorrelation which estimates the errors, or the residuals that have not been able to fit into the line of linear regression. It falls off to the noise very quickly because every time we get on to reducing the kind of error produced. No specific conclusion could be made through the plot obtained though.



7. Data Modification, for fitting into the Model

7. 1 Permutation Entropy

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.
The main features of Permutation Entropy include:

- Is non-parametric and is free of restrictive parametric model assumptions.
- Is robust with respect to noise, computationally efficient, flexible, and invariant with respect to non-linear monotonic transformations of the data.
- Relies on the notions of entropy and symbolic dynamics.
- Accounts for the temporal ordering structure (time causality) of a given time series of real values.
- Allows the user to unlock the complex dynamic content of nonlinear time series.

So, in order to find the permutation entropy, the first step was to partition the given dataset of the COVID Cases of Osmanabad, to current date data, spatial data as well as temporal data. So, accordingly, there were three columns formed for the whole dataset of "Osmanabad".

Dataset modified to calculate entropy:

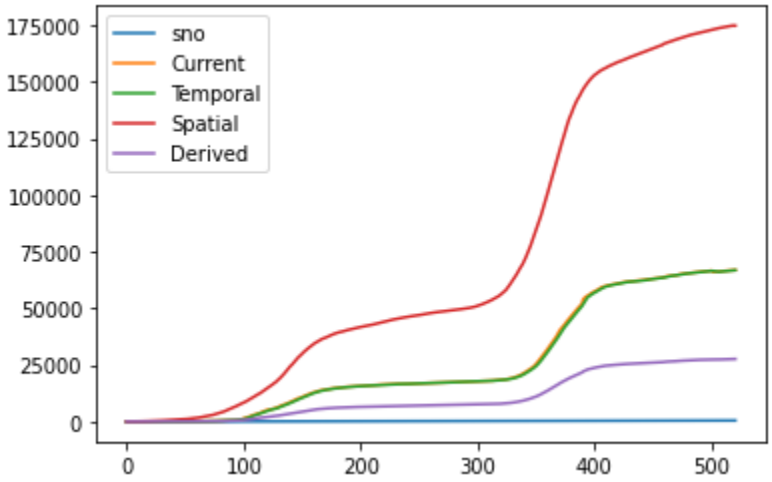
sno	Dates	Current	Temporal	Spatial
1	26-04-2020	3	3	74.36
2	27-04-2020	3	3	76.69231
3	28-04-2020	3	3	86
4	29-04-2020	3	3	88.42308
5	30-04-2020	3	3	92.34615
6	01-05-2020	3	3	99.15385
7	02-05-2020	3	3	105.5769
8	03-05-2020	3	3	111.4231
9	04-05-2020	3	3	148.0769
10	05-05-2020	3	3	154.9615
11	06-05-2020	3	3	164
12	07-05-2020	3	3	173.5
13	08-05-2020	3	3	179.5385
14	09-05-2020	3	3	187.0769
15	10-05-2020	3	3	217.9231
16	11-05-2020	3	3	229
17	12-05-2020	4	3.2	242.5385
18	13-05-2020	4	3.4	258.3077
19	14-05-2020	4	3.6	271.6154
20	15-05-2020	6	4.2	287.1538
21	16-05-2020	7	5	301.4231
22	17-05-2020	7	5.6	318.0385

7.2 ADF testing

Adf test is a tool used to check the stationarity of a dataset. It has been observed that stationary data provide better results in training models. Therefore we are observing here, the present trends in the data and if possible try to remove it, in order to make the data stationary.

```
ADF Statistic: -0.27925597101134114
n_lags: 17
p-value: 0.928391356920488
observations used: 503
Critical Values:
   1%: -3.4434175660489905
Critical Values:
   5%: -2.8673031724657454
Critical Values:
   10%: -2.5698395516760275
```

The values obtained from the ADF test shows that 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. Thus, providing strong evidence that the data is non-stationary.



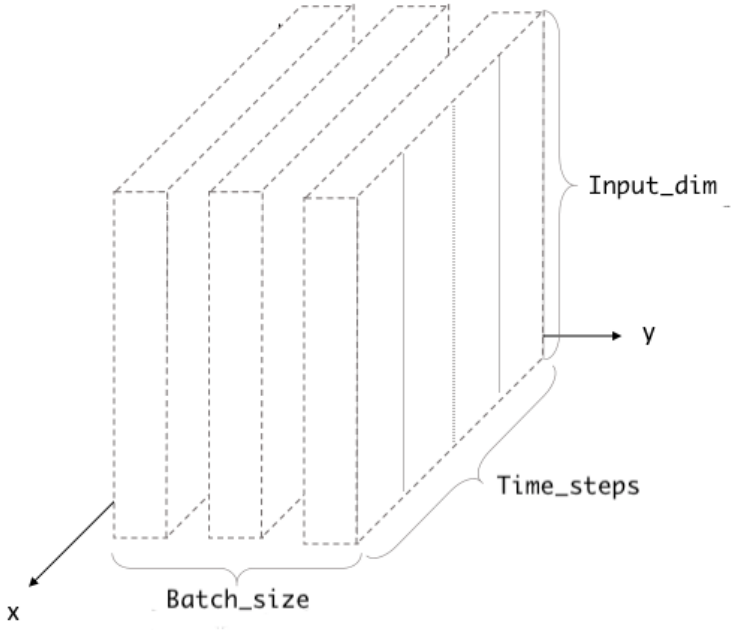
From this image, it is clear that the data has an increasing trend present, in spatial, temporal, and derived data. In order to make it stationary, differencing can be used but this makes some data as null, which makes the data non-useful.

Therefore we have chosen the LSTM model which does not explicitly require stationary data for the prediction purpose, because being an extension of recurrent neural networks, LSTMs can perform prediction on regular data. RNNs can learn complex patterns in data, unlike the basic regression models.

8. About the Model

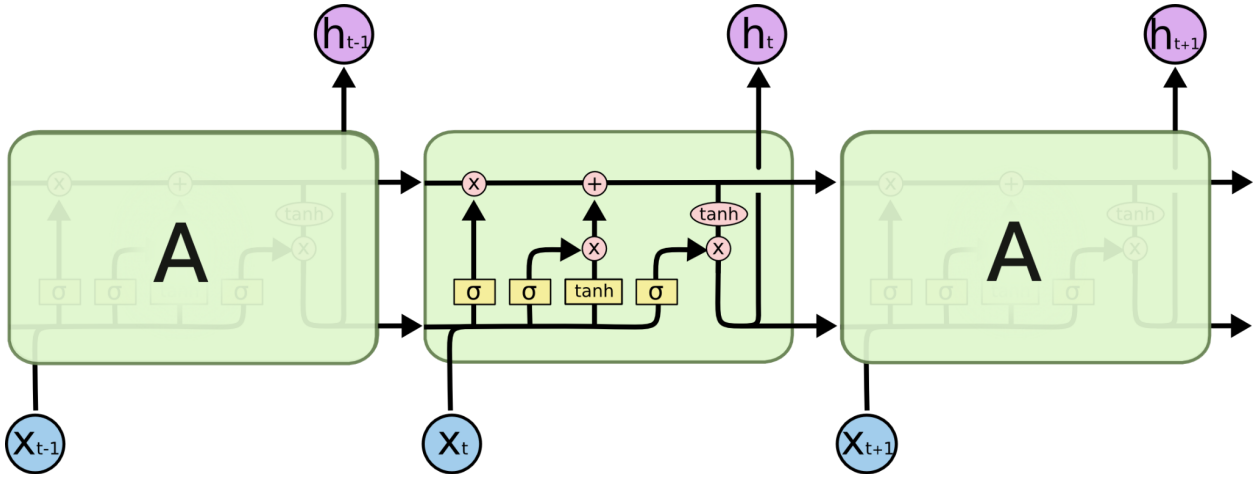
8.1 LSTM

The modified data which was shown above is feature scaled using Minmax scaling and fit transform. 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. In a nutshell, the input and output data in an LSTM model is a three-dimensional 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 (or input dimension)** 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]. Below is the illustration of LSTM input and output data shape.

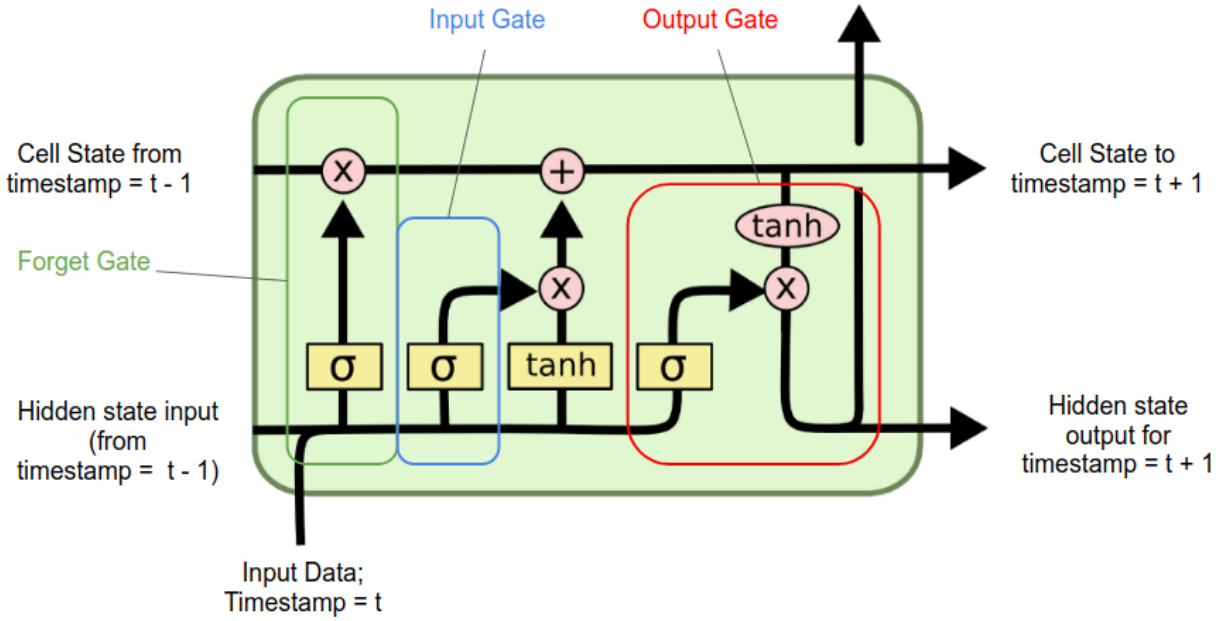


For the training and testing data, we have used an 80-20 split, with a batch of n value for lags as 64. Around 2 months for the prediction of the next day's cases.

Our hyperparameters here include, neurons, batch-size, and drop-out values.



There are three different gates in an LSTM cell: **a forget gate, an input gate, and an output gate**.



This is how LSTM works.

8.2 HyperParameter Tuning

A model hyperparameter is a configuration that is external to the model and whose value cannot be estimated from the data and a model parameter is a configuration variable that is internal to the model and whose value can be estimated from the given data.

In the other words, a hyperparameter is used to construct the structure of the model and cannot be learned from the data and its value is set before the learning process begins. Therefore, hyperparameters are like the settings of an algorithm that can be adjusted to optimize performance and prevent overfitting. This is exactly what we do in the hyperparameter tuning. We try to choose a set of optimal hyperparameters for a learning algorithm to enhance the performance of the model. There are two frequently used methods to perform hyperparameter tuning called 1)Grid Search and 2)Random Search.

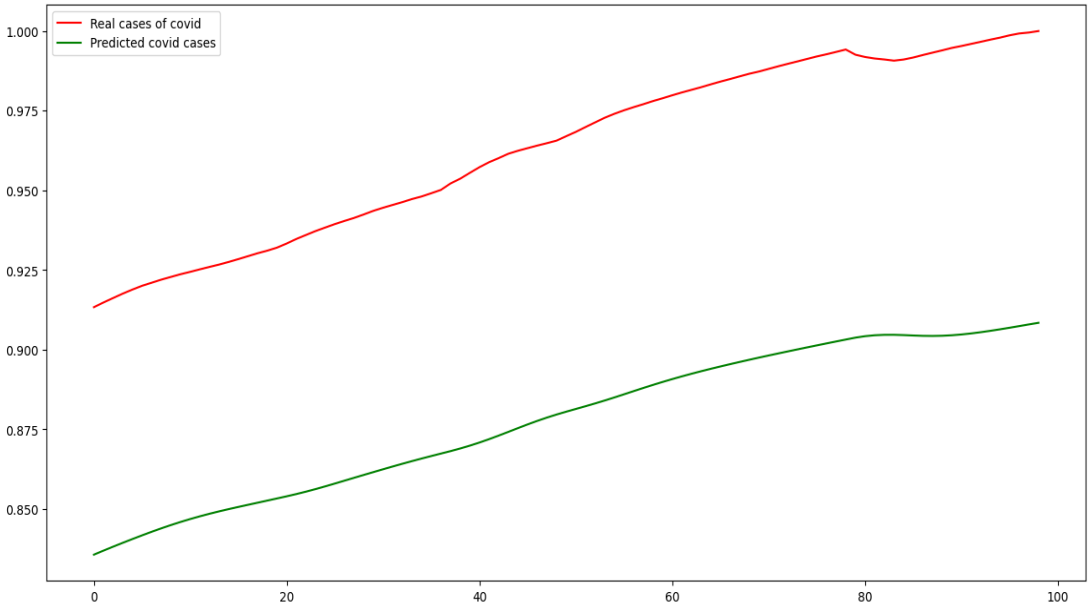
We have here done the tuning part using grid search because of simplicity and its powerful performance.

9. Milestones

OBJECTIVES
1. Downloading the required Dataset.
2. Perform Data Pre-processing (removing unwanted rows and columns)
3. Visualization of the covid cases of India
4. Finding the state with the maximum number of covid cases
5. Finding the district hotspot of the state which was detected as a hotspot of COVID cases.
6. Finding the entropy of the current, temporal, and spatial data of covid cases of the district.
7. Applying and creating actual data to feed the model
8. Testing its accuracy
9. Calculating errors
10. Future Scope

10. Results obtained

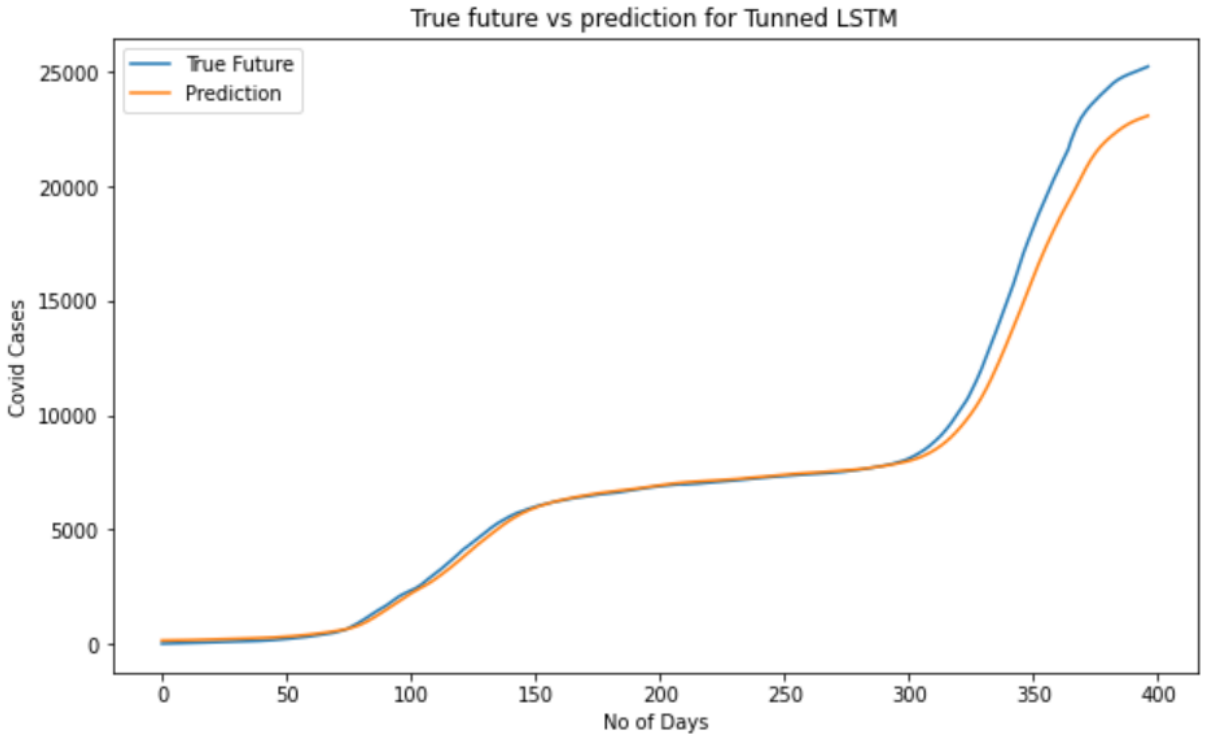
10.1 Result obtained



Here,
X-axis shows the testing data values
Y-axis shows the scaled values plot for testing data

We can see that the model is predicting closer values to the actual testing data set.
The best fit model is one with the configuration as follows:

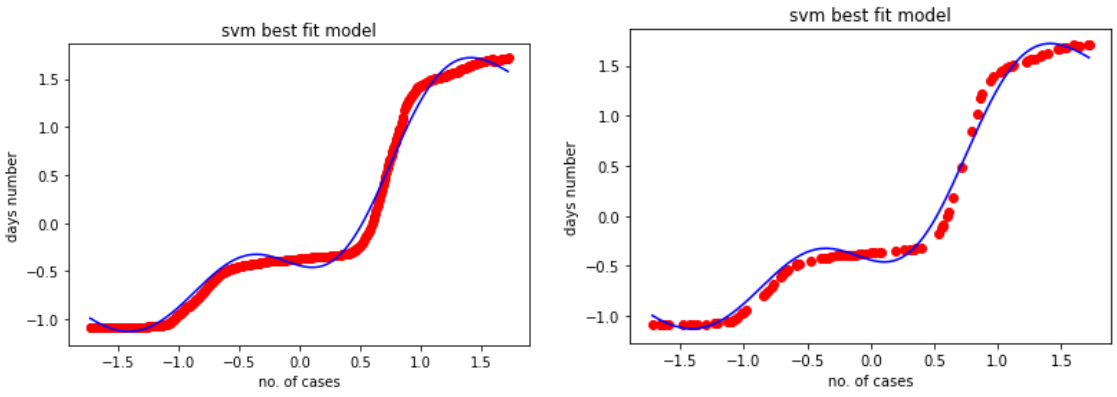
```
Best Combination:
 first_additional_layer = False
 second_additional_layer = False
 third_additional_layer = False
 n_neurons = 256
 n_batch_size = 64
 dropout = 0.1
```



Result for training data.

10. 2 Comparing results from SVM

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



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

11. 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 temporal data.

Bidirectional LSTM (BiLSTM):

In bidirectional LSTM we give the input from both the directions from right to left and from left to right. Bidirectional recurrent neural networks(RNN) are really just putting two independent RNNs together. This structure allows the networks to have both backward and forward information about the sequence at every time step.

Using bidirectional will run your inputs in two ways, one from past to future and one from future to past and what differs this approach from unidirectional is that in the LSTM that runs backward you preserve information from the future and using the two hidden states combined you are able in any point in time to preserve information from both past and future

Gated Recurrent Unit (GRU):

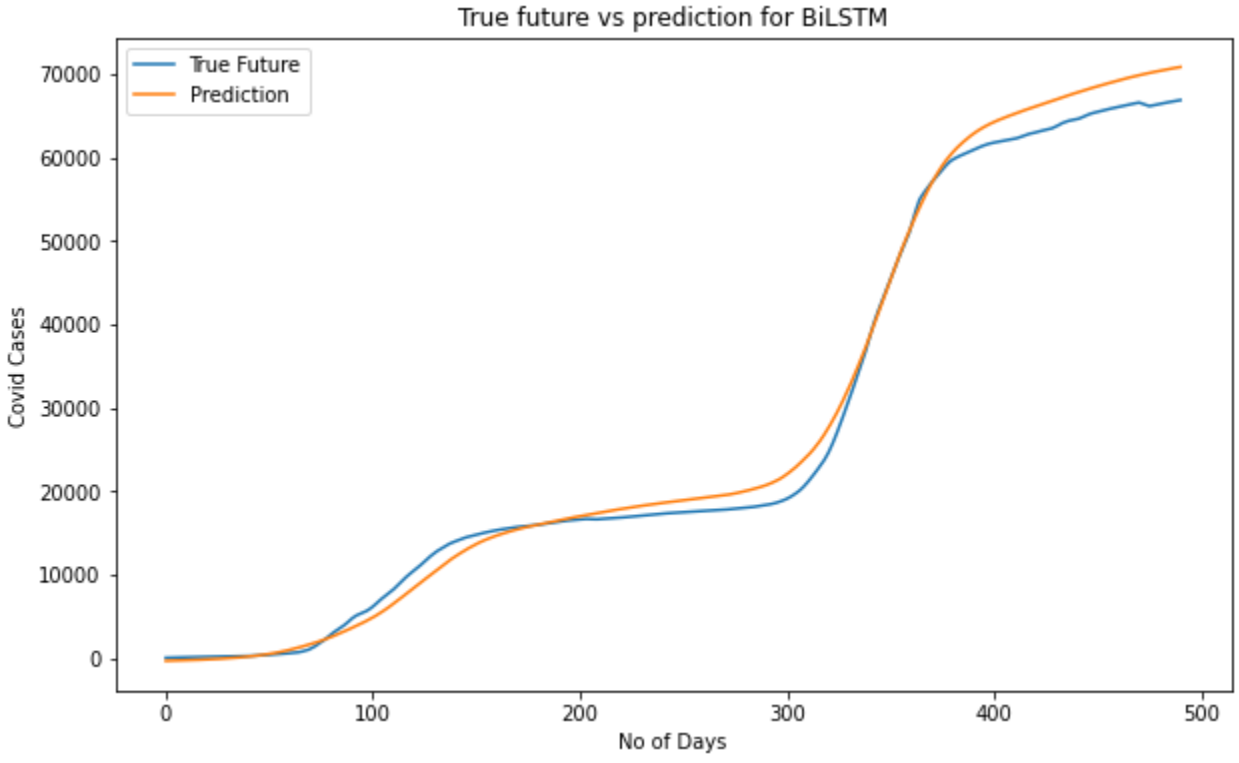
This was founded quite recently in 2014 where they reduced the number of parameters from LSTM, but just in case GRU doesn't work well, then we will have to roll back to LSTM.

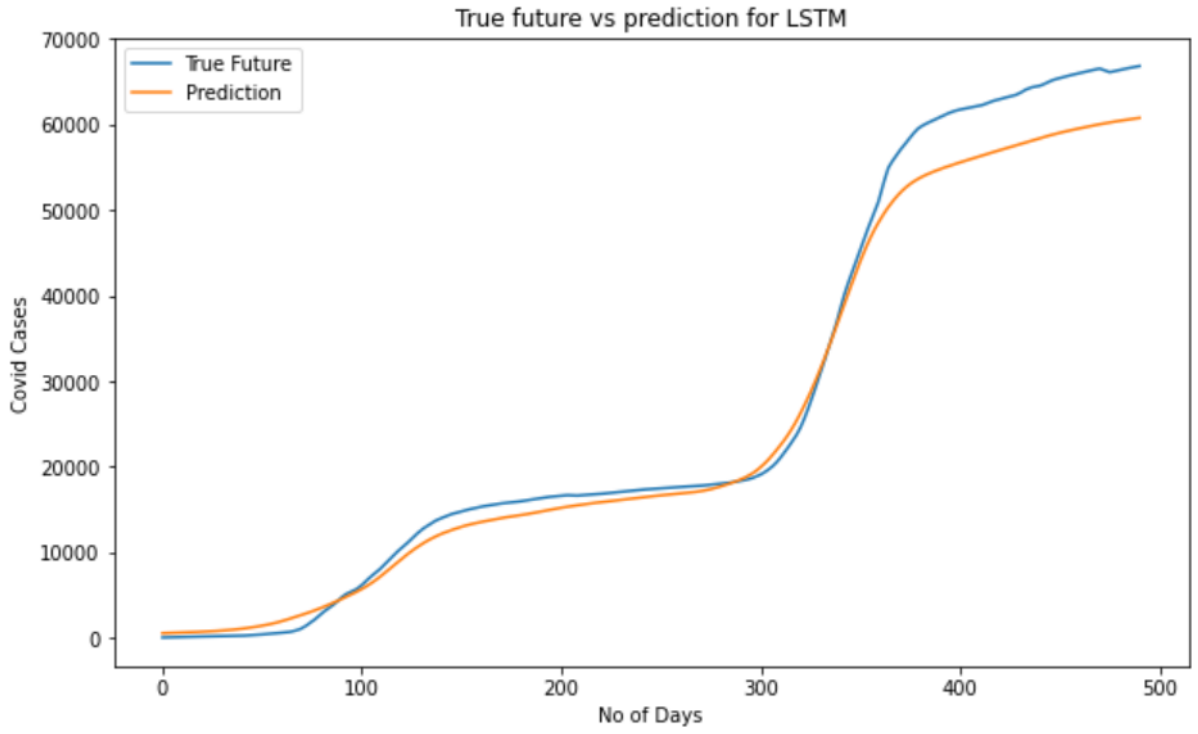
In GRU, there is no explicit memory unit. Memory unit is combined along with the network. The GRU is like a long short-term memory (LSTM) with a forget gate, but has fewer parameters than LSTM, as it lacks an output gate. They are both combined together and thus the number of parameters are reduced. The key difference between GRU and LSTM is that **GRU's bag has two gates that are reset and update** while LSTM has three gates that are input, output, and forget. GRU is less complex than LSTM because it has fewer gates. When comparing GRU with LSTM, it performs well but may have a

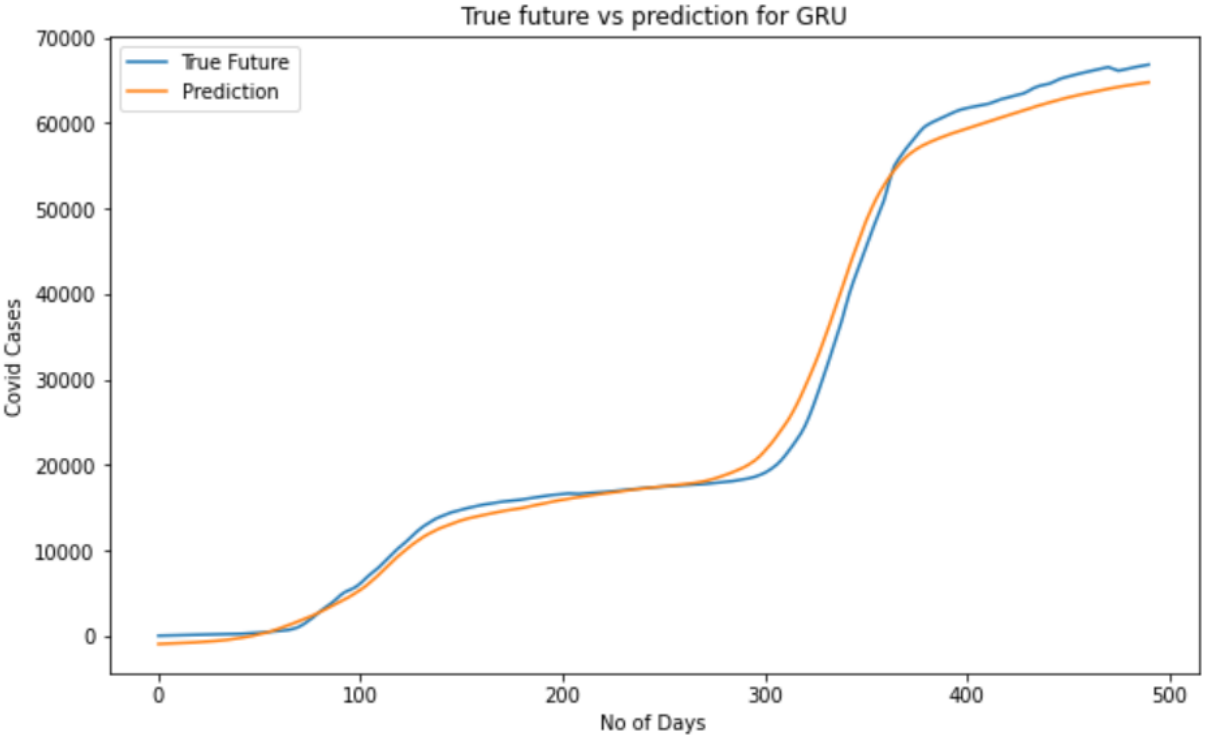
slight dip in the accuracy. But still, we have less number of trainable parameters which makes it advantageous to use.

Error comparison form 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 LSTM
True Future
Prediction
Covid Cases
No of Days
0
10000
20000
30000
40000
50000
60000
70000
0
100
200
300
400
500


True future vs prediction for GRU
True Future
Prediction
Covid Cases
No of Days
0
10000
20000
30000
40000
50000
60000
70000
0
100
200
300
400
500

12.Conclusions

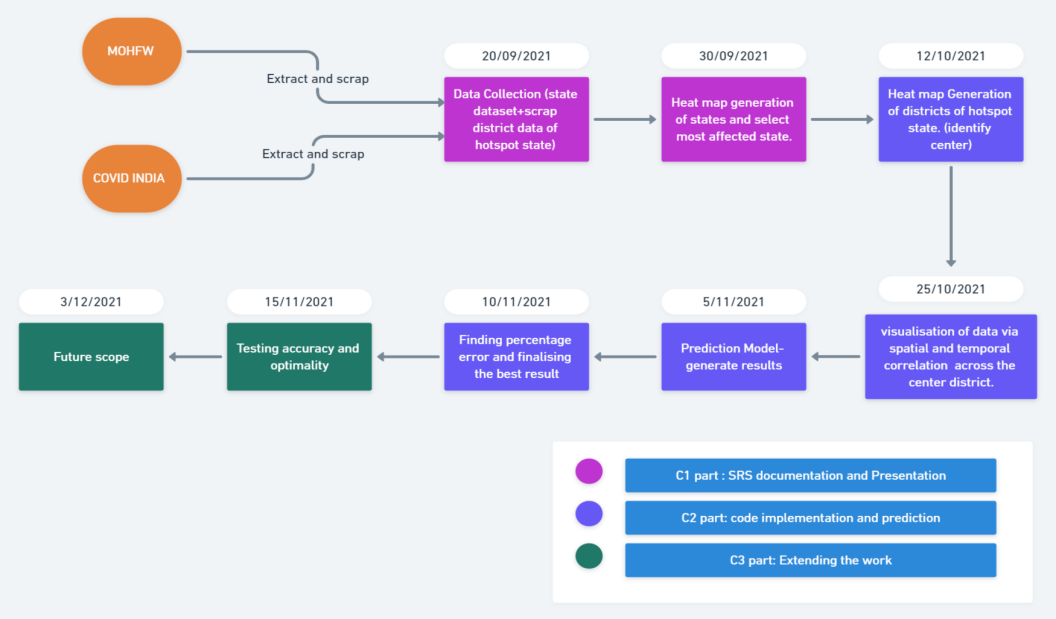
We have analyzed the spatial as well as the temporal spread of COVID 19 in India. We have taken the data from March 2020 to September 2021, and have observed the number of cases rise exponentially in different states of India especially during April-May 2020. With the infection growth and the doubling rate of COVID 19.

- We found that Maharashtra is the most affected region in India by covid. Since, on plotting the Correlation Matrix amongst the different states of India, Maharashtra was found to be the hotspot of COVID cases.
- In Maharashtra, district Osmanabad is most spatially temporally related to other neighbouring districts. Although pune contributes the most when it comes to the spreading of covid, due to its close proximity to Mumbai, highest daily covid cases, presence of international airports, it is also an industrial city. But for prediction, we choose Osmanabad here due to its spatial-temporal relation being highest resulting in the most accurate prediction.
- LSTM with spatial-temporal data is the most accurate model among all implemented models.
- If we apply our model to any district which is highly spatially temporally related to neighbouring districts, then our model is applicable to it and gives very accurate results.

Error comparisons of all models implemented-

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

13. Activity Diagram:



14. 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.

[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