Forecasting the Impact of COVID-19 on the States of India using ARIMA Algorithm

Varun Totakura
Department of Computer Science and Engineering,
Guru Nanak Institutions Technical Campus,
Hyderabad, Telangana, India.
E-mail: totakura.varun@gmail.com

E. Madhusudhana Reddy
Department of Computer Science and Engineering,
Guru Nanak Institutions Technical Campus,
Hyderabad, Telangana, India.
E-mail: e_mreddy@yahoo.com

Madhu Sake
Department of Computer Science and Engineering,
Guru Nanak Institutions Technical Campus,
Hyderabad, Telangana, India.
E-mail: madhu.sake@gmail.com

Abstract
As the effect of COVID-19 is increasing rapidly every day, it is becoming very difficult for the survival of many people and there is a high effect on the economic situations of every country which is effected with it. In the same way, it is effecting all states of India and causing economic crisis. This paper deals with the analysis of the impact caused by COVID-19 on each state in India and also gives an estimated date on which the effect will reduce and may even become zero along with the analysis report of overall India. For the forecasting of the effect we have used Auto Regressive Integrated Moving Averages (ARIMA) algorithm which has produced Root Mean Squared Error (RMSE) of around 5.89 for some of the states and other with 20.05 due to the data abnormality. The forecasted data for each state is project in the figure using the line plots. And the resulted graphs are explained clearly. The accuracy of the proposed model is around 94.6% to 96.8% for the states with good data and less RMSE and 80% for the sates with abnormal data and high RMSE value. From the produced results of the proposed methodology the dates of which the effect of COVID-19 will decrease is calculated for the states which has high number of cases.

Keywords- Coronavirus, COVID-19, Data Analysis, Virus Disease Outbreak, ARIMA, Forecast, RMSE, Visualization.

1. Introduction

The pandemic disease COVID-19 is out spreading rapidly throughout the world. As per the statistics in April, 2020 the number of cases reached around 3.27 million and no. of deaths has reached around 2.3 lakhs. In India, the no. of cases is above 35 thousand and no. of deceased persons are above 1150 and it is still growing. This disease has started in 2019 at Wuhan, China and had spread in almost all countries on the globe. The data about the cases in India is shown in a pie chart in figure – 1. The persons who is infected with virus will fall sick and experience breathing problem. It spreads from infected person to other people when the infected person sneezes, coughs or exhales. These droplets are too heavy to hang in the air, and quickly fall on floors or surfaces. The virus can live from 2 to 3 days on plastics, steel and on paper, wood it can live up to 5 days. As per the available data it was said that a person who was infected with mild COVID-19 on average can recover in 2 weeks and with severe infection they can recover in 3-6 weeks approximately. Along with the increasing of the cases the recovery rate is also increasing. Throughout the world there are more than 1 million who have recovered from this disease and in India the number crossed 9 thousand. The governments of respective countries and states are following some of the preventive measures that is helping to decrease the rate of spread. In India, the government has imposed lockdown period for some days to prevent the spread of virus. During the lockdown period, no person is allowed to commute from one place to another without proper permission and many organizations like IT industries, Schools, Colleges, Universities and many more are closed down. Only few organizations which are related to the essentials of the humans are permitted. Due to the improvement of lockdown there is a significant change in the spreading rate which is falling down. The following bar chart in Figure – 2 will show the number of confirmed cases in each state in India.

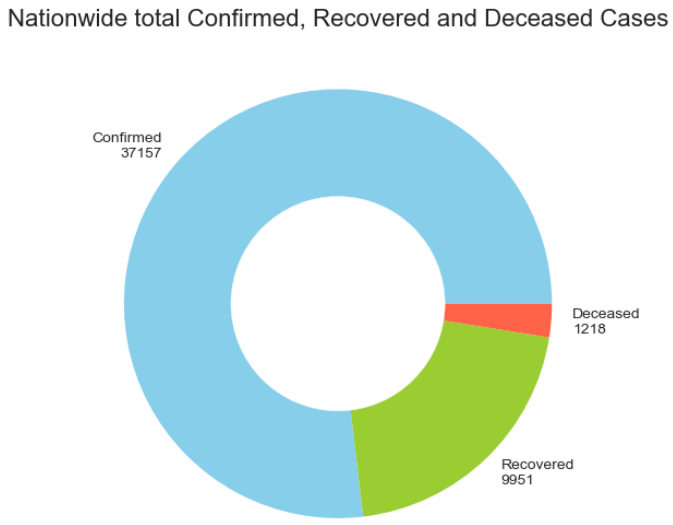


Figure 1. COVID-19 cases, recovered and deaths in India, as of 1st May 2020.

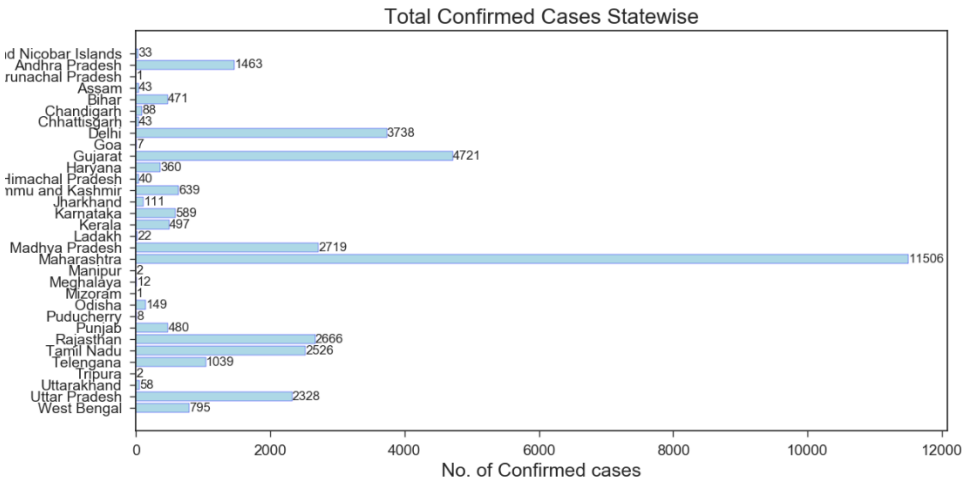


Figure 2. COVID-19 confirmed cases in the States of India, as of 1st May 2020.

At the places where the people are strictly following the rules and preventive measures imposed by the respective government the rate of spread of COVID-19 is falling down or very less, but at the places where the rules are broken the rate is very high. But there us a chance of having the economic crisis and the poor people will not be able to buy their essential goods. In keeping all these conditions as the priority there is a need for the analysis to predict the impact of virus on the people which helps to spread the awareness across the people in many ways so that the people start judging the situation in a right way and the government can increase or decrease the rules or preventive measures that are taken against the spread. If the people start to take care by maintain the social distancing, then the government can permit some more kind of organizations to start their work again which help to avoid the economic crisis. In keeping all the above reasons in mind we have proposed a statistical model to analyse the situation in each state in India.

The proposed statistical methodology will take the time-series data which contains the statistical information of no. of cases, recovered, deaths in each state in India as per the records on 25-April-2020. The time series data will provide an opportunity to forecast the future vales. Based on the previous data in time series and by using the forecasting statistical algorithms the values of the same parameter can be predicted for the future. But the model which was used should be accurate so that there are no wrong decisions are made. Forecasting is a most tremendous task which is generally performed by the higher officials of the organization for the growth of their business and even by the stock market investors to get the good amount of profits. The future trend of the stock or the sales will help the investors or the owners to make the right decisions. The same can be used for the prediction of the impact of Corona in India. We have used ARIMA model in our paper for the forecasting of impact of COVID-19 as it is one of the most used method specifically for the time series data. Forecasting a time series can be broadly divided into two types. If the previous values of the time series data are used to predict its future values; then it is called Univariate Time Series Forecasting. And if predictors other than the series to forecast it is called Multi Variate Time Series Forecasting. ARIMA is a forecasting algorithm based on the idea that the information in the past values of the time series can alone be used to predict the future values. And to confirm that the predicted data is accurate various error calculating methods are used.

We have used RMSE method for the calculation of the error rate of our proposed model. Generally, the RMSE is obtained from Mean Squared Error (MSE). The MSE is a commonly used error calculation method for the statistical data. The MSE of the unobserved quantity

measures the average of the squares of the errors which means that the average squared difference between the esteemed values and what is estimated. MSE is a risk function which is corresponding to the value of expectation of the squared error loss. There is a fact that MSE is almost always positive which means not zero because of randomness or because the estimator does not account for information that could produce a more accurate estimated value. RMSE is just the square root of the MSE. It is probably the most easily interpreted statistic. For the observations that are made by the data available the RMSE value has ranged from 5 to 22 approximately because of uncertain values. The RMSE value is low for the states for which the data has some kind of trend from the starting but it is high for states in which the count of the cases is abnormal. Further details about the proposed model and the respective calculations are described below.

2. Related Work

The ARIMA model was used in the prediction of COVID-19 cases (Perone March, 2020) in which it was mentioned that the ARIMA models can be used as the immediate tool for the prediction of the timer series data for the health monitoring system. It is a good model for short-term forecasting but there should be good procedure in the process of interpretation. The ARIMA interpreted that there will be around 200,000 cases in Italy.

The usage the ARIMA model is wide ranged. It can be used on the time-series data for the prediction of daily or monthly or even year average. The similar kind of work was performed by (Almasarweh et Al. 2018) on prediction of the banking stock market data. They have mentioned that the ARIMA model was very useful for the short term analysis with few time series of data. They have used MSE method in calculation of the error value to calculate the accuracy of the model. A graph was plotted for the dataset which tells about the comparison of the banking and index. The same kind of model was developed for the Nigerian Stock Exchange by (Ayodele et. Al 2014), (Alsharif et Al. 2019) The prediction of global solar radiation and by (Awajan et Al. 2014) in the prediction of the stock market data using EMD - HW bagging.

The generalized review about the ARIMA model for the prediction or forecasting of the time series data was represented in a paper published by (Chenghao et AL. 2016). They have proposed a novel online method using the ARIMA model. They have theoretically proved that their method has produced the most better results from previous fixed ARIMA methods. They have even compared their algorithm with the previous algorithms on ARMA and have mentioned that their model performed well.

A review work on forecasting the electricity price was performed by (Rafel Weron, 2014). The paper will give the information about the electric price forecasting study and also interprets on the directions of the price for a decade.

A generalized forecasting method was given by (Taylor et Al. 2017). The first system that they have made to forecast was performed on the data of Facebook. They have used a modular regression model for their first method. And the secondly they have performed a study on the tracking the forecasting accuracy of the model. The similar kind of methodology was also proposed by (J. Scott Armstrong, 2001) and also by (Manoj et Al. 2014) in prediction of the sugarcane production in India.

(Biljana Petrevska, 2017) The prediction of the tourism was performed using the ARIMA model. The paper has interpreted that the tourism in F. Y. R. Macedonia will play a major role in contribution for the country's economy. They have used ARIMA (1, 1, 1) model for the prediction and analysis of the international tourism at F. Y. R. Macedonia. The accuracy of the

model seems to be good but not perfect as there should be increase in the accuracy for the accurate prediction which will help in making the correct decisions for the tourism and economic growth.

3. Methodology

Auto Regressive Integrated Moving Average is a highly used statistical model specifically used for the time series analysis. It is generalized model of Auto Regressive Moving Average model. Basically, these two models are used for the understanding and forecasting of the time series data. ARIMA is mostly is also applied on non-stationary data. Because, it has an integral part which will help in removing the non-stationarity od the data. A stationary time series data is the data which the vales of the data will not depend on the time. The time series which exhibits trends or seasonality are considered to be non-stationary. ARIMA model is the combination of both Auto Regressive (AR) and Moving Average (MA) model. The equation of the ARIMA model can be written as:

$$y't = c + \phi 1 y't - 1 + \cdots + \phi p y't - p + \theta 1 \varepsilon t - 1 + \cdots + \theta q \varepsilon t - q + \varepsilon t$$

where y ′ t is the differenced series, and the "predictors" on the right hand side include both lagged values of y t and lagged errors. We call this an ARIMA (p, d, q) model, where p = order of the autoregressive part, d = degree of first differencing involved, q = order of the moving average part.

The term 'Auto Regressive' in ARIMA gives the interpretation that it is a linear regression model that uses its own lags as predictors. Generally, the Linear regression models will work best when there is no correlation with the predictors and also when they are independent to each other. To make the time series data stationary, the most common approach which is used is the differentiating method. That is, subtract the previous value from the current value. Sometimes, depending on the complexity of the time series data, more than one differencing is performed. Then the value of d will become the minimum number which will help to make the time series data stationary. And if the time series is already stationary, then d = 0. 'p' is the order of the 'Auto Regressive' (AR) term. It refers to the number of lags of Y to be used as predictors. And 'q' is the order of the 'Moving Average' (MA) term. It refers to the number of lagged forecast errors that should go into the ARIMA Model.

$$(1-\phi 1B-\cdots-\phi pBp)\,(1-B)\,dyt = c+(1+\theta 1B+\cdots+\theta qBq)\varepsilon t$$

$$\uparrow \qquad \uparrow \qquad \uparrow$$

AR(p) d differences MA(q)

After the calculation we get the estimated p, q, d values which should be given as the parameters to the ARIMA model as shown the figure – 3 to obtain the Akaike's Information Criterion (AIC) values.

```
Examples of parameter combinations for Seasonal ARIMA.
SARIMAX: (0, 0, 1) x (0, 0, 1, 12)
SARIMAX: (0, 0, 1) x (0, 1, 0, 12)
SARIMAX: (0, 1, 0) x (0, 1, 1, 12)
SARIMAX: (0, 1, 0) x (1, 0, 0, 12)
```

Figure 3. Examples of the Parameters p, d, q values given to ARIMA model.

AIC will be helpful in the process of selection of estimators and also to determine the order of the

model which will help in getting the accurate results. The AIC can be calculated by using the below equation:

$$AIC=-2\log(L)+2(p+q+k+1)$$

The example of the AIC values which are obtained by the ARIMA model when fitted with the COVID-19 “Andhra Pradesh” State data is as shown in figure – 4. Among the obtained values the least AIC value is taken and the corresponding parameters are given to the parameter list of the ARIMA model.

```
ARIMA(1, 1, 0)x(0, 0, 0, 12)12 - AIC:382.7930577094712
ARIMA(1, 1, 0)x(0, 0, 1, 12)12 - AIC:1349.8427995020402
ARIMA(1, 1, 0)x(0, 1, 0, 12)12 - AIC:302.45364070581377
ARIMA(1, 1, 0)x(0, 1, 1, 12)12 - AIC:184.47528661517313
ARIMA(1, 1, 0)x(1, 0, 0, 12)12 - AIC:291.3554690920168
ARIMA(1, 1, 0)x(1, 0, 1, 12)12 - AIC:1340.0250958845795
ARIMA(1, 1, 0)x(1, 1, 0, 12)12 - AIC:185.83037137423074
ARIMA(1, 1, 0)x(1, 1, 1, 12)12 - AIC:183.8762486009413
ARIMA(1, 1, 1)x(0, 0, 0, 12)12 - AIC:370.5255667788842
ARIMA(1, 1, 1)x(0, 0, 1, 12)12 - AIC:1217.0176398667315
ARIMA(1, 1, 1)x(0, 1, 0, 12)12 - AIC:291.8271786523849
ARIMA(1, 1, 1)x(0, 1, 1, 12)12 - AIC:174.77358103778263
```

Figure 4. Examples of the AIC values obtained by ARIMA model.

After obtaining the AIC values the ARIMA model is fitted or trained with the time series data which is the state COVID-19 data. And the results of the model can be represented as shown in figure – 5 which contains the vales of the Auto Regression, Moving Averages and the corresponding coefficient values, standard errors, and many more. And the values can be represented in the graphical mode as shown in figure – 6 which consists of the combination of four graphs which represents the values of the Standard residual, Histogram plus estimated density, Normal Q – Q, and Correlogram graphs. These two figures contain the results that were obtained by training or fitting the ARIMA model will the data of the Jharkhand State of India.

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.2576	1.165	0.221	0.825	-2.025	2.540
ma.L1	-0.9321	1.169	-0.797	0.425	-3.224	1.359
ma.S.L12	-0.7888	1.660	-0.475	0.635	-4.043	2.466
sigma2	22.6742	29.265	0.775	0.438	-34.685	80.033

Figure 5. Table which shows the result of ARIMA model.

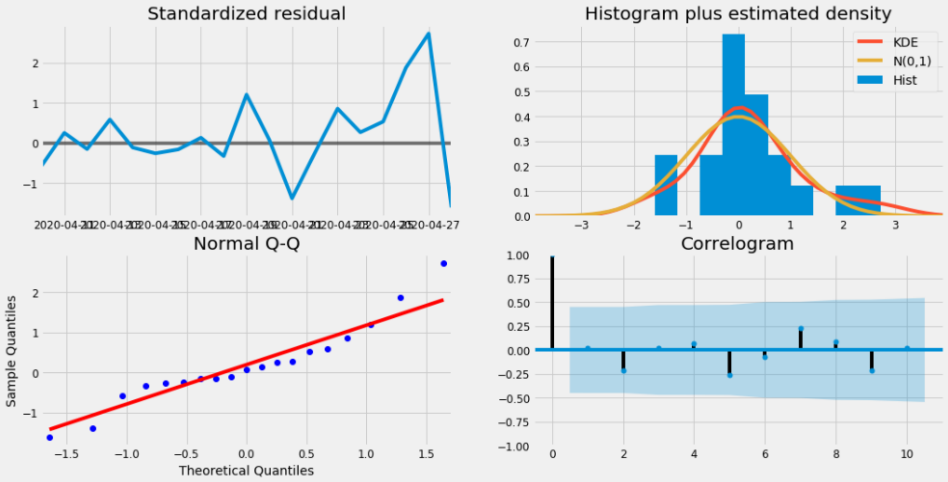


Figure 6. Graphs which shows the result of ARIMA model.

The predictions of the proposed Statistical model that is ARIMA model for each state can be seen in the results discussion section of this paper. Along with those details of the predicted date of which the impact of COVID-19 may decrease is also displayed using a table. And the error calculation and accuracy of the model is also displayed in the same results section.

4. Results and Discussion

The proposed ARIMA model has produced good results with the selected p, q and d values. But, as the accuracy of the model depends on the data that is given, for every state data that is given to the model its respective p, q, d values are used according to the least AIC value. The graph obtained using the predicted values of the number of cases and actual values by the proposed statistical model on the time series data of Jammu & Kashmir state was displayed in figure – 7. It proves that the model was finely fitted to that data as the predicted and the actual data seems to be equal. In the graph displayed below the blue line tells the actual or observed values whereas the red lines give the information of the predicted information.

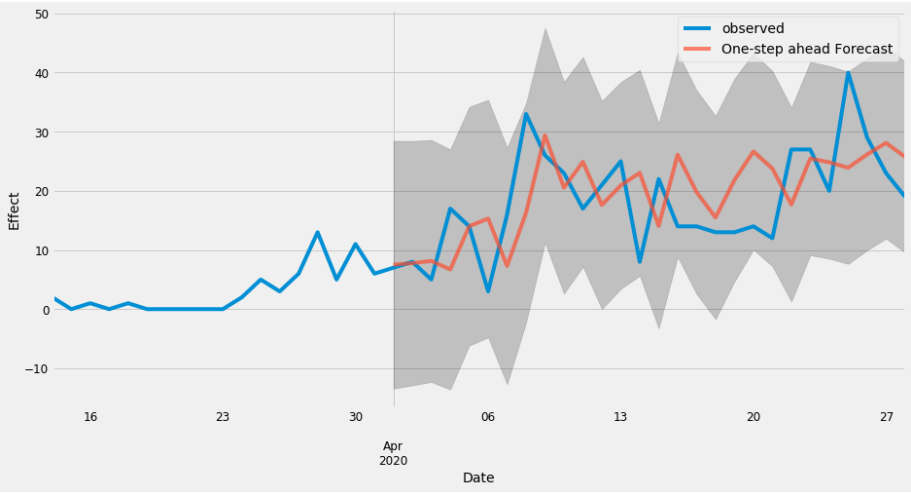


Figure 7. Prediction of number of Cases in Jammu & Kashmir State.

The prediction of number of recoveries by the proposed model in Kerala State using the data of the number of recoveries of the Kerala State is as shown in figure – 8. In the graph the blue line is

the actual vales of the number of recoveries with the day scale data and the red line is the predicted values which seems to be similar. Thus, we can say that the data of that state has fitted the model is very accurate way. But in every data there will be some noise and the actual values may become extreme and the predicted data will follow a trend line.

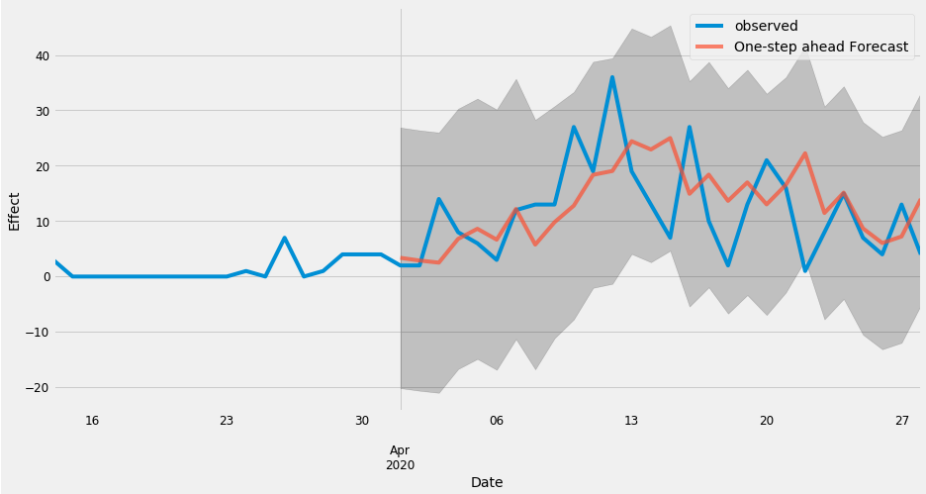


Figure 8. Prediction of number of recovered in Kerala State.

The predictions of the cases and the recovery in the future by the ARIMA model is shown in figure – 9. Here from the graph we can see that the actual recoveries values are displayed with yellow color line, actual cases are displayed with blue line, red line gives the information about the predicted number of cases in the future and the green line tells about the number of recoveries of Telangana State in the future. From the graph we can interpret that the Telangana seems to have less number of cases in future and the recoveries will be raising. This seems to be a better change where the COVID-19 spread may reduce gradually. In the table – 1, the information which was predicted by the model for other states are displayed. The states and the respective graph is arranged in a format.

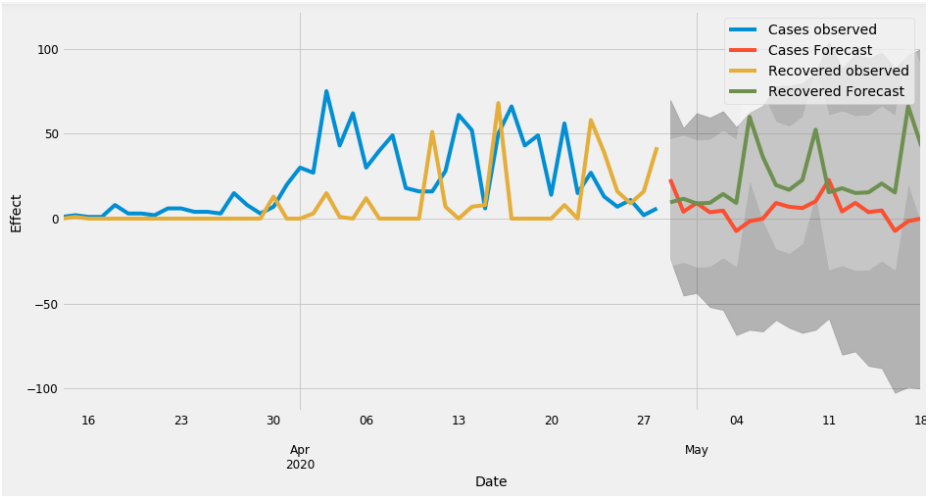


Figure 9. Prediction of number of Cases and Recovered in Telangana State.

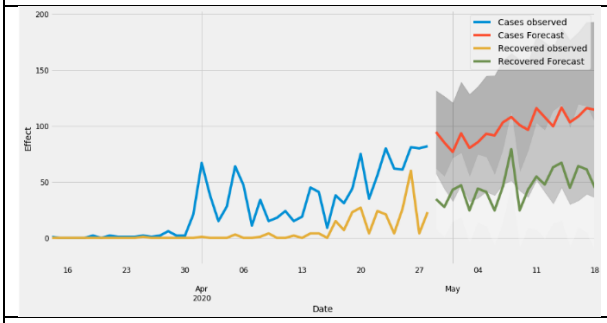
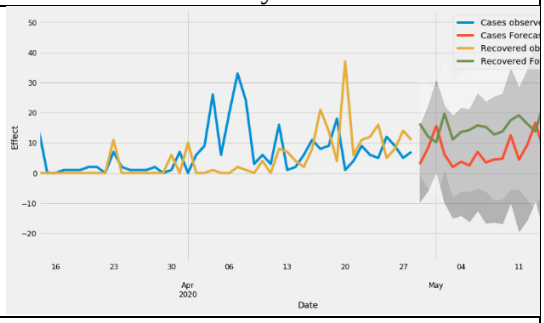
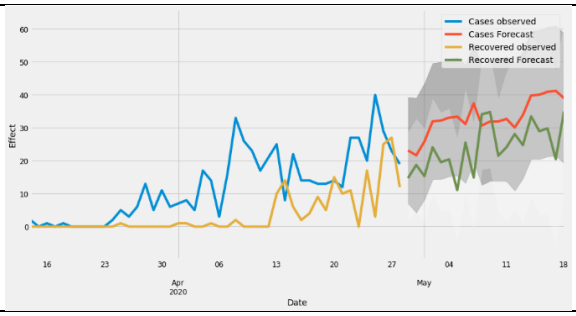
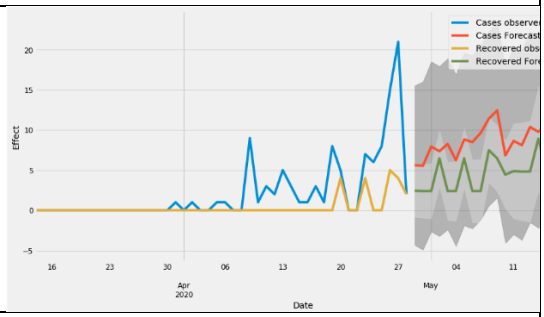
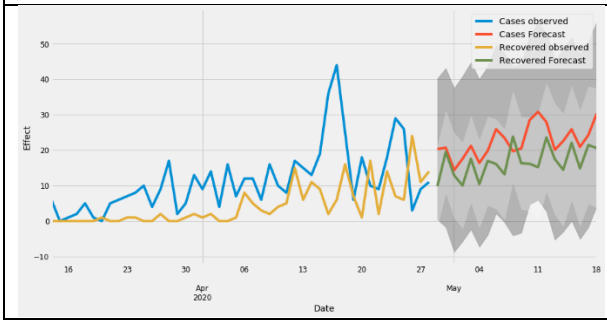
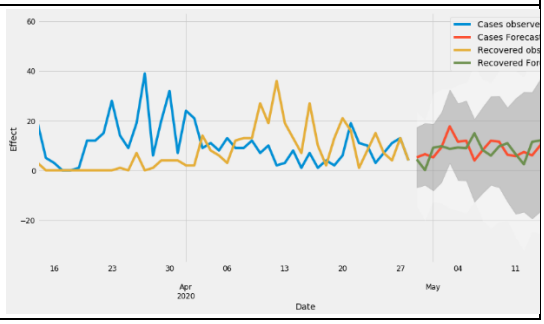
Andhra Pradesh	Haryana
	
Jammu & Kashmir	**Jharkhand**
	
Karnataka	**Kerala**
	

Table 1. Predictions of number of Cases and Recovered of various states is displayed in the form of graphs.

After the prediction of graphs, the data of the prediction produced by the ARIMA model is taken and interpreted to produce the estimated date on which the number of cases will reduce gradually form that day is displayed in the table – 2. The table consist of each state on which the model was trained by the respective state time series data was displayed along with the estimated date of recovery. The date which was shown in the model tells about the day from which the COVID-19 cases will reduce gradually in that state.

States of India	**Recovery Date (Predicted)**
Andhra Pradesh	07-07-2020
Bihar	08-05-2020 (Less Accurate)
Delhi	05-06-2020
Gujarat	10-06-2020
Haryana	07-06-2020
Jammu & Kashmir	13-06-2020
Jharkhand	25-06-2020
Karnataka	13-06-2020
Kerala	18-05-2020
Madhya Pradesh	04-07-2020

Maharashtra	18-06-2020 (Less Accurate)
Odisha	04-05-2020
Punjab	24-07-2020
Rajasthan	10-06-2020
Tamilnadu	14-07-2020
Telangana	05-05-2020
Uttar Pradesh	26-06-2020
West Bengal	18-05-2020 (Less Accurate)

Table 2. Predicted date when the number of Cases will reduce of various states is displayed.

In the table - 2, some of the dates are displayed with the caption "Less accurate" as the RMSE value of the model for that state data will comparatively high. Normally, all the other states data predictions have given the RMSE values in the range of 4 to 9 but Bihar, Maharashtra, and West Bengal has shown the RMSE values in the range of 20 to 30. So, the predicted information of those states is not as accurate and can change. The accuracy of the proposed model was predicted using the produced RMSE values. For the RMSE value which is less than 5 the accuracy was 95% and above, for the value less 10 it was 90% and above. But those states with has high RMSE value we have considered the accuracy to be 80%. Therefore, from all the values RMSE we can estimate that the accuracy of the model can be 90%.

5. Conclusion

The effect or impact of the pandemic virus COVID-19 in India was predicted using a statistical model ARIMA. The details about the ARIMA model was mentioned and the steps that were involved in the future prediction from the input data is also described. This paper also gives the graphical information about the COVID-19 effect on each state in India with 100 or more cases as per the records till 25-April-2020. As the accuracy of the model was 90% the predicted information may become true if the rules or precautions which were imposed by the government are strictly followed by the people of the respective states, then the effect may reduce from the predicted dates and even the economic condition will become normal else there will be a rapid increase in the number of cases and the economy of the India will collapse in the near future.

Conflict of Interest

The authors confirm that there is no conflict of interest to declare for this publication.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors sincerely appreciate the editor and reviewers for their time and valuable comments.

References

Perone., Gaetano. (2020). An Arima Model to Forecast the Spread and the final size of COVID-2019 Epidemic in Italy. SSRN Electronic Journal. DOI: 10.2139/ssrn.3564865.

Weron, Rafał. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. International Journal of Forecasting. 30. DOI: 10.1016/j.ijforecast.2014.08.008.

Kumar, Manoj & Anand, Madhu. (2014). An Application of Time Series Arima Forecasting Model for Predicting Sugarcane Production in India. Studies in Business and Economics. 9, 81 – 94.

Soumik Ray, Banjul Bhattacharyya (2015), Availability in Different Source of Irrigation in India: A Statistical Approach, International Journal of Ecosystem, 5(3A): 109-116.

Almasarweh, Mohammad & Alwadi, Sadam. (2018). ARIMA Model in Predicting Banking Stock Market Data. Modern Applied Science. 12. 309. 10.5539/mas.v12n11p309.

Awajan AM, Ismail MT, AL Wadi S (2018) Improving forecasting accuracy for stock market data using EMD-HW bagging. PLOS ONE 13(7): e0199582. https://doi.org/10.1371/journal.pone.0199582.

Fildes, Robert & Nikolopoulos, K & Crone, Sven & Syntetos, A. (2008). Forecasting and operational research: A review. Journal of The Operational Research Society - J OPER RES SOC. 59. 1150-1172. 10.1057/palgrave.jors.2602597.

Adebiyi, Ayodele & Adewumi, Aderemi & Ayo, Charles. (2014). Stock price prediction using the ARIMA model. Proceedings - UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, UKSim 2014. 10.1109/UKSim.2014.67.

Armstrong, J.. (2001). Evaluating Forecasting Methods. 10.1007/978-0-306-47630-3_20.

Chenghao Liu, Steven C. H. Hoi, Peilin Zhao, and Jianling Sun. 2016. Online ARIMA algorithms for time series prediction. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI'16). AAAI Press, 1867–1873.

Taylor, Sean & Letham, Benjamin. (2017). Forecasting at Scale. The American Statistician. 72. 10.1080/00031305.2017.1380080.

Petrevska, Biljana. (2017). Predicting tourism demand by A.R.I.M.A. models. Economic Research-Ekonomska Istraživanja. 30. 939-950. 10.1080/1331677X.2017.1314822.

Das, Soumitra & Ray, Soumik & Sen, Abhishek & Siva, G. & Das, Shantanu. (2019). Statistical Study on Modeling and Forecasting of Jute Production in West Bengal. International Journal of Current Microbiology and Applied Sciences. 8. 1719-1730. 10.20546/ijcmas.2019.807.204.

Alsharif, M.H.; Younes, M.K.; Kim, J. Time Series ARIMA Model for Prediction of Daily and Monthly Average Global Solar Radiation: The Case Study of Seoul, South Korea. Symmetry 2019, 11, 240.

Totakura, Varun & Devasekhar, V & Sake, Madhu. (2020). Prediction of Stock Trend for Swing Trades Using Long Short-Term Memory Neural Network Model. International Journal of Scientific & Technology Research. Volume 9. 1918-1923.

Data Source: https://api.covid19india.org/csv/.