Predictions of COVID-19 patientsraise, recovery and death rate in India by ARIMA model

Thadikamala Sathish[1], Anirban Ray[1], N. Nanda Gopal[2]
[1]Vaccines R&D, Cadila Pharmaceuticals, Dholka, Ahmedabad, Gujarat, India.
[2]National Institute of Biologicals, Ministry of Health and Family welfare, Noida, Uttar Pradesh.

Abstract
Background*: The recent outbreak of COVID-19 in different states of the India is a major concern for all the administrative units of our country. The effect of lockdown and social distancing plays a major role for COVID-19 spread and recovery percentage. Forecasting COVID-19 patients has recently become the focus of numerous researchers across the globe.*

Materials and Methods*: The current study aims to develop an auto-regressive integrated moving average (ARIMA) model to predict the COVID-19 patients raise, recovery and death in India based on the daily data obtained from the Indian Govt from 30th Jan 2020 to 15th May 2020. The autocorrelation function (ACF), partial autocorrelation function (PACF) and standardized residuals are used to analyze goodness of fit of the constructed model. Coefficient of determination (R^2) was used to evaluate the performance of the model.*

Results*: Forecasting was done by using the constructed models up to July 8th 2020. With the constructed model It was forecasted that there could be COVID-19 patients raise up to 2,30,000 in this around 1,00,000 patients recover and ~7000 effected people could die by 8th July 2020.*

Conclusion:*The constructed model mostly relies on the lock down data. However, changing the lockdown polices or restoring the normal conditions may increase the predicted number of patients.*

Key Words: *COVID-19, Forecasting, Lockdown, Time series modelling, ARIMA, Social distancing*

Date of Submission: 20-05-2020 Date of Acceptance: 05-06-2020

I. Introduction

The coronavirus disease 2019 (COVID-19) epidemic started in late December 2019 in Wuhan, the capital of central China's Hubei Province. Since then, it has rapidly spread across China and in other countries, raising major global concerns [1].By 15th May 2010 in India total number of COVID-19 patients raised up to 85,783 in this 30, 258 patients were recovered and 2,896 people were died and remaining people are under treatment. Person-to-person transmission of the virus has been documented in several countries outside of China, including large outbreaks in Iran, South Korea, and Italy [1]. Vaccines are proved to be the most effective and economical means to prevent and control infectious diseases. Several countries, companies, and institutions announced their programs and progress on vaccine development against the virus. While most of the vaccines are under design and preparation, there are some that have entered efficacy evaluation in animals and initial clinical trials [2].

Globally numerous researchers have been proposed various forecasting models on COVID-19 pandemic. These models mostly depend on complex statistical or artificial intelligence methods[3-7]. Normally the prediction models can be classified into three categories those are qualitative, quantitative techniques and Artificial neural networks (ANNs). Expert opinions and/or personal predictions are consider as a qualitative techniques. Quantitative techniques are based on mathematical models. Quantitative methods are divided as time series and causal forecasting techniques. Causal forecasting is rely on the relationship between dependent and independent variables. In current COVID-19 pandemic, finding the relationship between no of patients and other factors is difficult also fluctuation of regular patients data, causal forecasting models are not reliable. Time series forecasting models depends on the collection of data points over a designated period of time and then predicts future outputs based on previous data points[6]. The major limitation in time series forecasting is the deficiency of a deterministic cause [8]. ANNs are has wider application and these are used in COVID-19 predictions also [5]. The ANNs are consider as a "black-box" method, in this the working principle is unknown this makes ANNs predictions are lesser reliable. Some of the authors reported that time series auto-regressive integrated moving average (ARIMA) model performed better than ANNs in solar radiation prediction[9].Gupta et al [6] used the time series data to model the COVID-19 outbreak in India, they used the data from 22nd Mar 2020 to 8th April 2020. Therefore, in the present study we aimed to developing a time series forecasting technique for COVID-19 patients in terms of total confirmed cases, total recovery and total death cases. This forecast procedure isspeedy and standard way to generate predictions for all variables in a single step. The performance, accuracy, reliability and suitability of the constructed model could be investigated based on standardized residual, ACF (auto correlation factor), and PACF (Partially auto correlation factor).

II. Materials and Methodology

This aim of this study is to build a time series ARIMA model to forecast daily COVID-19 new patients, total recovery & death of patients in India, based on the daily data obtained from the Indian COVID-19 patients data over 160 days. The model performance was tested against established tests, such as standardized residual, ACF and PACF.

Data Collection

COVID-19 patient's data in India for over 160 days that obtained from the COVID-19 website [10]. Table 1 presents the basic statistics of the collected dataTo determine the outliers in the data the initially the overall box plot was drawn and later with along with the data box plot was drawn.

Table 1: Descriptive statistics of the total data

	Daily Confirmed	**Total Confirmed**	**Daily Recovered**	**Total Recovered**	**Daily Deceased**	**Total Deceased**
N of days	107	107	107	107	107	107
Mean	801.7103	12765.91	282.785	3445.766	25.71963	417.2523
Std. Error of Mean	115.4985	2107.773	49.81728	668.927	3.697707	69.27225
Median	74	497	5	25	1	9
Mode	0	3	0	3	0	0
Std. Deviation	1194.726	21802.98	515.3139	6919.435	38.24938	716.5578
Variance	1427370	4.75E+08	265548.4	47878579	1463.015	513455
Skewness	1.508662	1.878785	2.065488	2.315026	1.477347	1.85487
Kurtosis	1.093833	2.594971	3.581277	4.666784	1.066178	2.426225
Range	4311	85782	2277	30258	140	2752
Minimum	0	1	0	0	0	0
Maximum	4311	85783	2277	30258	140	2752

Trend analysis

To determine the total COVID-19 patients increase, recovery and death a trend analysis was done. The data was fitted with linear, exponential, decay and S-shape curves and analyzed which trend it follows.

ARIMA Forecasting Model

Time series forecasting is a multidisciplinary scientific tool used to solve prediction problems. Its implementation is easy and flexible because it only requires historical observations of the necessary variables [9]. Box and Jenkin in 1976 first presented the ARIMA model [11]. The general equation of successive differences at the d^{th}difference of Xt is as follows:

$$\varDelta^d X_t = (1 - B)^d X_t - - - (1)$$

Where d = difference order and is usually 1 or 2,

B = Backshift operator.

The successive difference at one-time lag equals to,

$$\varDelta^1 X_t = (1 - B)^d X_t = X_t - X_{t-1} - - - (2)$$

ARIMA model was developed, ACF and PACF data was calculated. Developed model performance was evaluated using RMSE (Root mean square error) and R^2(correlation coefficient) values.

III. Results and Discussion

In the present study the data was taken from the 30th Jan 2020, which was first COVID-19 case reported to 15th May 2020. This model doesn't count any lockdown effects, it is based on the patients no only. Table 1 depicts the basic statistics of the daily and total new patients, recovered and death patients. The high variance of the total confirmed, recovered and death cases is attributed due to many days no cases was observed and slow progression of disease in India at early stages. Further total cases, total recovery and total death cases only taken for further modeling. To observe the data variation a box plot was drawn fig 1. From this fig it was noticed that 4th May 2020 onwards the new COVID-19 patients are raising a lot which gives a dangerous alarm to the society.

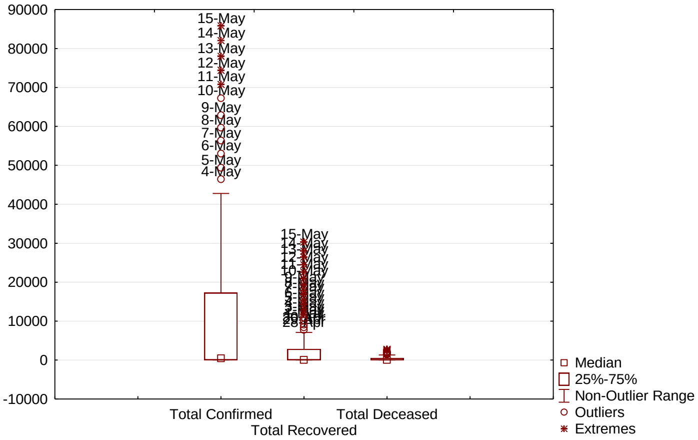

Fig 1: Box plot of total confirmed, total recovered and total deceased by COVID-19 in India from 30th Jan 2020 to 15th May 2020.

However the recovery rate in also raised from 24th April 2020, which gives a positive sign on COVID-19 patients. The total death cases are also rising however those are not incremental as new cases frequencies. Further to know the pattern of growth of COVID-19 patients trend analysis was performed. Fig 2 shows the selected three variables trend analysis. The analysis revels that COVID-19 total patients, recovery and death are not following the linear trend, they are following the quadratic trend which indicates that the increase the no of new patients as well as recovery and death are doubling along with the time, which indicates over the time the no COVID-19 victims are raising double as compared with previous days. The death cases also following the second order model which means in near future there is chances of more death cases in the COVID-19 infected patients.

Table 2: Model statistics

		Ljung-Box Q		
Variable	**R^2 value**	**Statistics**	**DF**	**Sig.**
Total Confirmed	0.729383	33.48779	14	0.002449
Total Recovered	0.811216	48.57543	14	1.06X 10^{-05}
Total Deceased	0.579722	18.69585	14	0.00176

To develop and use ARIMA model the data should be stationary time series. Therefore initially the data was analyzed for stationary time series. It was observed that the data it self is not a stationary time series, so further the data was processed, by applying the various no of difference order. First differences has sufficiently stationarize the data.

ARIMA is composed of two parts, one is AR (auto- auto-regressive) and second one is MA (moving average). ARIMA model is generally denoted to as ARIMA (p, d, q). In this, p and q are the order of AR and MA respectively, where d is the difference order. Usually p = 1–5, q = 1–3 and d=1. The appropriate ARIMA model selection is depends on the ACF and PACF of the stationary time series data. The ACF and PACF were teste for 16 lags to investigate the seasonality action. It was observe that in ACF all lags spiked out the limits. Where as in PACF the first lag for total confirmed cases, first and second spikes for total recovery cases and first, second and eleventh lag spiked out. Based on this ARIMA (1,1,1) & ARIMA (1,1,2) were tested. The ARIMA (1,1,1) was selected as a final model. Table 2 depicts the ARIMA (1,1,1) model statistics

Fig 3 shows the data along with the predictions up to the July 8th. The data that cross the horizontal line (blue colored line) was the predicted data. From this data, it was observe that up to 21stjune the cases could increase from that it reaches the stationary phase. As per the current lockdown conditions exists by the 8th July there could be 2,30,000 total COVID-19 confirmed cases and 1,00,000 recovery cases and ~7000 death cases could be recorded. All these predictions based on major data on lockdown period. However, these predicted values might exceed based on the lifting of lockdown condition in India.

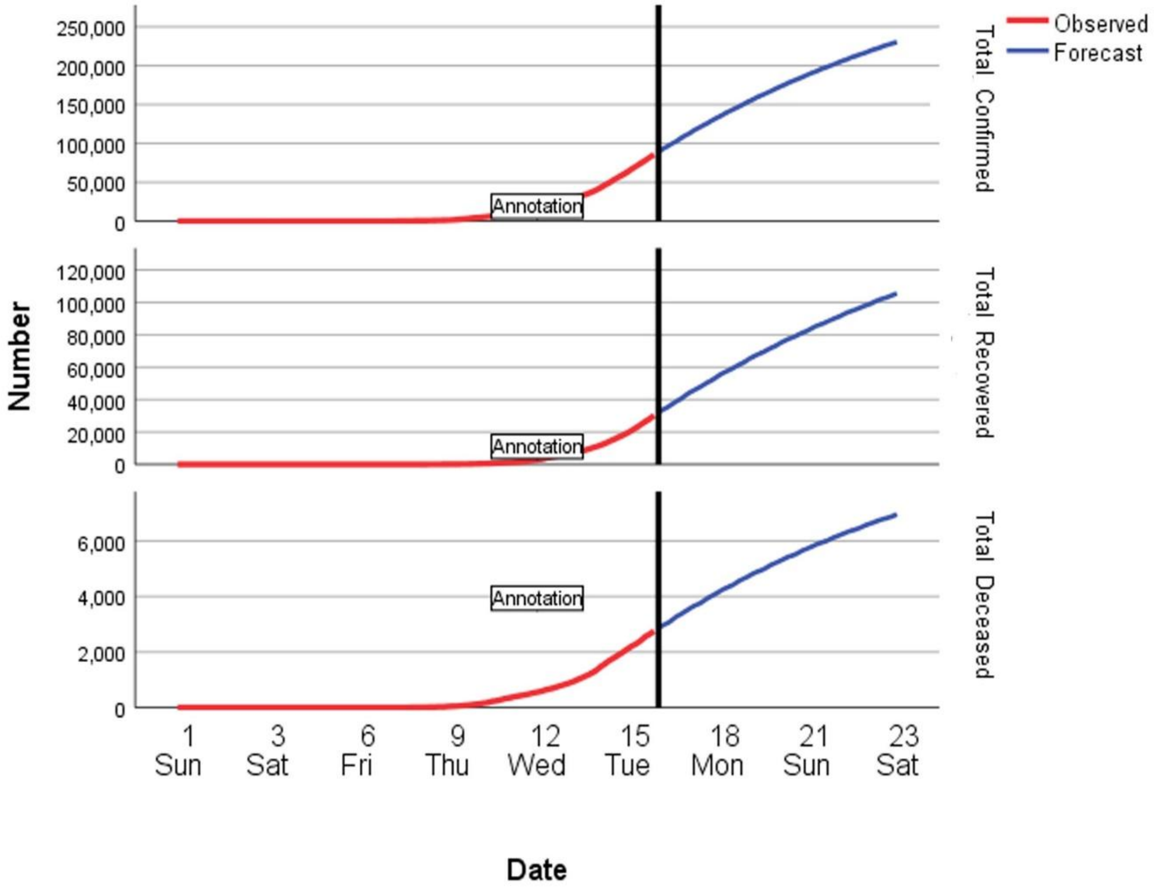

Fig 3: ARIMA model forecast of COVID-19 patients in India up to July 8th 2020.

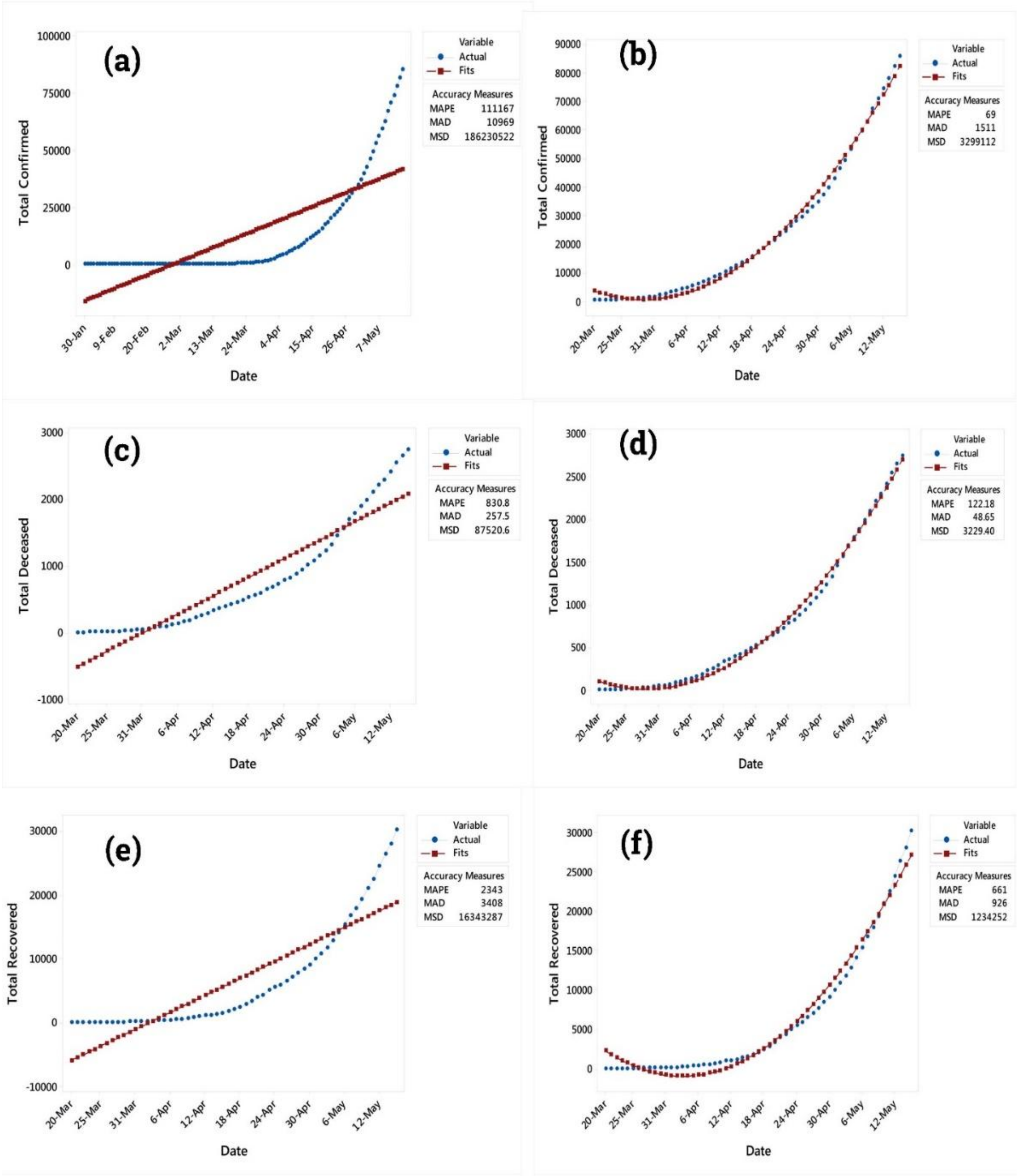

Fig 2: Trend analysis of the COVID-19 patients in India. (a) Linear trend analysis of total confirmed cases (b) Quadratic trend analysis of total confirmed cases (c) Linear trend analysis of total recovered cases (d) Quadratic trend analysis of total recovered cases (e) Linear trend analysis of total death cases (f) Quadratic trend analysis of total death cases

References

[1]. WHO Emergency Committee. Novel Coronavirus (2019-nCoV) Situation Report 50. 2020. Available at: https://www.who.int/docs/defaultsource/coronaviruse/situation-reports/

[2]. Zhang J, Zeng H, Gu J, Li H, Zheng L, and Zou Q. Progress and Prospects on Vaccine Development against SARS-CoV-2. Vaccines (Basel), 2020,29;8(2)

[3]. Agosto A , Giudici P. A Poisson autoregressive model to understand COVID-19 contagion dynamics. SSRN, 2020. (http://dx.doi.org/10.2139/ssrn.3551626)

[4]. Arti M.K., Modeling and Predictions for COVID 19 Spread in India. *medRxiv*, 2020

[5]. Chatterjee K, Chatterjee K, Kumar A, Shankar S. Healthcare impact of COVID-19 epidemic in India: A stochastic mathematical model. Medical journal armed forces India. 2020, 76(2), 147-155.

[6]. Gupta, R., Pal, S.K., Pandey, G. 2020. A Comprehensive Analysis of COVID-19 Outbreak situation in India. *medRxiv*, 2020.04.08.20058347

[7]. Rizk-Allah RM, Hassanien AE. COVID-19 forecasting based on an improved interior search algorithm and multi-layer feed forward neural network. 2020,arXiv

[8]. Beaumont, C.; Makridakis, S.; Wheelwright, S.C.; McGee, V.E. Forecasting: Methods and Applications. J. Oper. Res. Soc. 1984, 35, 79.

[9]. Alsharif MH, Younes MK, 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, 240-257

[10]. https://www.covid19india.org

[11]. Khashei, M.; Bijari, M. A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. Appl. Soft Comput. 2011, 11, 2664–2675.

Thadikamala Sathish, et. al. “Predictions of COVID-19 patientsraise, recovery and death rate in India by ARIMA model.” *IOSR Journal of Pharmacy and Biological Sciences (IOSR-JPBS)*, 15(3), (2020): pp. 05-10.