MedCrave
Step into the World of Research

Biometrics & Biostatistics International Journal

Research Article

Open Access

CrossMark

Analysis and forecast of COVID-19 in India, USA and Italy - an application of ARIMA Model

Abstract

Background: Corona virus disease (COVID-2019) is a severe ongoing novel pandemic that is spreading rapidly across the world. Analyzing and predicting COVID-19 prevalence will help governments to take necessary actions in controlling the spread. Time series analysis has proved to be efficient for estimating future impact in such circumstances.

Methods: Auto Regressive Integrated Moving Average (ARIMA) models were developed to predict the epidemiological trend of pandemic in India, The USA and Italy up-to 20 July 2020. A 15 day forecast were made to study the future scenario.

Results: We could identify that all the countries had a linear trend in the increase of daily confirmed COVID-19 cases. Our studies produced a 15 day forecast results for India, US and Italy. the 95% PI for 20 July 2020 are, (25821.72 , 32991.38), (44278.89 , 87095.68) and (-2935.20, 2685.14) respectively for the three countries.

Conclusion: The US and India will have a rise in number of daily counts of COVID-19. Italy will gradually slow the pace. It is supposed that the present prediction models will assist the government and medical personnel to be prepared for the upcoming conditions and have more readiness in healthcare systems.

Keywords: COVID-19, forecasting, ARIMA models, time series

Volume 10 Issue 2 - 2021

Elbin Siby,[1] Maria Joseph,[1] Aneena Thankachan,[1] K. K. Jose[1,2]

[1]Department of Biostatistics, St. Thomas College Palai, India
[2]School of Mathematics & Statistics, Mahatma Gandhi University, India

Correspondence: K. K. Jose, School of Mathematics & Statistics, Mahatma Gandhi University, India, Email kkj.smsda.mgu@gmail.com

Received: June 15, 2021 | **Published:** June 30, 2021

Introduction

The COVID-19 pandemic has spread rapidly across the world. It was initially identified in the Hubei province of China. COVID-19 has a dynamic structure and spreads more abruptly than SARS-CoV and MERS-CoV; two zoonotic viruses identified in the past decade.[1] COVID-19 has now, as of 06 July 2020, affected worldwide among 213 countries, territories and two international conveyances.[2] For people residing in or travelling to these countries, the risk of incidence is higher. Older people and those with pre-existing medical conditions are much likely to be affected.[3]

As on 1 September 2020, a total of about 11,563,000 confirmed cases and 536,800 deaths were reported globally.[2] At a point, Italy was the epicentre of the disease. The US is now, as of July 4th 2020, leading with over 3 million confirmed COVID-19 cases and about 1,29,000 deaths. India, having a different population dynamics compared to The US and Italy had a lower number of confirmed COVID-19 cases and deaths. The number of confirmed cases vary due to differences in epidemiological surveillance and healthcare capacities between countries. Up-to-date, no treatments or vaccination were found effective in curing the illness. In this situation, preventing the infection and preparations in health care services is of great importance.[3] Modelling and predicting confirmed COVID-19 cases will help the healthcare personnel to be prepared for the worst scenario. Statistical models that could predict the forthcoming situation and model the present situation will be of great help.

In this study, Auto Regressive Integrated Moving Average (ARIMA) models are used in modelling and forecasting confirmed COVID-19 cases in India, USA and Italy. This model has proved to be more efficient than some prediction models such as Wavelet Neural Network (WNN) and the Support Vector Machine (SVM) in prediction of natural disasters.[4]

How different are the pandemic trends in these countries? What will be the future scenario in these countries? So far no serious investigations were taken to analyse and compare the differences in COVID-19 status between these three countries. Therefore, the main aim of this paper is to study and make forecasts of confirmed COVID-19 cases. We also study the relation between COVID-19 mortality and older age group in these countries.

The remainder of the paper is unfolded as follows: In Section 2, materials and methods used for the study is briefly discussed. Section 3 lists out the relevant results of the current study along with tables and figures. In the final Section 4, methods and findings are discussed which ends with the conclusion of the study.

Materials and methods

a) Data

The data of cumulative confirmed cases, deaths and recoveries and testing details of India, the USA and Italy was obtained from the Kaggle repository which they derived from Centre for Systems Science and Engineering (CSSE) at Johns Hopkins University, and MS Excel was used to build a time series database. Data was filtered starting from the 50th confirmed case for each of the country till 05 July 2020. We used the ARIMA model with R programming (version 1.2.5042.0) and validated it using Akaike information criterion (AIC. New variables namely CFR (Case Fatality Rate) and CMR (Crude Mortality Rate) were calculated.

b) ARIMA models

The ARIMA class of models is an important forecasting tool, and is the basis of many fundamental ideas in time-series analysis. The acronym ARIMA stands for auto-regressive integrated moving average. The model includes an auto-regressive (AR) model, moving

average (MA) model and an integrated part that will account for the non-stationary behaviour of data.[5]

An ARIMA (*p, d, q*) model is given by

$$\Phi(B)(1-B)^d X_t = \theta(B) \quad (1)$$

where

$$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \varphi_p B^p \quad (2)$$

is a polynomial in B of order *p*, B is the backshift operator, such that $BX_t = X_{t-1}$ and X_t is the predicted number of variable,

$$\theta(B) = 1 + \theta_1 B + \ldots + \theta_q B^q \quad (3)$$

is a polynomial in B of order *q*. *p* is the order of the auto-regressive (AR) part of the ARIMA model, *q* is the order of the moving average (MA) part and d refers the degree of trend difference. The error terms are generally assumed to be independent identically distributed random variables (i.i.d.) sampled from a normal distribution with zero mean: ε_t follows $N(0, \sigma^2)$ where σ^2 is the variance.[6]

Before analyzing, time series must become stationary based on mean and variance. The Augmented Dickey-Fuller (ADF) is used in recognizing stationary behavior in the mean, the tests generally take the null hypothesis to be that there is a unit root (so that d = 1) and Box Cox test is used for testing whether the time series is stationary based on variance or not.[7]

After achieving stationarity, to determine *p*, *d*, *q* values, ACF and PACF of the differenced data is plotted. Once *p, d, q* are determined, the data can be fitted using the ARIMA *(p, d, q)* model. A conditional sum of squares method or method of maximum likelihood is used to fit the data. 5x5 matrices can be created for *p, q* values and all the models should be fitted. The model with the lowest AIC value whose residuals are white noise is selected as the best model. When the first 25 models were not appropriate, the matrices could be extended and the best model should be selected by AIC criterion.

$$AIC = -2\ ln\ (max.\ likelihood) + 2m \quad (4)$$

where *m* denotes the number of independent parameters estimated in the model.[6]

The most optimized and efficient model could thus be selected. The insignificant coefficients should be excluded. Model diagnostics is the next step. ACF plots are used to find if the residuals of fitted time series model are a white noise. Box-Ljung's test is used for the same. A significant *p* value suggests a lack of fit, implying the wrong model selection.[9]

Results

As a first step in ARIMA modelling, the stationarity was checked using ACF and PACF plots. To confirm the findings, ADF test was performed. The test showed non stationarity in the three countries with 95% confidence. All the *p* values were non-significant, indicating non stationary data. The first order differencing stabilized the mean of daily confirmed COVID-19 cases. The trends in data were thus eliminated. A significant *p* value by ADF test supported the finding. ARIMA models were determined according to ACF and PACF plots. In addition to the selected ARIMA models, several models were developed and those models with lowest AIC values were chosen. Minimum AIC values of the adopted models were $1838.1_{India,}$ $2426.72_{USA,}$ 1837.26_{Italy}. The AIC values of the competing models are given in Table 1. ARIMA $(9,1,9)_{India}$, ARIMA $(7,1,2)_{USA}$, ARIMA $(5,1,7)_{Italy}$ were the best fit models.

Table 1 AIC values of competed models in ARIMA modelling of daily confirmed cases in three countries

Model (India)	AIC	Model (USA)	AIC	Model (Italy)	AIC
ARIMA (9,1,7)	1843.57	ARIMA (4,1,3)	2428.6	ARIMA (4,1,7)	1846.46
ARIMA (9,1,9)	**1838.1**	**ARIMA (7,1,2)**	**2426.72**	**ARIMA (5,1,7)**	**1837.26**
ARIMA (9,1,8)	1841.47	ARIMA (6,1,5)	2431.54	ARIMA (5,1,8)	1841.23

The models fitted the data reasonably well. Box-Ljung's test was done to test the independence of residuals of the fitted model. All the *p* values were statistically non-significant indicating that the residuals are white noise.

The fitted models for India, USA and Italy are given below.

$$\Delta x_{t(India)} = C - 0.7030\ \Delta x_{t-1} - 0.1522\ \Delta x_{t-2} + 0.032\Delta x_{t-3} - 0.1477\ \Delta x_{t-4} - 0.0637\ \Delta x_{t-5} + 0.3183\ \Delta x_{t-6} + 0.6727\ \Delta x_{t-7} + 0.402\ \Delta x_{t-8} + 0.4946\ \Delta x_{t-9} + 0.7642\varepsilon_{t-1}$$
$$+0.2420\varepsilon_{t-2} - 0.1702\varepsilon_{t-3} + 0.0093\varepsilon_{t-4} - 0.0538\varepsilon_{t-5} + 0.3227\varepsilon_{t-6} + 0.169\varepsilon_{t-7} - 0.3191\varepsilon_{t-8} - 0.8093\varepsilon_{t-9} + \varepsilon_t \quad (6)$$

$$\Delta x_{t(USA)} = C + 0.6956\Delta x_{t-1} - 0.2044\ \Delta x_{t-2} - 0.2995\ \Delta x_{t-3}$$
$$+0.2667\ \Delta x_{t-4} 0.0153\ \Delta x_{t-5} 0.1060\Delta x_{t-6} + 0.2447\Delta x_{t-7} 1.3843\varepsilon_{t-1} 0.7011\varepsilon_{t-2} + \varepsilon_t \quad (7)$$

$$\Delta x_{t(Italy)} = C\ +0.1659\Delta x_{t-1} 0.2931\Delta x_{t-2} + 0.1289\Delta x_{t-3} - 0.4079\ \Delta x_{t-5}\ + 0.2441\ \Delta x_{t-5} - \varepsilon_{t-2} - 0.4250\varepsilon_{t-3} +$$
$$0.3522\ \varepsilon_{t-2} - 0.4250\varepsilon_{t-3} + 0.5738\ \varepsilon_{t-4} - 0.2148\varepsilon_{t-5} + 0.2025\varepsilon_{t-6} + 0.5026\varepsilon_{t-7} + \varepsilon_t - 0.04058\varepsilon_{t-4} + 0.04058\varepsilon_{t-4} + \varepsilon_t \quad (8)$$

Citation: Siby E, Joseph M, Thankachan A, et al. Analysis and forecast of COVID-19 in India, USA and Italy - an application of ARIMA Model. *Biom Biostat Int J.* 2021;10(2):75–79. DOI: 10.15406/bbij.2021.10.00333

Point as well as interval forecasts for the next 15 days were made. Forecasts are presented in Table 2.

Table 2 95% PI for India, the USA and Italy

Date	95% PREDICTION INTERVAL		
	India [Lower, Upper]	**USA** [Lower, Upper]	**Italy** [Lower, Upper]
05 July 2020	(22816.44, 24083.91)	(44400.70 , 58586.21)	(-523.71 , 805.77)
09 July 2020	(21615.63 , 24412.58)	(51264.20 , 68874.05)	(-1275.97 , 860.57)
14 July 2020	(23548.35 , 28345.53)	(45585.62 , 74077.82)	(-2024.82 , 1801.44)
17 July 2020	(23195.23 , 28982.92)	(48861.70 , 82919.44)	(-2508.58 , 2251.71)
20 July 2020	(25821.72 , 32991.38)	(44278.89 , 87095.68)	(-2935.20 , 2685.14)

Discussion

The first positive case was identified in the USA followed by India and the Italy. The three countries took different paths in the pandemic progress. India had a hike in the number of cases at mid-March. But this hike is not significant compared with the USA and Italy which had enormous increases in the pandemic incidence during the first weeks of March. The USA and Italy has just started flattening the curve during the last weeks of May 2020. As per forecasts on the three countries, Italy will have lower numbers of cases reported daily. But in India and The US, the situation will be just the opposite. Predictions made in India show a slow and steady increase in the number of daily cases during the 2nd and 3rd week of July 2020. A 95% PI and 80% PI also suggests the same results.

The high differences, through another point of view, may be because of the false number of total cases. A case is reported only when a test is done. The total number of cases hence depends directly on the number of tests done. The USA and Italy have the respective first and second places in the number of daily tests performed (Figure 1).

The USA, Italy and India have done 27.07, 42.44 and 1.17 tests per thousand people. India has a very lower number of daily tests and thus implicating the lower confirmed cases. The study suggests that the true cases in India are concealed. The undetected cases thus are larger in India and can cause a serious and dangerous outbreak than the USA and Italy.

Governments have adopted their own measures for prevention. This could result in changes in the future situation. For e.g. India has started taking NRI's. This could increase the number of cases. One of the main limitation of our study is that these couldn't be incorporated in ARIMA model.

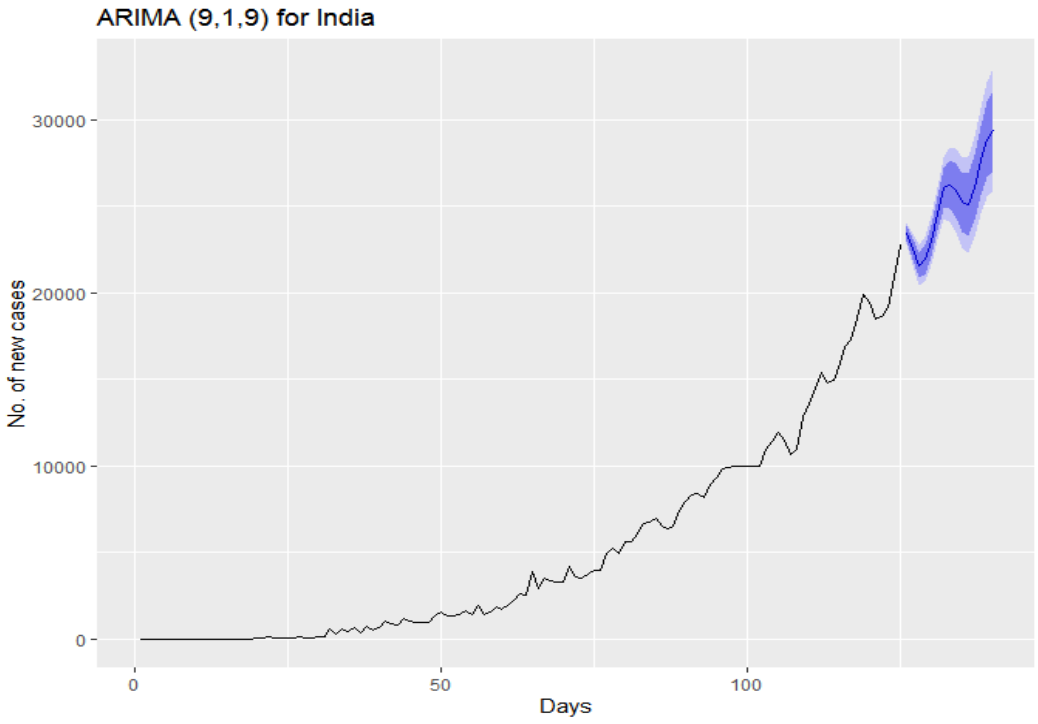

Figure 1 Forecast of daily confirmed cases in India.

Citation: Siby E, Joseph M, Thankachan A, et al. Analysis and forecast of COVID-19 in India, USA and Italy - an application of ARIMA Model. *Biom Biostat Int J.* 2021;10(2):75–79. DOI: 10.15406/bbij.2021.10.00333

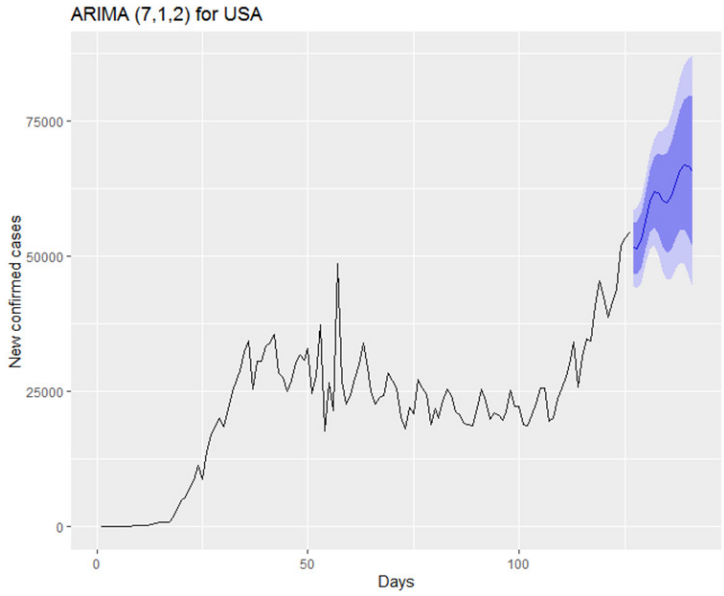

Figure 2 Forecast of daily confirmed cases in The USA.

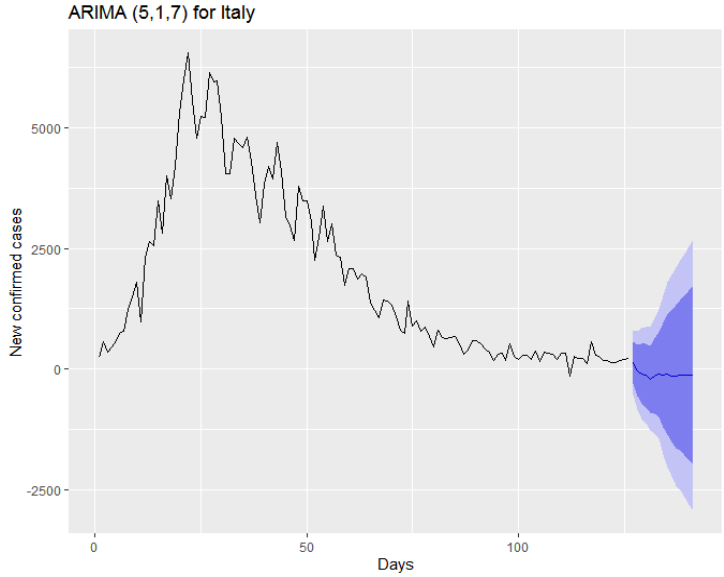

Figure 3 Forecast of daily confirmed cases in Italy.

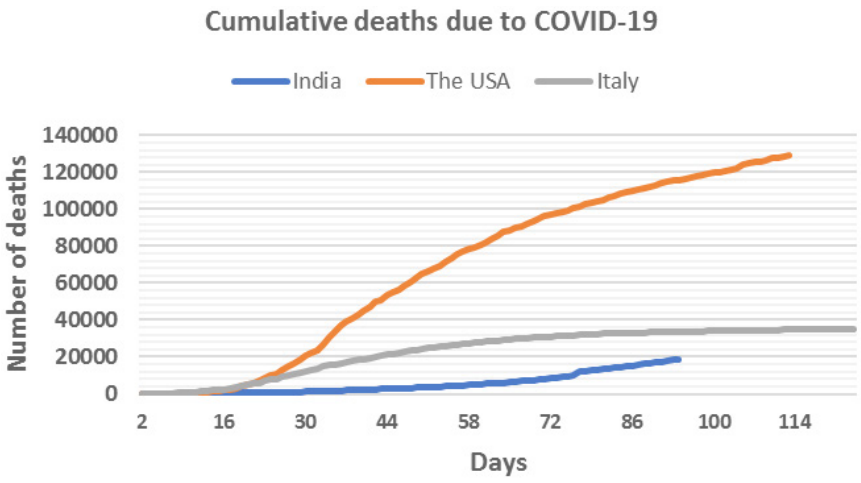

Figure 4 Total COVID-19 deaths in India, The USA and Italy.

Citation: Siby E, Joseph M, Thankachan A, et al. Analysis and forecast of COVID-19 in India, USA and Italy - an application of ARIMA Model. *Biom Biostat Int J.* 2021;10(2):75–79. DOI: 10.15406/bbij.2021.10.00333

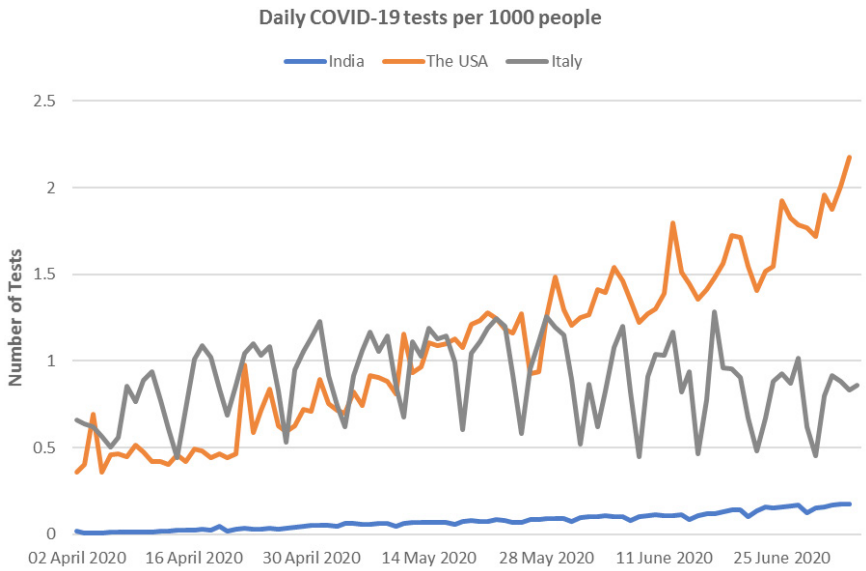

Figure 5 Daily COVID-19 tests per 1000 people.

Conclusion

The scenario will be in favour of Italy, whereas against USA and India. India should increase the number of tests and reveal the true number of affected patients. Italy has a total of 3.2 hospital beds per 1000 inhabitants. The USA has 2.8 and India has only 0.5 beds per 1000.[11] No. of hospital beds, ventilators and other facilities must be dealt with immediately. Italy even though is gaining control, should increase the pace or continue as of now in implementing measures to contain the disease, but never lessen the same. The continuation of the restrictive measures and the strict compliance with the rules, such as traffic and travel restriction, ban on gatherings, and closure of commercial activities, may mitigate the size of the pandemic.

Funding

This research did not receive any specific grant from any of the funding agencies in public, commercial or any other sector. It is part of PG Project Work.

Acknowledgments

The authors are grateful to Mr. Noel George, p-Value Solutions Ltd Pala for the valuable suggestions during various stages of this research.

References

1. Novel Coronavirus Pneumonia Emergency Response Epidemiology. The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China. *Zhonghua liu xing bing xue za zhi= Zhonghua liuxingbingxue zazhi*. 2020;41(2):145.
2. Novel Corona Virus 2019 Dataset Day level information on covid-19 affected case. 2020.
3. Yousef Alimohamadi, Maryam Taghdir, Mojtaba Sepandi. The estimate of the basic reproduction number for novel coronavirus disease (covid-19): A systematic review and meta-analysis. *Journal of Preventive Medicine and Public Health*. 2020;53(3):151–157.
4. Zhang Yuhu, Huirong Yang, Hengjian Cui, et al. Comparison of the Ability of ARIMA, WNN and SVM Models for Drought Forecasting in the Sanjiang Plain, China. *Natural Resources Research*. 2019;1–18.
5. Brockwell Peter J, Richard A Davis. Introduction to Time Series and Forecasting. Springer. 2016.
6. Chatfield Chris. Time Series Forecasting. *CRC Press*. 2000.
7. Box George EP, Gwilym M Jenkins, Gregory C. Reinsel, et al. Time series analysis: forecasting and control. *John Wiley & Sons*. 2015.
8. Metcalfe Andrew V, Paul SP Cowpertwait. Introductory time series with R. Springer-Verlag New York; 2009.
9. WHO regional Office for Europe. 2020.
10. Dicker Richard C, Fatima Coronado, Denise Koo, et al. Principles of epidemiology in public health practice; an introduction to applied epidemiology and biostatistics. 2006.
11. Hospital beds (indicator). 2020.