Time series analysis of COVID-19 cases

Kamalpreet Singh Bhangu, Jasminder Kaur Sandhu and Luxmi Sapra
Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Abstract

Purpose – This study analyses the prevalent coronavirus disease (COVID-19) epidemic using machine learning algorithms. The data set used is an API data provided by the John Hopkins University resource centre and used the Web crawler to gather all the data features such as confirmed, recovered and death cases. Because of the unavailability of any COVID-19 drug at the moment, the unvarnished truth is that this outbreak is not expected to end in the near future, so the number of cases of this study would be very date specific. The analysis demonstrated in this paper focuses on the monthly analysis of confirmed, recovered and death cases, which assists to identify the trend and seasonality in the data. The purpose of this study is to explore the essential concepts of time series algorithms and use those concepts to perform time series analysis on the infected cases worldwide and forecast the spread of the virus in the next two weeks and thus aid in health-care services. Lower obtained mean absolute percentage error results of the forecasting time interval validate the model's credibility.

Design/methodology/approach – In this study, the time series analysis of this outbreak forecast was done using the auto-regressive integrated moving average (ARIMA) model and also seasonal auto-regressive integrated moving averages with exogenous regressor (SARIMAX) and optimized to achieve better results.

Findings – The inferences of time series forecasting models ARIMA and SARIMAX were efficient to produce exact approximate results. The forecasting results indicate that an increasing trend is observed and there is a high rise in COVID-19 cases in many regions and countries that might face one of its worst days unless and until measures are taken to curb the spread of this disease quickly. The pattern of the rise of the spread of the virus in such countries is exactly mimicking some of the countries of early COVID-19 adoption such as Italy and the USA. Further, the obtained numbers of the models are date specific so the most recent execution of the model would return more recent results. The future scope of the study involves analysis with other models such as long short-term memory and then comparison with time series models.

Originality/value – A time series is a time-stamped data set in which each data point corresponds to a set of observations made at a particular time instance. This work is novel and addresses the COVID-19 with the help of time series analysis. The inferences of time series forecasting models ARIMA and SARIMAX were efficient to produce exact approximate results.

Keywords COVID-19, Time series analysis, ARIMA, SARIMAX, Forecasting

Paper type Research paper

1. Introduction

The coronavirus disease (COVID-19) emerged in Wuhan (China) in December 2019 (Yonar, 2020) and has now become a pandemic. Like many countries, India reported its first COVID-19 case on the 30th of January. When something like COVID-19 is on the spread, it goes through four stages; the first is the import stage of the disease and followed by the contact stage of the disease (Malki *et al.*, 2020). The third stage becomes the community spread and finally, the fourth stage is declared as a global pandemic. The outbreak of the COVID-19 is developing into a major international crisis and it is lending to influence the most important aspects of our daily lives. For example, places in most of the cities worldwide, where community gatherings occur have been shut down, such as malls, grocery shops, cinema halls and pubs. Apart from these, the global travel ban has impacted all aspects of life and majorly corporate travel has been immensely reduced.

The machine learning algorithms can be used to train from the COVID-19 data available publicly and then predict and infer useful information out of it (Lalmuanawma *et al.*, 2020). The extraction of different types of cases would help alert the government in making decisions of extension of lockdowns to safeguard the spread of this pandemic within their respective regions. This paper is our effort to make effective use of time series algorithms to predict and forecast COVID-19 cases accurately.

A time series is a time-stamped data set in which each data point corresponds to a set of observations made at a particular time instance (Nabi, 2020). An illustrative list of time series data across a variety of domains includes daily stock market figures, sales or revenue or profitability figures of companies by year and demographic information such as population, birth rate, infant mortality figures, literacy, per-capita income, school enrolment figures by year and also blood pressure and other body vitals by time units (Singh and Dhiman, 2018). The time series has a strong temporal (time-based) dependence – each of these data sets essentially consists of a series of time-stamped observations, i.e. each observation is tied to a specific time instance (Benvenuto *et al.*, 2020). Thus, unlike the regression,

The current issue and full text archive of this journal is available on Emerald Insight at: **https://www.emerald.com/insight/1708-5284.htm**

World Journal of Engineering
19/1 (2022) 40–48
 [ISSN 1708-5284]
[DOI 10.1108/WJE-09-2020-0431]

Received 10 September 2020
Revised 5 November 2020
Accepted 28 November 2020

the order of the data is important in a time series. Also, in a time series, the relationship between the response and explanatory variable is not taken into consideration because there is only one variable i.e. time (hours, minutes, seconds, month or year) in the model which is an independent variable (Abdulmajeed *et al.*, 2020). The cause behind the changes in the response variable is very much a black box.

In this study, the time series analysis of this outbreak forecast was done using the auto-regressive integrated moving average (ARIMA) model and also seasonal auto-regressive integrated moving averages with exogenous regressor (SARIMAX) and optimized to achieve better results. This work is organized as follows: Section 2 underlines the methodology of time series consisting of data set and model development. Section 3 discusses the stepwise evaluation, result and mean absolute percentage error (MAPE). Finally, Section 4 presents the conclusion.

2. Methodology

Time series data is the data measured at successive time period at uniform time interval. Some of the examples are closing prices of S&P, NSE, BSE and Dow Jones and quarterly revenue and profit of a company. Time series is usually univariate (one variable) and over time variable of interest changes with respect to univariate variable. The first step towards data exploration in time series is the trend analysis and is used for forecasting of the variable of interest such as profit and loss in most of the scenarios. These steps are covered under results and evaluation section and there are three variables of interest in this study that are confirmed, recovered and death cases of COVID-19.

2.1 Data set

Confirmed, recovered and death cases of COVID-19 infection are collected from the official website of Johns Hopkins University (https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html) for the time period from 22nd January 2020 to 12th August 2020. Table 1 shows the features of the data set.

2.2 Model development

It is worth mentioning about the components of the time series that comprise predictable and unpredictable components. The sub-components of predictable component consist of local ($\boldsymbol{X_t}$) and global (trend [T_t] and seasonality [S_t]) while unpredictable consist of noise (Z_t) component. The global component under predictable has trend (T_t) and seasonality (S_t) behaviours

Table 1 Description of features of data set

Feature name	Feature description
Observation date	The date when the COVID-19 case was observed
Province/state	Name of the province the case occurred
Country/region	Country in which the case occurred
Last update	The date when the data set recorded COVID-19 case
Confirmed	Number of confirmed cases on the observation date
Deaths	Number of death cases on the observation date
Recovered	Number of recovered cases on the observation date

(Fattah *et al.*, 2018). The trend (T_t) component refers to any pattern that talks about the overall increase or decrease in the values whereas seasonality (S_t) refers to a repeating pattern of values seen in the data (Dhiman and Kaur, 2018). The local (X_t) component is the sudden erratic influence or immediate effect and fluctuations on the dependent variable as compared to previous decision factors. The noise ($\boldsymbol{Z_t}$) is the error component and it cannot be predicted. So, in a time series, all the components are combined together i.e. time series (TS) = local (X_t) + trend (T_t) + seasonality (S_t) + noise (Z_t) and removing trend (T_t) and seasonality (S_t) from the original component gives left out components that is mixture of local (X_t) and noise (Z_t) (Li and Li, 2017). There are different ways in which these components of the time series can be related to each other (Dhiman and Kaur, 2019a). Their combination can either be additive or multiplicative. Additive model is given as $TS = T_t + S_t + X_t + N_t$ and multiplicative model is given as $TS = T_t * S_t * X_t * N_t$. When the magnitude of the seasonal pattern in the data increases with an increase in data values and decreases with a decrease in data values, the multiplicative model may be a better choice. When the magnitude of the seasonal pattern in the data does not directly correlate with the value of the series, the additive model may be a better choice.

2.2.1 Stationarity of a time series and its significance

A series is said to be "strictly stationary" if the marginal distribution of Y at time t, $p(Y_t)$, is the same as at any other point in time. Therefore, $p(Y_t) = p(Y_{t+k})$ and $p(Y_t, Y_{t+k})$ does not depend on t. (Here, $t > = 1$ and k is any integer). This implies that the mean, variance and covariance of the series Y_t are time invariant.

For a stationary time series, properties such as mean and variance will be the same for any two-time windows. A stationary time series will have no long-term predictable patterns such as trends or seasonality. Time plots will show the series to have a horizontal trend with the constant variance roughly. After removing the global pattern from the original data, the obtained series consists of left out components or residual series in terms of local ($\boldsymbol{X_t}$) and noise ($\boldsymbol{Z_t}$) components. Next is to find all the correlations of any two-time windows from the residual series. If any dependence of values on past values within these windows occurs, then that confirms the existence of local patterns and such predictions have to be modelled (Reddy *et al.*, 2017). In order words, stationary helps find the predictable i.e. local ($\boldsymbol{X_t}$) component and also called weak stationarity. Some of the tests to check the stationary are the histogram test and the Q-Q plot test. Autocorrelation function (ACF) and partial autocorrelation function (PACF) are the traditional tests used to determine stationary. If the data is not stationary, then it can be made stationary via differencing technique (Dhiman, 2020).

2.2.2 White noise

A series is called white noise if it is purely random in nature. It is a set of independent and uncorrelated values and has characteristics of Gaussian distribution with a zero mean and constant standard deviation. Time series is classified as white noise if it has mean equal to zero, standard deviation or volatility of time series over time is constant and the correlation between lags is zero, so that means that there should be no correlation between the time series and a lagged version of the

time series; in other words, there should be no correlation between lags.

Let $\{\epsilon_t\}$ denote such a series that it has zero mean $E(\epsilon_t) = 0$, has a constant variance $V(\epsilon_t) = \sigma^2$, is an uncorrelated $E(\epsilon_t \epsilon_s) = 0$ and random variable or auto-covariance is zero (Garg and Dhiman, 2020). Each observation is uncorrelated with other observations in the sequence and the scatter plot of such a series across time will indicate no pattern and hence forecasting the future values of such a series is not possible (Dhiman, 2019). The unpredictable component in time series is an error data called white noise (Z_t) and also called strong stationarity. If the time series data source shows white noise features, then there is no point performing time series.

2.2.3 Identifying stationary time series via autocorrelation function and partial autocorrelation function

Time series models such as ARIMA and SARIMAX are expressed in (p, d, q) form, where p, q and d are the parameters represented as auto regression, moving average (MA) and order of difference, respectively. These parameters combined to determine the order of the model and are used to optimize such models. The plots ACF graph and PACF graph are used to determine the initial values of p and q. In ACF, past values are not correlated with present values at all, making the series a sequence of independent, uncorrelated values, and in PACF, the influence of any of the intermediate values are nullified and thus isolates the direct correlation between the values. The ADF and KPSS are some other tests for checking stationarity (Chi, 2018). In the ADF test, the null hypothesis assumes that the series is not stationary; there is no white noise and predictable components exist. If the p-value given by this test is less than 5%, then the null hypothesis is rejected. In the KPSS test, the null hypothesis assumes that the series is stationary. The tests for normality (histogram/Q-Q plot) such as Shapiro test are also helpful for determining strong stationary (white noise).

2.2.4 Modelling predictable components

All the tests mentioned in the previous paragraph identify stationary in time series. The white noise component or strong stationary or unpredictable components can be neglected and no modelling can be performed. However, weak stationary or local components have to be modelled and the models prevalent for this purpose are the auto-regressive (AR) model. The equation of the AR model is represented as:

$$\mathbf{X_t} = \mathbf{a} + \mathbf{X_{t-1}} + \mathbf{Z_t}$$

where X_t is a predictable or local component and Z_t is a noise component. The previous lag values are effecting the present lag values in this model, for instance, in the equation $AR(2) = a + 2\,X_{t-1} + 3X_{t-2}$, previous two actual lag component values are contributing to the current X_t component value. Now, the predictable components or weak stationary may still have noise components and the model to predict the noise component is the MA model. The equation of MA model is represented as: $X_t = a + bZ_{t-1} + Z_t$. As noted in the previous paragraph; any time series model is expressed in (p, d, q) form and p in this model denotes AR(p) and q denotes MA(q) models which means p gives the number of previous actual component values contributing to the present original value and q gives the number of previous noise component values affecting the present original value (Nelson, 1998). Further, if both previous actual components and previous noise components are contributing to the present original value, then the model used is a combination of AR and MA model called ARMA (p, q) and the equation formed for this model is represented as $X_t = a + bX_{t-1} + cZ_{t-1} + Z_t$, where p and q parameters denote the previous lag values and previous lag noise values, respectively, and b and c are coefficients of the model estimated by likelihood methods as similar in the case to regression models. Also mentioned in the previous paragraph, using PACF, the value of p is determined and similarly using ACF, the value of q is determined.

2.2.5 Modelling stationary time series

It is a process wherein global predictions (trend and seasonality) get removed and the original series and the intermediate residue obtained is tested if a given time series is weekly stationary or not and if that time series is identified as weekly stationary then its equation is figured out.

2.3 Auto-regressive integrated moving average model

The extension of ARMA (p, q) to include differencing (d) in the model, ARIMA (p, d, q) is formed, where d is the differencing parameter. It is one of the most efficient algorithms in time series forecasting (Contreras *et al.*, 2003). This model is represented as: $ARIMA\ (p, d, q)$: $X_t = \alpha 1^* X_{t-1} + \alpha 2^\star X_{t-2} + \beta 1^\star Z_{t-1} + \beta 2^\star Z_{t-2} + Z_t$. To be specific in terms of COVID-19 data, X_t is the predicted number of confirmed, recovered and death cases on the particular (tth) day. In the present study, the trend of forthcoming incidences for the next 14 days is estimated from the previous observations after training the model over the period starting 21st January 2020 till 30th June 2020 and testing data ranged from 1st July 2020 to 12th August 2020. The forecasting trends of the future cases are recognized for dates staring 13th August 2020–27th August 2020.

2.4 Seasonal auto-regressive integrated moving averages with exogenous regressor model

It is often called as seasonal ARIMA model and seasonality component is also covered along with p, d and q components as in traditional ARIMA model. The SARIMAX model has four parameters as (p, d, q, s) where s is the seasonality behaviour that can be monthly or yearly as per the data set. In addition, other factors also influence the results in time series like stock price determination also depends on other independent variables other than time factor such as GDP of the country. The X denotes the external regressors and integrated with models such as ARIMA to form models such as SARIMAX and ARIMAX and thus they represent combination of ARIMA and linear regression (Soebiyanto *et al.*, 2010). The independent variables are passed with help of *exog* parameter in the ARIMAX and SARIMAX models (Arunraj *et al.*, 2016).

2.5 Steps for modelling time series model

The high-level steps in the process include: visualize the time series; recognize trend and seasonal component; apply regression to model the trend and seasonality; remove the trend and seasonal component from the series; what remains is the stationary part: a combination of the AR, and white noise;

model this stationary time series; combine the forecast of this model with the trend and seasonal component; find the residual series by subtracting the forecasted value from the actual observed value; and check if the residual series is pure white noise.

The parameters of the model are then estimated to optimize the ARIMA model. The model for forecasting future confirmed COVID-19 cases is represented as:

$$ARIMA\,(p, d, \mathrm{q}) : \mathbf{X_t}$$
$$= \boldsymbol{\alpha 1 X_{t-1}} + \boldsymbol{\alpha 2 X_{t-2}} + \boldsymbol{\beta 1 Z_{t-1}} + \boldsymbol{\beta 2 Z_{t-2}} + \mathbf{Z_t} \quad (1)$$

where:

$$\mathbf{Z_t} = \mathbf{X_t} - \mathbf{X_{t-1}} \quad (2)$$

Here, $\boldsymbol{X_t}$ is the predicted number of confirmed COVID-19 cases at tth day and $\alpha 1$, $\alpha 2$, $\beta 1$ and $\beta 2$ are parameters whereas $\boldsymbol{Z_t}$ is the residual term for tth day (Xue and Lai, 2018).

3. Result and evaluation

The original COVID-19 data are definitely an upward trend data and the plot of different case types – confirmed, recovered and death cases in Figure 1 – clearly shows the trend of an increasing number of those cases. In the mid of March onwards, the confirmed cases are seen to rise tremendously followed by recovered and death cases exhibiting trend and no seasonality.

3.1 Time series specific plots

This section presents the plots of the cases such as trends in the data, decomposition of data, ACF and PACF plots. In Figure 1, confirmed, recovered and death cases have upward trend and as time increases, the cases also increase. They also exhibit stationarity in the data. This plot justifies generation of the time series data as the feature – observation data is the time factor and the daily COVID-19 cases are dependent on this time factor feature.

The data is decomposed to separate trend, seasonality, residual and actual components, and the resultant plots are shown in Figure 2 and the additive model is used for this decomposition. The trend line in the figure is quite linear in nature along with that stationarity is present in the data in high

Figure 1 Upward trend of confirmed, recovered and death cases

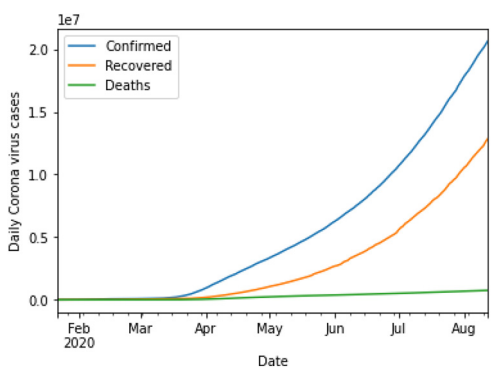
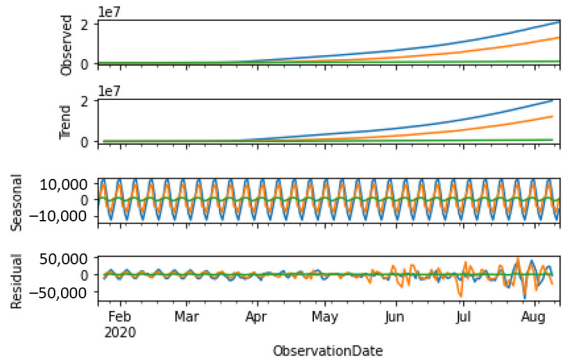


Figure 2 Decomposed into observed, trend, seasonal and residual data



scale. The residual component could be a local component or have only noise or irregularities in the data and needs to be checked for any weak stationary and strong stationary using ACF and PACF plots. To calculate the value of p (AR lag), PACF plot is used, and for the value of q, MA is calculated using ACF plot.

3.2 Partial auto correlation plot

Identification of an AR model is often best done with the PACF. For an AR model, the theoretical PACF "shuts off" past the order of the model. The phrase "shuts off" means that in theory, the partial autocorrelations are equal to 0 beyond that point. Putting in another way, the number of non-zero partial autocorrelations gives the order of the AR model. By the "order of the model", we mean the most extreme lag of x that is used as a predictor (Kaur *et al.*, 2020).

Partial correlation is conditional in nature and useful in detecting order of AR process. The correlation between observations at two time spots given that we consider both observations are correlated to observations at other time spots. PACF is used to determine the terms used in the AR model. Only significant terms will be chosen. The number of terms determines the order of the model.

The PACF plot of confirmed cases is shown in Figure 6, and in the AR component, p is of order 3 because the 4th lag lies within the confidence interval. This value of obtained p is used in ARIMA model for confirmed cases.

The PACF plot of recovered cases is shown in Figure 7, and in the AR component, p is of order 2 because the 3rd lag lies within the confidence interval. This value of obtained p is used in ARIMA model for recovered cases.

The PACF plot of death cases is shown in Figure 8, and in the AR component, p is of order 3 because the 4th lag lies within the confidence interval. This value of obtained p is used in ARIMA model for death cases.

3.3 Auto correlation plot

Identification of an MA model is often best done with the ACF rather than the PACF. For an MA model, the theoretical PACF does not shut off, but instead tapers towards 0 in some manner. A clearer pattern for an MA model is in the ACF. The ACF will have non-zero autocorrelations only at lags involved in the model (Dhiman and Kaur, 2019b).

Auto correlation is the correlation between the observations at the current time spot and the observations at previous time spots. It is the similarity between observations as a function of the time lag between them. ACF is used to determine the terms in the MA model.

The ACF plot of confirmed cases is shown in Figure 3 and clearly correlation values are predicted, thus the data exhibits weak stationary and MA component, and q is of order 1. This obtained value of q is used in ARIMA model for confirmed cases.

Also, ACF of recovered cases is shown in Figure 4 and clearly correlation values are predicted, thus the data exhibits weak stationary in the case of recovered cases and MA component, and q is of order 4. This obtained value of q is used in ARIMA model for recovered cases.

Next, the ACF plot of death cases is shown in Figure 5 and clearly correlation values again are predicted, thus the data exhibits weak stationary in death cases as well and the MA component, q is of order 1. This obtained value of q is used in ARIMA model for death cases.

3.4 Test and train data

The data is split into training for dates starting 21st January 2020 till 30th June 2020 and test data is reserved for starting date from 1st July 2020 to 12th August 2020. Training data helps in identifying and fitting right models and test data is used to validate the same.

In case of time series data, the test data is the most recent part of the series so that the ordering in the data is preserved.

Figure 3 ACF plot of confirmed cases

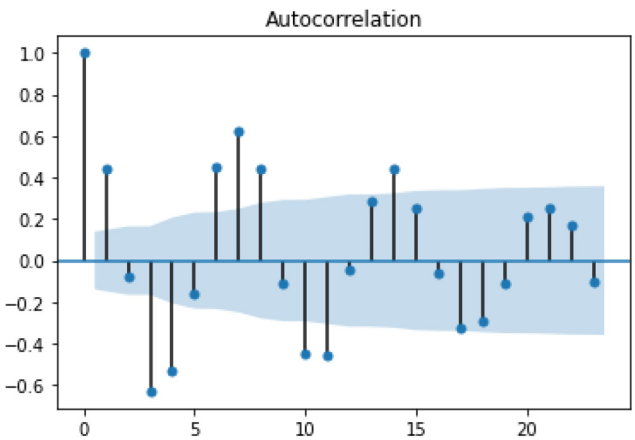


Figure 4 ACF plot of recovered cases

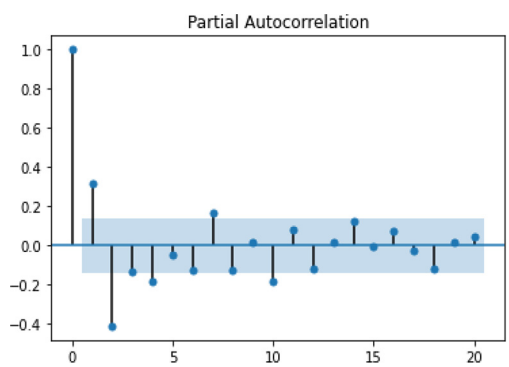


Figure 5 ACF plot of death cases

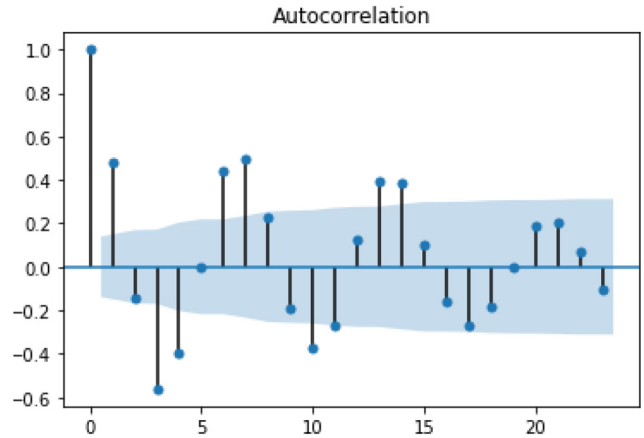


Figure 6 PACF plot of confirmed cases

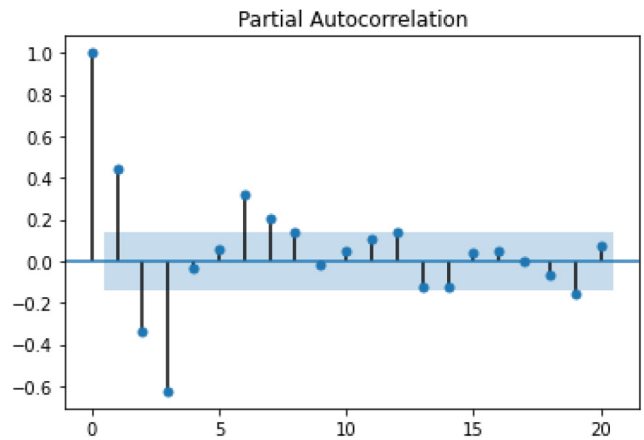


Figure 7 PACF plot of recovered cases

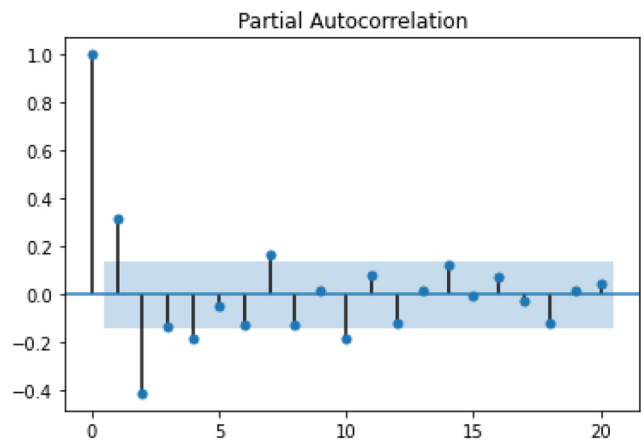
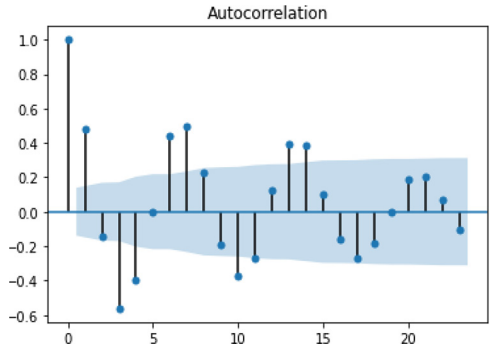


Figure 8 PACF plot of death cases



3.5 Auto-regressive integrated moving average model evaluation

The achieved values of p and q from ACF and PACF plots for all three cases were used to separately apply the ARIMA model with differencing (d) of 1. The values of likelihood and AIC determines the performance of the models. In general, the higher value of log-likelihood and the minimum value of AIC are considered better parameter values for the model evaluation. Once the predictions are made, one measure that is used widely is the MAPE.

The models were optimized by selecting appropriate values of p, d and q that resulted in minimum AIC and maximum log-likelihood values. As shown in Table 2, it is

Table 2 ARIMA model results of confirmed, recovered and death cases

Dep. variable	Model parameters	Log likelihood	AIC	*p*-value	Coefficients
D.Confirmed	ARIMA (2,1,0)	−1733.128	3474.256	0.00	ar.L1.D.Confirmed: 0.53
				0.00	ar.L2.D.Confirmed: 0.45
D.Recovered	ARIMA (2,1,0)	−1807.762	3623.523	0.00	ar.L1.D.Recovered: 0.52
				0.00	ar.L2.D.Recovered: 0.45
D.Deaths	ARIMA (2,1,2)	−1334.813	2681.626	0.00	ar.L1.D.Deaths: 1.23
				0.08	ar.L2.D.Deaths: −0.2
				0.00	ma.L1.D.Deaths: −0.54
				0.05	ma.L2.D.Deaths: −0.17

Figure 9 Standard residual, histogram, normal Q-Q and correlogram plots

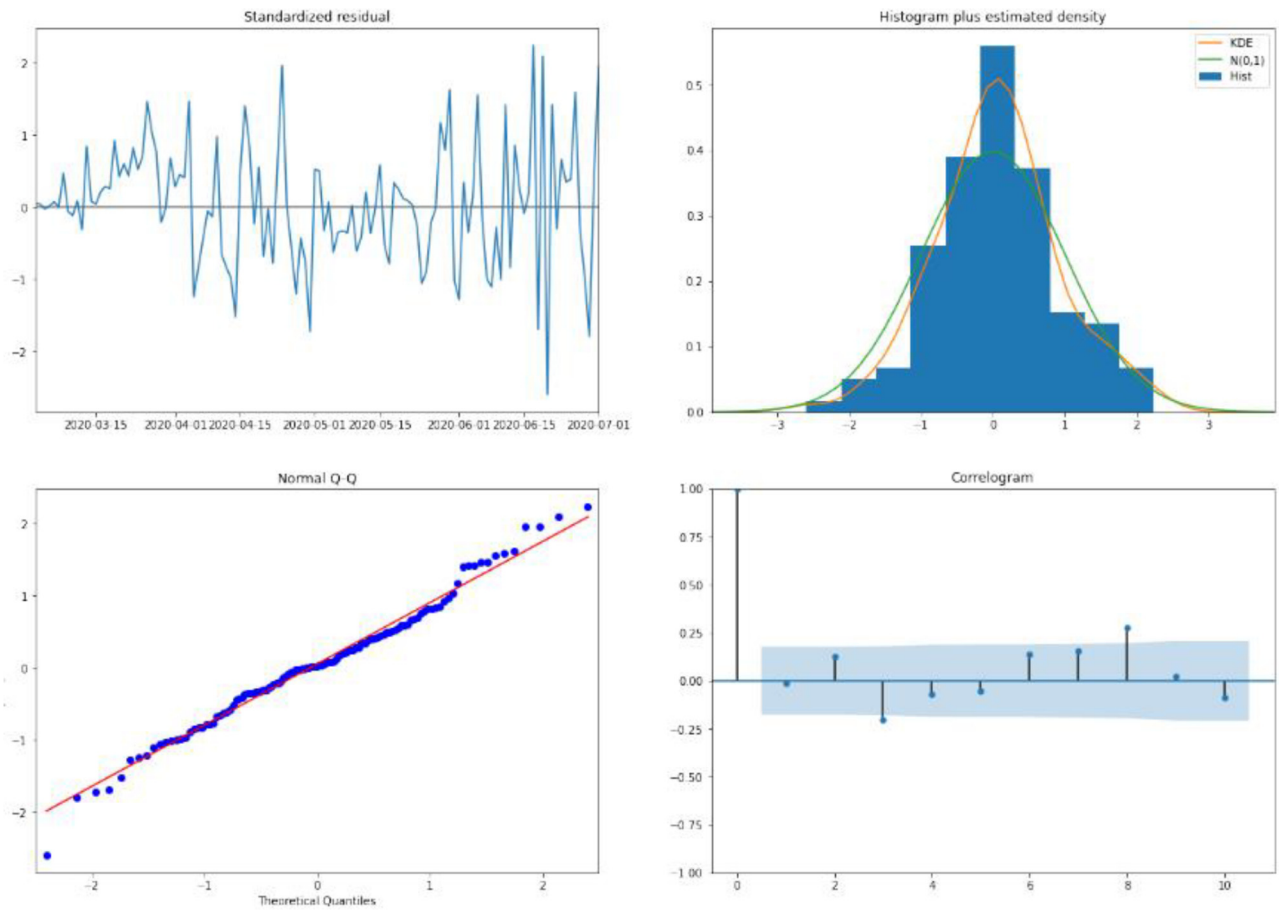
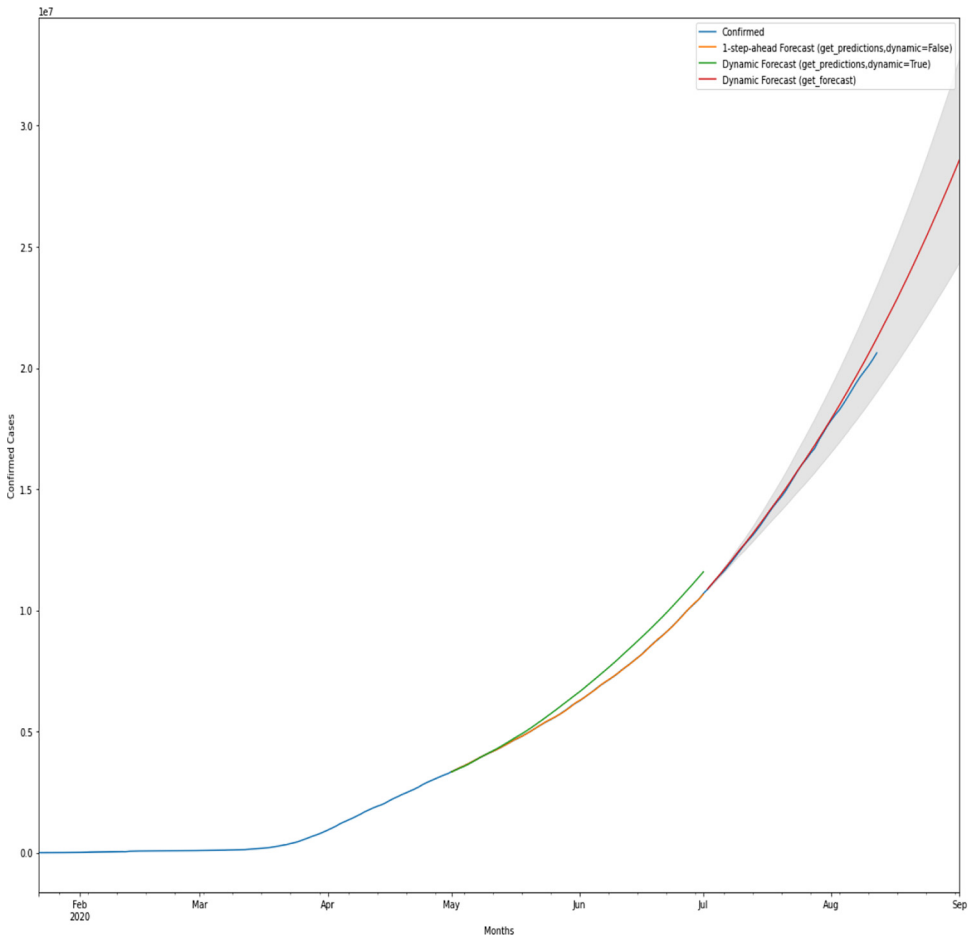


Figure 10 Standard residual, histogram, normal Q-Q and correlogram plots



observed that *p*-values of the AR and MA parameters of all types of cases imply that models have significant parameters. The smallest AIC value for confirmed cases came to be 3474.25 and ARIMA (2, 1, 0) is an optimized model for these cases. Similarly, the smallest AIC value for recovered cases came to be 3623.5225, and ARIMA (2, 1, 0) is an optimized model for recovered cases, and the smallest AIC value for death cases came to be 2681.62 and ARIMA (2, 1, 2) is an optimized model for these cases.

Thus, parameters set for the ARIMA model for each of the confirmed, recovered and death cases are optimized taking into account AIC values. Also, the values of AR and MA coefficients are depicted in the result summary figures above. The *p*-values of variables less than 5% decide the importance of the variable to the model. Hence, ARIMA (2, 1, 0) for confirmed, ARIMA (2, 1, 0) for recovered and ARIMA (2, 1, 2) for death cases are the best-optimized models that are selected based on AIC values.

3.6 Seasonal auto-regressive integrated moving averages with exogenous regressor model evaluation

Similar to the ARIMA model, the SARIMAX model is applied on the data set, the seasonality parameter is kept 12 and the rest of the parameters *p*, *d* and *q* are optimized for each of the cases. The smallest AIC is 2648.74 for confirmed cases and model SARIMAX (1, 1, 2) × (2, 1, 2, 12) i.e. *p*, *d*, *q* values as 2, 1, 2 is the best model. Once the model is fitted, the diagnostic plots of residual as shown in the figure are used to find any correlation in the data. The plots histogram, standardized residual and normal Q-Q plots depict if left out components have pure noise or any predictable components. There exists no noise and no correlation in the plots and hence, no local or predictable component exists and the noise component is the only left out component.

Time series analyses show significant statistic measures for COVID-19 data and Figure 9 checks if the meaningful correlation of confirmed, recovered and death cases exists and displays the residual plots for all types of cases. A deviation of residuals is observed in the plot and indicates that the errors are near to normal and thus assumption of normality is followed backed up with the residual histogram. The constant variance is satisfied by the model as shown in the graph of residuals and correlogram. The plots of residuals display the noncorrelation of the data

The plot below in Figure 10 outlines the prediction from 1st July 2020 to 12th August 2020, and forecasted data till 1st September 2020 for confirmed cases. The dynamic false and dynamic true lines are represented separately with orange and

green lines in the plot and the forecast lies within confidence interval as represented with a red line. The grey shade is the confidence level that determines that the cases will not exceed the confidence level for that particular forecasted time interval. The value of forecasted values is close to the actual values for the period until 12th August 2020 and also the model was able to forecast values for the period of two weeks. This similarity of forecasted data with actual data is clear from the plot and also the plot depicts an increasing trend suggesting a sharp rise in COVID-19 confirmed cases and death cases with time; however, the rate of recovery is higher than the death rate. Thus, a low mortality rate is expected in the upcoming months.

3.7 Mean absolute percentage error

The MAPE for the forecast on test data for confirmed cases is 2.37%, which is considered outstanding in the case of time series models. Similarly, the MAPE for the forecast recovered cases on test data is 4.21% and the MAPE for the forecast death cases on test data is 1.51% .

4. Conclusion

The inferences of time series forecasting models ARIMA and SARIMAX were efficient to produce exact approximate results. The forecasting results indicate that an increasing trend is observed and there is a high rise in COVID-19 cases in many regions and countries that might face one of its worst days unless and until measures are taken to curb the spread of this disease quickly. The pattern of the rise of the spread of the virus in such countries is exactly mimicking some of the countries of early COVID-19 adoption such as Italy and the USA. Further, the obtained numbers of the models are date specific so the most recent trend can be analysed with the help of time series models in real time.

The future scope of the study involves analysis of forecasting COVID-19 cases with models such as long short-term memory and then comparison of results with time series models covered in this paper such as ARIMA and SARIMAX that would help forecast more potential waves of pandemic. Another phase of the ongoing COVID-19 analysis is building COVID-19 classifier using PyTorch and determining key differences with time series approaches.

References

Abdulmajeed, K., Adeleke, M. and Popoola, L. (2020), "Online forecasting of COVID-19 cases in Nigeria using limited data", *Data in Brief*, Vol. 30, p. 105683, doi: 10.1016/j.dib.2020.105683.

Arunraj, N.S., Ahrens, D. and Fernandes, M. (2016), "Application of SARIMAX model to forecast daily sales in food retail industry", *International Journal of Operations Research and Information Systems*, Vol. 7 No. 2, pp. 1-21, doi: 10.4018/ijoris.2016040101.

Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S. and Ciccozzi, M. (2020), "Application of the ARIMA model on the COVID-2019 epidemic dataset", *Data in Brief*, Vol. 29, p. 105340, doi: 10.1016/j.dib.2020.105340.

Chi, W.L. (2018), "Stock price forecasting based on time series analysis", *AIP Conference Proceedings*, 1967(May). doi: 10.1063/1.5039106.

Contreras, J., Espinola, R., Nogales, F.J. and Conejo, A.J. (2003), "ARIMA models to predict next-day electricity prices", *IEEE Transactions on Power Systems*, Vol. 18 No. 3, pp. 1014-1020, doi: 10.1109/TPWRS.2002.804943.

Dhiman, G. (2019), "ESA: a hybrid bio-inspired metaheuristic optimization approach for engineering problems", *Engineering with Computers*, doi: 10.1007/s00366-019-00826-w.

Dhiman, G. (2020), "MOSHEPO: a hybrid multi-objective approach to solve economic load dispatch and micro grid problems", *Applied Intelligence*, Vol. 50 No. 1, pp. 119-137, doi: 10.1007/s10489-019-01522-4.

Dhiman, G. and Kaur, A. (2018), "Spotted hyena optimizer for solving engineering design problems", *Proceedings - 2017 International Conference on Machine Learning and Data Science, MLDS 2017, 2018-January*, pp. 114-119. doi: 10.1109/MLDS.2017.5.

Dhiman, G. and Kaur, A. (2019a), "A hybrid algorithm based on particle swarm and spotted hyena optimizer for global optimization", *Advances in Intelligent Systems and Computing*, Springer Singapore. doi: 10.1007/978-981-13-1592-3_47.

Dhiman, G. and Kaur, A. (2019b), "STOA: a bio-inspired based optimization algorithm for industrial engineering problems", *Engineering Applications of Artificial Intelligence*, Vol. 82 No. June 2018, pp. 148-174, doi: 10.1016/j.engappai.2019.03.021.

Fattah, J., Ezzine, L., Aman, Z., El Moussami, H. and Lachhab, A. (2018), "Forecasting of demand using ARIMA model", *International Journal of Engineering Business Management*, Vol. 10, pp. 1–9, doi: 10.1177/1847979018808673.

Garg, M. and Dhiman, G. (2020), "A novel content-based image retrieval approach for classification using GLCM features and texture fused LBP Variants", *Neural Computing and Applications*, Springer, London, 0123456789. 10.1007/s00521-020-05017-z.

Kaur, S., Awasthi, L.K., Sangal, A.L. and Dhiman, G. (2020), "Tunicate swarm algorithm: a new bio-inspired based metaheuristic paradigm for global optimization", *Engineering Applications of Artificial Intelligence*, Vol. 90 No. November 2019, pp. 103541, doi: 10.1016/j.engappai.2020.103541.

Lalmuanawma, S., Hussain, J. and Chhakchhuak, L. (2020), "Applications of machine learning and artificial intelligence for COVID-19 (SARS-CoV-2) pandemic: a review", *Chaos, Solitons and Fractals*, p. 139, doi: 10.1016/j.chaos.2020.110059.

Li, S. and Li, R. (2017), "Comparison of forecasting energy consumption in Shandong, China using the ARIMA model, GM model, and ARIMA-GM model", *Sustainability (Switzerland)*, No. 7, pp. 9, doi: 10.3390/su9071181.

Malki, Z., Atlam, E.S., Hassanien, A.E., Dagnew, G., Elhosseini, M.A. and Gad, I. (2020), "Association between weather data and COVID-19 pandemic predicting mortality rate: machine learning approaches", *Chaos, Solitons and Fractals*, Vol. 138, p. 110137, doi: 10.1016/j.chaos.2020.110137.

Nabi, K.N. (2020), "Forecasting COVID-19 pandemic: a data-driven analysis", *Chaos, Solitons and Fractals*, Vol. 139, p. 110046, doi: 10.1016/j.chaos.2020.110046.

Nelson, B.K. (1998), "Statistical methodology: v. Time series analysis using autoregressive integrated moving average (ARIMA) models", *Academic Emergency Medicine*, Vol. 5 No. 7, pp. 739-744, doi: 10.1111/j.1553-2712.1998.tb02493.x.

Reddy, J.R., Ganesh, T., Venkateswaran, M. and Reddy, P.R.S. (2017), "Forecasting of monthly mean rainfall in Rayalaseema", *International Journal of Current Research and Review*, Vol. 9 No. 24, pp. 20-27, doi: 10.7324/ijcrr.2017.9244.

Singh, P. and Dhiman, G. (2018), "A hybrid fuzzy time series forecasting model based on granular computing and bio-inspired optimization approaches", *Journal of Computational Science*, Vol. 27, pp. 370-385, doi: 10.1016/j.jocs.2018.05.008.

Soebiyanto, R.P., Adimi, F. and Kiang, R.K. (2010), "Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters", *PLoS One*, Vol. 5 No. 3, pp. 1-10., doi: 10.1371/journal.pone.0009450.

Xue, M. and Lai, C.H. (2018), "From time series analysis to a modified ordinary differential equation", *Journal of Algorithms & Computational Technology*, Vol. 12 No. 2, pp. 85-90, doi: 10.1177/1748301817751480.

Yonar, H. (2020), "Modeling and forecasting for the number of cases of the COVID-19 pandemic with the curve estimation models, the Box-Jenkins and exponential smoothing methods", *Eurasian Journal of Medicine and Oncology*, Vol. 4 No. 2, pp. 160-165, doi: 10.14744/ejmo.2020.28273.

Corresponding author

Kamalpreet Singh Bhangu can be contacted at: researcher.kamalpreet.bhangu@gmail.com