International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 5 ǁ 2022 ǁ PP. 07-14

Auto Regressive Integrated Moving Average Model (ARIMA) For Covid-19 Forecasting

Raghi R Menon[1], Abhimaneu K J[2], Akhil S Babu[3], Anandhu P M[4] and Arjun S[5]
Dept of CSE, Adi Shankara Institute of Engineering and Technology, Kalady,

Abstract—*World Health Organization in March 2020 declared COVID-19 as a world pandemic, and has infected over 40 crore people worldwide with over 60,00,000 deaths by early April 2022. Around the world many researchers and students are trying to search out a model to predict the spread of Covid-19 cases. Many have incorporated various prediction techniques and Auto Regressive Integrated Moving Average model is one such technique which may help to forecast Covid-19 cases. Real-time data has become a crucial aspect for understanding past, present, and future situations. Machine Learning (ML) provides a variety of algorithms like AR, MA, ARIMA, and SARIMA to understand the correlation between the given data, visualize the current scenario, and predict the future forecast which is that the most vital part. With many countries experiencing a resurgence in COVID-19 cases, it is important to forecast disease trends to enable effective planning and implementation of control measures. The prediction of the speed of infection of COVID-19 has become vital for decision and policy makers worldwide. it's important to estimate the rate as accurately as possible using reliable scientific techniques. A prediction of the amount of infections would assist policy makers in an exceedingly specific region to assess their current healthcare capacity and choose which measures must be taken to curb and control the spread of COVID-19. the aim of this study is to analyze the performance of ARIMA model for predicting the number of COVID-19 cases.*

Date of Submission: 12-05-2022 Date of acceptance: 26-05-2022

I. INTRODUCTION

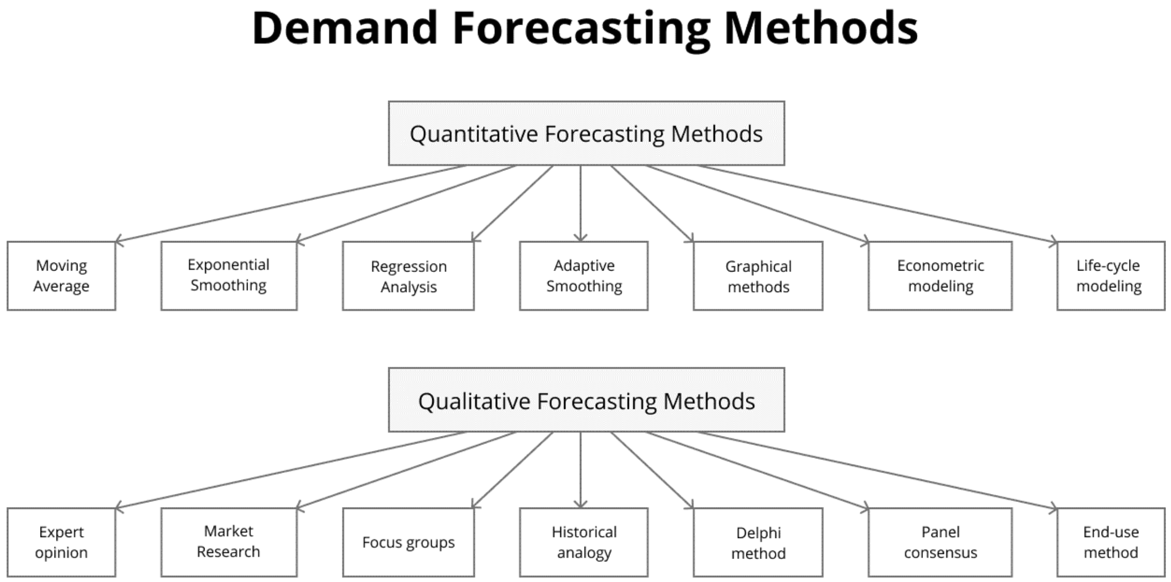
The novel coronavirus, also called COVID-19, was first detected in Wuhan, China on 31 December 2019. WHO was informed of cases of pneumonia of unknown cause in Wuhan City, China and a novel coronavirus was identified as the reason by the Chinese authorities on 7 January 2020 and was temporarily named as “2019-nCoV”. Within a brief period, this virus has spread globally affecting many countries worldwide. As a result, the WHO - World Health Organization declared COVID-19 a Public Health Emergency of International Concern (PHEIC) on 30 January 2020 and subsequently declared it a virulent disease pandemic on 11 March 2020. Since then, many countries globally have put into effect several non-pharmaceutical interventions (NPIs) to restrain the spread of COVID-19 and manage the outbreak. However, despite these measures, there are reports of a resurgence of cases leading to recurrent waves of COVID-19 infection worldwide. [1,2]

Coronaviruses (CoV) are a vast family of viruses that cause illnesses ranging from the common cold to more acute diseases. A novel coronavirus (nCoV) is a new strain that has not been formerly identified in humans. The new virus was later named the “COVID-19 virus”.[1]

The outbreak has resulted in a global epidemic with a large number of confirmed cases of COVID-19 being officially reported in China and abroad, and new cases are being diagnosed daily. In China, during the commencing stage of the outbreak, human-to-human transmission of COVID-19 was noticed and it included family clusters and healthcare settings. This human-to-human transmission led to the accelerated spread of COVID-19, ultimately resulting in intercity spread.[3]

Machine learning (ML) is a category of artificial intelligence (AI) that enables software applications to become more accurate at forecasting outcomes without being explicitly programmed to do so. Machine learning algorithms make use of historical data as input to predict new output values.

Forecasting is a technique of predicting the future based on the results obtained from the previous data. It involves a comprehensive analysis of past and present trends or events to predict future events. It makes use of statistical tools and techniques. Therefore, it is also called Statistical analysis. In other words, we can say that forecasting acts as a planning tool that helps companies to get ready for the uncertainty that can occur in the future. Forecasting begins with management's experience and knowledge sharing. To obtain the most numerous positives from forecasts, organizations must know the different forecasting methodsin detail.



II. RELATED WORKS

Forecasting is one of the crucial tasks performed by data science which has been of prime concern to perform numerous works in any organizations. One such related work is forecasting stock market using ARIMA model.

The important methodology is the analysts who can come up with high-quality results on forecasts which are quite rare because forecasting is a specialized skill that needs enormous expertise in data science. One cannot know where to invest and how much to invest because he doesn't know whether there will be profit or loss in their investment money. This gives out an interesting issue since many people in general eventually invest in any of the stock market sectors. The solution to this problem lets us, know more about stock market options and even helps in making much more precise decisions. Before working with non-stationary data, the Autoregressive Integrated Moving Average (ARIMA) Model transforms it into stationary data. It is one of the most widely used algorithmic models for forecasting linear time series data.[4]

The ARIMA model has been widely utilized in banking and economics since it is recognized to be reliable, efficient, and capable of predicting short-term share market movements.

In general, market datas are time-variant and are nonlinear in pattern, predicting the long run price of stock may be a challenging task. These forms of Predictions will provide the users with good information about the present running status of stock price. With some historical data, you'll be able to use forecasting tools to predict into the future a particular metric While there are several methods of completing a statistic model, you'll follow these general steps in a spreadsheet to estimate outcomes using information gleaned from recent analytical data.[5]

Future work can be considered as using standard time series analysis to achieve forecast prices of stock markets. It is an area of persistent research as investors and researchers strive to work with the market with the ultimate reason of acquiring higher returns. It is so unlikely that the new theoretical results will come out with the above-projected works. Once we've set up the forecasting model, we will then carry on to explicate it to formulate your finest estimation of the future.[6]

III. EXISTING METHODS

A. Auto Regressive Model

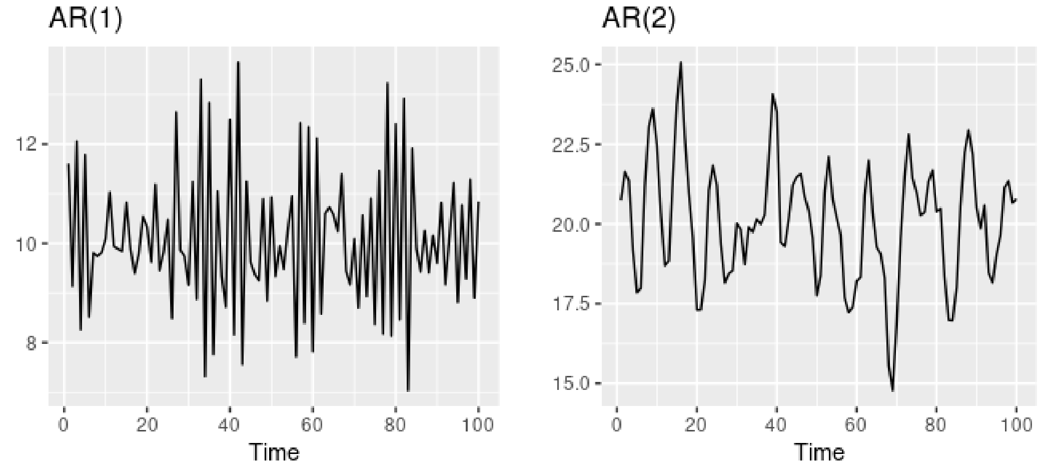
An AR model is one in which Yt depends only on its own past values
Yt-1,Yt-2,Yt-3, etc.

Thus,
Yt = f(Yt-1,Yt-2,Yt-3,….., Et)

A common representation of an autoregressive model where it depends on p of its past values called as AR(p) model is represented below:

$Yt = \beta 0 + \beta_1 \;\; Yt_{-1} \;\; + \beta_2 \;\; Y□_{-2} \;\; + \beta_3 \;\; Y□_{-3} \;\; + \ldots\ldots\ldots\ldots + \beta p \; Y□\text{-}p + Et$

Autoregressive models are remarkably flexible at handling a wide range of different time series patterns. The two series in the below figure show series from an AR(1) model and an AR(2) model. Changing the parameters results in different time series patterns. The variance of the error term Et will only change the scale of the series, not the patterns.



Two examples of data from autoregressive models with different parameters.

Left: AR(1) with $yt=18-0.8yt-1+\varepsilon t$.
Right: AR(2) with $yt=8+1.3yt-1-0.7yt-2+\varepsilon t$.

In both cases, εt is normally distributed white noise with mean zero and variance one.

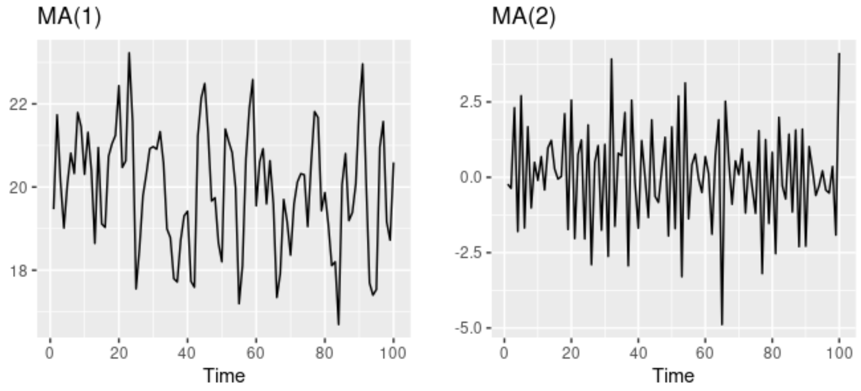
B. Moving Average Model

A moving average model is one when Y, depends only on the random error terms which follow a white noise process i.e.
Yt = f(Et, Et-1, Et-2, Et-3) .....)

A common representation of a moving average model where it depends on q of its past values is called MA(q) model and is represented below:

$Yt = \alpha_1 \;\; E□_{-1} \;\; + \alpha_2 \;\; E□_{-2} \;\; + \alpha_3 \;\; E□_{-3} \;\; + \ldots\ldots\ldots\ldots + \alpha q \; E□\text{-}q$

The below figure shows some data from an MA(1) model and an MA(2) model. Changing the parameters results in different time series patterns. As with autoregressive models, the variance of the error term ε t will only change the scale of the series, not the patterns.



Two examples of data from moving average models with different parameters.

Left: MA(1) with $y\,t = 20 + \varepsilon\,t + 0.8\,\varepsilon\,t - 1$.
Right: MA(2) with $y\,t = \varepsilon\,t - \varepsilon\,t - 1 + 0.8\,\varepsilon\,t - 2$.

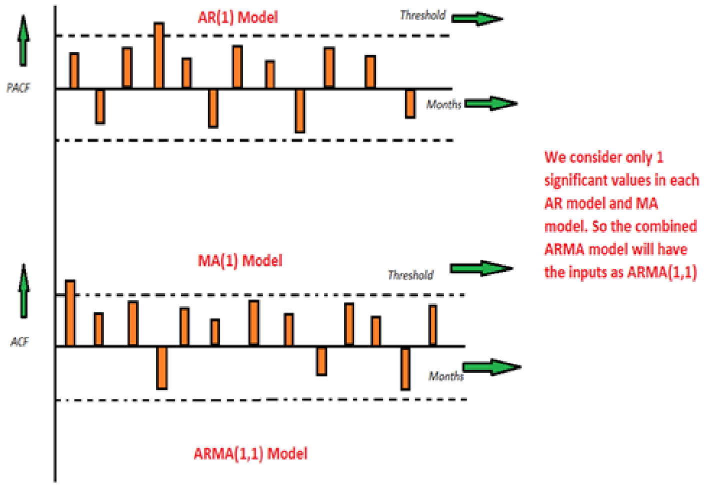
In both cases, $\varepsilon\,t$ is normally distributed white noise with mean zero and variance one.

C. Auto Regressive - Moving Average Model

There are situations where the time-series may be represented as a mix of both AR and MA models referred as ARMA (p,q).

The general form of such a time-series model, which depends on p of its own past values and q past values of white noise disturbances, takes the form:

$$Yt = \beta0 + \beta_1 \;\; Yt\text{-}_1 \; + \beta_2 \;\; Y\square\text{-}_2 \; + \beta_3 \;\; Y\square\text{-}_3 \; + \ldots\ldots\ldots\ldots + \beta p \; Y\square\text{-}p + Et$$
$$+ \alpha_1 \;\; E\square\text{-}_1 \; + \alpha_2 \;\; E\square\text{-}_2 \; + \alpha_3 \;\; E\square\text{-}_3 \; + \ldots\ldots\ldots\ldots + \alpha q \; E\square\text{-}q$$


AR(1) Model
Threshold
PACF
Months
We consider only 1 significant values in each AR model and MA model. So the combined ARMA model will have the inputs as ARMA(1,1)
MA(1) Model
Threshold
ACF
Months
ARMA(1,1) Model

IV. METHODOLOGY

In this study, the ARIMA time series analysis model was applied for the prediction. The Autoregressive Integrated Moving Average (ARIMA) model predicts an assumed variable's future value with several past observations and random errors with a linear function. The ARIMA processes are a combination of some stochastic processes used to analyze time series. To perform the time series in the ARIMA model, firstly the raw data should be ready and ran some tests with the data, such as the Dickey-Fuller test, to determine the trend and find the rolling statistics of the dataset. In this section, our proposed ARIMA model descriptions are presented with some general statistical methodology.[7]

A. STRATEGY

The general scheme is as follows:
Step 1: A class of models was formulated assuming certain hypotheses.
Step 2: The model parameters were estimated.
Step 3: Check the hypotheses of the model validation. If it satisfies the conditions, go to step 4; otherwise, go to step 1 to refine the model.
Step 4: The model was ready for forecasting.

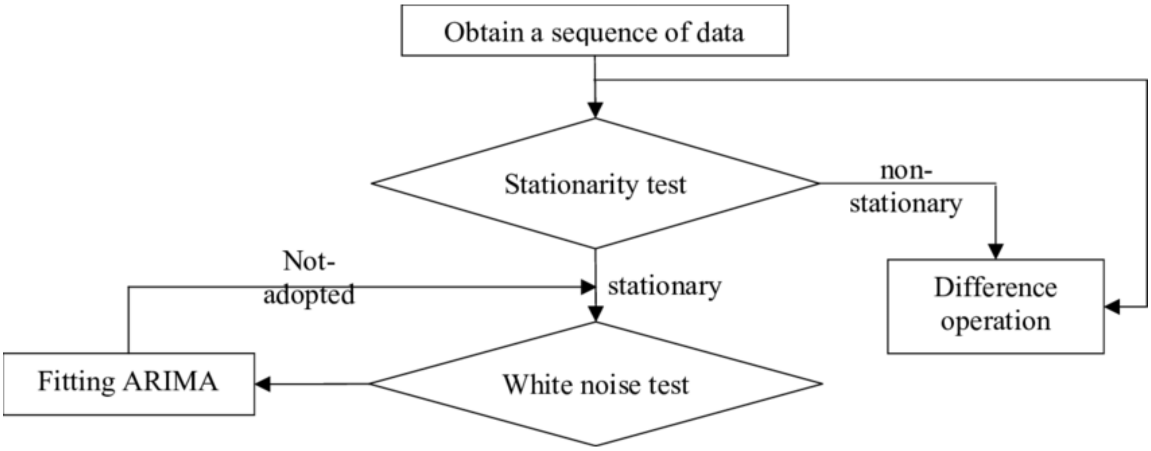
These steps were described briefly below.
Step 1: In this step, a general ARIMA formulation was selected to model the confirmed case data. This selection was carried out by careful inspection of the main characteristics of the daily confirmed case series. In this time series analysis, high frequency, nonconstant mean and variance, and multiple seasonality (similar to daily, weekly, and periodicity) were also considered.

Step 2: After formulating the functions of the model, the parameters of these functions must be estimated. Good estimators of the parameters were computed by assuming the data are observations of a stationary time series done by step 1 and by maximizing the likelihood concerning the parameters.

Step 3: In step 3, a diagnosis check was used to validate the model assumptions of step 1. This calculation checked if the hypotheses made on the residuals were true. Residuals should satisfy with zero mean, constant variance, uncorrelated process, and normal distribution.

Step 4: After that, the model was ready for the prediction. Now, go to step 2 and drive down to predict future values of confirmed cases. For this requirement, many difficulties arisen since the forecast lead time was more extensive.

B. FLOW DIAGRAM



Steps

– If it White Noise
Don’t do forecasting as we cannot predict.
– If it is stationary
Start with modelling.
– If it is not stationary
Make it stationary by differentiation.
And do ARIMA Modelling.

V. DESIGN

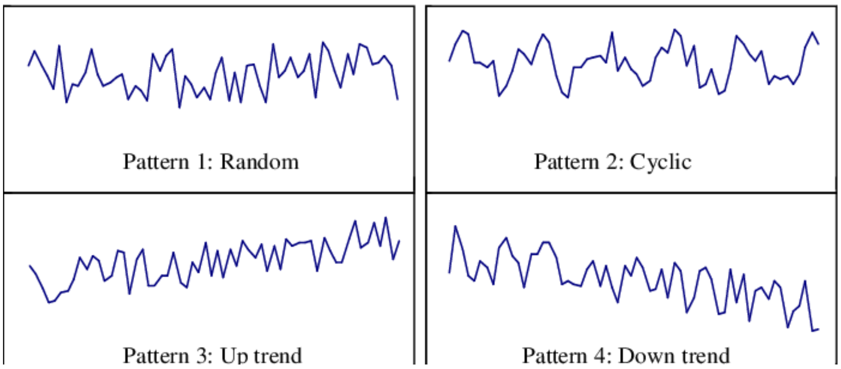
Time Series Analysis was done for the prediction, where some components of the analysis needed to consider for our model prediction.

A. Components of Time Series

Most of the time series have trend, seasonality, and irregularity associated with them. Moreover, some of these do have a cyclic order also. However, it is not compulsory to have a pattern in the time series model. So, let us discuss each one of them in detail. These components help find suitable forecasting methods for the short-term analysis.

B. Trend

The Trend is a movement of higher values and lower values over a long time. So, when the values are directing upward in a time series is known as an upward trend. Also, the trend exhibits lower patterns to the downward is known as downward trends. Moreover, if it does not show any trend, it will be called a horizontal or stationary trend.



C. Seasonality

Seasonality generally has upward or downward swings. Nevertheless, it is quite a different form of a trend that shows a repeated pattern for a fixed period.

D. Irregularity

Irregularity is also known as noise. It is the iritic nature of data and is also called residual. So, it happens for only a short duration and not repeated like the other.

E. Cyclic

Cyclic is the repeating up and down movement of a set of data in a graph. It means it can happen over more than a year and have no fixed pattern. They can repeat in one year, two years, or half of a year, and it is harder to predict. For predicting a time series analysis, it is crucial to consider the stationarity of the dataset. Time series requires data to be stationary, and it is a must for the analysis. The stationary has three components: the constant mean, constant variance, and autocovariance that does not depend on time. To check whether the dataset is stationary or not, two popular tests exist in python. The one is the rolling statistics and the Augmented Dickey-Fuller test (ADCF).

VI. ARIMA model - Autoregressive Integrated Moving Average Forecasting Model

ARIMA is one of the most effective models for time series data, which is a mixture of two models. The AR Model stands for the Auto Regressive part, and also the MA model stands for the Moving Average part. These two models are bound together by the integration part, indicated by" I" within the ARIMA Model. The ARIMA model has these parameters: "P" is the autoregressive lags, "Q" is the moving average, and "d" is the order of differentiation.[7] For predicting our dataset within the ARIMA model, firstly found the rolling statistics of the dataset. Then the Dickey-Fuller test was carried out and estimated the trends of the dataset. Again, the Dickey-Fuller test was executed linked to the trends. Then the Auto Correlation graph (ACF) and Partial Auto Correlation graph (PACF) were performed to see the worth of Q and P. After that, the AR model and also the MA model were performed, and they predicted the long run. Finally, we converted the dataset into the growing sum and plotted the graph of the output.[7]

A. Stationarity of a Time Series

A series is said to be “strictly stationary” if the
mean and variance at a period in a dataset
is same at every other period.

- Mean = same
- Variance = same

B. Differencing

A series which is non-stationary can be made stationary after differencing.
– The process of subtracting one observation from another.
X =[5, 4, 6, 7, 9, 12]

After 1 lag differencing,
X’ = [1, -2, -1, -2, -3]

A series which is stationary after being differentiated once is said to be integrated of order 1 and is denoted by I(1).

Therefore, a series, which is stationary without differencing, is said to be i(0).

VII. FORECASTING

A. Data Collection

The daily basis data of the confirmed cases worldwide can be obtained from the World Health Organization (WHO). Also, Covid-19 dataset from the GitHub data repository by Johns Hopkins University for Systems Science and Engineering. It contains Covid-19 total confirmed cases, recoveries and total deaths. The data contained over 198 countries, Covid-19 cases across the world.

B. Predicting Procedure

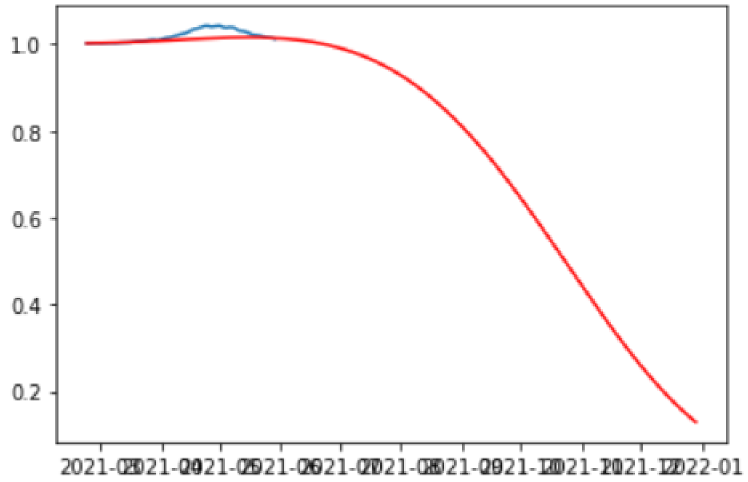
Jupyter Notebook and Google Colabwas used for the python coding. The dataset was plotted in graph. Then the fitting values from our ARIMA model analysis was calculated. Then the model converted the fitted values into the series format. So, the model determined the dates over the predictions and transferred our predicted data into our original format. For prediction, there was a function called predict in python, which helped us predict the number of confirmed COVID-19 cases.

VIII. RESULT

Pandemics are unpredictable by nature. This imposes a limitation on the use of ARIMA which belongs to a class of linear models. ARIMA cannot get the non-linear patterns hidden within time-series data. One major threat to the validity of our results is the lack of consistency in the number of COVID-19 tests conducted daily by the health authorities. This number was not made publicly available at the beginning.

If the number of daily tests was consistent, the actual number of confirmed cases could be significantly higher than what was announced. This, in turn, would affect the validity of our results obtained from the observed time period data set. Another factor that may have impacted the accuracy when comparing the predictions with the actual reported cases, is that initially, the COVID-19 virus was misdiagnosed and there were many uncertainties about the different symptoms associated with it. This could have caused an unknown number of false positives or false negatives in diagnosed cases.

The ARIMA model showed excellent accuracy in the time series analysis prediction which previous models could not achieve. The main advantage of ARIMA forecasting is that it works on most time series. Relatively robust especially when generating short-run forecasts. Better for relatively short series when the number of observations is not sufficient to apply more flexible methods. This avoids a problem that occurs sometimes with multivariate models.



Start='2021-02-22', End='2021-12-29'

The use of ARIMA for forecasting time series is essential with uncertainty as it does not assume knowledge of any underlying model or relationships as in some other methods. ARIMA essentially relies on past values of the series as well as previous error terms for forecasting. However, ARIMA models are relatively more robust and efficient than more complex structural models in relation to short-run forecasting.

IX. FUTURE WORK

ARIMA or Autoregressive Integrated Moving Average is a forecasting technique for univariate time series data. It braces both autoregressive and moving average elements. The integrated element refers to differencing allowing the technique to support time-series data with a trend seasonal.

The issue with ARIMA is that it does not support seasonal data which is time-series data with a repeating cycle that is seasonal. ARIMA expects data that is not seasonal or has the seasonal component removed that is seasonally adjusted via a technique such as seasonal differencing.

An alternate substitute is to use SARIMA or Seasonal Autoregressive Integrated Moving Average which is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component. SARIMA adds three new hyperparameters to specify the autoregression (AR), differencing (I) and moving average (MA) for the seasonal elements of the time series, as well as an additional parameter element for the period of the seasonality which will improve the results. [8]

REFERENCES

[1]. World Health Organization (WHO) (2020). Novel coronavirus China. Disease outbreak news. Geneva: WHO.

[2]. Forecasting COVID-19 Case Trends Using SARIMA Models during the Third Wave of COVID-19 in Malaysia. International Journal of Environmental Research and Public Health, 28 Jan 2022.

[3]. Outbreak of COVID-19 in a family, Wenzhou, China. Published online 2020 May 20.

[4]. Stock market forecasting using Time Series analysis With ARIMA model - Hardik KumarDhaduk July 18, 2021

[5]. The 4 Types of Forecasting Models with Examples-By Indeed Editorial Team: July 23, 2021

[6]. Stock Market Prediction for Time-series Forecasting using Prophet upon ARIMAJuly 2020 2020 7th International Conference on Smart Structures and Systems (ICSSS).

[7]. Time series analysis and predicting COVID-19 affected patients by ARIMA model using machine learning. Published online 2021 Dec 14.

[8]. A Gentle Introduction to SARIMA for Time Series Forecasting in Pythonby Jason Brownlee on August 17, 2018 in Time Series