International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Prediction & Analysis of Covid-19 Cases Using Autoregressive Integrated Moving Average (Arima)

Vedpal[1], Umesh Kumar[2]*, Harish Kumar[3], Akanshika Gandhi[4]

Submitted: 25/04/2023 **Revised**: 27/06/2023 **Accepted**: 07/07/2023

Abstract: For the past three years world is facing the pandemic COVID-19. For effective handling of COVID-19, accurate decisions should be taken. The accuracy of making any decision is totally dependent on the relevant data and the information. To determine the present situation of the COVID- 19 the collected data from the various states and Union Territories are processed and analyzed. The Collected data from the various resources also helps to forecast the expected confirmed cases in the future. In this paper, the prediction of positive cases of Covid-19 was carried out using the ARIMA time series model. The predictions made in this study were limited to the addition of positive cases of Covid-19 in India. The analysis and visualization of the data were performed using Python. The obtained results of the predictive analysis showed a trend of daily positive cases that tend to rise in the next 98 days from the data used.

Keywords: *Covid-19; ARIMA; time series; prediction*

1. Introduction

Throughout history, humanity has faced different pandemics which have decimated its population, currently, HIV is an active pandemic with the highest transmission and by 2021, 28.7 million people had access due to antiretroviral treatment [1], another pandemic that deeply affected the world was the Spanish flu in the 19th century, leaving hundreds of dead in its wake [2]. Humanity is currently facing a new pandemic that, since December 2019, has been advancing, claiming the lives of around 6,434,436 people worldwide [3] till 6 August 2022 called SARS-CoV-2 [4] having a very large global impact due to its rapid contagion and the measures implemented for its containment.

Statistics in the clinical context is a tool to comprehend the behavior of SARS-CoV-2 and use the SIR epidemiological model as the basis [5]. Independently studying its three variables, which group patients into three groups (infected, recovered, and deceased), the development of two Naïve Forecaster [6] and ARIMA time series predictive models will be implemented in parallel [7]. Finally, a board will be created to present the status of the vaccination plan and its impact on the result of the prediction.

At present, there are Machine Learning practices that permit predicting the behavior of the pandemic, as well as estimating its scope and thus finding an eradication route. In addition to the implementation of epidemiological mathematical models, this is a very helpful tool in the study of them [8]. These models consider the characteristics of spread, contamination, and immunization. The experimental results found during the infection studies help generate simulations and estimates of how the epidemic will behave so that measures can be taken to stop its progress.

In India, SARSCOV-2, till August 6, 2022, had left a result on the day of the preparation of this document 526,649 deaths, 43,465,552 recovered patients, and 134,793 active patients [9]. Being the second nation on the globe after the USA to experience the effects of the pandemic [9]. As per the Chinese Center for Disease Control (China CDC), although the virus spreads rapidly, 81% of those infected have no symptoms or mild symptoms of the disease, such as an acute respiratory infection calculated with fever, cough, secretion, nasal, general malaise; while 20% were hospitalized, 5% were seriously injured, and 2% required mechanical ventilation. The mortality rate reported by the CDC is 2.3%, and of those who die from the disease, the majority are of the age of 60 years or elder and/or have pre-existing medical conditions like hypertension, heart disease, diabetes, and cancer. [10]. The World Health Organization (WHO) stated that the SARSCOV-2 pandemic [11].

Several models regarding COVID-19 have been proposed by various researchers. Researchers have [12] forecasted the trends of COVID-19 in Saudi Arabia using four models, namely the Autoregressive Model, Moving Average, a combination of both (ARMA), and integrated ARMA (ARIMA). Researchers [13] examined the prediction of confirmed cases of COVID-19 in Karnataka province; in this study, the models used were the Multi-Layer Perceptron (MLP) and Artificial Neural Network (ANN) models. Using Johns Hopkins data, Researcher [14] modeled Covid-19 using the classic Susceptible, Infectious, or Recovered (SIR) and the development of the SIR with the addition of Infectious and Unconfirmed Recovered compartments to explore infection mortality rates and recovery rates in Covid-19 cases.

1,2,3,4 J C Bose University of Science & Technology, YMCA, Faridabad, Haryana, India
** Corresponding Author Email: umesh554@gmail.com*

The prediction of confirmed cases of Covid-19 was carried out by using the ARIMA time series model. The predictions made were limited to the addition of confirmed cases of Covid-19 in India. The training data used is the addition of positive cases on March 20, 2020, to May 5, 2021, while the validation data used is 98 days. The training data is used to create a state space representation based on the ARIMA model. While the validation data is used to calculate the accuracy of the positive case predictions generated.

The following are the objectives of this paper,

1. Obtain an appropriate state space model for data on the number of positive Covid-19 cases in India based on the ARIMA model.
2. Obtain predictions for the number of positive Covid-19 cases in India for the next seven days based on the ARIMA model.

2. Covid-19: Introduction and Background

In this section, a rewritten brief overview of the introduction and the background Technologies used to analyze the trends of COVID-19 has been presented

2.1 Covid-19 Overview

The World is facing a highly complex respiratory disease caused by a virus called COVID-19, which has been declared an epidemic, being spread from person to person, infections with this COVID-19 in humans in general cause respiratory signs, such as a runny nose, sore throat, cough and fever, due to being in direct contact with secretions or respiratory droplets that contain viruses. Given this, preventive security procedures to prevent the spread of COVID-19 will help reduce the level of infections worldwide [1]. In early December 2019, the primary cases of pneumonia from unidentified sources were recognized in Wuhan, China. The pathogen was recognized as a new RNA beta coronavirus that has now been named coronavirus 2 (SARS-CoV2), due to its similarity to SARS-CoV [2]. Given this, the WHO declared on March 11, 2020, the coronavirus disease 2019 (COVID-19) as a pandemic due to the shocking levels of transmission, sternness, and dithering [14-18].

According to the WHO, after an arduous investigation, it obtained quite strong evidence that the outbreak originated from exposures in a fish and seafood market in the city of Wuhan, where most of the infected cases were workers from the same wholesale fish market. and seafood, or handlers or regular visitors to the market, in this situation the market was closed on January 1, 2020, for environmental cleaning and disinfection, it should be noted that the sellers of the said market did not have the measures of prevention and protection for the sale of food supplies, in addition, it was evidenced that there was overcrowding and inadequate cleanliness among the vendors in the said market [14-18].

This COVID-19 disease has rapid transmission, which means big problems since even countries with advanced health systems have been overwhelmed by a large number of cases. This problem has 2 conditioned an enormous challenge for all national health systems, particularly in countries with medium and low resources [14-18].

Faced with this problem, the spread of the virus must be prevented and delayed so that large sectors of the population are not infected at the same time. For this reason, the WHO insisted on strengthening effective case observation, early recognition, isolation, and management of confirmed cases, contact tracing, and inhibition of the spread of the new virus [14-18].

2.2 Time Series

The time series contains the collected, recorded, or observed data over consecutive increments of time. For the analysis of time series of data, an immediate tendency is to try to explain or account for the behavior of the series [20-23].

2.2.1 Components of a Time Series

The Trend

It is a component of a time series that reflects its long-term evolution. It can be stationary or constant in nature (it is represented by a straight line parallel to the abscissa axis), linear, parabolic, or exponential in nature.

Cyclical Variations

It is an element of the series that contains periodic oscillations of amplitude for more than one year. These periodic oscillations are not regular and appear in economic phenomena when they occur alternately, periods of prosperity or depression. Seasonal variations are a component of the series that includes oscillations that occur around the trend, repeatedly and in periods equal to or less than one year [20-23].

2.2.2 Time Series Analysis

The analysis of the time series is dedicated to the study of the series. In general, the data of these series are independent, but they are correlated; it can be said that there is a relationship between contiguous observations. Time is usually the leading measurement of the data. They serve to establish the effectiveness of measures that affect population groups, considering the natural variations that may exist over time. They are very common in the evaluation of laws in the population. They allow a partial view of the cause-effect relationship, but cannot extrapolate population findings to specific individuals. Time series analysis consists of a description, generally mathematical, of the movements and components present [20-23].

According to [20] there are several objectives for which you want to analyze a time series:

Description: The first step in the analysis is to graph the data and find out simple expressive procedures of the properties of the series.

Explanation: When observations are taken on two or more variables, The variation in one series is used to explain the variation in the other series.

Prediction: Given a time series, try to forecast the future values of the series. This is the most frequent goal in time series analysis.

Control: If a time series is generated by quality measurements of a procedure, the objective of the analysis may be the control of the procedure.

2.2.3 Descriptive Classification of Time Series Stationary Series

A series is stationary when it is unchanging over time. This is imitated graphically showing the values of the series tend to oscillate around a persistent mean and the variability with respect to that mean also remains constant.

Not stationary - These are series in which the trend variability variation over time. Changes in the mean determine a long-term tendency to rise or decline. [20-23].

2.2.4 Model

A model is expressed, in symbols, in mathematical form. To build a good model it is necessary to have the observed data set. Experience, intuition, imagination, simplicity, and the ability to select the smallest subset of variables are also important. The first step is to establish the problem in a clear and logical way, delimiting its borders, then comes the collection and purification of data, the design of the experiment; contrast tests; model verification, and hypothesis validation. A model must be a good approximation of the real system, have important aspects, and be easy to understand and operate.

2.2.5 White Noise

White noise is a case of stochastic procedures, where the values are autonomous and identically distributed over time with zero mean and equivalent variance, denoted by $\boldsymbol{\varepsilon t}$.

$$\varepsilon t \sim N(0, \sigma^2), Cov\left(\varepsilon_{t_i}, \varepsilon_{t_j}\right) = 0 \; \forall \; t_i \neq t_j \; (eq. 2.1)$$

In addition to the definitions of time series and stochastic process, special emphasis is placed on time series as a recognition of a given stochastic method [19]. Although the objective of this Thesis focuses on the development of non-linear forecasting models, the underlying mathematical model of the time series [20] is fundamental when it comes to preprocessing the historical data, since, when subjected by the models proposed, characteristics are extracted that will be used in the models to make forecasts.

Currently, time series analysis is essential in many fields of science, such as engineering [21] and economics [22], that is, investigating how a variable of interest has evolved so far can be very useful in order to predict future behavior. However, if the time series has an analogous behavior is of great interest, particularly for decision-making in the area of precision agriculture for modeling. The time series is a set of observations on values at different moments in time. The data can behave in different ways over time [23], that is, it presents a trend, a cycle; not having a defined or random shape, seasonal variations (annual, monthly, etc.).

A time series can be constituted only by deterministic events, stochastic or a combination of both [20]. It is known that many time series show nonlinear dynamic behaviors, the density of which makes it intolerable to create a mathematical model. The model formulation problem is aggravated by the existence of outliers and structural changes.[20]

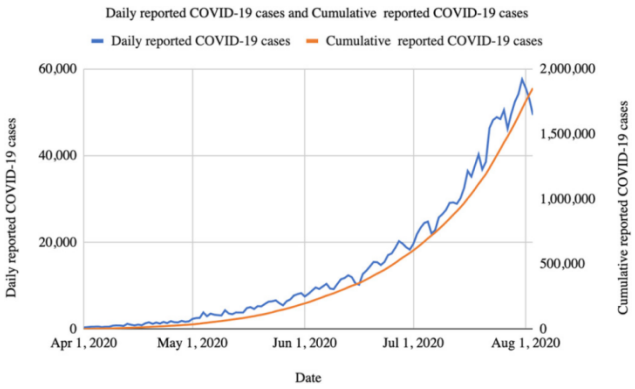

Fig 1. Time Series Elaboration

The traditional methods for the analysis of time series [20, 21, 23] are done through their decomposition into several parts. It is said that a time series can be divided into three components that are not directly observable, of which only estimates can be obtained. These three components are: Trend: represents the predominant behavior of the series. This can be informally defined as the change in the mean over an extended period.

Seasonality: it is a periodic movement that occurs within a short and known period. This component is determined, for example, by climatic factors.

Random: they are erratic movements that do not follow a specific pattern and are due to various causes. This component is practically unpredictable. These behaviors represent all types of movements in a time series that are neither trend nor seasonal variations nor cyclical fluctuations.

A significant amount of sample data is necessary for the analysis to be representative in the general population to which the series belongs. On the other hand, if known sufficient information to substantiate the causes of the behavior of a series, the analysis of they become accessory,

but not indispensable [20]. In general, by analyzing series time, the possibility of making a prediction is pursued.

2.2.6 Time Series Stationary

Stationary is essential in forecasting time series. Stationary time series are time series where the mean, variance, and covariance are constant over time. A time series is stationary if [23]

1. The expected value / mean time series is constant and limited in all periods:

$$E(y_t = \mu \; then \; |\mu| < \infty, t = 1,2 \ldots., T \; (eq. 2.2)$$

2. Variance and Covariance of a time series with time series itself is constant and limited in all periods:

$$Con(y_t, y_{t-s}) = \lambda, |\lambda| < \infty, t = 1,2, \ldots., T; s = 0 \mp 1, \mp, 2, \ldots. \mp T \; (eq. 2.3)$$

Whereas differentiation is evaluated as

$$y'_t = y_t - y_{t-1} = y_t - By_t = (1 - B)y_t \; (eq. 2.4)$$

In general, differentiation with level d is given by:

$$(1 - B)^{\,d} \, y_t \; (eq. 2.5)$$

where, B is the backshift operator:

$$By_t = y_{t-1} \; (eq. 2.6)$$

2.3 Autoregressive Integrated Moving Average (Arima)

ARIMA is a time series model that utilizes differentiation in ARMA model so that it can be used for modeling non-stationary time series. ARIMA uses time series differentiation with level d to generate constant time series. The ARIMA model (p, d, q) is given as [25-28]:

$$y'_t = c + \Phi_1 y'_{t-1} + \cdots + \Phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \; (eq. 2.7)$$

The three ARIMA components appear clearer when written using the backshift operator, $By_t = y_{t-1}$, as follows:

$$(1 - \Phi B - \cdots - \Phi_p B^p y_t = c + \left(1 + \theta_1 B + \cdots + \theta_q B^q\right)\varepsilon_t \quad (\text{eq. } 2.8)$$

P = order autoregressive

D = level differentiation

Q = order moving average

The various time series processes that stationary and non-stationary can be created by the ARIMA. The ARIMA model combines AR, MA and differentiation elements, Table 1 gives equivalence of time series processes with the ARIMA model (p, d, q).

Table 1. Time series processes in the form of ARIMA

Process	**ARIMA (p, d, q)**
White Noise	ARIMA (0,0,0)
Auto Regression	ARIMA (p, 0,0)
Moving Average	ARIMA (0,0, q)

3. Literature Review

In this section a brief rewritten overview of the work proposed by the various researchers in the direction of the forecasting and analyzing the trends of the COVID-19 epidemic.

Jyoti Arora et al. proposed model to predict the count of demises, recovered cases, daily new confirmed cases using the support vector regression framework. The data collected by the model was collected from March 1, 2020 to April 30, 2020. It was developed in order to investigate the data and predict values of various cases up to June 30, 2020. The model performed well in predicting the deaths, healthier cases, and the number of new cases. However, it did not perform well in forecasting the day-to-day new cases.

Fanelli et al. [30] examined the coronavirus infection pandemic that arose in China, Italy, and France in 2019 across the time period from 22 January to 15 March 2020. A basic mean-field framework may be utilized to acquire a measurable representation of the COVID-19 spreading, including the height, and timing of the peak of newly infected persons, according to an initial examination o day-lag maps that are given by the various institutions in the epidemic. The model predicts that the peak will occur in Italy around 21 March 2020, with a high of roughly 26,000 sick people and a total of about 18,000 deaths by the end of the outbreaks.

Devaraj et al. introduced a [32], time series forecasting of COVID-19 outcomes. They applied the PROPHET ARIMA, SLSTM, and LSTM models to evaluation the upcoming forecast of new, demise, and good health cases for the time period from 22nd Jan 2020 to 11th Nov 2020. The proposed technique is used to forecast both short-term and medium-term confirmed cases. The obtained outcomes of the investigation show that Stacked LSTM and LSTM models are efficient as compared with other existing frameworks in terms of exactness. proving their reliability in forecasting COVID-19 cases.

This study implements RNN and Auto-Regressive RNN on India's Covid-19 dataset of confirmed cases from May 8, 2020, to March 7, 2021 [33]. The MAPE and RMSE metrics were used to evaluate the obtained results. The research conducted was not effective in having a model that detects the complex trend of Covid-19, and it also has various

drawbacks.

Mahdavi at al. proposed [34] a prediction model based on data mining to forecast the production of oil. They consider the oil consumption history data. They performed the cleaning of the data to integrate the data and autoregression to obtain input models and pre-processing to upgrade ANFIS operations. Oil production prediction system with ANFIS Algorithm with clean and integrated data to delete data that is not appropriate so that the data obtained is valid. Kasih, Julianti at al. proposed [35] a prediction model to the final exam score before the test scores are issued by the examiner. They consider the performance of the students in the various subjects of their previous semester for prediction.

Wainana, Stephen et al. analyzed [36] the crime data in Kenya by using data mining approaches and R Software. They used the K-Means algorithm. The crimes such as robbery and theft have discrete clusters that have a very strong linear association. There are also types of crimes that are not strictly correlated and they create a group k, which does not have a linear connection. The APRIORI algorithm demonstrates that various crimes are connected. Ashabul Hoque et al. proposed [37] SEIATR compartmental model for forecasting and examination of the occurrence of covid 19 in most affected five countries in the World. The parameters of the model are resolute by using the Ranga and Kutta method. They used the data of the five countries for compared the confirmed and death cases due to epidemic.

Meenu Gupta et al. presented the AI enabled [38] prediction and analysis of the covid 19 epidemic. They concentrated on the analysis of states and Union Territories of the India. They also proposed a regression model to forecast the count of the confirmed and death cases due to covid19. The efficacy of the proposed prediction model is obtained by using Polynomial Regression (PR), Decision Tree Regression, and Random Forest (RF) Regression algorithms.

Abdullah Ali H. Ahmadini et al. used [39] the Kalman filter to forecast and analysis of the recover, death and confirmed cases of the covid19. They performed the analysis of the four most affected country by this epidemic. They concluded from the obtained result that the Kalman Filter based approach is able of trail of the real COVID-19 data in nearly all circumstances. Md. Shahriare Satu et al. proposed [40] a prediction model using the cloud-based machine learning for the Covid 19. They used the many regressions-based machine learning algorithm on the data of the confirmed cases of the COVID19 cases. The proposed method can accurately predict the number of infected cases daily. Yan Hao et al. proposed [41]a machine learning-based technique to forecast and analyzed the increasing trends and evolution of the covid 19. They applied Elman neural network and SVM on data from Wuhan and found that the Elman neural network and SVM can efficiently foresee the changing drift of cumulative deaths and cured cases. They also found that the LSTM model is not effectively predict confirmed cases, demises, and cured cases.

Ruifang Ma at al. combined [42] the LSTM and Markov methods with the objective to analyze the trends and prediction of the covid-19 cases. They used the confirmed data of US, Britain, Brazil and Russia to obtain the training errors of LSTM and determined the probability transfer matrix.

Song et al. tried to [43] analyse the COVID-19 virus in animals. They focused on understanding the epidemiology, pathogenesis, inhibition, and handling of SARS-Cov and MERS-Cov. They offered particulars on the important construction and role of spike proteins on the exterior of individually virus.

M. Farhan et al. used [44] deep learning to create a forecasting model that confirmed cases of the covid 19 in Pakistan. They determined the crucial features pattern for prediction and used them to specify the patterns of covid 19 in across the world.

Saud. Shaikh et al. introduced [45]a prediction model to foresee covid- 19 outbreak in India using machine learning. They implemented the linear and polynomial regression models and evaluated them using R-squared scores and error values error. They also employed the time series technique to predict the confirmed cases soon.

By critically reviewing the work presented by the various researchers, it has been observed that a lot of work has been done in the direction of the prediction and analyzing the data of the COVID -19. The researcher has used algorithms of the machine learning, neural network, time series etc to forecast the expected cases in near future by analyzing the available data from the various researchers. They have used the data of the most affected counter by the COVID-19 epidemic. In this paper, a ARIMA based approach to predict and analyze the covid 19 case has been presented.

4. Proposed Work

This research carried out several stages starting from pre-processing data, ARIMA Evaluation and Perform Evaluation using RMSE. Before the dataset is tested, it is preprocessed first. Perform type conversion data from nominal to numeric, changing the date data type from nominal to date replace missing values and normalize. The dataset is tested using the ARIMA algorithm by selecting confidence intervals of 10% and select Perform Evaluation, and also select periodicity. Based on the results of testing the ARIMA algorithm, it produces a Root Mean value Squared Error (RMSE). The below figure 3.1 depicts the scenario for ready reference.

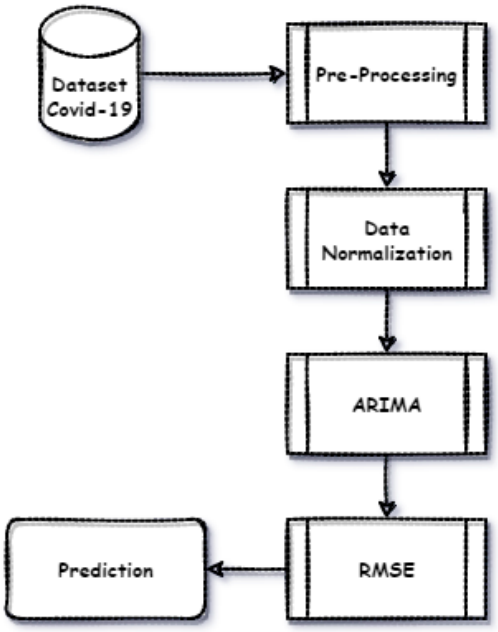

Fig 3.1: Proposed Model

4.1 Dataset Description

In this study, COVID-19 data for INDIA is used which were taken from the official website of the COVID-19 which is collected from https://www.mygov.in/corona-data/covid19-statewise-status/. However, the data is pertaining 511 rows for prediction comprising three columns namely (date, confirmed cases and mortality) below is the exemplary depiction for perusal.

Date	Confirmed	Deaths
05-08-21	1002735	13531
06-08-21	1002849	13533
07-08-21	1002958	13536
08-08-21	1003078	13539
09-08-21	1003154	13540
10-08-21	1003244	13540
11-08-21	1003356	13544

Fig 3.2: COVID-19 Dataset for India

4.2 Algorithm Arima

The ARIMA Prediction stages are:

1. *Enter the results of preprocessing COVID-19 data:* The data from the preprocessing of the COVID-19 data is entered into the program python so that algorithm can read the COVID-19 data that is inside excel.
2. *Create a distribution plot for COVID-19 data:* The data that has been entered into the python program is formed into a graph with the plot code in the python program. This is done to see the shape and pattern of the movement of COVID-19 cases in the past.
3. *Change COVID-19 data to univariate time series*: The COVID-19 data was a multivariate time series type which contained the mortality rate. Therefore, the segregated total positive cases are required for the prediction.
4. *Univariate time series data distribution plot:* COVID-19 past positive case data formed into a graph with the plot code in the program. This will show each movement up and down the level of COVID-19 positive cases in the past.
5. *Perform the differencing process*: By looking at COVID-19 data plot graph, it will be clear that there is a trend pattern in the graph. A trend is a sharp upward pattern in the data. This trend pattern must be removed so that the data looks stable.

$$Y_t' = Y_t - Y_{t-1} (eq. 3.1)$$

6. *Carry out the log transformation process*: By looking at each graph of the sales data plot, it will be clear that the data is not stationary in the variation of the data. The data must be made stationary with a log transformation.

$$Y_t^{new} = log_{10}(Y_t) (eq. 3.2)$$

7. *ACF and PACF plots to find AR and MA models:* After differencing process and log transformation, using the ACF and PACF to look at identifying patterns in the considered data that are stationary at both mean and variance. The idea is to determine the existence of components AR and MA..

Autocorrelation Function (ACF)

The observations in various period are often related or correlated When a variable is determined over time This correlation is determined on this basis of the coefficient of autocorrelation. Data patterns that include elements like trend and seasonality can be considered by means of autocorrelations. Patterns are determined inspecting the autocorrelation coefficients of a variable in different time delays. Equation 3.3 is the formula to calculate the autocorrelation coefficient k between the observations Y_t and Y_{t-k}, which are k periods apart.

$$p_k = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}, k = 0,1,2,3 \dots \dots (eq. 3.3)$$

p_k : It is the autocorrelation coefficient for a lag of k periods.

$\bar{Y}$: It is the mean of the values of the time series.

Y_t: Is the observation in the period t.

Y_{t-k}: Is the observation k previous periods or during the period $t - k$.

Partial Autocorrelation Function (PACF)

Partial autocorrelation is used to determine the correlation between two variables divided by k intervals when the dependencies created by the transitional lags between them is not accounted.

$\rho_{kk} = \text{corr}\,(Y_t Y_{t-k} | Y_{t-1} Y_{t-2} Y_{t-3}, \cdots, Y_{t-k-1})$ (eq.3.4)

ρ_{kk} : Is the partial autocorrelation coefficient.

Y_{t-k}: estimated value in period $t - k$

Y_{t-k}: Is the observation in the period t.

The autocorrelation coefficients can be used to answer the following questions about a Time series:

1. Data is white noise.

2. The data show a trend (they are non-stationary).

3. The data is stationary.

4. Data is seasonal.

Autoregressive Models (AR)

An autoregressive model of order p has the formula:

$$Y_t = \varphi_0 + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-1} + \cdots + \varphi_p Y_{t-p} + \varepsilon \; (eq.3.5)$$

where Y_t : is the response (or dependent) variable at time t $Y_{t-1}, \cdots, Y_{t-p}$: response variable at time lags $t - 1, t - 2, \cdots, t - p$, respectively, these *Y* play the role of self-determining variables.

$\varphi_1, \varphi_2, \varphi_3 \cdots, \varphi_p$: coefficients that will be calculated ε: error term at time t, which denotes the effects of variables.

Moving Average Models (MA)

A q-th order moving average model has the form

$$Y_t = \mu + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon \; (eq.3.6)$$

Where Y_t is the response (or dependent) variable at time t εt−1, ..., εt−q, Errors in previous periods for time t, are incorporated into the response, Yt , μ is constant mean of the process. θ1, θ2, ..., θ, q: Coefficients to be estimated.

ε_{t-1} is Error term, which represents the effects of the variables not explained by the model. The term moving average for the model in equation 3.6 is historical.

8. *Formation of the best ARIMA model process:* To construct the ARIMA model, standard error estimates can be used in the equation below:

$$y'_t = c + \Phi_1 y'_{t-1} + \cdots + \Phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \; (eq.3.7)$$

The three ARIMA components appear clearer when written using the backshift operator, $By_t = y_{t-1}$, as follows:

$$(1 - \Phi B - \cdots - \Phi_p B^p) y_t = c + (1 + \theta_1 B + \cdots + \theta_q B^q) \varepsilon_t \quad \text{(eq. 3.8)}$$

P = order autoregressive

D = level differentiation

Q = order moving average

Thereafter, error evaluation is executed using Root Mean Square Error.

$$RMSE = \sqrt{mean(e_i^2)} \; (eq.3.9)$$

9. *Residual test ACF and PACF plots*: Create graph plots of Autocorrelation Factor and Partial Autocorrelation Factor for see a spike in certain lag lags in the predicted data. If data contains spikes, then the data must be recalculated because it contains extracts information for prediction.

5. Implementation and Result

The prediction process using the ARIMA method will use with programming language python for implementation. In this research project Jupyter Notebook, an open-source web application will be used for performing the tasks of the structuring of the data set and for the implementation of the model of COVID-19 prediction using ARIMA.

5.1 Preprocessing and Data Normalization

This component performs data processing by performing indexing, rearranging, and range selection the correlation of relationships between data so that more valid data state for prediction.

5.2 Covid-19 Positive Case View

The following is the result of positive cases of COVID-19 in India: Therefore, the period from 20.3.2020 to 11.8.2021. In Figure 4.1 we can see the COVID-19 positive case escalation process the initial March 20, 2020 there was only 12 positive cases which is raised by 43529 active cases by August 11, 2021.

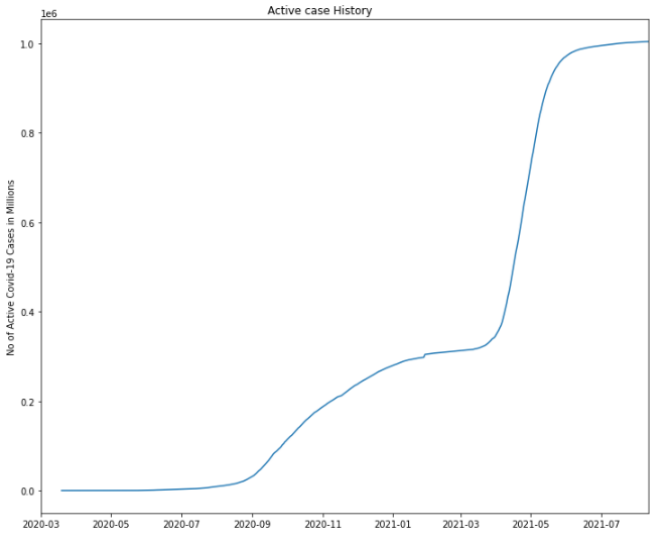

Fig 4.1: Rise in COVID-19 Cases

The above is the result of the Graph of Total Differencing COVID-19 positive cases in the period March 2020 to August 2021. After the different values of each data positive case are known, a differential graph plot is created to ensure that trend patterns are completely changed from September 2020 onwards. In Figure 4.1, it is evident that the positive case trend, namely the tendency of the pattern of the data graph is raised sharply and promptly.

5.3 Prediction Process

This component explains how to form the ARIMA method model, how to estimate the parameters, how to test the method model that has been obtained, and how to input the prediction system as well as how to process errors in testing the prediction.

SARIMAX Results

Dep. Variable:	y	No. Observations:	357
Model:	ARIMA(2, 0, 0)	Log Likelihood	-2746.727
Date:	Sat, 09 Jul 2022	AIC	5501.453
Time:	06:40:10	BIC	5516.964
Sample:	0	HQIC	5507.623
	- 357		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.197e+05	9.22e-11	1.3e+15	0.000	1.2e+05	1.2e+05
ar.L1	1.9198	0.009	218.433	0.000	1.903	1.937
ar.L2	-0.9198	0.009	-103.280	0.000	-0.937	-0.902
sigma2	2.718e+05	3.79e-08	7.16e+12	0.000	2.72e+05	2.72e+05

Ljung-Box (L1) (Q):	66.69	Jarque-Bera (JB):	114815.99
Prob(Q):	0.00	Prob(JB):	0.00
Heteroskedasticity (H):	196.51	Skew:	0.87
Prob(H) (two-sided):	0.00	Kurtosis:	90.84

Fig 4.2: Results by ARIMA

The following are the results of Total Differencing and log 10 of the volume of COVID-19 cases in the period March 2020 to August 2021. It is better if we process the difference between the results of the logarithm of 10 for each period to make the time series data to be stationary both on average (mean), as well as in the data variance. In Figure 4.2, we can see that the calculation of the difference in the logarithm of 10 between the periods March 2020 and August 2021 is -2746. Therefore, the difference between the outcome of the logarithm of 10 each period to make time series data become stationary in both the mean and data variance. Therefore, the Heteroskedasticity (H): score is 196.51 where the prediction has become stationary on the mean and variance of the data.

RMSE Score

The ACF and PACF values are calculated to find the AR and MA models of ARIMA in the period March 2020 to August 2021 by calling the auto. Arima (y = 2.) function log 10 the premium ARIMA prediction model obtained was ma 2 with a coefficient value ar. L1 is 1.9198 likewise ar. L2 is -0.9198 and sigma2 is 2.718e+05. Consequently, from the results obtained, the model obtained RMSE score 549390.96.

Prediction

After calculating the difference from the logarithm of 10 between a certain period, then do plot ACF and PACF graphs to see if there is a spike or not. If there is a spike it is concluded that the data contains residuals that are not random (random) which can be concluded that the data still contains the information component needed in the data the process of calculating predictions.

After calculating the difference from logarithm of 10 between a certain period, then the ACF and PACF calculations (Training Data - Blue Color) are carried out for see if there is a spike or not. Therefore, the spike exists in variance with sharp escalation, it is concluded that the data contains residuals that are not random (random) which can be concluded that the data is still contains information components needed in the process of calculating predictions In Figure 4.3 graph that there is a spike in the lag 4 to 9 (Red Line - Prediction). Therefore, the cases in the Green Line are predicted.

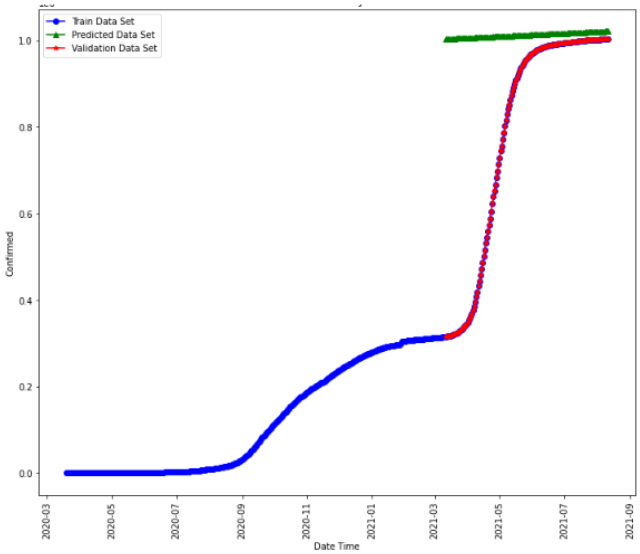

Fig 4.3: Prediction by ARIMA

6. Conclusion

Based on the analysis of the results of the research that has been carried out, it can be concluded that the data used is stationary with respect to variance after transformation. However, in the process of checking the stationarity of the mean using autocorrelation, it was found that more than six lags came out of the confidence interval line so that it was necessary to do differencing.

After differencing from value 3 of confirmed cases the result is that there are continues lags that come out of the interval confidence line. This shows that the data is stationary with respect to the mean. In the next step, checking the differencing data against Partial Autocorrelation, it was also found that there were no more than nine lags that came out of the interval confidence line. After performing the above steps, several ARIMA models were obtained. As a result, the ARIMA model was subjected to trial and error by examining the RMSE value. It was discovered that the best model for the total data on cases of positive COVID-19 patients was ARIMA (2, 0, 0). Prediction results that are close to the real data are generated based on the ARIMA model findings.

7. Future Scope

There are several suggestions for further researchers related to the development of Predictions covid-19 positive cases prediction needs using the ARIMA method include the following:

1. Prediction of covid-19 positive cases prediction using the ARIMA method that the author made is still using one method, for further development it is expected that combine with Two (2) methods, such as the GARCH method.

References

[1] https://www.who.int/news-room/fact-sheets/detail/hiv-aids

[2] Berche, Patrick. (2022). The Spanish flu. La Presse Médicale. 51. 104127. 10.1016/j.lpm.2022.104127.

[3] https://www.worldometers.info/coronavirus/

[4] Ahmed, Selmi & Ramazani, Ali. (2022). SARS-CoV-2. 10.22034/CHEMM.2022.335353.1462.

[5] Taghvaei, Amirhossein & Georgiou, Tryphon & Norton, Larry & Tannenbaum, Allen. (2020). Fractional SIR epidemiological models. Scientific Reports. 10. 10.1038/s41598-020-77849-7.

[6] CATAL, C., ECE, K., Arslan, B., & Akbulut, A. (2019). Benchmarking of Regression Algorithms and Time Series Analysis Techniques for Sales Forecasting. Balkan Journal of Electrical and Computer Engineering, 7(1), 20–26. https://doi.org/10.17694/bajece.494920

[7] Grégoire, Gérard. (2022). 3 - ARMA AND ARIMA TIME SERIES. 10.1051/978-2-7598-2741-1.c008.

[8] Syeda HB, Syed M, Sexton KW, Syed S, Begum S, Syed F, Prior F, Yu F Jr. Role of Machine Learning Techniques to Tackle the COVID-19 Crisis: Systematic Review. JMIR Med Inform. 2021 Jan 11;9(1):e23811. doi: 10.2196/23811. PMID: 33326405; PMCID: PMC7806275.

[9] https://covid19.who.int/region/searo/country/in.

[10] Response, Chinese. (2020). 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. 41. 145-151. 10.3760/cma.j.issn.0254-6450.2020.02.003.

[11] Cucinotta D, Vanelli M. WHO Declares COVID-19 a Pandemic. Acta Biomed. 2020 Mar 19;91(1):157-160. doi: 10.23750/abm.v91i1.9397. PMID: 32191675; PMCID: PMC7569573.

[12] Alzahrani, Saleh & Aljamaan, Ibrahim & Al-Fakih, Ebrahim. (2020). Forecasting the Spread of the COVID-19 Pandemic in Saudi Arabia Using ARIMA Prediction Model Under Current Public Health Interventions. Journal of Infection and Public Health. 13. 10.1016/j.jiph.2020.06.001.

[13] Shetty, Rashmi & Pai, P.. (2021). Forecasting of COVID 19 Cases in Karnataka State using Artificial Neural Network (ANN). Journal of The Institution of Engineers (India): Series B. 102. 10.1007/s40031-021-00623-4.

[14] Schaback, Robert. (2020). On COVID-19 Modelling. Jahresbericht der Deutschen Mathematiker-Vereinigung. 122. 10.1365/s13291-020-00219-9.

[15] Ghosh, Rakhi. (2022). Covid-19. 10.4324/9781003291527-20.

[16] Henry, Timothy & Garcia, Santiago. (2022). COVID-19. Cardiology Clinics. 40. i. 10.1016/S0733-8651(22)00036-4.

[17] Gong, Michelle & Martin, Gregory. (2022). COVID-19. Critical Care Clinics. 38. i. 10.1016/S0749-0704(22)00027-6.

[18] Adesanya-Davies, Funmilayo. (2022). COVID-19.

[19] Seidel, Karen. (2022). Modelling binary classification with computability theory. 10.25932/publishup-52998.

[20] Krispin, R. (2019). Hands-On Time Series Analysis with R: Perform time series analysis and forecasting using R. Packt Publishing.

[21] Lazzeri, F. (2020). Machine Learning for Time Series Forecasting with Python. Wiley.

[22] Nielsen, A. (2019). Practical Time Series Analysis: Prediction with Statistics and Machine Learning (1st ed.). O'Reilly Media.

[23] Shumway, R. H., & Stoffer, D. S. (2017). Time Series Analysis and Its Applications: With R Examples (Springer Texts in Statistics) (4th ed. 2017 ed.). Springer.

[24] Jones, R.H.. (2018). Autoregressive Moving Average Errors. 10.1201/9780203748640-6.

[25] Neusser, Klaus. (2016). Autoregressive Moving-Average Models. 10.1007/978-3-319-32862-1_2.

[26] Hoffmann, John. (2021). Homoscedasticity. 10.1201/9781003162230-9.

[27] Yang, K., Tu, J., & Chen, T. (2019). Homoscedasticity: an overlooked critical assumption for linear regression. General Psychiatry, 32(5). https://doi.org/10.1136/gpsych-2019-100148.

[28] Arora, Jyoti & Mahajan, Palvi & Singh, Trapti. (2019). SURVEY ON ARIMA (Autoregressive integrated moving average).

[29] Parbat, Debanjan & Chakraborty, onisha. (2020). A Python based Support Vector Regression Model for prediction of Covid19 cases in India. Chaos, Solitons & Fractals. 138. 109942. 10.1016/j.chaos.2020.109942.

[30] Fanelli, Duccio & Piazza, Francesco. (2020). Analysis and forecast of COVID-19 spreading in China, Italy and France. Chaos, Solitons & Fractals. 134. 109761. 10.1016/j.chaos.2020.109761.

[31] Ma, R., Zheng, X., Wang, P., Liu, H., & Zhang, C. (2021). The prediction and analysis of COVID-19 epidemic trend by combining LSTM and Markov method. Scientific Reports, 11(1), 17421. https://doi.org/10.1038/s41598-021-97037-5.

[32] Devaraj, J., Madurai Elavarasan, R., Pugazhendhi, R., Shafiullah, G. M., Ganesan, S., Jeysree, A. K., Khan, I. A., & Hossain, E. (2021). Forecasting of COVID-19 cases using deep learning models: Is it reliable and practically significant? Results in Physics, 21, 103817. https://doi.org/10.1016/j.rinp.2021.103817.

[33] Bouhaddour, S., Saadi, C., Bouabdallaoui, I., Guerouate, F., & Sbihi, M. (2022). Recurrent Neural Network and Auto-Regressive Recurrent Neural Network for trend prediction of COVID-19 in India. ITM Web of Conferences, 46, 02007. https://doi.org/10.1051/itmconf/20224602007.

[34] Zahra Mahdavi , Maryam Khademi Prediction of Oil Production with: Data Mining, Neuro-Fuzzy and Linear Regression International Journal of Computer Theory and Engineering, Vol. 4, No. 3, June 2012

[35] Julianti Kasih, Mewati Ayub, Sani Susanto Predicting students' final passing results using the Apriori Algorithm World Transactions on Engineering and Technology Education 2013 WIETE Vol.11, No.4, 2013

[36] Stephen Mangara Wainana, Joseph Njuguna Karomo, Rachael Kyalo, Noah Mutai Using Data Mining Techniques and R Software to Analyze Crime Data in Kenya International Journal of Data Science and Analysis 2020; 6(1): 20-31 ISSN: 2575-1883 (Print); ISSN: 2575-1891 (Online)

[37] Hoque, A., Malek, A. & Zaman, K.M.R.A. Data analysis and prediction of the COVID-19 outbreak in the first and second waves for top 5 affected countries in the world. *Nonlinear Dyn* **109**, 77–90 (2022). https://doi.org/10.1007/s11071-022-07473-9

[38] Meenu Gupta, Rachna Jain, Simrann Arora, Akash Gupta , Mazhar Javed Awan , Gopal Chaudharyand Haitham Nobanee "AI-Enabled COVID-19 Outbreak Analysis and Prediction: Indian States vs. Union Territories" Computers, Materials & Continua, CMC, 2021, vol.67, no.1 , DOI:10.32604/cmc.2021.014221

[39] Abdullah Ali H. Ahmadini , Muhammad Naeem, Muhammad Aamir , Raimi Dewan, Shokrya Saleh A. Alshqaq and Wali Khan Mashwan "Analysis and Forecast of the Number of Deaths, Recovered Cases, and Confirmed Cases From COVID-19 for the Top Four Affected Countries Using Kalman Filter" Frontiers in Physics, August 2021 | Volume 9, doi: 10.3389/fphy.2021.629320

[40] Satu, M.S.; Howlader, K.C.; Mahmud, M.; Kaiser, M.S.; Shariful Islam, S.M.; Quinn, J.M.W.; Alyami, S.A.; Moni, M.A. Short-Term Prediction of COVID-19 Cases Using Machine Learning Models. Appl. Sci. 2021, 11, 4266. https://doi.org/ 10.3390/app110942

[41] Hao Y, Xu T, Hu H, Wang P, Bai Y (2020) Prediction and analysis of Corona Virus Disease 2019. PLoS ONE 15(10): e0239960. https://doi.org/10.1371/journal.pone.0239960

[42] Ruifang Ma1, Xinqi Zheng, PeipeiWang, Haiyan Liu,Chunxiao Zhang, The prediction and analysis of COVID-19 epidemic trend by combining LSTM and Markov method Scientifc Reports | (2021) 11:17421 | https://doi.org/10.1038/s41598-021-97037-5

[43] Song, Z.; Xu, Y.; Bao, L.; Zhang, L.; Yu, P.; Qu, Y.; Zhu, H.; Zhao, W.; Han, Y.; Qin, C. From SARS to MERS, Thrusting Coronaviruses into the Spotlight. Viruses 2019, 11, 59. https://doi.org/10.3390/v11010059

[44] M. Farhan, S. Jabbar and M. R. Shahid, "Prediction and Analysis of Covid-19 Positive Cases Using Deep Learning Model," *2021 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)*, Quetta, Pakistan, 2021, pp. 1-6, doi: 10.1109/ICECube53880.2021.9628335.

[45] S. Shaikh, J. Gala, A. Jain, S. Advani, S. Jaidhara and M. Roja Edinburgh, "Analysis and Prediction of COVID-19 using Regression Models and Time Series Forecasting," *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, 2021, pp. 989-995, doi: 10.1109/Confluence51648.2021.9377137.

[46] Andrew Hernandez, Stephen Wright, Yosef Ben-David, Rodrigo Costa, David Botha. Risk Assessment and Management with Machine Learning in Decision Science. Kuwait Journal of Machine Learning, 2(3). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/196

[47] Andrew Hernandez, Stephen Wright, Yosef Ben-David, Rodrigo Costa, David Botha. Intelligent Decision Making: Applications of Machine Learning in Decision Science. Kuwait Journal of Machine Learning, 2(3). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/197

[48] Dhabliya, D. (2021). Delay-tolerant sensor network (DTN) implementation in cloud computing. Paper presented at the Journal of Physics: Conference Series, , 1979(1) doi:10.1088/1742-6596/1979/1/012031 Retrieved from www.scopus.com