Research on Biomedical Engineering
https://doi.org/10.1007/s42600-020-00105-4

ORIGINAL ARTICLE

Application of machine learning time series analysis for prediction COVID-19 pandemic

Vikas Chaurasia[1] · Saurabh Pal[1]

Received: 12 June 2020 / Accepted: 13 October 2020

Abstract

Purpose Coronavirus disease is an irresistible infection caused by the respiratory disease coronavirus 2 (SARS-CoV-2). It was first found in Wuhan, China, in December 2019, and has since spread universally, causing a constant pandemic. On June 3, 2020, 6.37 million cases were found in 188 countries and regions. During pandemic prevention, this can minimize the impact of the disease on individuals and groups. A study was carried out on coronavirus to observe the number of cases, deaths, and recovery cases worldwide within a specific time period of 5 months. Based on this data, this research paper will predict the future spread of this infectious disease in human society.

Methods In our study, the dataset was taken from WHO "Data WHO Coronavirus Covid-19 cases and deaths-WHO-COVID-19-global-data". This dataset contains information about the observation date, provenance/state, country/region, and latest updates. In this article, we implemented several forecasting techniques: naive method, simple average, moving average, single exponential smoothing, Holt linear trend method, Holt-Winters method and ARIMA, for comparison, and how these methods improve the Root mean square error score.

Results The naive method is best suited as described over all other methods. In the ARIMA model, utilizing grid search, we recognized a lot of boundaries that delivered the best-fit model for our time series data. By continuing the model, future predictions of death cases indicate that the number of deaths will increased by more than 600,000 by January 2021.

Conclusion This survey will support the government and experts in making arrangements for what is about to happen. Based on the findings of instantaneous model, these models can be adjusted to guide long time.

Keywords COVID-19 · SARS-CoV-2 · WHO · Forecasting techniques · ARIMA

Introduction

So far, coronavirus, which has killed millions of people throughout, is constantly taking people under its arrest. Washing hands, covering your face, isolating hygiene, and staying away from the community may be a way to prevent this communicable disease, but it is not enough (Nussbaumer-Streit et al. 2020). As per the World Health Organization (WHO), there are neither immunizations nor explicit antiviral medicines for COVID-19 (Q&A on corona viruses (COVID-19). World Health Organization (WHO) 2020). As like Middle East respiratory syndrome (MERS) and severe acute respiratory syndrome (SARS), coronaviruses are an enormous group of infections which may cause ailment in creatures or people. In people, a few coronaviruses are known to cause respiratory contaminations going from the basic virus to increasingly extreme infections. The most as of late found coronavirus causes coronavirus infection, COVID-19. The basic symptoms for COVID-19 include fever, exhaustion, shortness of breath, and loss of smell and taste. Further there will be progress in severe respiratory disease (ARDS), multiple organ failure, septic syncope, and blood clots. The starting of side effects is usually about 5 days, but it may be increase from 2 to 14 days (Symptoms of coronavirus. U.S. Centers for Disease Control and Prevention (CDC) 2020; Hopkins and Kumar 2020; Velavan and Meyer 2020). COVID-19 spreads principally when individuals are in close contact, and one individual breathes in little droplets created by a contaminated individual sniffling, talking, or singing. Airborne

✉ Saurabh Pal
drsaurabhpal@yahoo.co.in

Vikas Chaurasia
Chaurasia.vikas@gmail.com

[1] Department of Computer Applications, VBS Purvanchal University, Jaunpur, India

Published online: 24 October 2020

transmission is unique to transmission because it implies the existence of organisms in the core of the droplet. These particles are usually regarded as particles with a width of less than 5 μm, which can remain for a long time and can be separated from others. It is worth noting that people should communicate more than 1-m distance (Q and A on COVID-19 2020; Hamner et al. 2020). Some clinical strategies are responsible and result in the infection being transmitted more effectively than typical. As pointed out by the Centers for Disease Control and Prevention, the spread of these infections can be controlled. From the basic principles of hand cleaning to a group approach based on a group-based comprehensive safety plan, there are simple, easy-to-use, and financially savvy systems that can help prevent pollution. Hand hygiene, environmental cleanliness, screening and grouping of patients, vaccination, surveillance, antibiotic management, care coordination based on evidence, appreciation of all departments that contribute to the infection avoidance plan, a safety plan based on a comprehensive unit are the methods used to control and prevent the infectious Covid-19 disease.

In the healthcare industry, there is a lot of evidence that machine learning algorithms can provide effective models to solve problems in order to identify patients. To date, there is no vaccine or antibiotic that can cure infected people and avoid this pandemic disease. There are right now more than 169 COVID-19 immunization up-and-comers being worked on, with 26 of these in the human preliminary stage. WHO is working in a joint effort with researchers, business, and worldwide well-being associations to accelerate the pandemic reaction. At the point when a sheltered and powerful antibody is found, COVAX (group of WHO, GAVI, and CEPI) will encourage the impartial access and dispersion of these immunizations to secure individuals in all nations. Individuals most in danger will be organized (WHO 2020). Many researchers and scientists related to machine learning are also involved in solving this situation. In order to understand the patterns and characteristics of virus attacks, many data scientists may make the right decisions and take specific actions.

The purpose of this study is as follows:

1. Using time series to predict imminent deaths worldwide
2. Comparing the root-mean-square value of each model using time series predictive modeling through several methods.
3. Finding a method suitable for prediction on the COVID-19 dataset
4. Using ARIMA model for future forecasting of death cases worldwide

Literature review

The following are some studies that present evidence on the use of machine learning or mathematical models in modeling COVID-19 and other infectious diseases.

The autoregressive integrated moving average (ARIMA) model is used to estimate the number of normal COVID-19 cases per day in Saudi Arabia in the next month. They implemented four different forecasting models. Autoregressive model moving average (combined use of the two) and combined ARMA (ARIMA) determine the best model fit, and they found that the ARIMA model is superior to other models. The evaluation results indicate that if strict prudence and control measures are not taken to limit the spread of Saudi Arabia, Saudi Arabia's model will continue to evolve, possibly adding up to 7668 new cases per day, and 127,129 cases per day in just 1 month. New Coronavirus Pneumonia (COVID-19) indicates that pilgrimages to the two paradise cities of Mecca and Medina in Saudi Arabia. This indicates that pilgrimages to the two paradise cities of Mecca and Medina in Saudi Arabia, which are scheduled to be performed by nearly 2 million Muslims in mid-July, may be suspended. In order to prevent this situation, many unconventional prevention and control measures have been proposed (Alzahrani SI, Aljamaan IA, Al-Fakih EA., at el.) (Alzahrani et al. 2020). Aslam M. proposed to use a more logical method of the Kalman channel in combination with the autoregressive integrated moving average (ARIMA) model in order to obtain a more accurate estimation of the prevalence, dynamic cases, active cases, and death cases of COVID-19 pandemic identification in Pakistan (Aslam 2020). A basic econometric model is proposed by Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, and Ciccozzi M. et al., which may be very valuable for predicting the spread of COVID-2019. They carried out autoregression integrated moving average (ARIMA) model prediction on Johns Hopkins epidemiological information to predict the epidemiological model of the prevalence and incidence of COVID-19. For further inspection or future consideration, the case definition and information classification must be maintained continuously (Benvenuto et al. 2020). By Duan X and Zhang X. et al., the newly confirmed cases of COVID-19 in Japan and South Korea from January 20, 2020, to April 26, 2020, are grouped daily. We are familiar with the fact that automatic regression integrated moving average (ARIMA) model, studied two pieces of information, collected and predicted that there will be new confirmed cases every day during the 7 days from April 27, 2020 to May 3, 2020. Similarly, the determination result and two types of information collection are also provided (Duan and Zhang 2020). Ilie OD, Cojocariu RO, Ciobica A, Timofte SI, Mavroudis I, and Doroftei B. proposed an autoregressive integrated moving average (ARIMA) model was established and used to predict the epidemiological patterns of COVID-19 in the three countries of Ukraine, Romania, the Republic of Moldova, Serbia, Bulgaria, Hungary, the USA, Brazil, and India. In order to improve the accuracy, the main daily information of COVID-19 from March 10, 2020, to July 10, 2020, was collected from the official sites of the Romanian government (GOV.RO), the World Health Organization (WHO), and the European Center for Disease Control and Prevention (ECDC) website. Some ARIMA models have various ARIMA boundaries. ARIMA (1, 1, 0), ARIMA (3, 2, 2), ARIMA (3, 2, 2), ARIMA (3, 1, 1), ARIMA (1, 0, 3), ARIMA ((1, 2), 0), ARIMA (1, 1, 0), ARIMA (0, 2, 1), and ARIMA (0, 2, 0) models are selected as the best models, depending on their average absolute percentage error (MAPE) are reduced the most)), which are Ukraine, Romania, Republic of Moldova, Serbia, Bulgaria, Hungary, the USA, Brazil, and India (4.70244, 1.40011, 2.67551, 2.16373, 2.98154, 2.11139, 3.21569, 4.10596, and 2.78051). This survey shows that the ARIMA model is reasonable for expectations in current emergencies and provides ideas for the epidemiological stage of these regions (Ilie et al. 2020). KırbaşI, SözenA, Tuncer AD, and Kazancıoğlu F. et al. confirmed that the COVID-19 instances in Denmark, Belgium, Germany, France, the UK, Finland, Switzerland, and Turkey have passed the autoregression integrated moving average (ARIMA), non-linear autoregression neural network (NARNN), and long-term memory (LSTM) methods. Perform metrics using six models to select the most accurate model (MSE, PSNR, RMSE, NRMSE, MAPE, and SMAPE). As shown by checking the after effects of the initial steps, LSTM was found to be the most accurate model. In the second stage of the investigation, the LSTM model was given, and expectations were made in 14 days. This is yet to be understood. The results of the second step of the test showed that in many countries, the increase in the absolute total number of cases has only marginally decreased (Kırbaş et al. 2020) (Pourghasemi HR, Pouyan S, Farajzadeh Z, Sadhasivam N, Heidari B, Babaei S, etc.). The focus is to decompose the risk factors of COVID incidents, to distinguish areas with high pollution risks, and to evaluate disease behavior in Iran's Fars province. Through AI calculations (MLA) and support vector machines (SVM) based on geographic data framework (GIS), the risk of COVID-19 outbreak in Iran's Fars province has been assessed, even though people are tainted with it every day in polynomial and autoregressive coordination case studies were conducted in the moving normal (ARIMA) model to examine the spread of infections in the region and Iran. The consequences of the Iran infection incident were compared and provided information for Iran and the world. Sixteen successful features were selected. The approval results show that SVM has completed the AUC estimates of 0.786 (March 20), 0.799 (March 29), and 86.6 (April 10), showing high expectations for the location of the accident hazard change. The detailed cross-check polynomial and ARIMA model results in this area reveal an extended pattern with proof of turning, indicating that the extensive isolation has been successful. The overall pattern of infection spread in Iran and Fars province is comparable, although the development of infected cases in the region is more unstable. The squeal of this research may help to better program the anticipation and control of COVID-19 disease, and obtaining various prior capabilities will have broad advantages (Pourghasemi et al. 2020). Sahai AK, Rath N, Sood V, and Singh MP studied the timing information of the top five countries affected by COVID-19 to estimate the spread. From the online dataset, the time schedule information from February 15, 2020, to June 30, 2020, of all contaminated cases from five main countries (especially the USA, Brazil, India, Russia, and Spain) was collected. The detailed information of the ARIMA model was evaluated using Hannan and Rissanen calculations. The ARIMA model was used to process the test chart for the next 77 days. The estimated values for the 18 days in early July were compared, and within a satisfactory understanding range, the true information and graphic accuracy obtained by MAD and MAPE were found. The real track gauge infographic shows that although Russia and Spain have become the focus of the epidemic, the USA, Brazil, and India are still in exponential bending. The survey shows that India and Brazil will reach 1.38 million and 2.47 million marks, respectively, while the USA will reach 4.29 million marks on July 31 (Sahai et al. 2020) (Singh RK, Rani M, Bhagavathula AS, Sah R, Rodriguez-Morales AJ, Kalita H, etc.) In confirmed cases, there is a correlation between transit and recuperation with the recognized top 15 countries and autoregressive coordinated movement. The normal (ARIMA) model is used to predict the direction of the spread of COVID-19 infection in the next 2 months. Conversations were held among the 15 most contaminated countries, and the latest news about confirmed cases, causes of death, and recuperation in the past 3 months were considered. The ARIMA model is used to predict future information that depends on timing information. In the main 15 countries, through the confirmed cases of COVID-19, past and recuperation conditions, the correlation between the total number of late cases and the expected cases has been completed. The researchers comprehensively considered the top 15 recognized countries from the COVID-19 case, inheritance, and recuperation. The disease has spread in the USA, Britain, Turkey, China, and Russia. Rehabilitation rates are fast in China, Switzerland, Germany, Iran, and Brazil. Rehabilitation rates are moderate in the USA, the UK, the Netherlands, Russia, and Italy. Italy and the UK have high mortality rates, while Russia, Turkey, China, and the USA have lower death rates. The ARIMA model is used to predict the estimated confirmed cases, death, and recovery rates in the best 15 countries/regions from April 24 to July 7, 2020. Its value is measured by 95%, 80%, and 70% certainty of stretch quality. The approval of the ARIMA model was completed using the Akaike data metric; for the combined confirmed cases of COVID-19, the quality of passing and recuperation was approximately 20, 14, and 16, respectively, indicating that the results are satisfactory. The expected quality of observations indicates that in all the countries observed, except for China, Switzerland, and Germany, the number of positive cases, inheritance, and recuperation will double. It was also found that the fatality rate and recovery rate were faster compared with the confirmed cases in the following 2 months. The related death rates will be much higher in the USA, Spain, and Italy, followed by France, Germany, and the UK. An estimated survey of the elements of COVID-19 shows an alternative picture of the entire world, which looks even more terrifying than

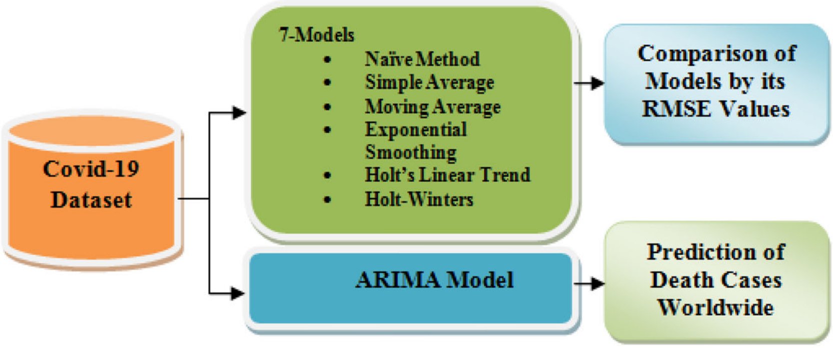


Fig. 1 Structure of experiment

expected, but by July 7, 2020, recovery rates are beginning to be encouraging (Singh et al. 2020) (Singh S, Parmar KS, Kumar J, Makkhan SJS, et al.) The proposed hybrid technique includes the use of cautious wavelet attenuation of the dataset due to COVID-19, dividing the information into segments, and then applying the targeted segmentation. Appropriate econometric models are arranged in each segment to predict future death cases. The ARIMA model is a well-known econometric guidance model. When applied to the wavelet decay time schedule, it can produce accurate estimates. The information dataset contains daily cases in the five countries/regions most affected by COVID-19, which are provided to the crossover model for approval, and death cases are expected 1 month in advance. These expected values are compared with the expected values obtained from the ARIMA model to evaluate the forecast results. Regardless of the prudent measures taken by the competent authorities of these countries, expectations are that the death toll will rise sharply (Saboia 1977). During this period, Yang Q Wang J, Wang J, Ma H, and Wang X. et al. established a daily scheduling ARIMA model for new cases and new deaths. In addition, these models have been used in Italy with similar population conditions and confinement in Italy, in order to predict Italy's scourge in the next 10 days and provide hypothetical premises for the improvement of epidemics in certain countries in the future (Yang et al. 2020). Chaurasia, V. and Pal, S. et al. show the current pace of death of the entire subject, using attribute correlation coefficients (MAE, MSE, RMSE, and MAPE), in which the normal maximum interest rate incorrectly approved the model A reduction of 99.09%. The ARIMA model is used to generate auto_arima SARIMAX results, auto_arima residual maps, ARIMA model results, and related expectation maps on the prepared dataset. This information indicates that the number of deaths is decreasing. By applying the recurrence model, the coefficients created by the recurrence model are evaluated, and actual death cases and expected eligible cases are considered and studied. It was found that the expected mortality rate has decreased after May 2, 2020. This will provide support to the legislature, and experts will prepare for upcoming plans. Given short-term expectations, these strategies can be used to speculate on a wide range of mortality rates (Chaurasia and Pal 2020).

Methodology

The World Health Organization (WHO) time series data has been used for empirical analysis. The data period is from January 22, 2020, to May 28, 2020. The dataset contains 8 attributes and 30,883 instances. The first column of the attribute contains the serial number, so this column is not needed, so we delete it. We used 80% of the dataset for training purposes and 20% of the data for testing purposes to minimize the impact of data differences and better understand the characteristics of the model. Data includes confirmed cases, deaths, and recovered cases from all countries (https://data.humdata.org/dataset/coronavirus-covid-19-cases-and-deaths 2020). This article focuses on the data used for analysis and prediction of COVID-19 in India and around the world to confirm the diagnosed patients, those who died and recovered. For analysis and forecast quantity in patients with COVID-19 in India and worldwide, the following time series analysis have been used.

We are provided with 5 months of data (January 2020–May 2020), and using this data, we will forecast the number deaths for future. The overall structure of the experiment is shown in Fig. 1.

Data preprocessing

For creating training and test files for modeling:

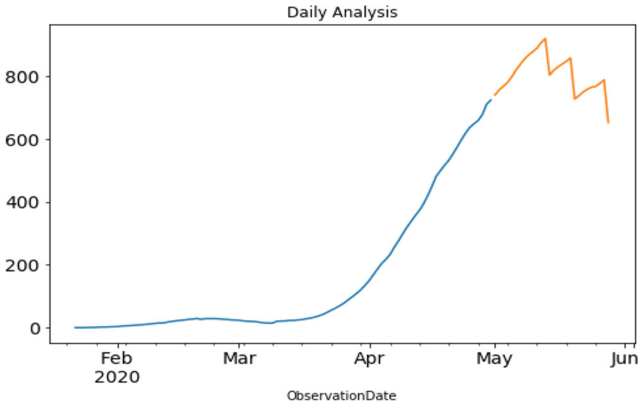
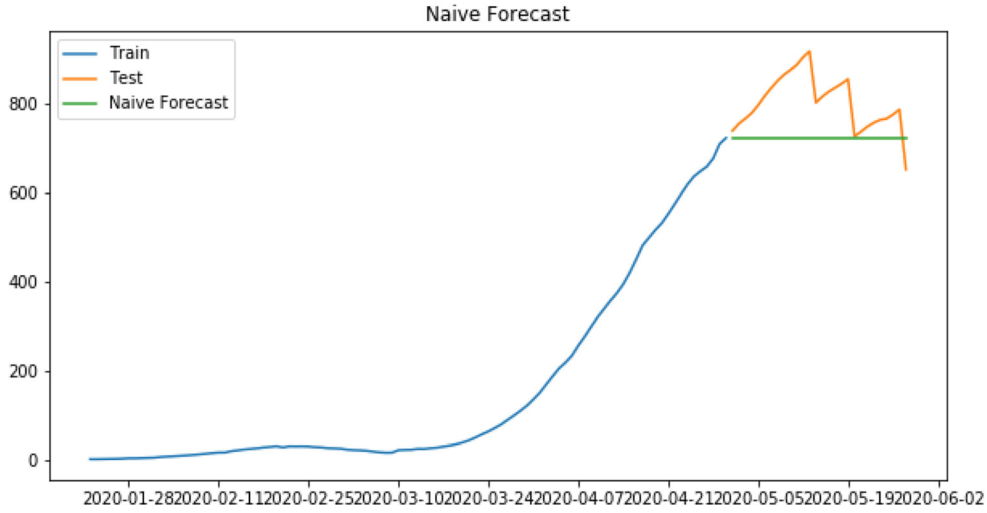


Fig. 2 Distribution of training and testing dataset over time period

Fig. 3 Naïve forecast at test dataset



- The first 4 months (January 2020 to April 2020) are used as training data, and the next 1 month (May 2020) is used as test data.
- The dataset is summarized on daily basis.

The training and testing of the dataset is different during the time period shown in the Fig. 2 below.

Naïve method

When we use naive methods to predict the next day, we can get the value of the last day; it is estimated that the value is the same the next day (Frey and Weck 1983). This prediction technique is called the naive method, and we assume that the next expected point is equal to the last observed point, i.e.:

$$\widehat{y}_{t+1}=\widehat{y}t$$

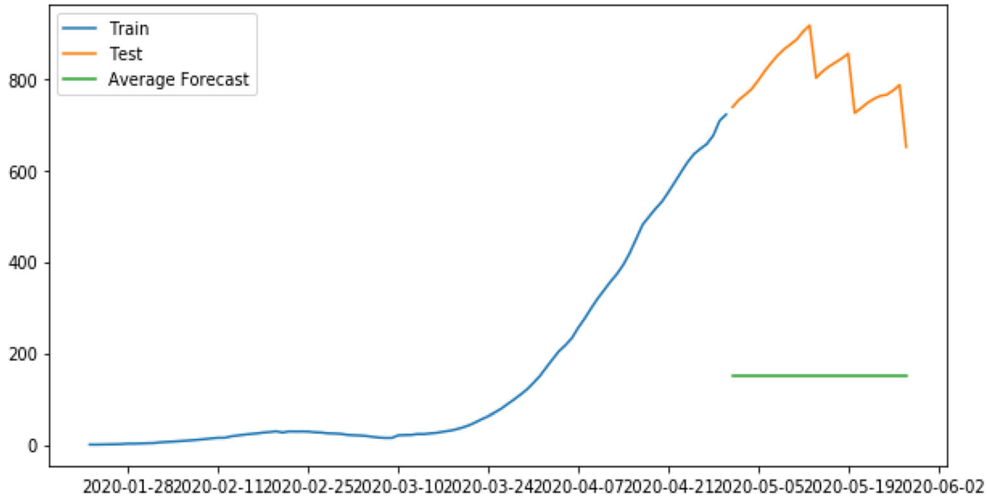
where $\widehat{y}_{t+1}$ is a short-hand for the estimate of $\widehat{y}_{t+1}$ based on the data $\hat{y}_{1,}\ \hat{y}_{2}\dots\ \hat{y}_{t.}$

Now we will perform a naive method to predict "death" worldwide observed in the test dataset. In Fig. 3, the *y*-axis depicts the deaths of a patients, and *x*-axis depicts the time (months).

Simple average method

In some cases, the numbers in the data are increasing and decreasing randomly with small amplitude; the average value is kept constant. Although the dataset has a small change in the entire time period, the average value of each time period remains unchanged (Genre et al. 2013). In this case, we can predict the number of the next day, which is similar to the average of the past few days. This prediction technique in which the predicted expected value is equal to the average value of all previously observed points is called simple averaging technique. We take all previously known values of order n, $M\hat{y}(n)$ for a period $t+1$, and calculate average, and use it as the next value, i.e.:

Fig. 4 Simple average forecast at test dataset



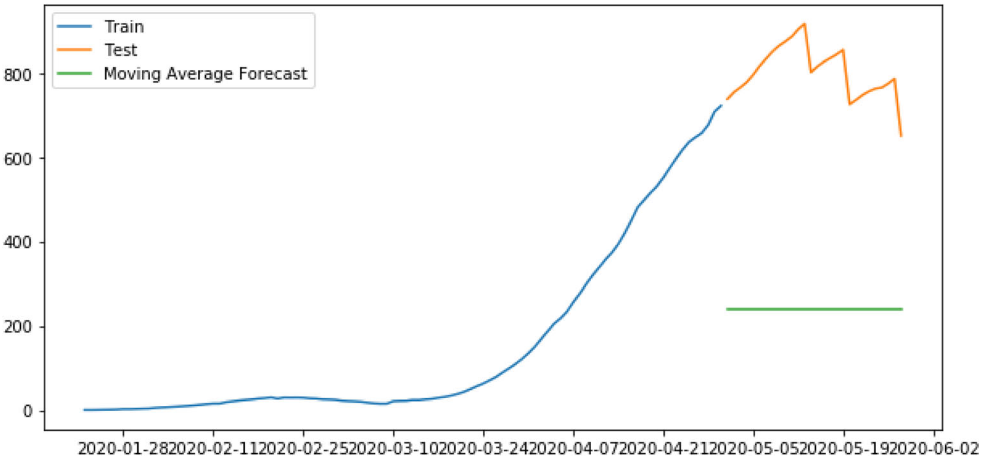


Fig. 5 Moving average forecast at test dataset

$$M\widehat{y}_{t+1} = \frac{1}{n}\sum\widehat{y}_t$$

where n is the number of observations used in calculation and time period $t+1$ is the forecast for all future time periods.

In Fig. 4, the y-axis depicts the deaths of a patients, and x-axis depicts the time (months).

Moving average method

In numerous multiple times, we are given a dataset, in which the quantity of passing increased/decreased pointedly some timeframes back. So as to utilize the past average technique, we need to utilize the mean of all the past information.

Utilizing the demise number of the underlying timeframe would profoundly influence the conjecture for the following timeframe.

In this manner, as an improvement over basic normal, we will take the normal of the passing for last scarcely any time spans as it were. Obviously, the theory here is that continuous characteristics are important. Such anticipating strategy which utilizes window of timeframe for ascertaining the normal is called moving average method (Williams et al. 2011). Computation of the moving normal includes what is here, and it is called a "sliding window" of size n.

Utilizing a basic moving normal model, the h-step-a-head forecast $F\ (t+h)$ is:

$$F(t+h) = M(t) + \left[h + \frac{n-1}{2}\right]F(t)$$

where *F(t)* is the smoothed series adjusted for any local trend and *M(t)* is the moving average smoothing of order n.

In Fig. 5, the y-axis depicts the deaths of a patients, and x-axis depicts the time (months).

Simple exponential smoothing method

It might be reasonable to append bigger loads to later perceptions than to perceptions from the removed past. The method which works on this principle is called simple exponential smoothing (Ostertagova and Ostertag 2012).

Forecasts are determined by utilizing the weighted midpoints where the weights decline exponentially as perceptions

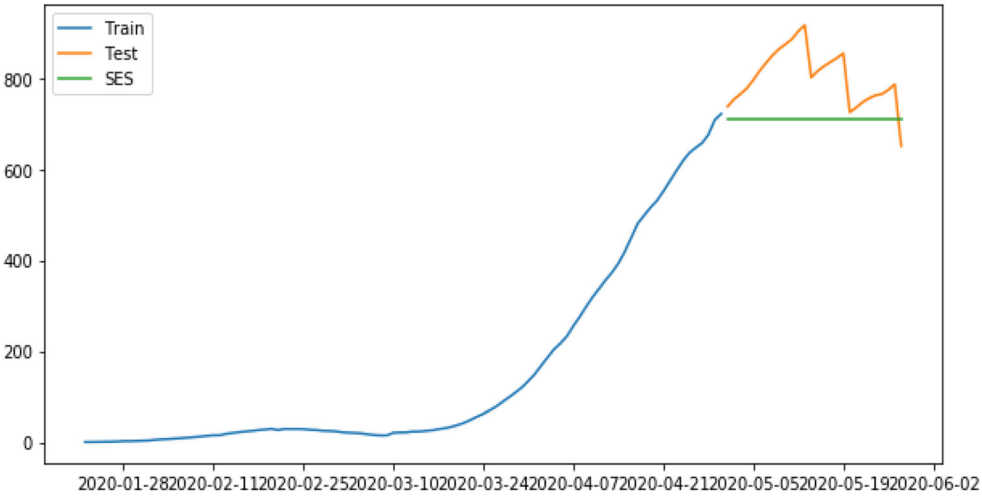


Fig. 6 Simple exponential smoothing forecast at test dataset

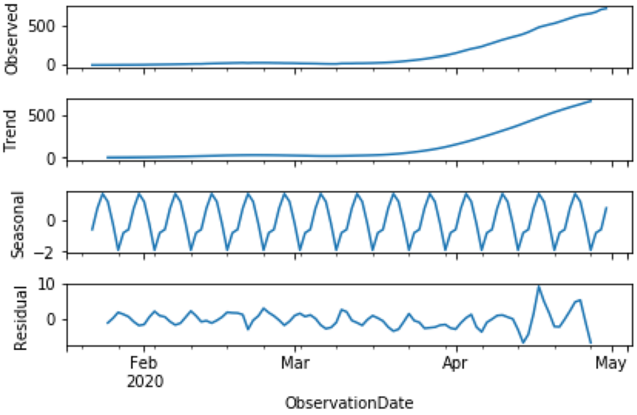


Fig. 7 Holt's pattern to estimate the future trend

originate from further before; the smallest weights are related with the most seasoned perceptions:

$$F_{t+1} = \alpha D_t + (1-\alpha)F_t$$

where D_t is the actual value; F_t is the forecasted value; α, the weighting factor, ranges from 0 to 1; and t is the current time period.

In Fig. 6 the y-axis depicts the deaths of a patients, and x-axis depicts the time (months).

Holt's linear trend method

We need a strategy that can delineate pattern precisely with no presumptions. Such a strategy that considers the pattern of the dataset is called Holt's linear trend technique (Yapar et al. 2018). Each time arrangement dataset can be disintegrated into its components which are trend, irregularity, and residual. Any dataset that follows a pattern can utilize Holt's direct pattern technique for forecasting.

We can see from Fig. 7 that this dataset follows an expanding pattern. Henceforth, we can utilize Holt's direct pattern to estimate the future trend.

Holt stretched out straightforward exponential smoothing to permit estimating of information with a pattern. It is just exponential smoothing applied to both level (the normal incentive in the arrangement) and pattern. To communicate this in numerical documentation, we currently need three conditions: one for level, one for the pattern, and one to consolidate the level and pattern to get the normal forecast y_t

For Forecast $F_t = \alpha y_t + (1-\alpha)F_t$

For Level $T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$

For Trend $F_{n+k} = L_n + k.T_n$

where α and β are two smoothing parameters, T_t is the trend at time t, y_t is a non-seasonal series but shows the trend, and k is the period of time.

The qualities we anticipated in the above algorithms are called level. In the over three conditions, we have added level and pattern to create the forecast condition.

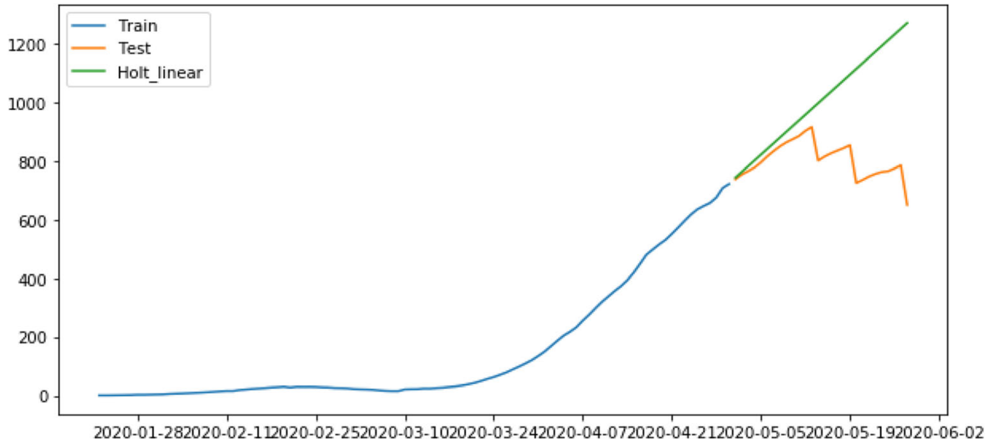
Likewise with straightforward exponential smoothing, the level condition here shows that it is a weighted normal of perception and the inside example one-step ahead (Fig. 8) the pattern condition shows that it is a weighted normal of the assessed pattern at time t dependent on $l(t) - l(t-1)$ and $b(t-1)$, the past estimate of the pattern.

Holt-Winters method

Holt's winter method, the triple exponential smoothing (Holt's winter), is to apply exponential smoothing to the occasional segments not withstanding level and pattern (Archibald and Koehler 2003).

Holt's winter technique utilizes the irregularity factor. The Holt-Winters occasional strategy contains the conjecture condition and three smoothing conditions —one

Fig. 8 Holt's pattern to forecast at test dataset



for the level L_t, one for pattern T_t, and one for the occasional segment meant by S_t, with smoothing parameters α, β, and γ.

$$\text{For Level} \quad L_t = \alpha(L_{t-1} + T_{t-1}) + (1-\alpha)y_t/S_{t-s}$$

$$\text{For Trend} \quad T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$$

$$\text{Seasonal} \quad S_t = \gamma S_{t-1} + (1-\gamma)y_t/L_t$$

where s is the length of the seasonal period, for $0 \le \alpha \le 1$, $0 \le \beta \le 1$, and $0 \le \gamma \le 1$.

In the following (Fig. 9), the level condition shows a weighted normal between the occasionally balanced perception and the non-occasional forecast for time t. The pattern condition is indistinguishable from Holt's direct strategy. The occasional condition shows a weighted normal between the ebb and flow occasional file and the occasional list of a similar season a year ago (i.e., s timeframes prior).

Root mean squared error (RMSE)

In regression line prediction, it is necessary to predict the average y value associated with a given x value and obtain a measure of the distribution of y values around this average value. To construct the RMS error first, we need to determine the residual error. The residual is the difference between the actual value and the predicted value (Barnston 1992). The RMS error may be positive or negative because the predicted value is lower or exceeds the actual value. To extract the RMS error, we must first square the residual, and then average the squared value, and finally taking the square root of the average value. Then, we use RMS error as a measure of the distribution of y values relative to the predicted y values:

$$\text{RMS Errors} = \sqrt{\sum_{i=1}^{n} \left(\widehat{y}_i - y_i\right)^2 / n}$$

where $\hat{y}_i$ observed value for i^{th} observation and y_i predicted value and n number of observations.

Table 1 Comparison of models by RMSE values on test data

Model	RMSE
Naïve method	99.9844
Simple average	655.4500
Moving average	565.8570
Simple exponential smoothing	110.0948
Holt's linear trend	277.1642
Holt's winter	236.4859

We can compare above models based on their RMSE scores in the following Table 1.

ARIMA

ARIMA involves techniques used to analyze timing information in order to obtain important insights and different attributes of information. ARIMA uses models to predict future values that depend on the most recently observed value. ARIMA has typical transient requirements. This makes the timeline survey unquestionable for cross-sectional inspections, in which there is no common requirement for perception. Scheduled surveys are especially special for spatial information surveys. In spatial information surveys, perception is usually consistent with geological regions. The stochastic model of periodic arrangements will generally reflect a way that time close to each other is more strongly related than time separated from each other. More importantly, the ARIMA model will periodically utilize the public single direction request of time, so the value for a given period of time to be delivered based on past value rather than future value. The ARIMA model is suitable for evidence that the information shows non-stationarity, and the basic difference step can be applied at least multiple times to eliminate non-stationarity (Kwiatkowski et al. 1992). Figure 10 shows the experimental design of the ARIMA model.

Fig. 9 Holt-Winters forecast at test dataset

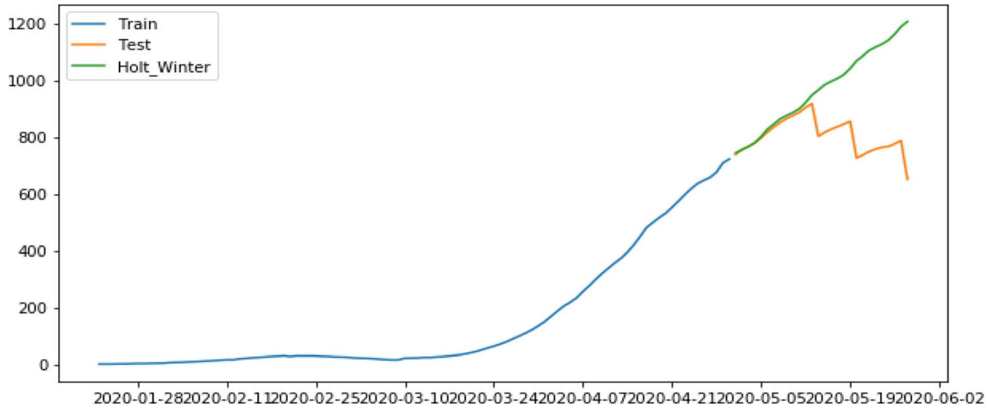
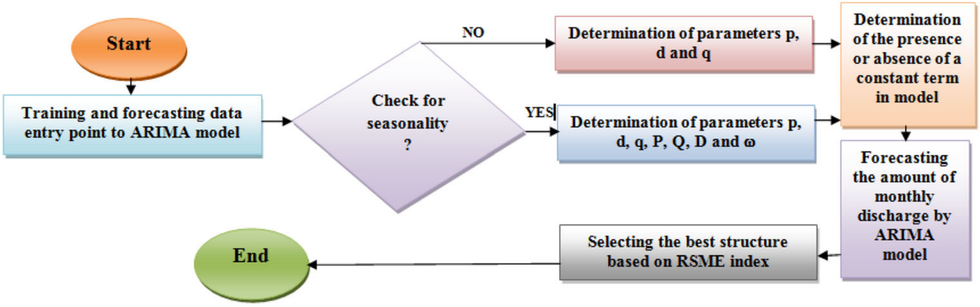


Fig. 10 Flowchart of calculation steps of ARIMA model


Start
Training and forecasting data entry point to ARIMA model
Check for seasonality ?
NO
YES
Determination of parameters p, d and q
Determination of parameters p, d, q, P, Q, D and ω
Determination of the presence or absence of a constant term in model
Forecasting the amount of monthly discharge by ARIMA model
Selecting the best structure based on RSME index
End

Autoregressive integrated moving average (ARIMA): while exponential smoothing models depended on a description of pattern and irregularity in the data, ARIMA models plan to depict the connections in the data with one another (De la Torre et al. 2003). An improvement over ARIMA is seasonal ARIMA. It considers the irregularity of dataset simply like Holt's winter strategy.

Expressed by y, the general prediction equation of ARIMA is:

$$\widehat{y}_t = \mu + \Phi_1 \, y_{t-1} + \ldots + \Phi_p y_{t-p} - \theta_1 e_{t-1} - \ldots - \theta_q e_{t-q}$$

The moving average parameters (θ) are defined here so that their sign is negative in the equation. The parameters are represented there by AR (1) and MA (1) in Table 2. Stationary series may still have autocorrelation errors, which indicates that certain number of AR items ($p \geq 1$) and/or some MA items ($q \geq 1$) are also required in the prediction equation.

The summary characterized that outcome from the yield of SARIMAX restores a lot of data on the table of coefficients (Tarsitano and Amerise 2017). The *coef* segment shows the weight (for significance) of each component and how everyone affects the time arrangement. The $P > |z|$ segment illuminates us regarding the significance of each component weight. Here, each weight has a p value lower or near 0.05, so it is sensible to hold every one of them in our model.

The following (Fig. 11) produces display and examine for any unusual conduct.

The diagnostic model above shows that the model residuals are based on the following normal distribution:

- In the histogram plus estimated density graph, the red KDE line immediately follows the $N(0,1)$ line, which is the standard symbol of the normal distribution with an average value of 0 and a standard deviation of 1. This indicates that the residuals are normally distributed.
- The QQ plot shows that the ordered distribution of residuals (blue dots) follows the linear trend of samples taken from the standard normal distribution with $N(0, 1)$. This strongly indicates that the residuals are normally distributed.
- There is no obvious seasonal variation in the standardized residuals over time; it seems to be white noise.

Table 2 SARIMAX results

	coef	Std. err	z	$P > \|z\|$	[0.025	0.975]
ar.L1	1.9024	5.909	0.322	0.747	− 9.679	13.484
ma.L1	65.0254	2.84e+04	0.002	0.998	− 5.55e+04	5.56e+04
sigma2	9.1957	8028.624	0.001	0.999	− 1.57e+04	1.57e+04

Despite the fact that we have a satisfactory fit, a few parameters of our seasonal ARIMA model could be changed to improve our model fit.

Forecasting visualization

In the last step, we portray in Fig. 12 our seasonal ARIMA time series model to forecast future values (Alquisola et al. 2018).

Both the numbers (forecasts and associated confidence interval) we generated and the associated confidence intervals can be used to further understand the time series. Our predictions indicate that we rely on time series to maintain a consistent growth rate.

As we further develop the future, it is normal to lose faith in our values. The confidence interval generated by our model reflects this, and as we move toward a farther future, the confidence interval will become larger and larger.

Discussion

The current model shows that the upcoming next few months will hard happen for the world. The control system adopted by the different national governments is indeed very strict and works well. In addition, adopting the direct mode can effectively supervise the recovered patients and also control the case fatality rate. If the government does not take strict control measures to its residents, the findings of this research may explode. The arrangement of emergency clinics and the improvement of the clinical office should be carried out as soon

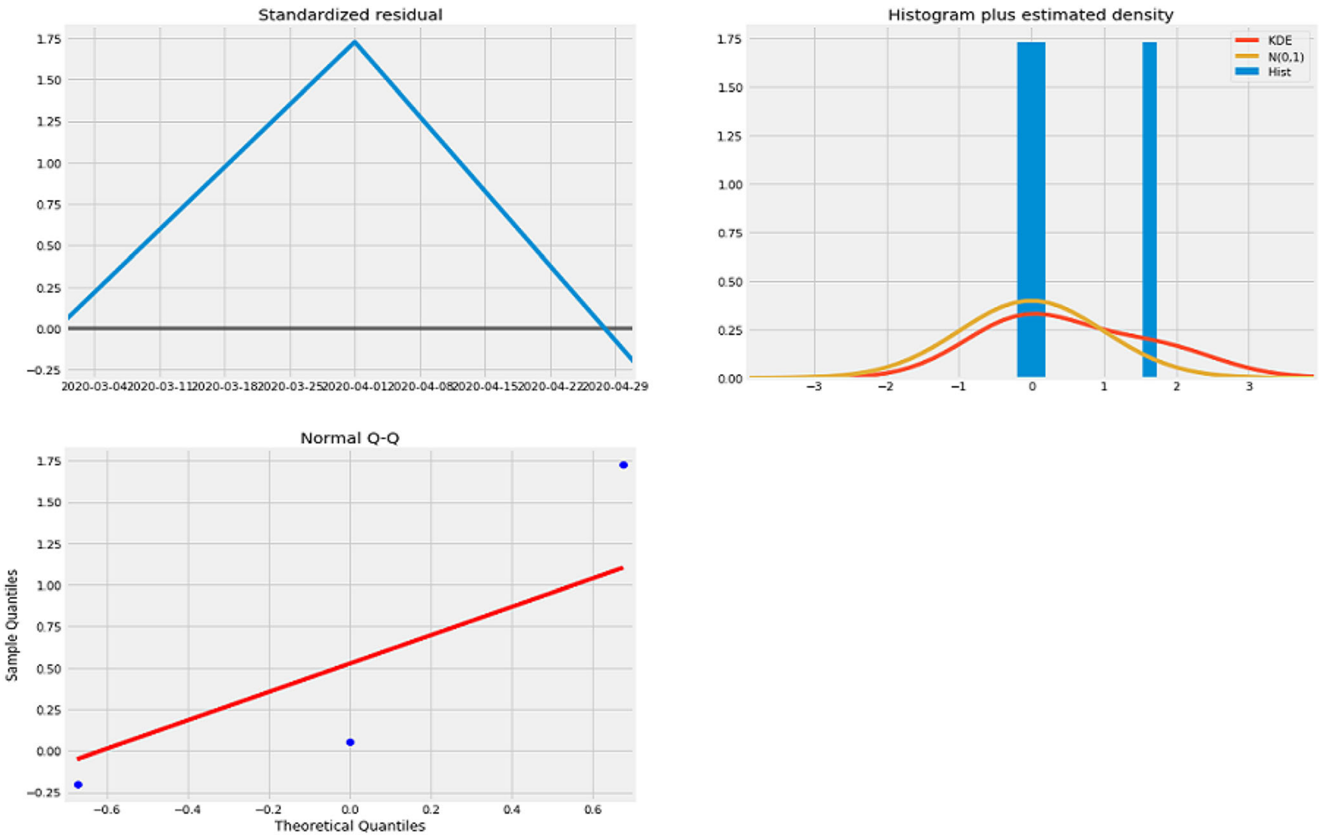


Fig. 11 Seasonal ARIMA diagnostics on dataset

as possible to establish an exponential development of the country to prevent this from happening.

Table 3 below mentioned that some researchers used the ARIMA model to describe some experimental models describing the COVID-19 pandemic. The basic motto of this comparison is to show that whether the predictions of these models are contemporary and fit our model.

In this table, we can see various methods and models proposed by different authors for predicting or forecasting the death cases, infection cases, and recovery cases of the global COVID-19 pandemic. These models have their own advantages and disadvantages in predicting global death cases. We have used several methods to observe deaths due to the COVID-19 pandemic. The data is unstable; it also shows that the number of deaths has increased exponentially since mid-March 2020. Another issue facing the study is insufficient training data. Four months of (January 2020 to April 2020) data are used for training purposes, 29 days of verification data, based on which the number of deaths can be determined expected in the coming months. There are very few training data for machine learning to train itself. Moreover, the number of infected people changes rapidly; the case occurred in mid-March.

By looking at the figure number (2–9), it is difficult to prove which method is suitable for this time series dataset in future

Fig. 12 Future values forecasts 2021 and beyond

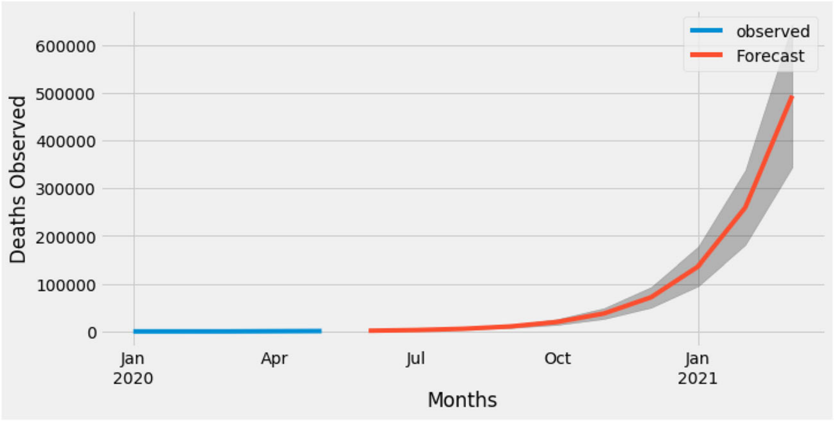


Table 3 Comparison of various models

Author	Used Model	Motive	Time duration/prediction	Best model	Prediction
Alzahrani et al. (2020)	Autoregressive model, moving average, a combination of both (ARMA), and integrated ARMA (ARIMA)	Forecasting COVID-19 cases on daily bases	March 2, 2020, to April 20, 2020/4 weeks	ARIMA model	New cases, 7668/day Cumulative daily cases, over 127,129/day
Ceylan (2020)	Autoregressive integrated moving average (ARIMA)	Predicting the prevalence of COVID-19	21 February 2020 to 15 April 2020	ARIMA model	Confirmed cases decreased in Italy–over 2000–4500 in last 10 days, deaths in Spain, 18,056 in the past 5 days; no downward trend in new confirmed cases in France
Chintalapudi et al. (2020)	ARIMA (1,2,0) for registered cases, ARIMA (3,2,0) for recovered cases	Forecasting of COVID cases	15 February 2020 to 31 March 2020/60 days	ARIMA (1,2,0) registered, 93.75% accuracy ARIMA(3,2,0) recovered, 84.4% accuracy	Rising in infected case Range (105,732–182,757) Increasing in recovered cases Range (16,742–81,635)
Chakraborty and Ghosh (2020)	Autoregressive integrated moving average model and wavelet-based forecasting model	Forecasts of the future COVID-19, risk assessment novel COVID-19	For India, January 30, 2020–April 4, 2020 For UK, January 31, 2020–April 4, 2020 For Canada, January 20, 2020–April 4, 2020 For France, January 25, 2020–April 4, 2020 For South Korea, January 26,2020–April 4, 2020/10 days ahead	Hybrid ARIMA-WBF model	Hybrid ARIMA-WBF model, superior for Canada, France, and the UK. ARIMA superior for India and South Korea
Gupta and Pal (2020)	ARIMA model, Holt's linear trends	Trend analysis and forecasting of COVID-19 outbreak	30th January 2020–24th March 2020/30 days	ARIMA and Holt's linear	Infected cases, 700 thousands in 30 days, 3 million infected case by Holt's linear

predictions. To overcome this situation, we describe the RMSE value of each method in Table 1. Compared with other methods, the naive method has a lower RMSE score of 99.98. Therefore, the naive method is suited in described all other methods. In the ARIMA model, using grid search, we identified a set of parameters that produced the best-fit model for our time series data. By continuing the model, future predictions of death cases indicate that the number of deaths will increase by 500,000 to more than 600,000 by January 2021 and beyond. But weather conditions, national geographic distribution, state-level population, and governance parameters may be affected the prediction. It can further improve the model prediction rate.

Conclusion

In this study, some AI models were used to decompose and predict the worldwide adjustment of COVID-19 mortality. We investigated this information and found that the number of deaths continued to increase from mid-March 2020. The results obtained from this inspection were taken from the information as of May 29, 2020. In addition, according to the ARIMA model, the number of death cases will definitely increase. Experts, welfare workers, and people who provide basic assistance types must be ensured according to the recommended clinical standards. Due to people's frivolous behavior, the disease spread later, just as infected people can multiply the number of cases. However, there may be different factors affecting any epidemic, among which population, geographical conditions, social distance not considered, lack of diagnostic facilities, lack of doctors and clinical staff, the most important thing is the willpower of the head of the country and the right decision made at the time. The peak has not yet arrived, so the government must be vigilant and insist on stringent measures. In addition, the arrangements for clinical clinics across the country must be greatly improved.

In the future, it should be ensure to create computerized calculations to provide information within a standard range and naturally predict the number of cases daily and every week. According to these principles, the government and emergency clinics can also keep a clear responsibility and provide flexible clinical help/services for new patients.

Authors' contribution Vikas Chaurasia developed the theoretical formalism, performed the analytic calculations, and performed the analysis under the supervision of Saurabh Pal.

Data availability Dataset is available on WHO website.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Code availability Code is available with author.

References

Alquisola GLV, Coronel DJA, Reolope BMF, Roque JNA, Acula DD. Prediction and visualization of the disaster risks in the Philippines using discrete wavelet transform (DWT), autoregressive integrated moving average (ARIMA), and artificial neural network (ANN). In 2018 3rd international conference on computer and communication systems (ICCCS) (pp. 146-149). IEEE. 2018.

Alzahrani SI, Aljamaan IA, Al-Fakih EA. Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions. J Infect Public Health. 2020;13(7):914–9.

Archibald BC, Koehler AB. Normalization of seasonal factors in winters methods. Int J Forecast. 2003;19:143–8.

Aslam M. Using the Kalman filter with ARIMA for the COVID-19 pandemic dataset of Pakistan. Data in Brief, 105854. 2020.

Barnston AG. Correspondence among the correlation, RMSE, and Heidke forecast verification measures: refinement of the Heidke score. Wea Forecast. 1992;7:699–709.

Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, Ciccozzi M. Application of the ARIMA model on the COVID-2019 epidemic dataset. Data in brief. 2020; 105340.

Ceylan Z. Estimation of COVID-19 prevalence in Italy, Spain, and France. Sci Total Environ. 2020;279:138817. https://doi.org/10.1016/j.scitotenv.2020.138817.

Chakraborty T, Ghosh I. Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: a data-driven analysis. Chaos Solitons Fractals. 2020;135:109850. https://doi.org/10.1016/j.chaos.2020.109850.

Chaurasia V, Pal S. COVID-19 pandemic: ARIMA and regression model-based worldwide death cases predictions. SN Comput Sci. 2020;1(5):1–12.

Chintalapudi N, Battineni G, Amenta F. COVID-19 disease outbreak forecasting of registered and recovered cases after sixty day lockdown in Italy: a data driven model approach. J Microbiol Immunol Infect. 2020;53(3):396–403.

De la Torre S, Conejo AJ, Contreras J. Simulating oligopolistic pool-based electricity markets: a multiperiod approach. IEEE Trans Power Syst. 2003;18(4):1547–55.

Duan X, Zhang X. ARIMA modelling and forecasting of irregularly patterned COVID-19 outbreaks using Japanese and South Korean data. Data in brief. 2020; 105779.

Frey BS, Weck H. Estimating the shadow economy: a "naive" approach. Oxf Econ Pap. 1983;35(1):23–44.

Genre V, Kenny G, Meyler A, Timmermann A. Combining expert forecasts: can anything beat the simple average? Int J Forecast. 2013;29(1):108–21.

Gupta R, Pal SK. Trend analysis and forecasting of COVID-19 outbreak in India. medRxiv. 2020.

Hamner L, Dubbel P, Capron I, Ross A, Jordan A, Lee J, et al. "High SARS-CoV-2 attack rate following exposure at a choir practice—Skagit County, Washington, march 2020" (PDF). MMWR Morb Mortal Wkly Rep. 2020;69(19):606–10. https://doi.org/10.15585/mmwr.mm6919e6.

Hopkins C, Kumar N. Loss of sense of smell as marker of COVID-19 infection. The Royal College of Surgeons of England: British Rhinological Society. 2020.

https://data.humdata.org/dataset/coronavirus-covid-19-cases-and-deaths. 2020.

Ilie OD, Cojocariu RO, Ciobica A, Timofte SI, Mavroudis I, Doroftei B. Forecasting the spreading of COVID-19 across nine countries from Europe, Asia, and the American continents using the Arima models. Microorganisms. 2020;8(8):1158.

Kırbaş İ, Sözen A, Tuncer AD, Kazancıoğlu FŞ. Comperative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches. Chaos Solitons Fractals. 2020; 110015.

Kwiatkowski D, Phillips PCB, Schmidt P, Shin Y. Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root? J Econ. 1992;54(1–3):159–78. https://doi.org/10.1016/0304-4076(92)90104-Y.

Nussbaumer-Streit B, Mayr V, Dobrescu AI, Chapman A, Persad E, Klerings I, et al. Quarantine alone or in combination with other public health measures to control COVID-19: a rapid review. Cochrane Database Syst Revi. 2020;4:CD013574.

Ostertagova E, Ostertag O. Forecasting using simple exponential smoothing. Acta Electrotech Informatica. 2012;12:62–6.

Pourghasemi HR, Pouyan S, Farajzadeh Z, Sadhasivam N, Heidari B, Babaei S, et al. Assessment of the outbreak risk, mapping and infection behavior of COVID-19: application of the autoregressive integrated-moving average (ARIMA) and polynomial models. PLoS One. 2020;15(7):e0236238.

Q & A on COVID-19. European Centre for Disease Prevention and Control. Retrieved 30 April 2020.

Q&A on corona viruses (COVID-19). World Health Organization (WHO). Archived from the original on 14 May 2020. Retrieved 14 May 2020. 2020.

Saboia JLM. Autoregressive integrated moving average (ARIMA) models for birth forecasting. J Am Stat Assoc. 1977;72(358):264–70.

Sahai AK, Rath N, Sood V, Singh MP. ARIMA modelling & forecasting of COVID-19 in top five affected countries. Diabetes Metab Syndr Clin Res Rev. 2020;14(5):1419–27.

Singh RK, Rani M, Bhagavathula AS, Sah R, Rodriguez-Morales AJ, Kalita H, et al. Prediction of the COVID-19 pandemic for the top 15 affected countries: advanced autoregressive integrated moving average (ARIMA) model. JMIR Public Health Surveill. 2020;6(2):e19115.

Symptoms of coronavirus. U.S. Centers for Disease Control and Prevention (CDC). Archived from the original on 30 January 2020. 2020.

Tarsitano A, Amerise IL. Short-term load forecasting using a two-stage sarimax model. Energy. 2017;133:108–14.

Velavan TP, Meyer CG. The COVID-19 epidemic. Tropical Med Int Health. 2020;25(3):278–80. https://doi.org/10.1111/tmi.13383. PMC7169770.

WHO The push for a COVID-19 vaccine, https://www.who.int/emergencies/diseases/novel-coronavirus-2019/covid-19-vaccines; accessed on 22 sep. 2020.

Williams B, Lacy PS, Yan P, Hwee CN, Liang C, Ting CM. Development and validation of a novel method to derive central aortic systolic pressure from the radial pressure waveform using an N-point moving average method. J Am Coll Cardiol. 2011;57(8):951–61.

Yang Q, Wang J, Ma H, Wang X. Research on COVID-19 based on ARIMA modelΔ-Taking Hubei, China as an example to see the epidemic in Italy. J Infect Public Health. 2020; 13(10):1415–8. https://doi.org/10.1016/j.jiph.2020.06.019.

Yapar G, Capar S, Selamlar HT, Yavuz I. Modified holt's linear trend method. Hacettepe J Math Stat. 2018;47(5):1394–1403.