

# Covid-19 Pandemic: ARIMA and Regression Model based Worldwide Death Cases Predictions

**Vikas Chaurasia**

Department of Computer Applications, Veer Bahadur Singh Purvanchal University, Jaunpur  
<https://orcid.org/0000-0002-3304-7433>

**Saurabh Pal** (✉ [drsaurabhpal@yahoo.co.in](mailto:drsaurabhpal@yahoo.co.in))

Department of Computer Applications, Veer Bahadur Singh Purvanchal University, Jaunpur  
<https://orcid.org/0000-0001-9545-7481>

---

## Research Article

**Keywords:** Covid-19, epidemic, humanity, breath, ARIMA model, Regression model, RMSE.

**Posted Date:** July 29th, 2020

**DOI:** <https://doi.org/10.21203/rs.3.rs-49697/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

**Version of Record:** A version of this preprint was published on August 29th, 2020. See the published version at <https://doi.org/10.1007/s42979-020-00298-6>.

# Abstract

Covid-19 has now taken a frightening form. As the day passes, it is becoming more and more widespread and now it has become an epidemic. The death rate, which was earlier in the hundreds, changed to thousands and then progressed to millions respectively. If the same situation persists over time, the day is not far when the humanity of all the countries on the globe will be endangered and we yearn for breath. From January 2020 till now, many scientists, researchers and doctors are also trying to solve this complex problem so that proper arrangements can be made by the governments in the hospitals and the death rate can be reduced. The presented research article shows the estimated mortality rate by the ARIMA model and the Regression model. This dataset has been collected precisely from DataHub-Novel Coronavirus 2019 - Dataset from 22<sup>nd</sup> January to 29<sup>th</sup> June 2020. In order to show the current mortality rate of the entire subject, the correlation coefficients of attributes (MAE, MSE, RMSE and MAPE) were used, where the average absolute percentage error validated the model by 99.09%. The ARIMA model is used to generate auto\_arima SARIMAX results, auto\_arima residual plots, ARIMA model results, and corresponding prediction plots on the training data set. These data indicate a continuous decline in death cases. By applying a regression model, the coefficients generated by the regression model are estimated, and the actual death cases and expected death cases are compared and analyzed. It is found that the predicted mortality rate has decreased after May 2, 2020. It will learn help the government and doctors prepare for the next plans. Based on short- period predictions these methods can use forecast for long-period.

## Introduction

As indicated by the World Health Organization, the CoViD-19 virus is a communicable disease that spreads from one person to another. Personal contact and small droplets in the breath of infected person can cause the virus to spread to others and cause severe acute respiratory syndrome [1]. In 2020 of January, WHO initially informed the humanity regarding pneumonia for obscure reasons, and the world came to know that this disease has spreading from person to person [2]. And the incident of this mysterious disease started from a city called Wuhan of China [3]. The status report of WHO say that till date from January 15 2020 to July 1, 2020; 507,435 people have lost their lives worldwide and this number is continuously increasing [4]. Group of RNA viruses are called corona virus [5]. In humans, it causes respiratory tract infections. It has SARS, MERS and COVID-19 deadly varieties. There are no immunizations or antiviral medications to protect from corona contamination in people yet [29]. Common symptoms include high fever, tiredness, cough, shortness of breath, loss of taste and smell, and complications include pneumonia and acute breathing infection [6]. The confirmation of the first death case from the Asia to Europe was: initial confirmed casualty in Wuhan, China on 9<sup>th</sup> of January, 2020, initial confirmed decease in Philippines outside the China on February 1, 2020, and on February 14, 2020 the first confirmed death in European country France [7]. The ratio of death rate is 5.4% till 16 June 2020 against 437,283 deaths for 8,051,732 cases. This number may vary from time to time and region to region [8].

The motto of our research is to predict the future of death cases based on machine learning regression model as well as time series ARIMA model. Both models are used to predict the future values. No eye-catching and extensively tested antibody against CoViD-19 has been designed more importantly; the key part of the subsequent response to this pandemic is to reduce the top of the plague, or to smooth the pandemic curve. The work of information researchers and information mining analysts is to coordinate relevant information and it is more likely to innovate to understand the infection and its quality, which helps to make the right choices and solid activities. This will prompt people to take stronger measures to establish frameworks, medicine, antibodies and control of comparable plagues with greater prospects. The aim of the present investigation are as follows.

1. The current mortality rate and corresponding measures are displayed on the graph.
2. Extracting different statistical measures from these graphs.
3. Using ARIMA and regression models to predict and find future deaths worldwide.

Except for the first part of the introduction, the structure of the remaining research papers is as follows:

Section II: Background, Section III: Methodology, Section IV: Dataset description and analysis, Section V: Results and discussion, Section VI: Conclusion.

## Background

Although it has been about 6 months since the Covid-19 pandemic has spread, many researchers have done a lot of work on it and it is being worked on continuously. The following is a description of some of the researches presented.

Benvenuto D., Giovanetti M., Vassallo L., Angeletti S., Ciccozzi M. [9] Carried out an ARIMA model forecast on COVID-2019 information gathered from Johns Hopkins epidemiological of the predominance and rate. For additional correlation or from a future point of view, case definitions and information assortment must be kept up progressively.

Zeynep Ceylan [10] Comprehensive information related to CoViD-19 was collected from WHO website from Feb. 21 to April 15, 2020. Some ARIMA models with different ARIMA boundaries were selected. Which includes ARIMA (0,2,1) for the lowest MAPE (4.7520) for Italy, similarly for Spain and France were selected separately with ARIMA (1,2,0) and ARIMA (0,2,1) and lowest MAPE (5.58486) and (5.6335) respectively. This test shows that ARIMA modal is appropriate to understand the effect of CoViD-19. The aftereffects of the examination can reveal insight into understanding the patterns of the episode and give a thought of the epidemiological phase of these locales.

MHDM R., Silva R.G., Mariani V.C., Coelho L.S. [11] For the purpose of time series analysis, different models are used like ARIMA, CUBIST, RF, RIDGE, SVR and stacking-ensemble method are assessed. The created models can produce exact forecasting, accomplishing mistakes in a scope of 0.87%–3.51%, 1.02%–5.63%, and 0.95%–6.90% in one, three, and six-days-ahead, separately. The positioning of models,

from the best to the most noticeably terrible with respect to precision, in all situations is SVR, stacking-gathering learning, ARIMA, CUBIST, RIDGE, and RF models.

Pandey, G.; Chaudhary, P.; Gupta, R.; Pal, S. [12] In this inspection, until March 30, 2020, this suffering scene in India has been meticulous, and the number of cases in the next 14 days was evaluated. Taking into account the data accumulated from the Johns Hopkins University depository in the period from January 30, 2020 to March 30, 2020, the SEIR model and the regression model were used. RMSLE evaluated the introduction of the model, and the data of the SEIR model were 1.52 and 1.75, respectively. For the rear slip model. The RMSLE tightening rate between the SEIR model and the regression model is 2.01. In addition, the estimation of  $R_0$  as the diffusion of pollution was analyzed to 2.02. It is foreseeable that in the next 14 days, the number of cases may rise among 5000-6000.

Chakraborty T., Ghosh I. [13] Collect the data as of April 4, 2020, it has caused a pandemic flare-up with in excess of 11,16,643 affirmed diseases and in excess of 59,170 revealed passings around the world. The fundamental focal point of this paper is two-overlap: (a) producing present moment (constant) estimates of things to come COVID-19 cases for various nations; (b) chance evaluation of the novel COVID-19 for some significantly influenced nations. To take care of the primary issue, they introduced a half breed approach dependent on ARIMA model and Wavelet-based forecasting model that can create present moment (ten days ahead) conjectures of the quantity of day by day affirmed cases for Canada, France, India, South Korea, and the UK. They applied an ideal relapse tree calculation to discover basic causal factors that altogether influence the case casualty rates for various nations.

Chintalapudi N, Battineni G, Amenta F. [14] From mid-February to the end of March, CoViD-19, which experienced tolerance data, deleted cases registered and restored on-site by the Italian Ministry of Health. Appointment of the accidental ARIMA vision group using R real model was completed. The accuracy of the enrollment case model reached 93.75%, and the accuracy of the recovery case model reached 84.4%. At the end of May, a decision of a serious patient may occur, estimated at 182,757, and the recovered case may include an estimated 81,635. Their findings indicate that it is possible to reduce enrollment cases by approximately 35% and improve recovery cases by approximately 66%.

Vardavas CI, Nikitara K. [15] From March 18, 2020, a total of 194909 COVID-19 representatives participated, including 7876 passes, a large part of which were in China (3242) and Italy (2505). In their multivariate key backslip test, chronic diseases of smoking are risk factors for disease development (OR = 14.28; 95% CI: 1.58-25.00;  $p = 0.018$ ). In the scattered data, they found that smokers are increasingly 1.4 events. Possible (RR = 1.4, 95% CI: 0.98-2.00) with abnormal symptoms of COVID-19, usually 2.4 events will inevitably be sent to the ICU, requiring mechanical ventilation or passage, which is different from non-smokers (RR = 2.4, 95% CI: 1.43-4.04).

Yan CH, Faraji M, Prajapati DP, Boone CE. [16] The calculated relapse rate is used to sense signs associated with CoViD-19 positive. Between March 3, 2020 and March 29, 2020, a total of 1,480 patients with influenza-like reactions underwent the CoViD-19 test. Our assessment yielded 59 out of 102 (58%) CoViD-19 positive patients and 203 out of 1378 COVID-19 negative patients (15%). CoViD-19 positive

subjects accounted for 68% (40/59) and 71% (42/59) of odor and taste events, respectively, and 16% (33/203) and 17% (35/203) of subjects different. CoViD-19 negative patients ( $p < 0.001$ ). In addition, odor incontinence and COVID-19 motives persisted without restraint (insomnia: adjusted possibility range [aOR] 10.9; 95% CI, 5.08-23.5; age imbalance: aOR 10.2; 95% CI, 4.74-22.1 ), but the sore throat is related to the COVID-19 enemy (aOR 0.23; 95% CI, 0.11-0.50). Of the patients who reported loss of olfaction associated with COVID-19, 74% (28/38) of the patients found insomnia with clinical goals.

## Methodology

In this section, we collected data from DataHub-Novel Coronavirus 2019-Dataset. The data set includes information on patients with COVID-19 dated from January 22, 2020 to June 20, 2020. The data set has the attributes of globally confirmed cases, rehabilitation cases, death cases and COVID-19 prevalence. There are basically two methods for analyzing the outbreak of a pandemic. Both ARIMA and regression models are used to predict future value. In this sense, we have basically analyzed the correlation between mortality and all precious attributes.

### ARIMA Model

Since the administrator needs to carefully consider the time of sick leave, this exploratory paper proposes an inspection of the autoregressive merged moving normal model. The ARIMA model is additionally utilized as a proficient device to design assets, for example, pandemic and groups for the crisis department [17, 18]. Another relevance of the ARIMA model is to foresee and contemplate the impact of COVID-19 [19-21].

Time Series Forecasting based specific sort of forecasting strategy called ARIMA modeling. ARIMA or "Auto Regressive Integrated Moving Average" is really a class of models that clarifies a given time arrangement dependent on its own past qualities, that is, its own slacks and the slacked forecast errors, with the goal that condition can be utilized to figure future values. Mathematically non seasonal ARIMA model define as:

An ARIMA model is portrayed by 3 terms-  $p, d, q$

Where,

$p$  – Order for the Auto Regressive expression

$q$  – Order for the Moving Average expression

$d$  – Number of differencing required making the time arrangement fixed.

The estimation of  $d$  is the base number of differencing expected to make the differencing fixed. What's more, on the off chance that the time differencing is now fixed, at that point  $d = 0$ .

“p” is the request for the AR term. It alludes to the quantity of slacks of Y to be utilized as indicators. Furthermore, “q” is the request for the MA term. It alludes to the quantity of slacked forecast errors that ought to go into the ARIMA Model.

An unadulterated AR model is one where  $Y_t$  relies just upon its own slacks. That is,  $Y_t$  is an element of the 'slacks of  $Y_t$ '.

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_p Y_{t-p} + \epsilon_t$$

Where,  $Y_{t-1}$  is the lag1 of the arrangement,  $\beta_1$  is the coefficient of lag1 that the model evaluations and  $\alpha$  is the block term, additionally assessed by the model.

Moreover an unadulterated MA model is one where  $Y_t$  relies just upon the slacked forecast errors.

$$Y_t = \alpha + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \dots + \phi_q \epsilon_{t-q}$$

Where, the error terms are the errors of the autoregressive model of the particular slacks. The mistakes  $\epsilon_t$  and  $\epsilon_{t-1}$  are the errors from the accompanying conditions:

$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_0 Y_0 + \epsilon_t$	AR Model
$Y_{t-1} = \beta_1 Y_{t-2} + \beta_2 Y_{t-3} + \dots + \beta_0 Y_0 + \epsilon_{t-1}$	MA Model

An ARIMA model is one where the time series was differenced in any event once to make it fixed and we consolidate the AR and the MA expressions. So the condition becomes-

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_p Y_{t-p} + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \dots + \phi_q \epsilon_{t-q}$$

There are some huge differences in the arrangement of the explanation model, which is the premise of at least 50 kinds of recognition. In order to reduce the burden on officials, almost no large amounts of data are needed before the critical month of mediation. After that, the model can place inevitable models, which may interfere with the entire range of boundary activities [22].

## Regression Model

Linear regression is a prescient measurable methodology for displaying connection between a dependent variable with a given arrangement of autonomous factors. It is a direct way to deal with displaying the connection between a dependent variable and at least one independent variable. At the point when we have just a single independent variable it is as called straightforward linear regression. For more than one independent factor, the procedure is called multiple linear regressions. This investigation has utilized linear regression and multiple regressions for expectation of CoViD-19 cases [23].

The linear regression description includes a linear condition that adds a specific information literacy particular arrangement  $x$ , whose response is the predictable return  $y$  of the data particular arrangement ( $y$ ). The linear condition gives each information value or part a scale factor, called the coefficient, which is represented by the Greek word Beta " $\beta$ ". Including an additional coefficient in the same way provides additional degrees of freedom for the line and is repeatedly called the intercept or offset coefficient.

In a straightforward regression issue, the type of the model would be:

$$y = \beta_0 + \beta_1 x$$

Where,

$\beta_0$  – intercept

$\beta_1$  – coefficient

$x$  – independent variable

$y$  – dependent variable

In higher estimates, when we have multiple information  $x$ , the line is called a plane or hyperplane. Described in this way are the kinds of conditions and specific characteristics for the coefficients ( $\beta_0$  and  $\beta_1$ ).

The General condition for a multiple linear regression with  $n$ , independent factors is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon$$

Where,

$\beta_0, \beta_1, \beta_2, \dots, \beta_n$  – coefficients

$x_1, x_2, \dots, x_n$  –  $x$ -variable

$y$  –  $y$ -variable

$\epsilon$  – random error "noise"

## Dataset description and analysis

The Covid-19 data set is taken from the DataHub-Novel Coronavirus data set from January 22, 2020 to June 29, 2020. It contains five independent attributes, such as date, confirmed cases, rehabilitation cases, death and growth rate, and 160 instances. As we have seen in the data set, the death toll has increased over time until June 29. This is confirmed by the following figure 2.

# Results And Discussion

The earliest Covid-19 patients were recorded in the data set on January 22, 2020. We have taken examples from January 22, 2020 to June 29, 2020. It consists of 160 instances and five attributes. These attributes have information about the date of recording, confirmed cases, recovered cases, deaths, and growth rates related to CoViD-19 patients. The following estimates are made from the data set to explore and extract useful information.

## Correlation coefficients

The statistical measure correlation coefficient is the strength of the relationship between the relative motions of two variables. The range is defined as -1 to +1. Incorrect correlation measurement occurs when values greater than +1 and less than -1. The correlation measurement at -1 is completely negative, the correlation measurement at +1 is positive, and the value at 0.0 is the nonlinear relationship between the two variables [24].

Related statistics can be used to define the relationship between different attributes of the disease. A correlation coefficient can be calculated to determine the correlation level between the confirmed cases and the recovered cases under the current pandemic situation and the rate of increase in deaths and mortality, as shown in Table 1 and Figure 3. We found that in Covid-19 confirmed case and recovered case the correlation between these two variables is highly positive.

Table 1: Correlation Coefficients of attributes

	Confirmed	Recovered	Deaths	Increase rate
Confirmed	1.000000	0.986051	0.988177	-0.378478
Recovered	0.986051	1.000000	0.950569	-0.337027
Deaths	0.988177	0.950569	1.000000	-0.401742
Increase rate	-0.378478	-0.337027	-0.401742	1.000000

## ARIMA Model Results

In the ARIMA model, we choose the parameters  $p, d, q$  [28]. For this reason, even without drawing graphics, we use `auro_arma` to find the appropriate parameters. The `auro_arma` work works by directing differencing tests like Kwiatkowski–Phillips–Schmidt–Shin, Augmented Dickey-Fuller or Phillips– Perron to decide the request for differencing,  $d$ , and afterward fitting models inside scopes of characterized  $start_p, max_p, start_q, max_q$  ranges [25]. In the event that the occasional discretionary is empowered, `auto_arma` likewise tries to distinguish the ideal  $P$  and  $Q$  hyper-boundaries in the wake of directing the Canova-Hansen to decide the ideal request of occasional differencing,  $D$ . The following figure 4 shows the parameters obtained by the `auro_arma` model.

When viewing the residual plot from the `auto_arma` model, as shown in Figure 5.

The output of the auto\_arma model is explained as follows:

**Standardized residual:** The error of the residual is near the mean of the zero line and has a uniform variance.

**Histogram and density plot:** In the figure below, the density plot shows the equal distribution around the zero line average.

**QQ-plot:** In the QQ chart, all blue dots (ordered distribution of residuals) are on the red line, and any deviations will be skewed by the line. It is usually distributed along  $N(0, 1)$  and is considered to be uniformly distributed.

**Correlogram:** Correlogram or ACF plots show that the residual error isn't autocorrelated. Any autocorrelation implies that Residual error.

The optimal values of  $p$ ,  $d$ , and  $q$  obtained by the auto\_arima model are 1, 2, and 2, respectively. Now, using the best parameters obtained (1, 2, 2) to create an ARIMA model, the results are shown in figure 6.

Figure 6 above shows the importance of the ARIMA model. In this figure, we will focus on the coefficient table. The coef section shows the weight of each element and how each element affects the time series.  $P > |z|$  this section provides advice on the importance of the weight of each element. Here, the p-value of each weight is less than or close to 0.05, so it is wise to include each weight in our model.

These views make us think that our model can create a good fit, which can help us understand time series information and calculate future value. Although we have a reasonable fit, we can occasionally change some limitations of the ARIMA model to improve the model's aggressiveness. We have obtained a model for the time series and can now use it to create estimates [26]. We first compare the predicted value with the actual estimated value of the time series, which will help us understand the accuracy of the prediction. The numbers and associated confidence intervals we have now created can now be used to additionally understand time series and predict what to store. Our data shows that relying on time series can maintain a consistent growth rate.

As our predictions for the future say, it is normal to be less optimistic about our values. This is reflected by the deterministic interval generated by our model, as we further develop, the deterministic interval will become larger and larger. We start predicting death cases in a test data set that maintains 95% confidence. Figure 7 below shows the prediction results.

In the figure below, the actual death of the training data set is shown by the blue line, and the predicted death is shown by the red line. The prediction of death on the red line has dropped, which means that in the future, the incidence of deaths will become shorter and shorter, as more and more people recovered quickly, and people maintained the social distance in this pandemic situation.

By using statistical data, we created summary metrics that classify and collect residuals into single value, which are related to the model's a predictive ability.

In order to judge the prediction results, let us apply commonly used accuracy indicators, the results are shown in table 2.

Table 2: Correlation Coefficients of attributes

Measures of Accuracy	Value
Mean Absolute Error (MAE)	0.12044588473307338
Mean Squared Error (MSE)	0.023012953284359018
Root Mean Squared Error (RMSE)	0.15170020858376898
Mean Absolute Percentage Error (MAPE)	0.009196691386663233

The MAE of our model is 0.1204, which is quite small suppose our data death case starts at 0.01.

For MSE, the value 0.0230 is less than MAE. We found this to be the case: MSE is an order of magnitude smaller than MAE.

The value 0.1517, of RMSE is similar to standard deviation and is a measure of how much the residual distribution is.

Around 0.91% MAPE implies the model is about 99.09% accurate in predicting the test set observations.

## Regression Model Results

In order to find out which factor has the most significant influence on the forecasted output and how the various factors identify each other, we will consider different input functions such as "confirmation case", "recovered case" and "increase rate". Based on these characteristics, we will predict the deaths of Covid-19 patients. The data set splited into 80%:20% as training and testing respectively.

In multiple linear regression, then regression the model has selected the best coefficients for all attributes [27]. The coefficients of the regression model are shown in Table 3 below.

Table 3: coefficients of regression model

Attributes	Coefficient
Confirmed	0.103305
Recovered	-0.100568
Increase rate	69.616876

From the table 3, it is clear that if increase in “recovered case” by 1 unit, there is decrease of “death case” by 0.1005 units vice versa. Similarly, increase in “confirmed case” and “increase rate” by 1 unit, there is increase in “death case” by 0.1033 units and 69.6168 units respectively.

Now we predict the test data to check the difference between the actual value and the predicted value in Table 4 below.

Table 4: Difference between the actual value and predicted value

Instance Number	Actual Value	Predicted Value
110	286697	221975.301362
112	297539	286646.565236
143	430047	423127.482077
7	133	-6528.684075
44	3459	-2713.950271
101	244129	236968.993751
122	342565	329894.990367
66	31990	47224.597929
85	148157	160515.287829
86	157022	167041.159151
133	386298	376198.729391
92	193926	198189.689192
26	1868	-1385.556916
146	443685	438945.896459
119	328483	318945.015040
62	19026	25233.066196
51	5411	808.770349
97	221109	221511.564448

128	365380	355638.073651
90	180475	187102.115303

When plotting and comparing the actual value and the predicted value, as shown in Figure 8.

As shown in the multiple regression model shown in Table 4 and Figure 8, the initial predicted number of deaths has increased compared with actual deaths, but as we progress in the data table, compared with actual deaths, the predicted deaths the number has decreased from the month of May 2<sup>nd</sup> 2020.

Overall, this study shows that the reduction in deaths worldwide is a good sign for human society.

## Conclusion

In this study, two AI models ARIMA and regression models were used to decompose and predict changes in the spread of CoViD-19 infection. We have investigated this information and predicted that the number of deaths will be reduced compared to the overall situation. The decline shown in the ARIMA model graph (Figure 7) indicates that the future mortality rate will decrease (based on the current situation). The training data set verified by the mean absolute percentage error (MAPE = 99.09%) indicates the accuracy of the model. The regression model also indicated an increase in the initial number of deaths, but over time, it predicted fewer deaths than actual deaths (Table 4 and Figure 8) from 2<sup>nd</sup> May 2020.

Based on the above results and discussion, through ARIMA and regression models, we can conclude that there is a possibility of reducing deaths worldwide and should be reduced. Over time, there must be new opportunities to deal with this pandemic. Many researchers, scientists, doctors, nurses, medical support staff, and government agencies are all playing their roles. However, we ourselves have a responsibility to follow the guidelines provided by these agencies. If we do not maintain social estrangement, gather in public places, and do not keep the neighborhood clean, how can we overcome the CoViD-19 pandemic?

## Declarations

### Competing interests:

The authors declare no competing interests.

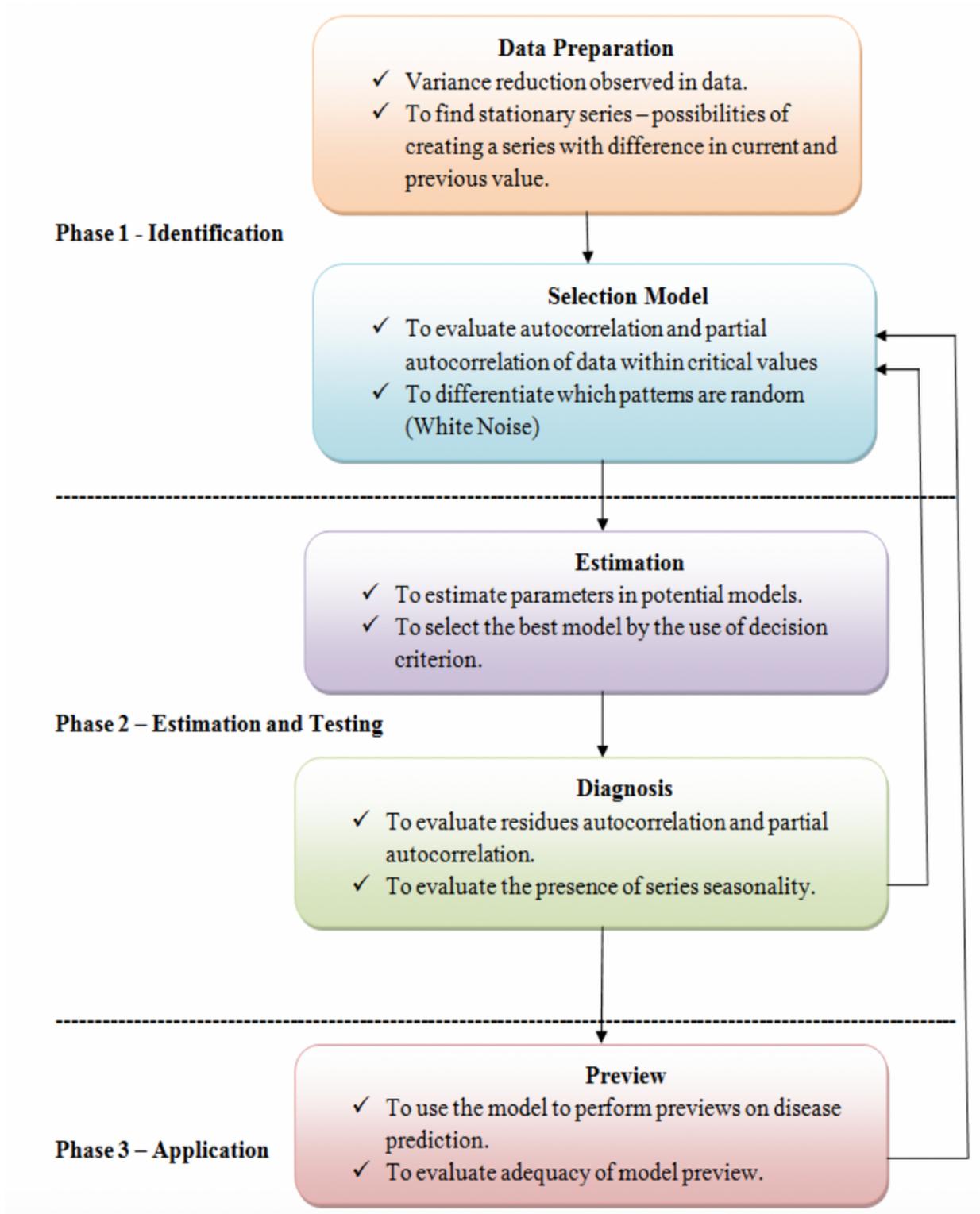
## References

1. World Health Organization. Coronavirus disease 2019 (COVID- 19): situation report, 67, 2020.
2. Kessler, Glenn. "Trump's false claim that the WHO said the coronavirus was 'not communicable'". The Washington Post. Archived from the original on April 17, 2020. Retrieved April 17, 2020.

3. "WHO: Pneumonia of unknown cause – China". WHO. Archived from the original on 7 January 2020. Retrieved 9 April 2020.
4. Coronavirus disease (COVID-19). Situation Report – 147 Data as received by WHO from national authorities by 10:00 CEST, 15 June 2020.
5. Woo PC, Huang Y, Lau SK, Yuen KY. "Coronavirus genomics and bioinformatics analysis". *Viruses*. 2 (8): 1804–20. doi:3390/v2081803, (August 2010).
6. "Coronavirus Disease 2019 (COVID-19)—Symptoms". U.S. Centers for Disease Control and Prevention(CDC). 20 March 2020. Retrieved 21 March 2020.
7. "Coronavirus Live Updates: First Death Outside Asia Reported in France". *The New York Times*. 15 February 2020. Retrieved 15 February 2020.
8. "COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)". ArcGIS. Johns Hopkins University, 2020.
9. Benvenuto D., Giovanetti M., Vassallo L., Angeletti S., Ciccozzi M. Data in brief application of the ARIMA model on the COVID- 2019 epidemic dataset. *Data Br.* 2020;29 doi: 10.1016/j.dib.2020.105340, 2020.
10. Zeynep Ceylan , "Estimation of COVID-19 Prevalence in Italy, Spain, and France", PMID: 32360907 PMCID: PMC7175852 , DOI: 10.1016/j.scitotenv.2020.138817, 2020.
11. MHDM R., Silva R.G., Mariani V.C., Coelho L.S. Short-term forecasting COVID-19 cumulative confirmed cases: perspectives for Brazil. *Chaos Solitons Fractals*. doi: 10.1016/j.chaos.2020.109853, 2020.
12. Pandey, G.; Chaudhary, P; Gupta, R.; Pal, S. SEIR and Regression Model based COVID-19 outbreak predictions in India. *arXiv* 2020, arXiv:2004.00958.
13. Chakraborty T., Ghosh I. Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: a data-driven analysis. *Chaos Solitons Fractals*. 2020 doi: 10.1016/j.chaos.2020.109850.
14. Chintalapudi N, Battineni G, Amenta F. COVID-19 virus outbreak forecasting of registered and recovered cases after sixty day lockdown in Italy: A data driven model approach. *J Microbiol Immunol Infect* [Internet]. 2020 Apr 13
15. Vardavas CI, Nikitara K. COVID-19 and smoking: a systematic review of the evidence. *Tob Induc Dis.* 2020;18:20. doi:18332/tid/119324.
16. Yan CH, Faraji M, Prajapati DP, Boone CE. Association of chemosensory dysfunction and Covid-19 in patients presenting with influenza-like symptoms. *Int Forum Allergy Rhinol.* (in press). Epub 12 April 2020. <https://doi.org/10.1002/alr.22579>.
17. Sun Y, Heng B, Seow Y, Seow E. Forecasting daily attendances at an emergency department to aid resource planning. *BMC Emerg Med.* 2009;9:1–1.
18. Rathlev NK, Chessare J, Olshaker J, Obendorfer D, Mehta SD, Rothenhaus T, et al. Time series analysis of variables associated with daily mean emergency department length of stay. *Ann Emerg Med.* 2007; 49(3):265–271.

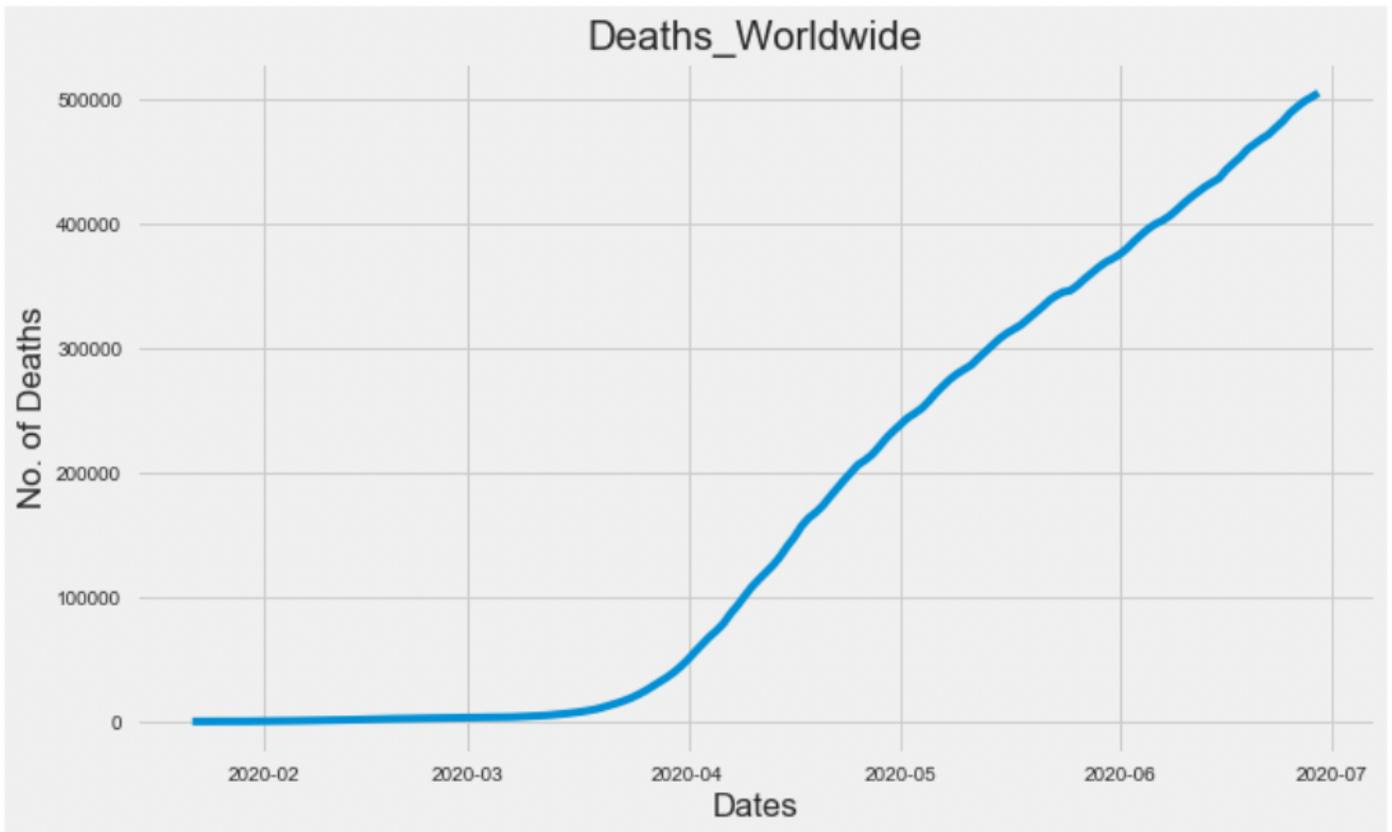
19. López-Lozano JM, Monnet DL, Yagüe A, Burgos A, Gonzalo N, Campillos P, et al. Modelling and forecasting antimicrobial resistance and its dynamic relationship to antimicrobial use: a time series analysis. *Int J Antimicrob Agents*. 2000; 14(1):21–31.
20. Hsueh PR, Chen WH, Luh KT. Relationships between antimicrobial use and antimicrobial resistance in Gram-negative bacteria causing nosocomial infections from 1991–2003 at a university hospital in Taiwan. *Int J Antimicrob Agents*. 2005; 26(6):463–472.
21. Aldeyab MA, Monnet DL, López-Lozano JM, Hughes CM, Scott MG, Kearney MP, et al. Modelling the impact of antibiotic use and infection control practices on the incidence of hospital-acquired methicillin-resistant *Staphylococcus aureus*: a time series analysis. *J Antimicrob Chemother*. 2008; 62(3):593–600.
22. Linden A, Adams JL, Roberts N. Evaluating disease management program effectiveness: an introduction to time series analysis. *Dis Manag*. 2003; 6(4):243–255.
23. Pandey, G.; Chaudhary, P.; Gupta, R.; Pal, S. SEIR and Regression Model based COVID-19 outbreak predictions in India. *arXiv* 2020, arXiv:2004.00958.
24. Meng, X., Rosenthal, R., & Rubin, D. B.. Comparing correlated correlation coefficients. *Psychological Bulletin*, 111, 172–175, 1992.
25. Manoj K, Madhu A. An Application Of Time Series Arima Forecasting Model For Predicting Sugarcane Production In India[J]. *Studies ITI Business and Economics*, 9(1): 81-94, 2014.
26. Calheiros, R. N., Masoumi, E., Ranjan, R., and Buyya, R. Workload Prediction Using ARIMA Model and its Impact on Cloud Applications' QoS. *IEEE Transactions on Cloud Computing* 3, 4 (2015).
27. Catalina T, Iordache V, Caracaleanu B. Multiple regression model for fast prediction of the heating energy demand. *Energy and Buildings*. 57:302-12; 2013 Feb 28.
28. Chaurasia, V., Pal, S., Covid-19 Pandemic: Application of Machine Learning Time Series Analysis for Prediction of Human Future. Preprint at <https://www.researchsquare.com/article/rs-39149/v1> (2020).
29. Singh, P, Chaurasia, V. Era of Covid-19 Pandemic: Yoga contemplation and medical mystery. *Turkish Journal of Kinesiology*, 6(2), 88-100. (2020). DOI: 10.31459/turkjin.745955.

## Figures



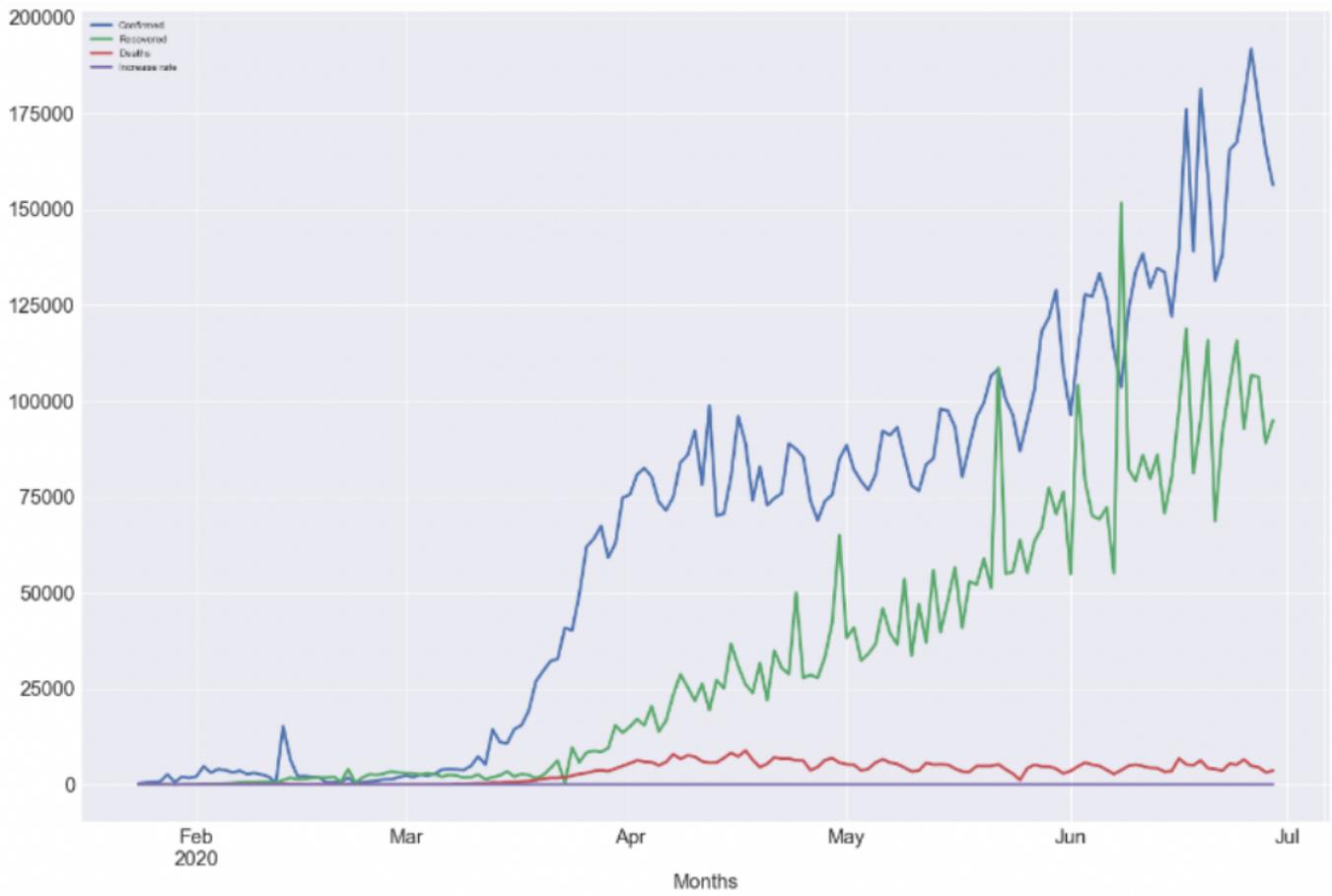
**Figure 1**

shows a schematic diagram of the ARIMA strategy estimation model. The general association of non-incident models is ARIMA (p, d, q).



**Figure 2**

Confirmed death cases worldwide increasing continuously



**Figure 3**

Visual presentation of correlated coefficients

```

Performing stepwise search to minimize aic
Fit ARIMA(0,2,0)x(0,0,0,0) [intercept=True]; AIC=-407.745, BIC=-401.876, Time=3.646 seconds
Fit ARIMA(1,2,0)x(0,0,0,0) [intercept=True]; AIC=-453.845, BIC=-445.042, Time=0.758 seconds
Fit ARIMA(0,2,1)x(0,0,0,0) [intercept=True]; AIC=-476.274, BIC=-467.471, Time=0.898 seconds
Fit ARIMA(0,2,0)x(0,0,0,0) [intercept=False]; AIC=-409.562, BIC=-406.628, Time=1.134 seconds
Fit ARIMA(1,2,1)x(0,0,0,0) [intercept=True]; AIC=-478.707, BIC=-466.969, Time=0.965 seconds
Fit ARIMA(2,2,1)x(0,0,0,0) [intercept=True]; AIC=-484.396, BIC=-469.723, Time=1.859 seconds
Fit ARIMA(2,2,0)x(0,0,0,0) [intercept=True]; AIC=-480.718, BIC=-468.980, Time=0.466 seconds
Fit ARIMA(3,2,1)x(0,0,0,0) [intercept=True]; AIC=-477.908, BIC=-460.301, Time=1.025 seconds
Fit ARIMA(2,2,2)x(0,0,0,0) [intercept=True]; AIC=-487.464, BIC=-469.857, Time=0.931 seconds
Fit ARIMA(1,2,2)x(0,0,0,0) [intercept=True]; AIC=-507.846, BIC=-493.174, Time=1.393 seconds
Fit ARIMA(0,2,2)x(0,0,0,0) [intercept=True]; AIC=-480.974, BIC=-469.237, Time=1.149 seconds
Fit ARIMA(1,2,3)x(0,0,0,0) [intercept=True]; AIC=-486.881, BIC=-469.274, Time=0.906 seconds
Fit ARIMA(0,2,3)x(0,0,0,0) [intercept=True]; AIC=-494.706, BIC=-480.033, Time=0.720 seconds
Fit ARIMA(2,2,3)x(0,0,0,0) [intercept=True]; AIC=-504.708, BIC=-484.167, Time=0.735 seconds
Near non-invertible roots for order (2, 2, 3)(0, 0, 0, 0); setting score to inf (at least one inverse
root too close to the border of the unit circle: 0.999)
Total fit time: 17.398 seconds

```

#### SARIMAX Results

```

=====
Dep. Variable:          y      No. Observations:      141
Model:                 SARIMAX(1, 2, 2)  Log Likelihood      258.923
Date:                  Wed, 01 Jul 2020  AIC                -507.846
Time:                  11:27:05         BIC                -493.174
Sample:                0               HQIC              -501.884
                    - 141
Covariance Type:      opg
=====

```

	coef	std err	z	P> z	[0.025	0.975]
intercept	-0.0006	0.001	-0.825	0.410	-0.002	0.001
ar.L1	0.7770	0.072	10.762	0.000	0.635	0.918
ma.L1	-1.8055	0.128	-14.068	0.000	-2.057	-1.554
ma.L2	0.9540	0.132	7.212	0.000	0.695	1.213
sigma2	0.0014	0.000	7.885	0.000	0.001	0.002

```

=====
Ljung-Box (Q):          33.09  Jarque-Bera (JB):      6983.75
Prob(Q):                0.77   Prob(JB):              0.00
Heteroskedasticity (H): 0.01   Skew:                 -3.52
Prob(H) (two-sided):    0.00   Kurtosis:             37.00
=====

```

Figure 4

auto\_arima parameters obtain on train dataset

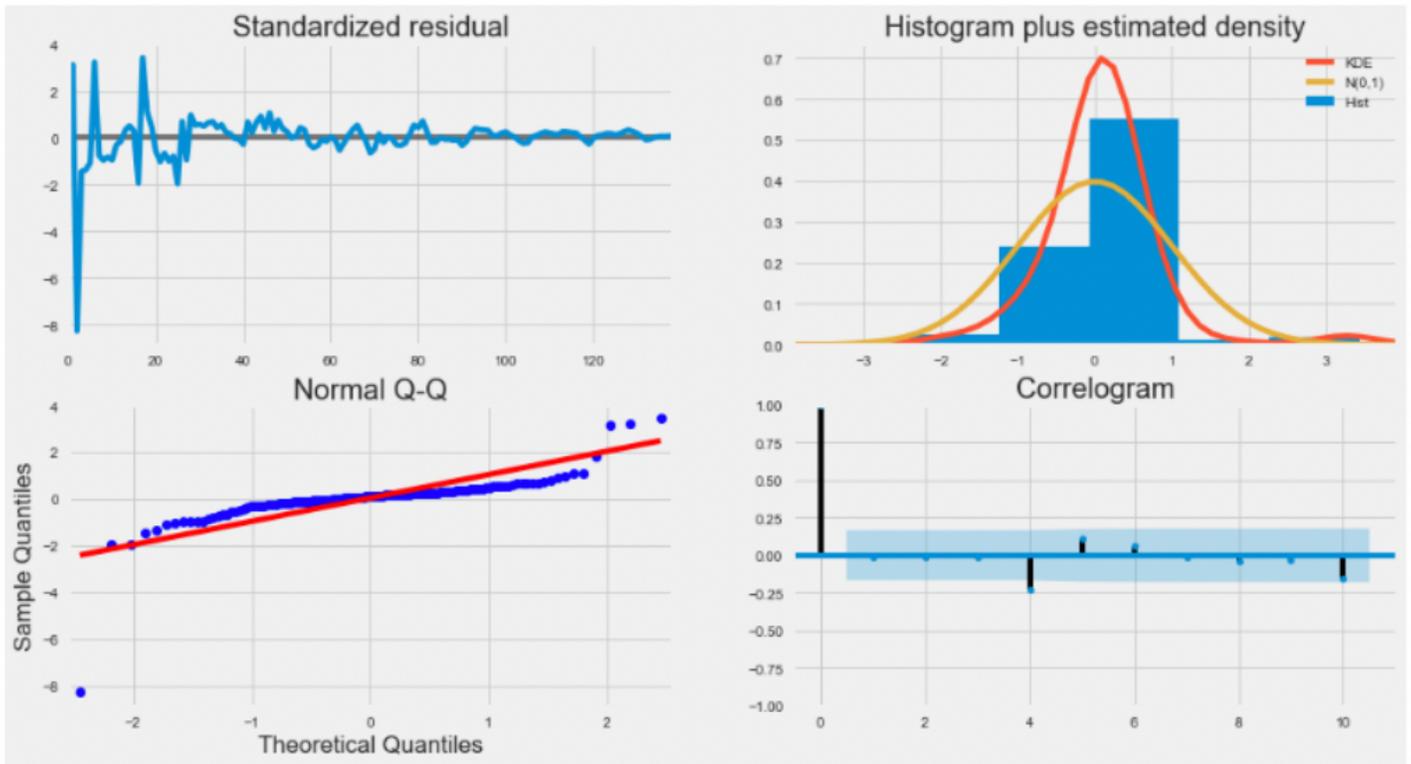


Figure 5

residuals plot by auto\_arima

### ARIMA Model Results

```

=====
Dep. Variable:          D2.Deaths      No. Observations:          139
Model:                 ARIMA(1, 2, 2)  Log Likelihood             259.190
Method:                css-mle        S.D. of innovations        0.036
Date:                  Wed, 01 Jul 2020 AIC                         -508.380
Time:                  11:27:15       BIC                         -493.708
Sample:                01-27-2020     HQIC                        -502.418
                    - 06-13-2020
=====

```

	coef	std err	z	P> z	[0.025	0.975]
const	-0.0036	0.003	-1.327	0.184	-0.009	0.002
ar.L1.D2.Deaths	0.8632	0.058	14.928	0.000	0.750	0.977
ma.L1.D2.Deaths	-1.8761	0.057	-32.995	0.000	-1.988	-1.765
ma.L2.D2.Deaths	0.9996	0.058	17.113	0.000	0.885	1.114

#### Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.1584	+0.0000j	1.1584	0.0000
MA.1	0.9385	-0.3460j	1.0002	-0.0562
MA.2	0.9385	+0.3460j	1.0002	0.0562

Figure 6

Results of ARIMA model

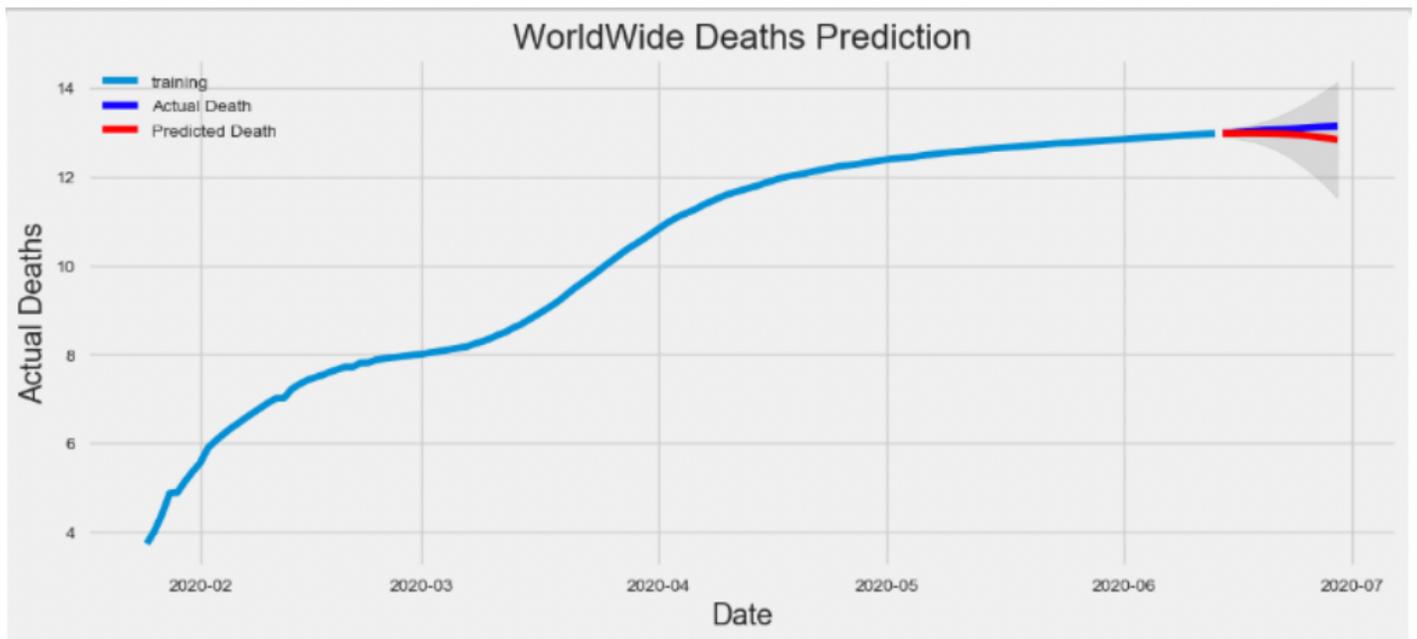
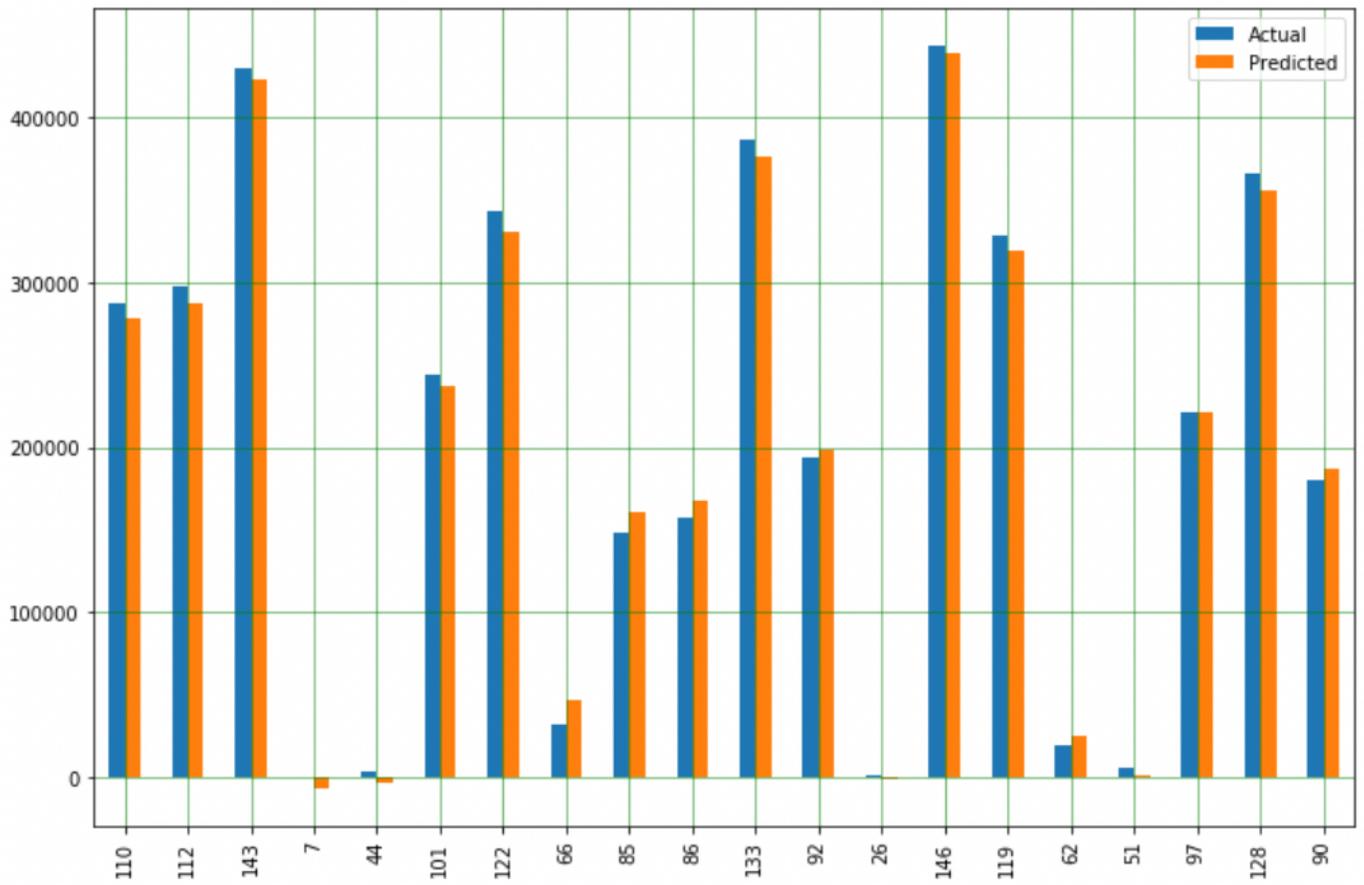


Figure 7

# Prediction of death cases in compare to actual death cases over training data set



**Figure 8**

Comparison plot of actual and predicted values of death cases