COVID-19 Forecasting based on SARIMA*

Boqing Zheng, Yuan Shan, and Minyue Qiu

Duke Kunshan University

Abstract. Most countries and areas are reopening or considering loosening stringent prevention such as lockdowns. Consequently, daily COVID-19 cases will increase significantly. While in some areas, excessive COVID-19 prevention still exists and will bring inconvenience to people's health as well as negative impacts on economic growth. There is a tremendous necessity for supervising and estimating COVID-19 cases in order to help countries and policymakers prepare their healthcare systems and reassess related policies and preventions in a timely manner. In this project,time-series models — Seasonal Auto-Regressive Integrated Moving Average (SARIMA) are used to forecast the epidemiological trends of the COVID-19 pandemic based on Global Confirmed Cases and Global Deaths data by finding the optimized parameters based on the best fit between prediction and the test data. And Auto-Regressive Integrated Moving Average (ARIMA) is used for comparison. Analytical tools Auto-Correlation function (ACF), Partial Auto-Correlation Function (PACF), and Bayesian Information Criterion (BIC) were used to assess the reliability of the models. Evaluation metrics Mean Absolute Error (MAE) is used as criteria. The results show that predictions from the SARIMA model are more accurate than ARIMA model, and reflect the seasonality of COVID-19 cases.

Keywords: COVID-19 Forecasting; ARIMA; SARIMA; Duke Kunshan

* Supported by Prof. Kaizhu Huang

1 Introduction

In 2020, COVID-19 was announced as a global pandemic by the World Health Organization (WHO) By December 9,2022, there have been 643,875,406 confirmed cases of COVID-19, including 6,630,082 deaths, reported to WHO. Due to the rapid infection rate, many countries have adopted several precautionary measures to slow down the spreading of the disease and not overwhelm healthcare capacity, including quarantine, vaccination, or even lockdown. However, the mortality risk among patients hospitalized primarily for COVID-19 decreased to 4,9% from April to June 2022, which is lower than any previous time in the pandemic [1], indicating that the mortality rate of COVID-19 now is not as severe as before, and the precautions could be adjusted or loosened. Accurately judging the trend of the epidemic by scientific means will help policymakers to formulate the most appropriate preventive measures and prevent damage to people's mental health and the negative impact on the economy caused by excessive epidemic prevention and control.

The Autoregressive Integrated Moving Average (ARIMA) and Seasonal-ARIMA (SARIMA) are the most widely used statistical models for many time-series forecasting problems [2]. ARIMA has been used to predict several disease outbreaks such as Hand–Foot–Mouth Disease (HFMD) [3], Hepatitis-B [4], as well as COVID-19 [5]. Given the historical application of ARIMA/SARIMA models, and some cases that ARIMA/SARIMA models outperformed deep learning models [6], we decide to apply ARIMA/SARIMA to develop our model.

2 Background

Since the emergence of COVID-19 in late 2019, many mathematical modeling methodologies have been used to study the rate and pattern of its infection by researchers. In [7], a prediction model is developed using Artificial Neural Network to estimate the future situation by the use of geo-location and numerical data from past 2 weeks. It compared predicted numbers produced by their model with actual values and found that they are closely matched. In [8], a model that combines an improved adaptive neuro-fuzzy inference system (ANFIS) using an enhanced flower pollination algorithm (FPA) by the salp swarm algorithm (SSA)is developed, which improves the performance of ANFIS by determining the parameters of it using FPA and SSA. Some compartmental models, such as the susceptible, infected, recovered model (SIR) and the susceptible, exposed, infected, recovered model (SEIR) are used to model the spread of infectious diseases like Human Immunodeficiency Virus (HIV), Middle East Respiratory Syndrome (MERS), and Severe Acute Respiratory Syndrome (SARS) [9]. In these models, the population will be divided and placed in one of the 3 or 4 compartments listed above. People could be in one compartment at a certain time, or could also be shifted between these compartments given their infectious status. In some studies, ARIMA and SARIMA are both applied in forecasting cumulative COVID-19 cases in 16 countries [10], with the results showing that

SARIMA model is more realistic than ARIMA model, which confirms the seasonality in COVID-19 data. The SARIMA model takes both overall trends and seasonal changes into account, which is widely used in model time series [11]. It is performed over a time series in an automated manner to maximize prediction accuracy. The idea is to find a best-fit model early on and make the prediction for an extended period of time. We use the actual reported infections and deaths, then compare them to the results produced by our best-fit prediction model.

3 Problem

To make the COVID prediction more accurate and valuable, it is necessary to select more valuable predictors and more accurate prediction models. Therefore, the focus of this project is how to select appropriate prediction targets and learning models based on the characteristics of the epidemic (infectious disease).

First of all, for the prediction of infectious diseases, the daily number of confirmed cases, recovered cases, and deaths cases should be the key measurement, but considering the Completeness of the original data, this project finally chooses the daily number of confirmed and deaths cases as predictors. Second, learning models that are biased towards autocorrelation should be prioritized. This is because of the characteristics of infectious diseases, that it can be considered that the number of confirmed cases in a day depends largely on the number of confirmed cases in the previous few days (because most of the patients on that day were infected by the patients in the previous few days). Finally, considering the mutation speed of the COVID virus and the speed of policy changes, it is considered that the forecasting period of 30-60 days is more appropriate. Therefore, models with better short-term performance should be considered first. Based on this, the ARIMA/SARIMA model is selected for forecasting.

4 Methodology

This section details the algorithms and methodologies used in the data analysis process. Specifically, for time series, decomposition is first implemented for exploratory data analysis. By decomposing, raw data can be intuitively understood. Then, for the prediction of specific time series, the ARIMA algorithm is adopted.

4.1 Decomposition

Through decomposition, the change with raw data over time can be decomposed into three parts - trend, seasonal, and residual components. Additive decomposition can be expressed by the following equation:

$$X_t = T_t + S_t + Y_t \tag{1}$$

Where, X_t represents the value of original data, T_t is the trend component S_t is the seasonal component, and Y_t is the residual component. The trend

component describes the general trend of the original data over time, and the change of T_t over time should be a smooth curve. Excluding the influence of the trend, the Seasonal part describes the part of the original data that changes periodically over time. The Residual part is the remaining value after excluding the trend and seasonal components.

4.2 ARIMA Model

Auto-Regressive Integrated Moving Average (ARIMA) model is used to predict a time series. ARIMA model is a combination of Auto Regression (AR) model and Moving Average (MA) model.

The AR model describes the relationship between the current value and the historical value and uses the historical data of the variable to predict itself. The general P-order AR model:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + ... + \alpha_p X_{t-p} + u_t \tag{2}$$

If the random disturbance term is a white noise (i.e. $u_t = \varepsilon_t$), it is called a pure AR(p) process, recorded as:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + ... + \alpha_p X_{t-p} + \varepsilon_t \tag{3}$$

It can be seen that the AR model first needs to determine a parameter p, which represents which historical values are used to predict the current value. In addition, the time series data applied by the AR model must have stationarity. Therefore, differential preprocessing is often required for the original data to make the predicting data stationary.

The MA model assumes that the current value of the time series has no relationship with the historical value, but only depends on the linear combination of historical white noise. That is:

$$X_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + ... + \beta_q \varepsilon_{t-q} \tag{4}$$

ARIMA model is a combination of AR, MA models and differencing. Therefore, there are 3 parameters in ARIMA model that need to be determined – (p,q,d), where d is the order by which the data need to be differentiated to get stationary data. So, in the ARIMA model, after d^{th} order differencing, the following formula is used to predict the current value:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + ... + \alpha_p X_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + ... + \beta_q \varepsilon_{t-q} \tag{5}$$

Since the ARIMA model essentially uses historical data to predict current data, the data is required to have autocorrelation to itself. The autocorrelation function can be used to measure the degree of correlation for the same event between two different time periods, which is:

$$ACF(k) = \Sigma_{t=k+1}^{n} \frac{(Z_t - \bar{Z})(Z_{t-k} - \bar{Z})}{\Sigma_{t=1}^{n}(Z_t - \bar{Z})^2} \tag{6}$$

ACF will be used to test whether the original data has the feasibility of ARIMA prediction, and ACF can also be used to help determine the value of p, q. In the following, there will be a detailed explanation of the operation level.

4.3 SARIMA

SARIMA is an extended version of ARIMA, which takes into account the seasonality existing in the time series. In this COVID forecasting, the seasonality of the raw data clearly exists. The reason for seasonality may be the natural characteristics of infectious diseases and the periodically changing prevention and control policies.

In addition to ARIMA's three non-seasonal parameters, SARIMA adds four seasonal parameters. See the table below for detailed parameter explanations:

Table 1. Table captions should be placed above the tables.

Symbol	Abbreviation
p	Trend Auto Regression order
q	Trend moving average order
d	Trend difference order
P	Seasonal Auto Regression(AR) order
Q	Seasonal Moving Average (MA) order
D	Seasonal difference order
s	The number of time steps for a single seasonal period (periodicity)

4.4 Model Evaluation

Bayesian Information Criterion(BIC) is introduced to evaluate the performance of ARIMA/SARIMA model. The BIC is formally defined as:

$$BIC = kln(n) - 2ln(L) \tag{7}$$

Where L is the maximized value of the likelihood function, k is the number of parameters estimated by the model, and n is the number of samples. The smaller the BIC, the better the model performance. Specifically, a model with a better performance prefers smaller p,q, and a larger maximum likelihood function.s

4.5 General Analysis Process

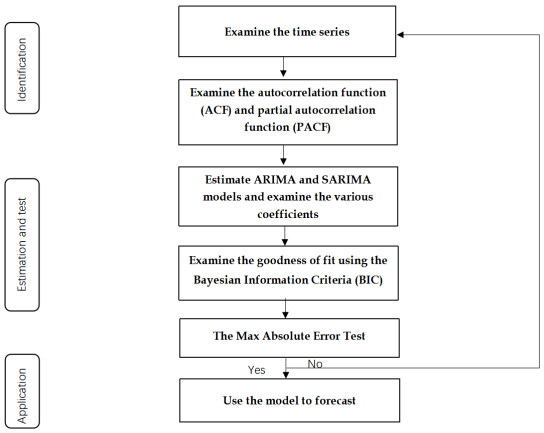


Fig. 1. Problem Analysis Procedure

5 Computational Results

This section describes how we can implement the ARIMA model using Python code. This section includes how to process the data, how to determine the parameters, how to optimize the model, and the final prediction results.

5.1 Data Process

The data used for this project comes from the JHU GovEx GitHub repository [1]. There are three data sets used by the model, which are confirmed cases, recovered cases and death cases respectively. Figure 2 shows the Pandas Dataframe for the data set of confirmed cases, which contains geographic information, dates, and cumulative confirmed cases for each country or region. The other two data sets are similar to this figure.

[1] https://github.com/CSSEGISandData/COVID-19

	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	12/1/22	12/2/22	12/3/22	12/4/22	12/5/22
0	NaN	Afghanistan	33.93911	67.709953	0	0	0	0	0	0	...	206073	206133	206145	206206	206273
1	NaN	Albania	41.15330	20.168300	0	0	0	0	0	0	...	333360	333381	333391	333408	333413
2	NaN	Algeria	28.03390	1.659600	0	0	0	0	0	0	...	271096	271100	271102	271107	271113
3	NaN	Andorra	42.50630	1.521800	0	0	0	0	0	0	...	47219	47219	47219	47219	47219
4	NaN	Angola	−11.20270	17.873900	0	0	0	0	0	0	...	104676	104676	104676	104676	104750

5 rows × 1058 columns

Fig. 2. Dataframe of confirmed cases

In order to adapt to the data format of model training and the prediction requirements of the project, we processed the three data sets respectively,in the format of row for time and column for a number of cases. Then, the two adjacent numbers of cases are subtracted to get the global daily number of cases. Finally, the line plots of the three processed data sets are shown in Figure 3. However, according to the figure of recovered cases, in the second half of the recovered cases, daily increased cases are 0 and did not change. Therefore, it is possible that the data set of recovered cases may not be updated, and this model will not predict recovered cases.

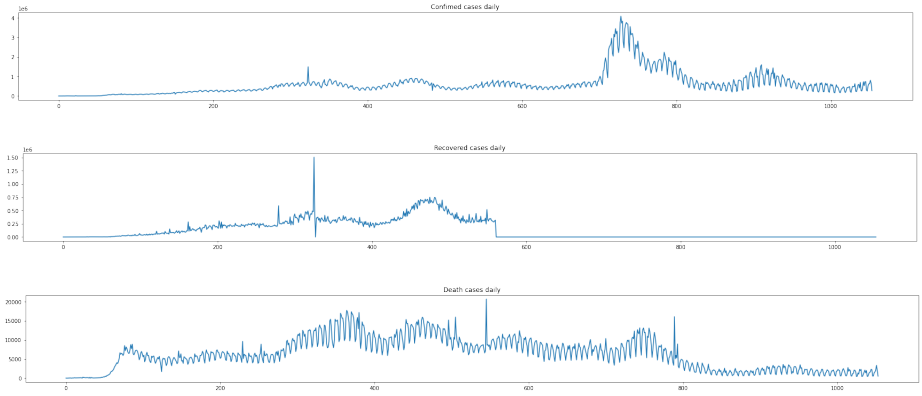


Fig. 3. Line plot of 3 data sets

5.2 Data Attributes and Model Determination

For time series data, it has two properties, namely seasonality and stationarity. Stationarity refers to whether the data fluctuates around a fixed mean. Seasonality refers to whether the data show a repeating pattern. By analyzing these two properties in our data sets, the model can be better determined and have better prediction results. Through decomposition method, the data attributes of the data sets of confirmed cases and death cases are obtained, namely Trend, Seasonality and Residue. Taking the confirmed cases as an example, as shown

in Figure 4, it can be observed that the data set has a period of 28 days. At the same time, the data do not fluctuate around a stable average. Therefore, both data sets have non-stationarity and seasonality with a period of 28 days. The same is true of death cases data set.

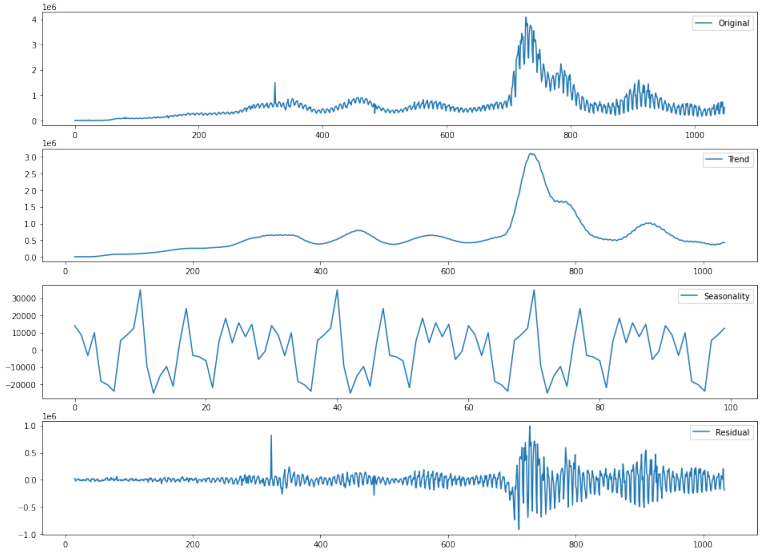


Fig. 4. Decomposition of confirmed cases

Thus, a category of ARIMA, called SARIMA, was adopted. Its parameters are shown in Figure 5. This is because SARIMA can better handle seasonality in the time series data, allowing for more accurate forecasts [12].

$$SARIMA \underbrace{(p,d,q)}_{non-seasonal} \underbrace{(P,D,Q)_m}_{seasonal}$$

Fig. 5. SARIMA parameters

5.3 Intuitive Determination of Parameters

This subsection introduces an intuitive method of parameter determination. First, two parameters by eliminating seasonality and making the data stationary. Next, the remaining parameters can be intuitively determined by looking at autocorrelation and partial autocorrelation graphs of the data.

In the first step, the confirmed cases and death cases data sets are tested for stationarity. Here, the Augmented Dickey-Fuller (ADF) Test is used. The ADF checks for the presence of a unit root in the dataset and returns a value of p-value. If the p-value is less than 0.05, it is considered that there is no unit root and the data is stable. If the value is greater than 0.05, the data is considered not stationary, and then the difference will be used to make the data stationary. As shown in Figure 6, the data of death cases is not stationary, while the data of confirmed cases is basically stationary.

```
Death data after diff 1: 4.574645141595912e-05
Death data after diff 2: 1.4848224858589865e-05
Confirmed data after diff 1: 2.704774752236141e-08
Confirmed data after diff 2: 5.960612057026131e-09
```

Fig. 6. ADF results of two data sets in Python

In order to make the two sets of data more stationary, one-time difference and two-time difference are respectively made for both sets of data. As shown in Figure 7, both sets of data are more stationary afterwards. Since the difference between one-time difference and two-time difference is not significant, making difference once is determined to be final for both data sets. In the second step, to eliminate seasonality, the data will be differentiated 28 times, which is 1 period.

```
Death data after diff 1: 4.3868654367625706e-05
Death data after diff 2: 1.4117255155760136e-05
Confirmed data after diff 1: 2.35200531239551e-08
Confirmed data after diff 2: 5.166876127491306e-09
```

Fig. 7. ADF results after difference in Python

Figure 8 shows the data sets before and after processing, where it can be clearly observed that data sets have been stationary and non-seasonal.

Fig. 8. Death and confirmed cases data before and after difference

At this point, m, d and D for the confirmed cases and death cases model can be determined. m = 28, d(confirmed)=1, D(confirmed)=1, d(death)=1, D(death)=1. Next, according to the processed data, autocorrelation graphs and partial autocorrelation graphs of confirmed cases and death cases will be obtained, as shown in Figure 9 and Figure 10. Therefore, the p, q, P, Q parameters of the two data set models can be obtained. q (death) = 1, p (death) = 6, P (death) = 1, Q (death) = 1 q (confirmed) = 1, p (confirmed) = 0, P (confirmed) = 1, Q (confirmed) = 1

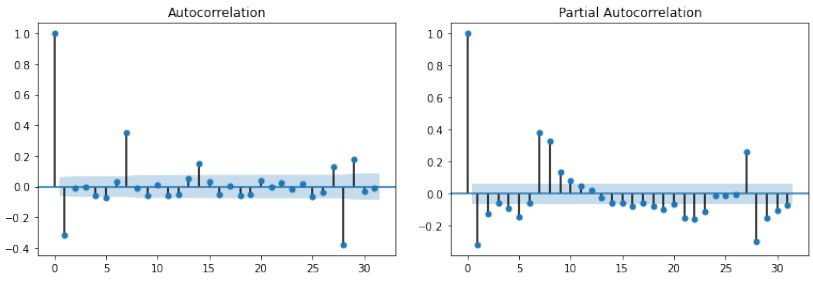


Fig. 9. Autocorrelation and partial autocorrelation for confirmed cases

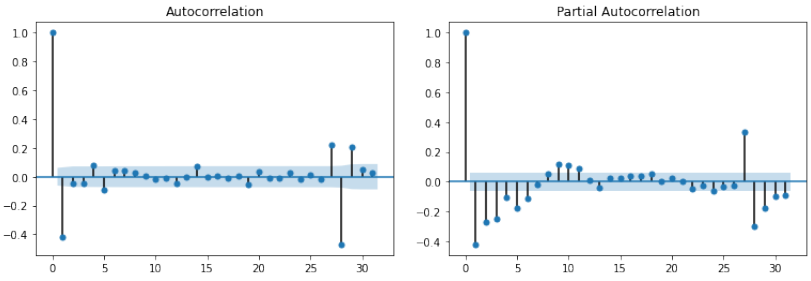


Fig. 10. Autocorrelation and partial autocorrelation for death cases

5.4 Brute-force Determination of Parameters

The intuitive judgment of parameters is often not accurate, so a brute-force method is taken to determine the parameters. In this method, several possible p, q, P, Q parameter cases will be tested and measured by Bayesian Information Criterion(BIC). The lower the BIC is, the better the performance will be. As Figure 11 shows, the brute-force method outputs some relatively good parameter values. The values of these relatively excellent parameters were modeled and the following model parameters were finally determined.

For the confirmed cases model, p = 3, d = 1, q = 2, P = 3, D = 1, Q = 2, m = 28

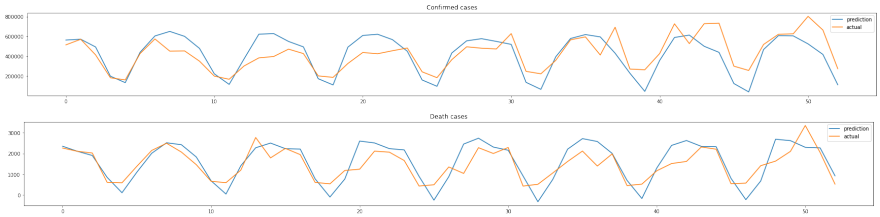
For the death cases model, p = 6, d = 1, q = 1, P = 1, D = 1, Q = 1, m = 28

```
p: 0 q: 0 P: 3 Q: 1
Bic: 15.222612188617115
p: 0 q: 2 P: 3 Q: 1
Bic: 21.31165706406396
p: 1 q: 0 P: 3 Q: 0
Bic: 15.222612188617115
p: 1 q: 1 P: 2 Q: 2
Bic: 21.31165706406396
p: 1 q: 1 P: 2 Q: 3
Bic: 24.356179501787384
p: 1 q: 1 P: 3 Q: 3
Bic: 27.400701939510807
p: 1 q: 2 P: 3 Q: 1
Bic: 24.356179501787384
p: 1 q: 3 P: 3 Q: 3
Bic: 33.489746814957655
p: 2 q: 1 P: 3 Q: 1
Bic: 24.356179501787384
p: 3 q: 0 P: 2 Q: 1
Bic: 21.31165706406396
p: 3 q: 2 P: 3 Q: 1
Bic: 30.44522437723423
p: 3 q: 2 P: 3 Q: 2
Bic: 33.489746814957655
p: 4 q: 0 P: 2 Q: 2
Bic: 27.400701939510807
```

Fig. 11. Brute-force method output (partially)

5.5 Prediction Results and Comparison

By bringing the training set into the model, the number of cases over the next 50 days was predicted. Figure 12 shows the prediction in the next 50 days of the model of the confirmed cases data set and the prediction in the next 50 days of the model of the death cases data set respectively. The mean absolute error of the prediction is shown. The same p, d, and q parameters and data sets were put into the ARIMA model, and the predicted results as shown in Figure 13 were obtained. By comparison, the SARIMA model predicted the overall trend and seasonal trend better than the ARIMA model with a lower MAE. Although the MAE of ARIMA is not bigger than that of SARIMA significantly, this can be explained by the fact that the ARIMA model's forecast in the last 10 days tends to be around the average as a line. This further indicates that the ARIMA model could not reflect the seasonality of diseases well.




```
Mean absolute error of confirmed cases is:  69067.88070863715
Mean absolute error of confirmed cases is:  455.67010226730616
```

Fig. 12. Prediction results of two models based on SARIMA


```
Mean absolute error of confirmed cases is:  131425.84797865537
Mean absolute error of death cases is:  470.0988629487032
```

Fig. 13. Prediction of two models based on ARIMA

6 Conclusion

In this project, the SARIMA model was adopted to make 50-day predictions for both the death cases and confirmed cases data sets. In terms of parameters, intuitive and brute-force methods are combined, and the models of the two data sets are finally determined through multiple attempts of model parameters.

For the confirmed cases model, p = 3, d = 1, q = 2, P = 3, D = 1, Q = 2, m = 28; for the death cases model, p = 6, d = 1, q = 1, P = 1, D = 1, Q = 1, m = 28.

Based on the final predictions, SARIMA is able to predict the overall and seasonal trends in cases and predict the inflection point of each increase or decrease in cases with accuracy. In comparison with other methods, ARIMA is compared. By bringing the same parameter configuration into the ARIMA model, the model's predictions for the two data sets did not reflect their seasonal trends and had larger errors. As a result, ARIMA does not perform as well as SARIMA.

However, for SARIMA, the forecast values will not necessarily equate to the actual values observed for the same time period. This can be due to several factors such as the different restrictions that are mandated by authorities to curb the spread and the degree to which the public adheres to these restrictions. As different variants of viruses keep emerging, the infection rate and its reproduction rate will keep changing, which also affects the difference between the prediction value and the actual value. Age, health facilities, and vaccination rates also play a vital role in the rapid spread of the COVID-19 pandemic.

In terms of future model improvement study, the brute-force method will test more parameters such as d, D to train SARIMA model more accurately. Second, more models will be used, such as LSTM. This is because SARIMA may not be able to account for human factors in the data, as mentioned above. So other methods will be taken, and COVID cases will be predicted in a multi-model way.

References

1. Y. Lin, Z. Hu, Q. Zhao, H. Alias, M. Danaee, and L. P. Wong, "Understanding COVID-19 vaccine demand and hesitancy: A nationwide online survey in china," vol. 14, no. 12, pp. 1–22, publisher: Public Library of Science. [Online]. Available: https://doi.org/10.1371/journal.pntd.0008961
2. W. Anne and S. Jeeva, "ARIMA modelling of predicting COVID-19 infections," publisher: Cold Spring Harbor Laboratory Press _eprint: https://www.medrxiv.org/content/early/2020/04/23/2020.04.18.20070631.full.pdf. [Online]. Available: https://www.medrxiv.org/content/early/2020/04/23/2020.04.18.20070631
3. L. LIU, R. S. LUAN, F. YIN, X. P. ZHU, and Q. LÜ, "Predicting the incidence of hand, foot and mouth disease in sichuan province, china using the arima model – corrigendum," *Epidemiology and Infection*, vol. 144, no. 1, p. 152–152, 2016.
4. Y.-w. Wang, Z.-z. Shen, and Y. Jiang, "Comparison of ARIMA and GM(1,1) models for prediction of hepatitis b in china," vol. 13, no. 9, pp. 1–11, publisher: Public Library of Science. [Online]. Available: https://doi.org/10.1371/journal.pone.0201987
5. Z. Ceylan, "Estimation of COVID-19 prevalence in italy, spain, and france," vol. 729, p. 138817. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0048969720323342
6. K. ArunKumar, D. V. Kalaga, C. Mohan Sai Kumar, M. Kawaji, and T. M. Brenza, "Comparative analysis of gated recurrent units (GRU), long short-term memory (LSTM) cells, autoregressive integrated moving average (ARIMA), seasonal autoregressive integrated moving average (SARIMA) for forecasting COVID-19 trends," vol. 61, no. 10, pp. 7585–7603. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1110016822000138
7. M. Wieczorek, J. Siłka, D. Połap, M. Woźniak, and R. Damaševičius, "Real-time neural network based predictor for cov19 virus spread," vol. 15, no. 12, pp. 1–18, publisher: Public Library of Science. [Online]. Available: https://doi.org/10.1371/journal.pone.0243189
8. M. A. A. Al-qaness, A. A. Ewees, H. Fan, and M. Abd El Aziz, "Optimization method for forecasting confirmed cases of COVID-19 in china," vol. 9, no. 3, p. 674, number: 3 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2077-0383/9/3/674
9. K. O. Kwok, A. Tang, V. W. Wei, W. H. Park, E. K. Yeoh, and S. Riley, "Epidemic models of contact tracing: Systematic review of transmission studies of severe acute respiratory syndrome and middle east respiratory syndrome," vol. 17, pp. 186–194. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2001037018301703
10. J. Cong, M. Ren, S. Xie, and P. Wang, "Predicting seasonal influenza based on SARIMA model, in mainland china from 2005 to 2018." vol. 16, no. 23, place: Switzerland.
11. K. E. ArunKumar, D. V. Kalaga, C. M. Sai Kumar, G. Chilkoor, M. Kawaji, and T. M. Brenza, "Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: Auto-regressive integrated moving average (ARIMA) and seasonal auto-regressive integrated moving average (SARIMA)." vol. 103, p. 107161, place: United States.
12. S. Wang, J. Feng, and G. Liu, "Application of seasonal time series model in the precipitation forecast," *Mathematical and Computer Modelling*, vol. 58, no. 3, pp.

677–683, 2013, computer and Computing Technologies in Agriculture 2011 and Computer and Computing Technologies in Agriculture 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S089571771100639X