Predicting the Pandemic COVID-19 using ARIMA Model

Nguyen Quoc Duong[1,*], **Le Phuong Thao**[1], **Dinh Thi Quynh Nhu**[1],
Cao Thi Ai Loan[2], **Phung Thi Hong Diem**[2], **Le Thanh Binh**[2]

[1]*Faculty of Education, Quy Nhon University, 170 An Duong Vuong, Quy Nhon, Binh Dinh, Viet Nam*
[2]*Faculty of Mathematics and Statistics, Quy Nhon University, 170 An Duong Vuong, Quy Nhon, Binh Dinh, Viet Nam*

Abstract: Coronavirus disease 2019 (COVID-19) has been recognized as a global threat, and several studies are being conducted using various mathematical models to predict the probable evolution of this epidemic. The main objective of this study is to apply autoregressive integrated moving average (ARIMA) model with the objective of monitoring and short-term forecasting the total confirmed new cases per day all over the world. The data are extracted from daily report of World Health Organization from 21st January 2020 to 16th March 2020. Akaike's Information Criterion (AIC) and Ljung-Box test were used to evaluate the constructed models. To assess the validity of the proposed model, the Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) between the observed and fitted of COVID-19 total confirmed new cases was calculated. Finally, we applied "forecast" package in R software and the fitted ARIMA model to predict the infections of COVID-19. We found that the ARIMA(1,2,1) model was able to describe and predict the epidemiological trend of the disease of COVID-19. The MAPE and RMSE for the training set and validation set respectively, which we found was reasonable for use in the forecast. Furthermore, the model also provided forecast total confirmed new cases for the following days. ARIMA model applied to COVID-19 confirmed cases data are an important tool for COVID-19 surveillance all over the world. This study shows that accurate forecasting of the COVID-19 trend is possible using an ARIMA model. Unless strict infection management and control are taken, our findings indicate the potential of COVID-19 to cause greater outbreak all over the world.

Keywords: COVID-19, coronavirus, ARIMA, Box-Jenkins Methodology, time series.

1. Introduction

Coronaviruses (CoV) are a large family of viruses that cause illness ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS-CoV) and Severe Acute Respiratory Syndrome (SARS-CoV). The 2019 novel coronavirus is now named severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) while the disease associated with it is referred to as COVID-19.

In response to the spread of COVID-19 all over the world, several countries have demonstrated the ability to reduce, or stop transmission of the COVID-19 virus. However, the impact of control efforts remains difficult to measure due to the inherent complexities of COVID-19 as a disease: multiple viral strains with identified genetic polymorphisms, complex disease manifestation, and multiple routes of transmission [1]. Infectious diseases have certain characteristic features that lead themselves to modeling, such as: speed of pathogen variation, accumulation of susceptible hosts, and environmental indices [2]. Thus, epidemic modeling and forecasting can be essential tools to prevent and control COVID-19. Recently, statistical methods including linear regression, correlation coefficients, back propagation artificial neural network model have been used for prediction of COVID-19 disease. The variation of COVID-19 diseases, which is influenced and constrained by diversified factors, is characterized by tendency and randomicity. These statistical tools are inappropriate for analyzing the randomicity of COVID-19. Autoregressive integrated moving average (ARIMA) models, which take into account changing trends, periodic changes, and random disturbances in time series, are very useful in modeling the temporal dependence structure of a time series. In epidemiology, ARIMA models have been successfully applied to predict the incidence of infectious diseases, such as influenza mortality [3], malaria incidence [4], as well as other infectious diseases [5,6]. This study aimed to develop a univariate time series model for the COVID-19 disease; specifically, a stochastic ARIMA model, for short-term forecasting of total confirmed new cases of COVID-19 infections for the following days.

Besides mathematical, software tools today also play an important role in forecasting. There are many software tools for highly effective data analysis such as SPSS, Eviews, Python, etc. In this study, we use R statistical software to analyze the COVID-19 data. The advantages of R programming are open source programming language, providing exemplary support for data organization, package arrays, quality plotting and graphing, highly compatible, platform independent reporting and machine learning activity.

2. Methods

The data are extracted from daily report of World Health Organization from 21st January 2020 to 16th March 2020 (https://www.who.int). All COVID-19 cases were initially diagnosed by clinical symptoms. In all the world, COVID-19 is a notifiable disease and hospital physicians must report every case of COVID-19 to the national health authority. Later report daily COVID-19 case totals for World Health Organization with surveillance purposes. Due to mandatory reporting, it is believed that the degree of compliance in disease notification over the study period was consistent.

We used the Box-Jenkins approach to ARIMA (p, d, q) modeling of time series [7-10]. This model-building process is designed to take advantage of associations in the sequentially lagged relationships that usually exist in periodically collected data. The following were the parameters selected when fitting the ARIMA model: p, the order of autoregression; d, the degree of difference; q, the order of moving average.

The daily data used in this study did not show seasonal pattern, so the series was differenced at the non-seasonal level to induce stationarity. Autocorrelation function (ACF) graph and Partial autocorrelation function (PACF) graph were used to identify the order of moving average (MA) and autoregressive (AR) terms included in the ARIMA model. Estimates of the model is parameters were obtained by the conditional least squares method. Diagnostic checking including residual analysis and the Akaike Information Criterion (AIC) was used to compare the goodness of fit among ARIMA models.

$$AIC = -2\log(L) + 2(p + q + k + 1),$$

where L is the likelihood of the data, $k = 1$ if $c \neq 0$ and $k = 0$ if $c = 0$ [9]. The Ljung-Box test was used to measure the ACF of the residuals. In addition, we used the mean absolute percentage error (MAPE), root mean squared error (RMSE) and fitting effect diagram to assess forecast accuracy [7-10].

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{x_t - \hat{x}_t}{x_t}\right|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_t - \hat{x}_t)^2}$$

where x_t and $\hat{x}_t$ denote observed and fitted values at time point t. The MAPE and RMSE value was calculated based on observed values and fitted values from 21st January 2020 to 14th March 2020. A lower AIC, MAPE and RMSE values indicates a better fit of the data. Finally, the fitted ARIMA model was used for forecasting total confirmed new cases of COVID-19 for the next five days.

3. Results

Our data have a daily series with 56 observations (56 days) and the goal is to forecast the next five days. The corresponding command packages and libraries for model prediction are *forecast*, *readxl*, *tseries*, *TSstudio* and *ggplot2*. Let us load the total confirmed new cases series from file *newcase.xlsx*.

Let us plot the series with the *autoplot* function and review the main characteristics of the series:

```
datats <- ts(newcase)
autoplot(datats) + ylab("Total confirmed new cases of COVID-19") +
xlab("From 21/1/2020 to 16/3/2020")
```

We attain the output as shown in Figure 1:

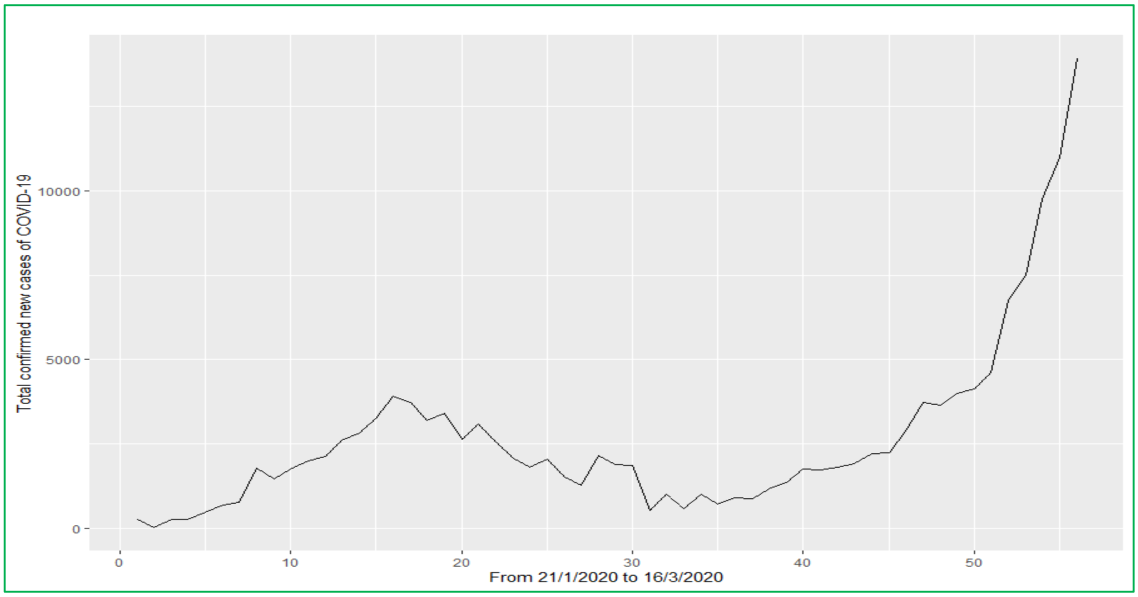

Figure 1. Total confirmed new cases of COVID-19 from 21st January 2020 to 16th March 2020

Observe the picture above, the series is trending up, so we can already conclude that the series is not stationary and some differencing of the series is required. We would use the first 56 observations for training and test the performance using the last 2 observations. Creating partitions in R can be done manually with the *ts_split* function from the stats package. For instance, let us split the *datats* series into partitions, leaving the last 2 observations of the series as the testing partition and the rest as training:

```
x<- ts_split(datats, sample.out = 2)
train <- x$train
test <- x$test
```

Before we start the training process of the ARIMA model, we will conduct diagnostics in regards to the series correlation with the ACF and PACF functions. Since we are interested in viewing the relationship of the series with its lags, we will increase the number of lags to calculate.

```
par(mfrow=c(1,2))
acf (ts (train), main="ACF For Time Series", col="blue", lwd = 4)
pacf (ts (train), main="PACF For Time Series ", col="coral", lwd = 4)
```

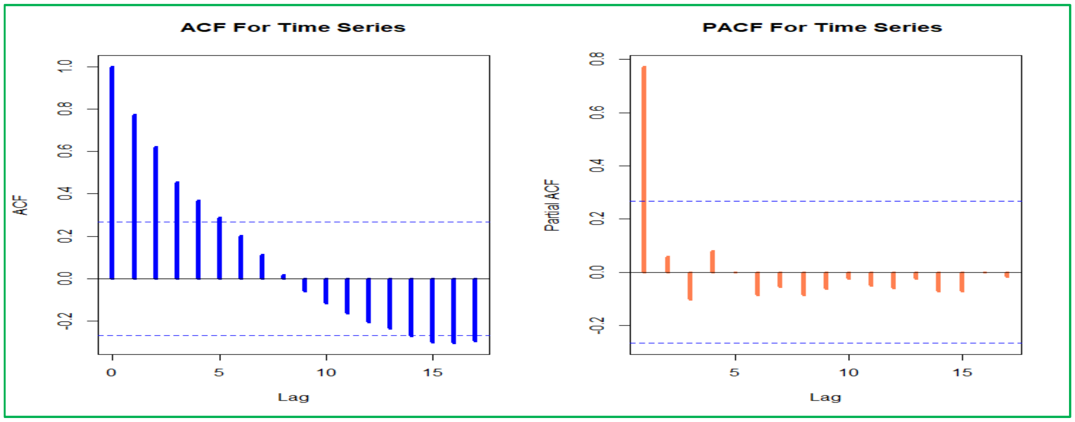

Figure 2. ACF and PACF plot of total confirmed new cases of COVID-19

As well as looking at the time plot of the data, the ACF plot is also useful for identifying non-stationary time series. For a stationary time series, the ACF will drop to zero relatively quickly, while the ACF of non-stationary data decreases slowly. Therefore, differencing can help stabilize the mean of a time series by removing changes in the level of a time series, and therefore eliminating (or reducing) trend. Consequently, we will take a first difference of the data. The first differenced data are shown in Figure 3.

```
diff1 <- diff(train, differences = 1)
par(mfrow=c(1,2))
acf(ts(diff1),main="ACF For First Order Difference", col="green", lwd = 4)
pacf(ts(diff1),main="PACF For First Order Difference", col="red",lwd = 4)
```

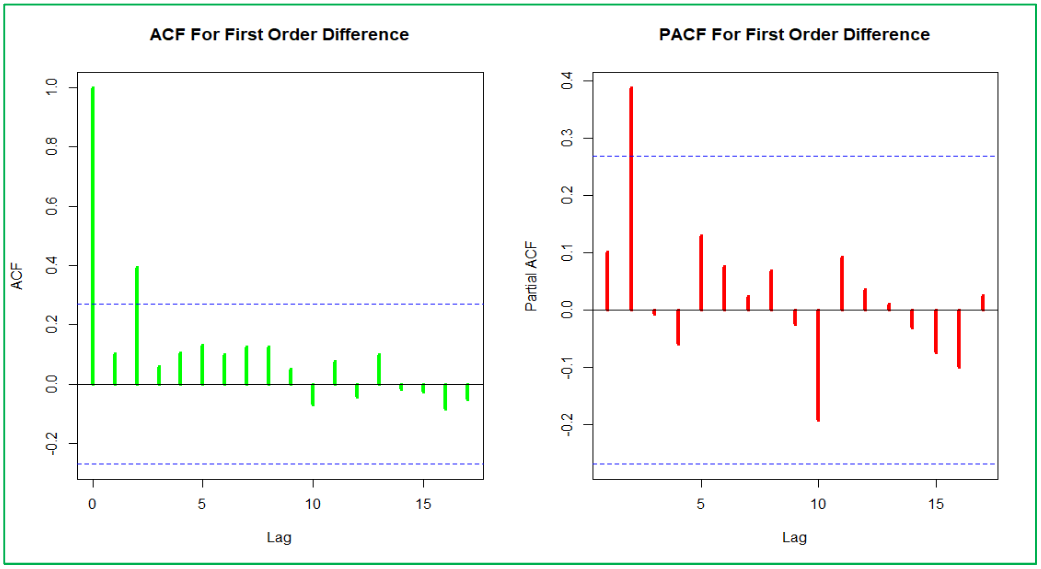

Figure 3. ACF and PACF plot for first order differencing total confirmed new cases of COVID-19

Observing Figure 3, the time series look stationary. However, we will check stationary by using ADF (Augmented Dickey-Fuller) test. The ADF test is a formal statistical test done to ensure stationarity. In ARIMA modeling using R the univariate data is converted into time series data format. The graph follows an overall upward trend with some outliers in terms of sudden lower values. The ADF is unit root test for stationarity [7].

```
adf.test (diff1, alternative = "stationary")

## Augmented Dickey-Fuller Test
## data:  diff1
## Dickey-Fuller = -0.85086, Lag order = 3, p-value = 0.9518
## alternative hypothesis: stationary
```

After taking the first order differencing, the p-value is 0.9518 and is more than 0.05. Therefore, the first order differencing is non-stationary. We necessary to difference the data a second time to obtain a stationary series.

```
diff2 <- diff(train, differences = 2)
par(mfrow=c(1,2))
acf(ts(diff2),main="ACF For Second Order Difference", col="green", lwd = 4)
pacf(ts(diff2),main="PACF For Second Order Difference", col="red",lwd = 4)
```

We then have the output as shown in Figure 4:

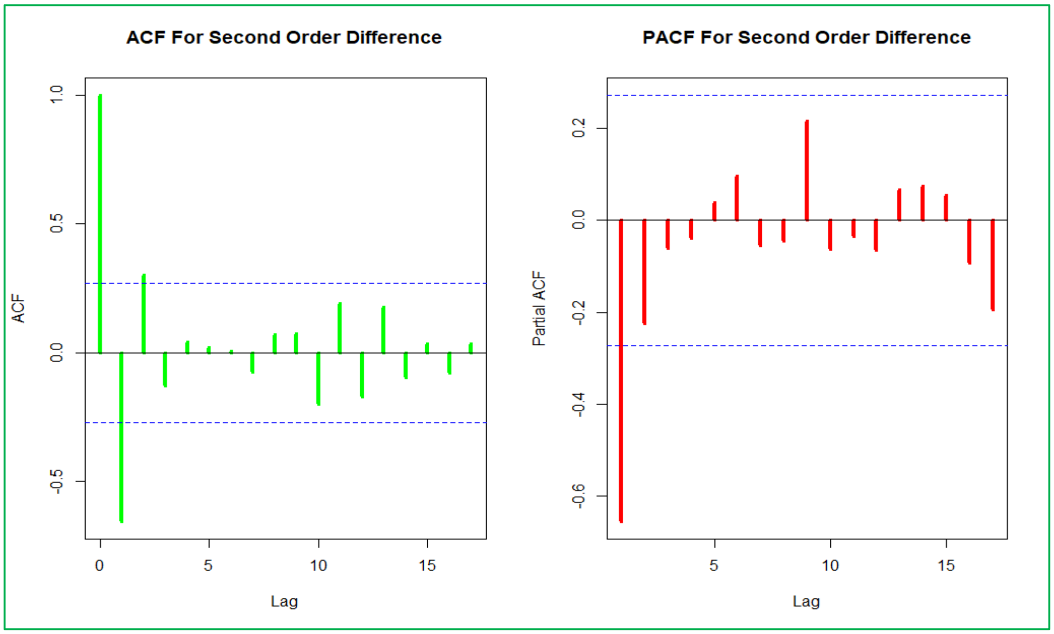

Figure 4. ACF and PACF plot for second order differenced total confirmed new cases of COVID-19

We will check stationary of second-order differencing data by using ADF test.

```
adf.test (diff2, alternative = "stationary")

##      Augmented Dickey-Fuller Test
## data:  diff2
## Dickey-Fuller = -5.2815, Lag order = 3, p-value = 0.01
## alternative hypothesis: stationary
autoplot(diff2) + ylab("diff2") + xlab("Time")
```

Since the p-value after differencing is 0.01 and is less than 0.05 the null hypothesis is rejected and the data does not have a unit root and is stationary. The second order differenced data are shown in Figure 4.

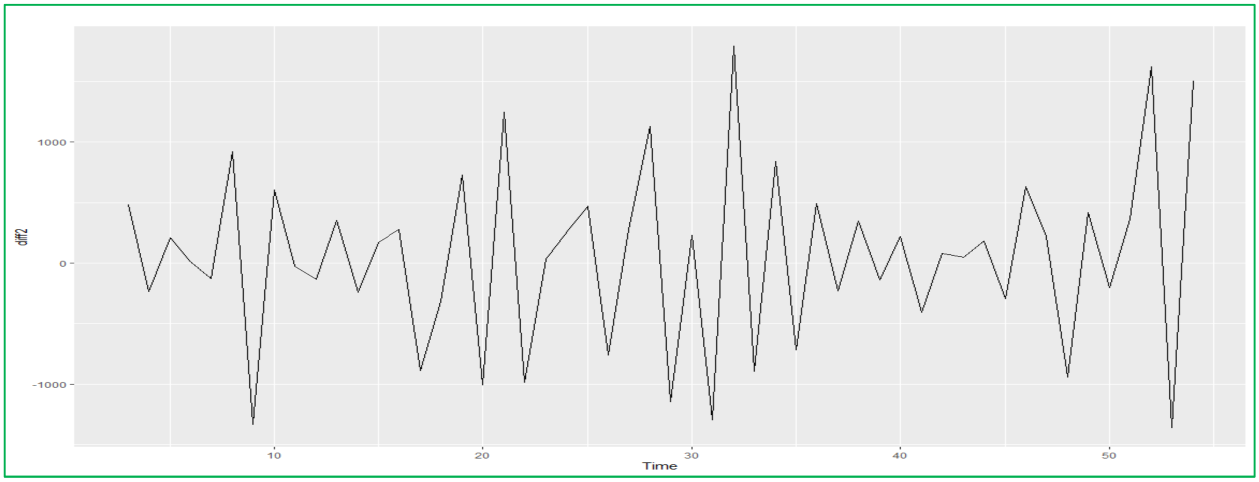

Figure 5. Time series after taking the second order differencing

Our aim now is to find an appropriate ARIMA model based on the ACF and PACF plot of Figure 4. The significant spike at lag 1 and lag 3 in the ACF suggests ar(2) component, and the significant spike at lag 2 in the ACF suggests ma(1) component. Consequently, we begin with an ARIMA(2,2,1) [AIC = 806.57]. The PACF shown in Figure 4 is suggestive of an ar(1) model. So an initial candidate model form PACF plot is an ARIMA(0,2,1) [AIC = 808.76]. Base on the AIC smallest, we chose an ARIMA(2,2,1) model to fit along with some variations on it, compute the AIC values and test set evaluation shown in the Table 1. The best is the ARIMA(1,2,1) model (i.e., it has the smallest AIC, MAPE and RMSE values).

Table 1. AIC, MAPE and RMSE values for various ARIMA models applied for total confirmed new cases of COVID-19

Models	AIC	MAPE	RMSE
ARIMA(2,2,1)	806.57	22.40385	508.6987
ARIMA(3,2,1)	808.50	22.08726	508.2025
ARIMA(3,2,2)	809.75	22.56819	511.7869
ARIMA(2,2,2)	808.54	22.27889	508.4950
ARIMA(1,2,1)	805.15	21.66519	508.0969
ARIMA(1,2,0)	806.65	24.24934	530.3712

Next step, we check the residuals from ARIMA(1,2,1) model by plotting the ACF of the residuals.

```
best.md <- Arima(train, order = c(1,2,1))
checkresiduals(best.md)
```

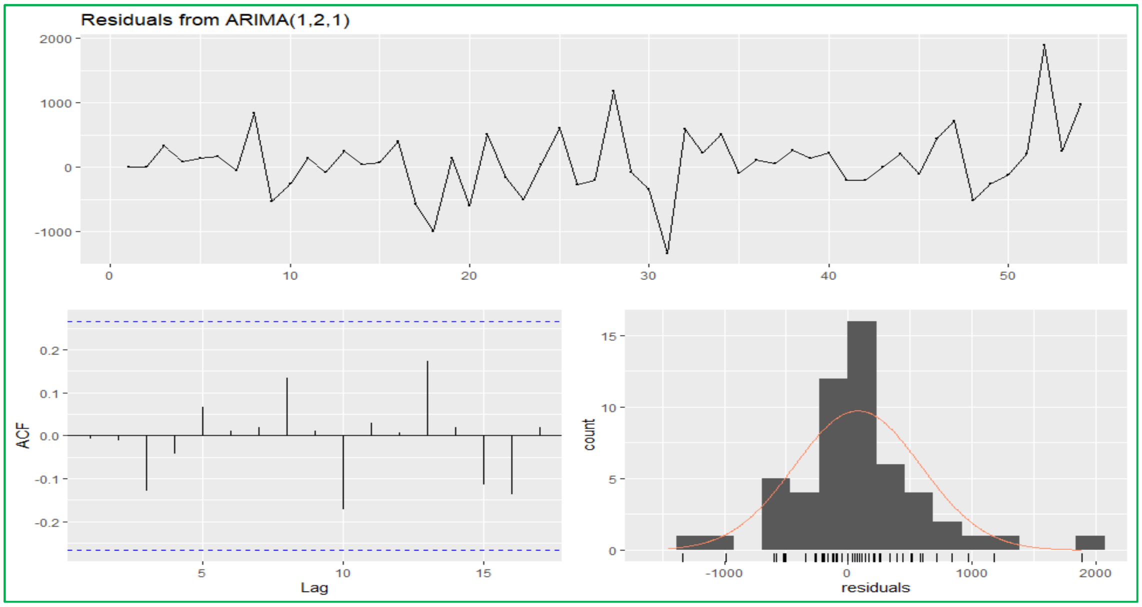

Figure 6. Residuals from the ARIMA(1,2,1) model applied to total confirmed new cases of COVID-19

The output of the Ljung-Box test suggested that the residuals of the model are white noise:

```
##      Ljung-Box test

## data:  Residuals from ARIMA(1,2,1)
## Q* = 4.644, df = 8, p-value = 0.7949
## Model df: 2.   Total lags used: 10
```

By looking at the preceding residuals plot, you can see that the residuals are white noise and normally distributed. Furthermore, the Ljung-Box test confirms that there is no autocorrelation left on the residuals with a *p*-value of

0.7949, we cannot reject the null hypothesis that the residuals are white noise. Thus, we now have a seasonal ARIMA model that passes the required checks and is ready for forecasting.

Table 2. Summary of ARIMA(1,2,1)

ARIMA(1,2,1)		
ARIMA	**ar1**	**ma1**
Coefficients	-0.4715	-0.4471
s.e.	0.1816	0.2125

Let us use the `best.md` trained model to forecast the corresponding observations of the testing set:

```
test_fc <- forecast(best.md, h = 2)
```

Table 3. Assess performances of ARIMA(1,2,1) model

Date	Forecast	Actual	Lo 80	Hi 80	Lo 95	Hi 95	FE
15/3/2020	10882.19	10982	10200.43	11563.95	9839.532	11924.85	0.9%
16/3/2020	12541.07	13903	11536.90	13545.24	11005.325	14076.82	9.79%

According to Table 3, we see the predicted values had given by the reference model at two days are in the confident interval with respect to level 80% and 95%. This shows that the model could be fitting and statistical significance. In particular, actual values are near the upper bound of the 95% confidence interval.

Now, we will use the *test_forecast* function to get a more intuitive view of the model is performance on the training and testing partitions:

```
test_forecast(datats,forecast.obj = test_fc,test = test)
```

We get the following output:

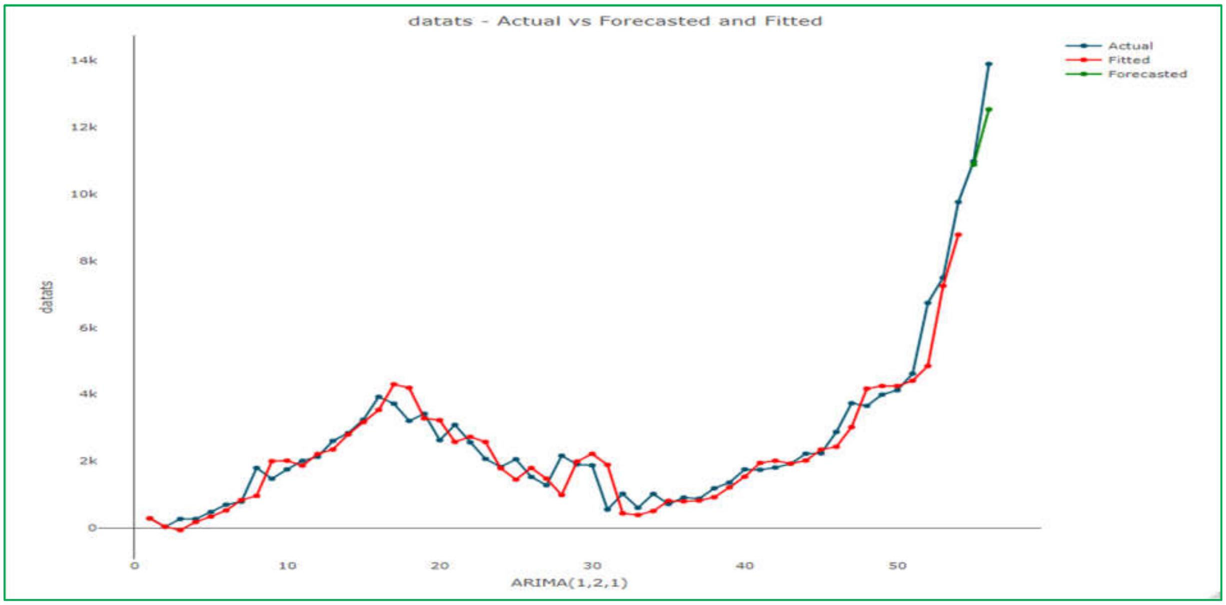

Figure 7. datats - Actual versus Forecasted and Fitted.

Now that we have satisfied the preceding conditions, we can move on to the last step of the forecasting process and generate the final forecast with the selected model. We will start by retraining the selected model on all the series:

```
model <- Arima(datats, order = c(1,2,1))
```

The main goal of the forecasting process is to minimize the level of uncertainty around the future values of the series. Although we cannot completely eliminate this uncertainty, we can quantify it and provide some range around the point estimate of the forecast. The confidence interval is a statistical approximation method that is used to express the range of possible values that contain the true value with some degree of confidence (or probability). Let us use the forecast function to forecast the next five days of the *datats* series:

```
forecast_virus <- forecast(model, h = 5)
```

The output is as Table 4:

Table 4. Assess performances of ARIMA(1,2,1) model

Date	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
17/3/2020	15491.89	14791.70	16192.08	14421.04	16562.73
18/3/2020	17868.12	16820.01	18916.23	16265.17	19471.07
19/3/2020	19778.99	18165.76	21392.22	17311.77	22246.22
20/3/2020	21964.92	19791.75	24138.08	18641.34	25288.49
21/3/2020	23988.27	21157.58	26818.96	19659.10	28317.44

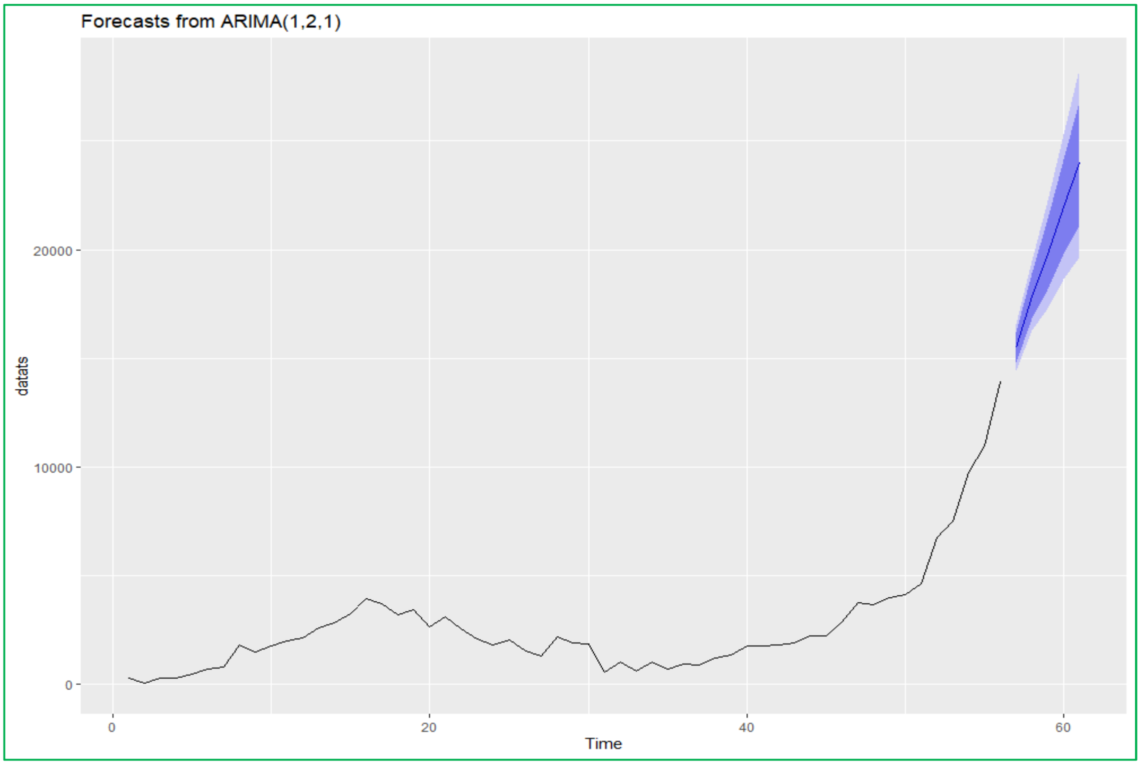

Figure 8. Forecasts of the total confirmed new cases of COVID-19 using the ARIMA(1,2,1) model.

According to the results in the Figure 7, the confirmed new cases of COVID-19 tends to increase rapidly in the next 5 days. The prediction model will help the government and medical workforce to be prepared for the upcoming situations and have more readiness in healthcare systems.

4. Discussion

Time series analysis of surveillance data on spread of various infections is very helpful in developing hypotheses to explain and anticipate the dynamics of the observed phenomena and subsequently in the establishment of a quality control system and reallocation of resources. ARIMA model is one of the most widely used time-series forecasting techniques because of its structured modeling basis and acceptable forecasting performance. In this paper, we applied an ARIMA(p, d, q) model to analyze the surveillance data of COVID-19 infection all over the world. Disease monitoring by public health department entails ongoing data collecting, processing, and updating. However, the World Health Organization is the appropriate level of organization for the implementation of an ARIMA predictive model, because reported data is continually received and updated. We found that model predictions are further improved by the assured availability of the Health Department data. In this study, we have obtained an ARIMA model that closely fits for the spread of COVID-19 all over the world. The autoregression and moving average parameters of our model imply the number of infection COVID-19 in a day can be estimated by the residual occurring one day prior. According to the results above, the conducted model is reliable with a high validity. Once a satisfactory model has been obtained, it can be used to forecast expected numbers of cases for a given number of future time intervals. The forecast results suggest that the total confirmed new cases of COVID-19 all over the world will experience a strong growth in the next five days (17th March 2020 to 21st March 2020). Therefore, knowledge of COVID-19 forecasts is necessary to prompt countries to strengthen surveillance systems and give appropriate solutions. Moreover, the fitted ARIMA(1,2,1) model can be used to predict the total number of COVID-19 infection for the following days.

Several studies have used ARIMA model to fit and predict changing trends in infectious disease. L.LIU et al applied an ARIMA$(1,0,1)\times(0,1,0)_{12}$ model to predict the incidence of hand, foot and mouth disease in Sichuan province, China [11] and found that ARIMA models were useful tools for monitoring hand, foot and mouth incidence. Earnest et al indicated that ARIMA models provided useful tools for administrators and clinicians in planning for real-time bed capacity during infectious diseases outbreaks such as SARS [12]. Li et al have applied an ARIMA model to monthly incidence of HFRS (Hemorrhagic Fever with Renal Syndrome) in Linyi City, China to predict HFRS incidence, and found that the ARIMA model could be used to predict HFRS incidence with high predictive precision in the short-term [13]. In the present study, we further confirmed the consensus that ARIMA model is a useful tool in monitoring and predicting changing trends in infectious diseases.

To the best of our knowledge, this is the first study to apply ARIMA model to fit the COVID-19 incidence in all the world with as many as 56 observations at day level. Some previous study [14,15] used ARIMA model to fit and forecast COVID-19 incidence of a region, but the problem that the number of observations was not enough, which led to the instability of their forecast results. In order to conduct a stable and effective ARIMA model, we have to collect at least 50 observations [7]. Thus, parameter estimates of the fitted model would be more robust. The longer the series, the better; however, the series should not extend so far into the past as to include periods during which a different case definition was applied or in which any other reporting artifact resulted in a mean number of cases per interval that differs from the mean of recent intervals. As mentioned above, for adequate ARIMA modeling, a time series should be stationary with respect to mean and variance. If the mean increases or decreases over time, or if the variance does, the series may need to be transformed to make it stationary, before being modeled. Otherwise, the prediction effect of the model will be poor.

In order to improve the model, updating the forecasts is very important. A model without seasonal terms will need to be updated frequently. Confidence intervals that widen rapidly as time increase from the starting point of the forecasts also indicate a model that needs frequent updating. Generally speaking, there are two ways to implement the updating. The model can be reapplied to the original series with extra observations added at the end to give forecasts based on a later starting point. Alternatively, a new model can be fitted to the longer series. This is probably preferable, since fitting a model is quick, especially when the old model is used as a guide, and it makes better use of the additional observations.

In the last part, we want to emphasize and explain of the the large number of cases in China on February 17th: You will notice a dramatic increase in the number of confirmed cases on 17th February 2020. As the WHO notes in its Situation Report 27, this is the result of a change in reporting methodology to include all confirmed cases including both laboratory-confirmed as previously reported, and those reported as clinically diagnosed. This change in methodology only affected figures in the Hubei province in China, but due to the large number of cases in this region it had a significant impact on global figures too. The possible biases in disease reporting and potential underreporting of COVID-19 cases might influence the precision of our analysis.

5. Conclusion

There is an urgent need for monitoring and predicting COVID-19 incidence to reduce the substantial morbidity and mortality caused by this disease [16]. ARIMA models applied to historical COVID-19 infection data are an important tool for COVID-19 surveillance. Accurate forecasting the number of infections of COVID-19 is possible. Our modeling approach can be used to monitor and predict the total confirmed new cases of COVID-19 infection all over the world for the following days. The ARIMA model could be used to optimize COVID-19 prevention by providing estimates on COVID-19 infection trends all over the world.

Acknowledgements

The authors would like to thank the referees for the valuable suggestions and comments, which improve the paper considerably. Funding was provided by Quy Nhon University.

References

[1]. Zhang Y, The epidemiological research status and problems and prospects of hemorrhagic fever with renal syndrome in China, Chin J Vector Bio & Control, (2002), 13 (2): 85-88.

[2]. Guan P, Huang DS and Zhou BS, Forecasting model for the incidence of hepatitis A based on artificial neural network, World J Gastroenterol, (2004), 10 (24): 3579-3582.

[3]. Reichert TA, Simonsen L, Sharma A, Pardo SA, Fedson DS and Miller MA, Influenza and the winter increase in mortality in the United States, 1959-1999, Am J Epidemiol, (2004), 160 (5): 492-502.

[4]. Gaudart J, Toure O, Dessay N, Dicko AL, Ranque S, Forest L, Demongeot J and Doumbo OK, Modelling malaria incidence with environmental dependency in a locality of Sudanese savannah area, Mali, Malaria Journal, (2009), 8: 61-10.

[5]. Luz PM, Mendes BV, Codeco CT, Struchiner CJ, Galvani AP, Time series analysis of dengue incidence in Rio de Janeiro, Brazil, Am J Trop Med Hyg, (2008), 79 (6): 933-939.

[6]. Yi J, Du CT, Wang RH, Liu L, Applications of multiple seasonal autoregressive integrated moving average(ARIMA) model on predictive incidence of tuberculosis, Chinese Journal of Preventive Medicine, (2007), 41 (2): 118-121.

[7]. Box GE, Jenkins GM, Time Series Analysis: Forecasting and Control. Rev. ed, San Francisco: Holden-Day, (1976).

[8]. Rami Krispin, Hands-On Time Series Analysis with R, Packt Publisher, UK, (2019).

[9]. Rob J. Hyndman and George Athanasopoulos, Forecasting: Principles and Practice, 2nd edition, OTexts Publisher, (2018).

[10]. Rob J. Hyndman and Yeasmin Khandakar, Automatic Time Series Forecasting: The forecast Package for R, Journal of Statistical Software, (2008) Volume 27, Issue 3.

[11]. Luz PM, Mendes BV, Codeco CT, Struchiner CJ, Galvani AP, Time series analysis of dengue incidence in Rio de Janeiro, Brazil, Am J Trop Med Hyg, (2008), 79 (6): 933-939.

[12]. Earnest A, Chen MI, Ng D, Leo YS, Using autoregressive integrated moving average(ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore, BMC Health Services Research, (2005), 5: 36-10.

[13]. Li XJ, Kang DM, Cao J, Wang JZ, A time series model in incidence forecasting of hemorrhagic fever with renal syndrome, Journal of Shandong University (Health Sciences), (2008), 46 (5): 547-549.

[14]. Domenico Benvenuto, Marta Giovanetti, Lazzaro Vassallo, Silvia Angeletti, Massimo Ciccozzi, Application of the ARIMA model on the COVID-2019 epidemic dataset, Data in Brief, Published by Elsevier Inc, (2020).

[15]. Xinguang Chen and Bin Yu, First two months of the 2019 Coronavirus Disease (COVID-19) epidemic in China: realtime surveillance and evaluation with a second derivative model, Global Health Research and Policy, Published by BMC, (2020).

[16]. Website: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public.

Contact information with corresponding author

*Corresponding author: **Nguyen Quoc Duong**

Author’s address: Department of Education, Quy Nhon University, Viet Nam

170 An Duong Vuong, Quy Nhon City, Binh Dinh Province, Viet Nam

Email: nguyenquocduongqnu1999@gmail.com

Phone number: 0375113359