Advances in Social Science, Education and Humanities Research, volume 550
Proceedings of the 1st International Conference on Mathematics and Mathematics Education (ICMMED 2020)

ARIMA Model for Forecasting COVID-19 in East Java

E S Nugraha* A H Ulya

School of Business, Study Program of Actuarial Science, President University, Cikarang, Bekasi 17550, Indonesia
**E S Nugraha. Email: edwin.nugraha@president.ac.id*

ABSTRACT
Coronavirus is a group of viruses that can cause disease in both humans and animals. The newly discovered coronavirus triggers COVID-19 disease. COVID-19 is now a pandemic that is emerging in many countries around the world, including Indonesia. Several sectors have been affected as a result of this pandemic, such as medical, economics, government, industry, etc. By using ARIMA model, we try to predict the daily cases of COVID-19 that occur in East Java. We obtained this model with the help of R software. The best model we obtained was the ARIMA model (7,1,7), which we used to predict the next 14 days from November 1, 2020, to November 14, 2020. The results of forecasting obtained by comparing real data with a 99% confidence interval, we obtained that the forecasting results are close to the real data that has occurred until November 14, 2020. This prediction is expected to help various sectors affected by this pandemic, such as government, economy, health especially in East Java.

Keywords: *COVID-19, Statistical Modelling, Time Series Analysis, ARIMA.*

1. INTRODUCTION

Based on government data seen by the South China Morning Post (SCMP), the first case of Coronavirus Disease occurred on November 17, 2019, which infected a 55-year-old person from Hubei province [1]. Furthermore, this disease is known as COVID-19 which stands for Coronavirus Disease 2019. Until now, almost every country in the world still fight with the COVID-19 disease, it was recorded that up to October 31, 2020, there were 46,414,696 cases and 1,202,167 deaths in the world [2]. Indonesia has been facing the COVID-19 disease since the first case in Indonesia which occurred on March 2, 2020, was announced directly by President Joko Widodo, in which 2 people contracted the COVID-19 virus, namely a 64-year-old mother and 31-year-old daughter [3]. As of October 31, 2020, there were 410,088 cases and 13,869 deaths in Indonesia [4]. In dealing with and preventing the increase in COVID-19, several forms and policies have been taken by the government, including implementing social distancing and large-scale social restrictions (LSSR). Social distancing means that everyone must stay away from formal forms of association, keep a distance from other people, and avoid meetings that involve many people. Large-scale social restrictions (LSSR) are restrictions imposed by local governments, but include measures such as closing public places, restricting public transportation, and restricting travel to and from restricted areas [5]. The government is taking this action because when someone who is infected with COVID-19 coughs, sneezes or talks, the COVID-19 virus will mainly spread from person to person through droplets from the nose or mouth that come out. This splash will stick around individuals, such as tables, doorknobs, and handrails, to other items and surfaces. When they inhale or touch items that an infected person has splashed and then touch their eyes, nose or mouth, that person can become infected with COVID-19.

Several models have been developed for forecasting COVID-19, for example understanding the trend of the outbreak and providing an overview of the epidemiological stage of COVID-19 in Italy, Spain and France. Prediction results for the next 10 days, April 16, 2020 to April 25, 2020, show that new cases will range between 196,520-229,147 in Italy, 204,755-257,497 in Spain, and 140,320-159,619 in France [6]. Estimated baseline reproductive number (R_0), and mortality rate and infection recovery per day for COVID-19 based on the Susceptible Infectious-Recovered-Dead (SIDR) model. There are two scenarios studied, namely using exact numbers for confirmed cases, and using x20 contaminated data x40 data recovered from confirmed cases. Prediction results for Hubei province on February 29, 2020, using the first scenario showed that the total number of infected cases was around 45,000-760,000, recovered 22,000-170,000, and died 2,700-34,000 [7]. Optimization Method for Forecasting Confirmed Cases of COVID-19 in China using an adaptive neuro-fuzzy

inference system (ANFIS) model. The prediction results for the next 10 days, February 19, 2020 to February 28, 2020, show that the outbreak will reach its highest point on February 28, 2020, with the average percentage increase over the estimated period of 10%, with the highest percentage occurring on February 28, 2020. at 12%, and the lowest on February 19, 2020 at 8.7% [8]. Developed the ARIMA model based on time series analysis and then used it to estimate future COVID-19 cases in India. The prediction results for the next 20 days, April 14, 2020 to May 3, 2020, show an upward trend with the lowest limit is 11,070.1, and the highest limit is 37,728.2 [9].

There are several reports in the literature that forecast the distribution of COVID-19 in many countries around the world. Indonesia as the 4th most populous country in the world with a population of 267,026,366 people is certainly very vulnerable to the spread of COVID-19 [10]. East Java is in the second position with the most total COVID-19 cases in Indonesia with a total of 52,465 cases, and 3,768 people died as of October 31, 2020 [11]. With the implementation of the LSSR in dealing with the spread of COVID-19, East Java as the province with the second largest population in Indonesia will certainly be greatly affected. Therefore, we created an Auto-Regressive Integrated Moving Average (ARIMA) model to predict daily cases in the next 14 days with the help of R software. This model is expected to support medical professionals, policy makers, the general public, and make it easier us to predict how best to deal with this pandemic in the future.

2. METHODOLOGY

Box-Jenkins models or also known as Auto-Regressive Integrated Moving Average (ARIMA) model is the development of parametric models for certain types of non-stationary time series cases. Here, we follow the ARIMA procedure in [12]. Components that contained in ARIMA model, such as Autoregressive Model (AR), Integrated (I), and Moving Average (MA).

Autoregressive Models indicates that the changing vector of interest is reduced on its own lagged values. The general formula for autoregressive model can be written as follows:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + e_t \quad (1)$$

Integrated means that the data has been substituted by the difference between their values and the previous values. This separation process may have been carried out more than once. The main goal of both of these features is to ensure that the model matches the data as best as possible.

Moving Average indicates that the regression error is the amount by which an observation differs from its expected value. In truth, it is a linear combination of error terms where values have existed simultaneously and at different times in the past. The general formula for moving average model can be written as follows:

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots - \theta_q e_{t-q} \quad (2)$$

Several steps need to be done when creating an ARIMA model, which can be seen in the flow chart in Figure 1.

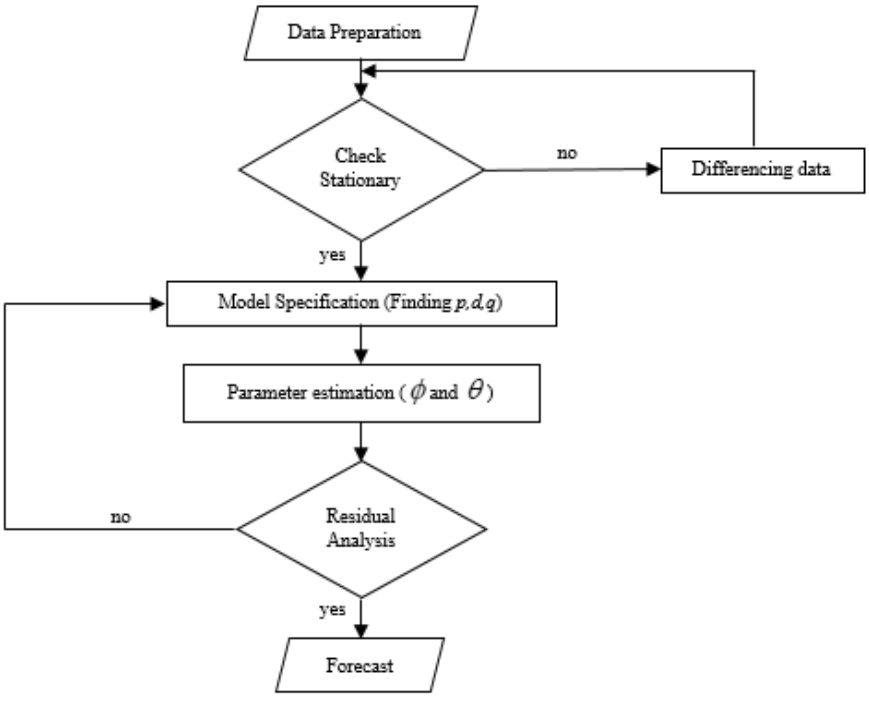

Figure 1. Flow chart of creating ARIMA model

2.1. Data Preparation

Prepare the data that will be analyzed to obtain the ARIMA model. The first step of data preparation is the data collection that can be found from reliable sources. Then, do data profiling and cleansing to prevent missing data and other issues. The last one is structuring, transforming, and validating the data so that it fits the format that we can use in data modelling.

2.2. Check Stationary

Stationary indicates that the structural properties of a time series process do not change over time or if it has constant mean and variance, and covariance is independent of time. We can judge that data is stationary or not using R software, there are many functions that can be performed, which is:

- *ACF (Autocorrelation function)*: if the graph goes to zero then the data is stationary, and vice versa.
- *ADF (Augmented Dickey-Fuller) test*: if the p-value is less than 0.05, then the data is stationary, and vice versa.

2.3. Identification

At this stage we will look for or determine p, d, and q. determining p and q can be done with the help of the ACF and partial auto correlation function (PACF). Meanwhile d is determined from the order of differencing.

- Sample autocorrelation function.

$$r_k = \frac{\sum_{t=k+1}^{n} \left(Y_t - \bar{Y}\right)\left(Y_{t-k} - \bar{Y}\right)}{\sum_{t=1}^{n} \left(Y_t - \bar{Y}\right)^2} \quad \text{for } k = 1,2,... \qquad (3)$$

- Sample partial autocorrelation function.

$$\phi_{kk} = \frac{\rho_k - \sum_{j=1}^{k-1} \phi_{k-1,j}\rho_{k-j}}{1 - \sum_{j=1}^{k-1} \phi_{k-1,j}\rho_j} \qquad (4)$$

2.4. Parameter Estimation

If we already got p, d and q from identification, then we can make the parameter estimation by some methods, there are:

2.4.1. The method of moments estimator

The procedure consists of equating the sample moments with the related theoretical events and implementing the resulting equations in order to obtain estimates of some uncertain parameters.

2.4.2. The Least Squares Estimator

Since the timing approach is unsatisfactory for certain models, we need to consider other estimation approaches. At this point, we are adding a potential nonzero mean in our stationary models and considering it as another parameter to be calculated by the least squares.

2.4.3. Full Maximum Likelihood Estimators

The benefit of this approach is that all the details in the data is used instead of either the first or second moments, as is the situation with the least squares. Another benefit is that certain large-scale effects are defined under very general conditions. One weakness is that for the first time, we need to focus explicitly on the joint probability density function of the method. Maximum probability or likelihood estimators are then categorized as those values for which the results currently observed are most probably to be the values that maximize the probability function.

2.5. Diagnostic Check (Residual Analysis)

From the residuals, a good model can be seen. If the residual is white noise, the model can be assumed to be nice and vice versa. White noise can be checked using the ACF and PACF residual correlations. If the ACF and PACF insignificant, it indicates residual white noise meaning the model is suitable.

2.6. Forecasting

After the best model is obtained, the next process is forecasting which allows us to know what will happen in the future. It is important for preparing an effective strategy in the future. So that we can better anticipate the consequences if predictions do not match expectations. A model that has produced fairly good forecasting results, of course, still needs to be developed, by following the data development in order to produce better forecasts.

3. ANALYSIS AND EVALUATION

3.1. Data Preparation

We use data on daily COVID-19 cases that have occurred in East Java from 1st April to 31th October which is available at kawalcovid19.id [4]. A summary statistics of COVID-19 cases can be seen in the Table 1, as well as the COVID-19 chart can be seen in the Figure 2. To make us more understand the trend and mean value estimation in time series analysis we can use MA, in this case, we use MA(7) and MA(30). We use moving averages since moving averages are one form of basic

smoothing and are widely used for time series analysis and time series forecasting. By allowing this smoothing happen, we expect to remove noise and further show the signal of the cause - effect processes [13].

Table 1. Summary of COVID-19 cases

Type	Date	Cases
Min.	04/01/2020	47.0
1st Qu.	05/24/2020	123.0
Median	07/16/2020	123.0
Mean	07/16/2020	490.0
3rd Qu.	09/07/2020	893.5
Max.	10/31/2020	1398.0

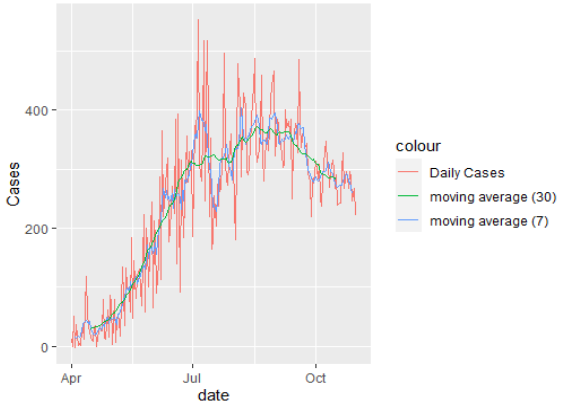

Figure 2. East Java daily cases.

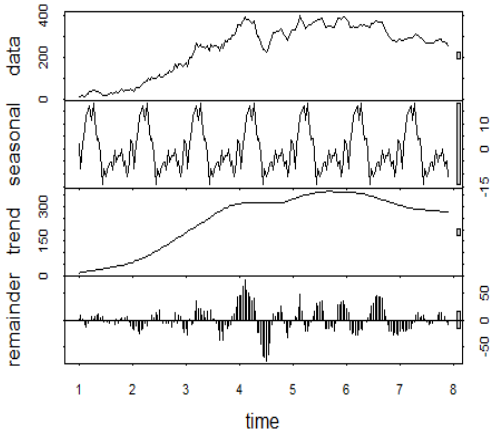

Figure 3. Decompose

Figure 3 demonstrate that our data has a trend and some seasonal components. So, we should build our data for time series and removes the most extreme observation as an outlier by its internal algorithm, so it made smoothen the portion of the peak in the curve. We can do this easily with the help of R software. The plot of the time series for the COVID-19 model in East Java can be seen in Figure 4.

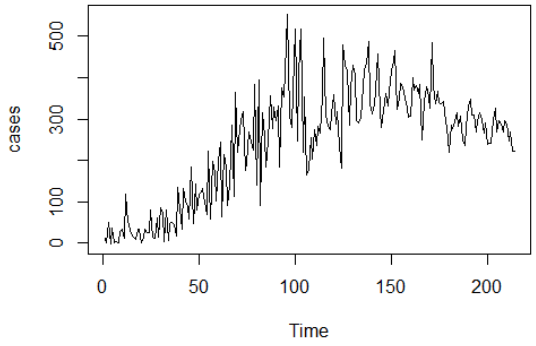

Figure 4. Time series plot

3.2. Check Stationary

By using the ADF function on R, it can be seen that this data is not stationary because the p-value is > 0.05. Therefore, differencing data is done to produce stationary data, in this data, stationary data can be obtained in the first differencing. The difference in the ADF test before and after differentiation can be seen in Figure 5 and Figure 6, respectively.

```
        Augmented Dickey-Fuller Test

data:  cases_ma
Dickey-Fuller = -0.91459, Lag order = 5,
p-value = 0.9499
alternative hypothesis: stationary
```

Figure 5. Before differencing.

```
        Augmented Dickey-Fuller Test

data:  cases_d1
Dickey-Fuller = -5.4205, Lag order = 5,
p-value = 0.01
alternative hypothesis: stationary
```

Figure 6. After differencing.

3.3. Identification and Parameter Estimation

To determine the correct model, we can use the auto.arima function on R. In this case, we got ARIMA(1,1,2) as our best model. But to make it better we can find the Model with a lower Akaike Information Criterion (AIC) value by seeing the ACF and PACF of ARIMA(1,1,2). Eventually, our best model is ARIMA(7,1,7). The plot ACF and PACF of model ARIMA(7,1,7) shown in the Figure 7 and Figure 8, respectively.

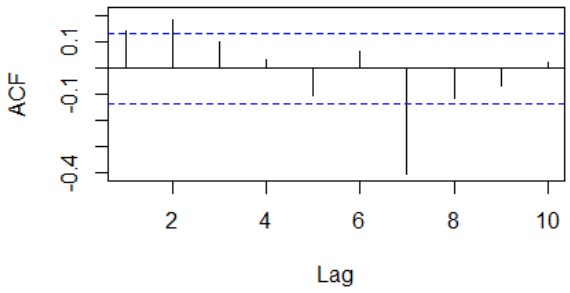

Figure 7. ACF of differenced data

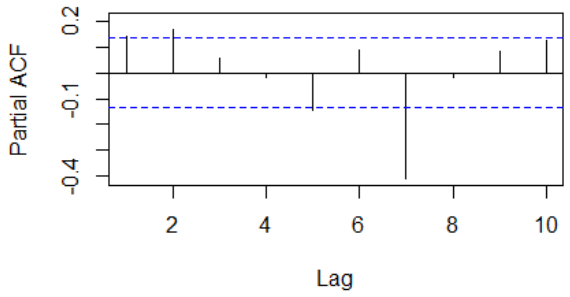

Figure 8. PACF of differenced data

We can see the difference of AIC value between ARIMA(1,1,2) and ARIMA(7,1,7) from Figure 9 and Figure 10.

```
Coefficients:
          ar1      ma1     ma2   drift
      0.4933  -0.3918  0.1430  1.2174
s.e.  0.2012   0.2069  0.0713  1.2156

sigma^2 estimated as 143:  log likelihood
=-805.4
AIC=1620.81   AICc=1621.11   BIC=1637.47
```

Figure 9. ARIMA(1,1,2)

```
Coefficients:
          ar1     ar2     ar3      ar4
      -0.0329  0.1561  0.0210  -0.0284
s.e.   0.1499  0.1317  0.0993   0.0886
          ar5     ar6     ar7      ma1
      -0.1078  0.0179  0.0202   0.1971
s.e.   0.0881  0.0921  0.0816   0.1355
          ma2     ma3     ma4      ma5
       0.1406  0.2258  0.1688   0.2897
s.e.   0.1350  0.1250  0.1277   0.1270
          ma6     ma7
       0.2080  -0.7690
s.e.   0.1414   0.1373

sigma^2 estimated as 77.09:  log likelihood
= -752.55,  aic = 1535.09
```

Figure 10. ARIMA(7,1,7)

3.4. Diagnostic Check (Residual Analysis)

The residual analysis can be proven by looking at the QQ plot, histogram, Shapiro test, and Ljung Box function. As we can see from the QQ plot in Figure 11, a small deviation of the residue from the straight line can be seen from the graph. This implies that the errors are nearly standard with a few outliers. The assumption of normality is then followed. This statement is backed up by the residual histogram as we can see from Figure 12, Shapiro test that has p-value more than 0.05 as shown in Figure 13, and also plot of model residual ARIMA(7,1,7) that shown in Figure 14. The Ljung Box as seen in Table 2 indicates that the p-value of every lags is higher than the significance amount (0.05), which implies that there is no breach of an independent assumption. The model equation of ARIMA(7,1,7) is:

$$\begin{aligned} Y_t = {} & -0.0329Y_{t-1} + 0.1561Y_{t-2} + 0.0210Y_{t-3} - \\ & 0.0284Y_{t-4} - 0.1078Y_{t-5} + 0.0179Y_{t-6} + \\ & 0.0202Y_{t-7} + 0.1971e_{t-1} + 0.1406e_{t-2} + \\ & 0.2258e_{t-3} + 0.1688e_{t-4} + 0.2897e_{t-5} + \\ & 0.2080e_{t-6} - 0.7690e_{t-7} + e_t \end{aligned} \quad (5)$$

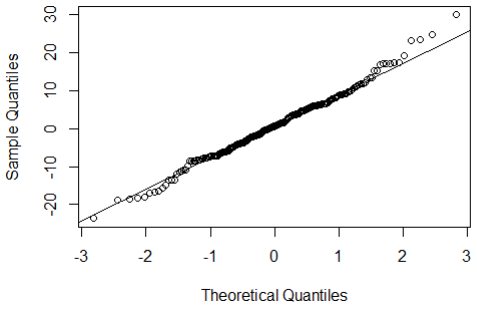

Figure 11. Normal QQ Plot

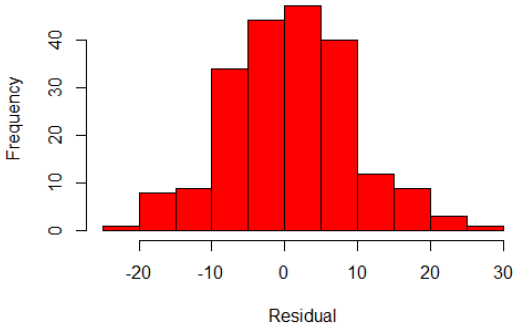

Figure 12. Histogram

```
        Shapiro-Wilk normality test

data:  fit2$residuals
W = 0.99193, p-value = 0.3058
```

Figure 13. Shapiro test

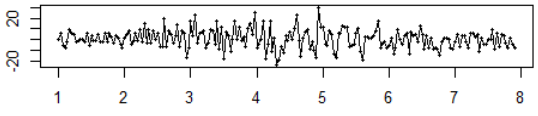

Figure 14. ARIMA(7,1,7) model residual

Table 2. Ljung-Box

X^2	df	p-value
0.26871	7	0.9999
11.953	14	0.6101
19.724	21	0.5388
36.378	28	0.1332

3.5. Forecasting

The forecasting of COVID-19 confirmed cases and their lower and upper limits for 14 days, November 1, 2020 to November 14, 2020, with 99% confidence interval (CI) can be seen in Table 3, followed by forecasting plot shown as blue line in Figure 15.

Table 3. Comparison between forecasting COVID-19 cases and real data

Date	Forecast	Lower limit	Upper limit	Real data
11/1/2020	262.4969	239.6583	285.3355	253
11/2/2020	256.4356	221.3344	291.5367	284
11/3/2020	253.2447	204.9949	301.4946	272
11/4/2020	256.9856	194.9229	319.0483	239
11/5/2020	252.9956	177.4712	328.5199	278
11/6/2020	255.4209	166.0911	344.7508	289
11/7/2020	261.2948	157.5129	365.0766	269
11/8/2020	261.3800	151.5588	371.2011	282
11/9/2020	261.8753	146.0491	377.7015	234
11/10/2020	262.3598	141.6683	383.0513	272
11/11/2020	261.9985	136.7857	387.2114	168
11/12/2020	261.4238	131.9294	390.9182	270
11/13/2020	261.5275	127.4212	395.6339	239
11/14/2020	261.4797	122.9993	399.9601	256

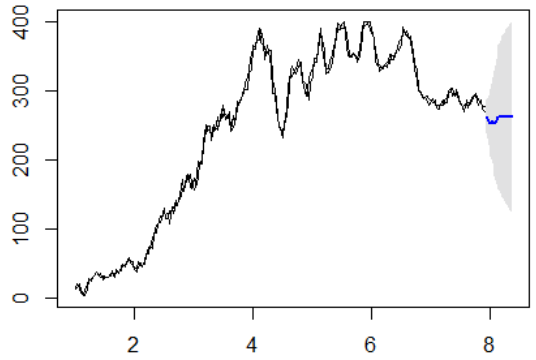

Figure 15. Forecasts from ARIMA(7,1,7)

4. CONCLUSION

In this work, we found the best ARIMA model for predicting COVID-19 over the next 14 days. The dataset used is daily COVID-19 cases in East Java from April 1, 2020 to October 31, 2020 [4]. The results showed that the ARIMA(7,1,7) is the best model for this case where the mathematical model are expressed below.

$$\begin{aligned} Y_t = &-0.0329Y_{t-1} + 0.1561Y_{t-2} + 0.0210Y_{t-3} - \\ &0.0284Y_{t-4} - 0.1078Y_{t-5} + 0.0179Y_{t-6} + \\ &0.0202Y_{t-7} + 0.1971e_{t-1} + 0.1406e_{t-2} + \\ &0.2258e_{t-3} + 0.1688e_{t-4} + 0.2897e_{t-5} + \\ &0.2080e_{t-6} - 0.7690e_{t-7} + e_t \end{aligned} \quad (6)$$

Forecast results for 14 days from November 1, 2020 to November 14, 2020, using the model are compared with the actual data display the fairly decent result, where all the actual data is included in the forecast interval. These research results will be good input in helping medical professionals, regulators, and the general public, especially in the East Java area to deal with this pandemic in the future.

AUTHORS' CONTRIBUTIONS

A H Ulya contributed to data collection, visualization making, ARIMA programming, and analysis using R. E S Nugraha provides an analysis of R programming, and perfecting the overall manuscript. All authors revised the paper critically and approved the final manuscript.

ACKNOWLEDGMENTS

The author would like to thank the anonymous referee who carefully read the original paper and gave a lot valuable comments and suggestions to improve the quality of this manuscript.

REFERENCES

[1] J. Ma. "Coronavirus: China's First Confirmed Covid-19 Case Traced Back to November 17." scmp.com. https://www.scmp.com/news/china/society/article/3074991/coronavirus-chinas-first-confirmed-covid-19-case-traced-back (accessed Nov. 12, 2020).

[2] "Coronavirus Worldwide Graphs." worldometers.info. https://www.worldometers.info/coronavirus/worldwide-graphs/#total-cases (accessed Nov. 1, 2020).

[3] J. Akbar. "Perjalanan Pandemi Covid-19 di Indonesia." kompas.com. https://www.kompas.com/tren/read/2020/07/28/060100865/perjalanan-pandemi-covid-19-di-indonesia-lebih-dari-100.000-kasus-dalam-5?page=all (accessed Oct. 31, 2020).

[4] Kawal Covid19, Nov 2020, "Kasus Covid19 di Indonesia." kawalcovid19.id. [Online]. Available: https://experience.arcgis.com/experience/bf4eb5d76e98423c865678e32c8937d4 (accessed Nov. 1, 2020).

[5] "Pembatasan Sosial Berskala Besar." kemenkopmk.go.id. https://www.kemenkopmk.go.id/pembatasan-sosial-berskala-besar (accessed Nov. 2, 2020).

[6] Z. Ceylan, "Estimation of COVID-19 prevalence in Italy, Spain, and France," Sci. Total Environ., vol. 729, no. 138817, pp. 1-7, Aug. 2020.

[7] C. Anastassopoulou, L. Russo, A. Tsakris, and C. Siettos, "Data-based analysis, modelling and forecasting of the COVID-19 outbreak," PloS one, vol.15, no. 3, pp. 1-21, Mar. 2020. Accessed on Nov. 12, 2020. [Online]. Available: doi: https://doi.org/10.1371/journal.pone.0230405.

[8] Al-qaness Mohammed A. A., A. A. Ewees, H. Fan, and M. Abd El Aziz, "Optimization Method for Forecasting Confirmed Cases of COVID-19 in China," Journal of Clinical Medicine,vol. 9, no. 3, p. 674, Mar. 2020. Accessed on Nov. 12, 2020. [Online]. Available: http://dx.doi.org/10.3390/jcm9030674.

[9] H. Tandon, P. Ranjan, T. Chakraborty, and V. Suhag, "Coronavirus (COVID-19): ARIMA based time-series analysis to forecast near future," 2020. [Online]. Available: arXiv:2004.07859.

[10] "U.S. Census Bureau Current Population." census.gov. https://www.census.gov/popclock/print.php?component=counter (accessed Nov. 13, 2020).

[11] R. Austen. "Statistik Covid19 per Provinsi." kawalcovid19.id. https://kawalcovid19.id/ (accessed Nov. 13, 2020).

[12] J. D. Cryer and K. S. Chan, "Models for Non Stationary Time Series" in Time series analysis: with applications in R, second Ed. NY, New York, USA: Springer Science and Business Media, 2008, pp. 87-102.

[13] K. Strom, "Introduction to construction statistics using Excel." sciencedirect.com. https://www.sciencedirect.com/topics/engineering/moving-average (accessed Nov. 11, 2020).