ARFIMA Model for Short Term Forecasting of New Death Cases COVID-19

Puspita Kartikasari*, *Hasbi* Yasin, *Di* Asih I Maruddani

Department of Statistics, Faculty of Science and Mathematics, Diponegoro University, Semarang – Indonesia

Abstract. COVID-19 is an infectious disease that can spread from one person to another and has a high potential for death. The infection of COVID-19 is spreading massive and fast that causes the extreme fluctuating data spread and long memory effects. One of the ways in which the death of COVID-19 can be reduce is to produce a prediction model that could be used as a reference in taking countermeasures. There are various prediction models, from regression to Autoregressive Fractional Integrated Moving Average (ARIMA), but it still shows shortcomings when disturbances occur from extreme fluctuations and the existence of long memory effects in the form of analysis of a series of data becomes biased, and the power of statistical tests generated for identification become weak. Therefore, the prediction model with the Autoregressive Fractional Integrated Moving Average (ARFIMA) approach was used in this study to accommodate these weaknesses because of their flexible nature and high accuracy. The results of this study prove that ARFIMA (1,0,431.0) with an RMSE of 2,853 is the best model to predict data on the addition of new cases of patients dying from COVID-19.

Keywords. ARFIMA; Prediction; Death; COVID-19.

1. Introduction

COVID-19 is a respiratory infection that can spread from one person to another. This disease is caused by Novel Viruscorona (SAR-Cov-2) which was first identified in Wuhan China in December 2019 and has spread in various countries [1]. At the beginning of identification, sufferers of this disease have a high potential for death [2]. The World Health Organization (WHO) has designated the disease as a global pandemic due to its rapid and massive transmission over time [3]. One way to reduce the mortality rate caused by COVID-19 is to produce a prediction model so that it can be a reference in taking countermeasures.

COVID-19 which occurred in Indonesia has infected 1528 people in a period of less than one month from March 2, 2020, 2 confirmed cases were reported. On March 29, 2020, this

* Corresponding author: puspitakartikasari@live.undip.ac.id

case increased to 1,285 cases in 30 provinces, the five highest provinces including Jakarta, West Java, Banten, East Java, and Central Java [4, 5].

Time series is one of the most popular methods in statistics for making prediction models. This is because the method is simple but able to solve more complex problems, if the case under study is affected by time [6]. Several studies using this method in predicting an event both in the fields of kymiatology, agriculture or health. Studies that examine these include [7-12].

There are various models in the time series including decomposition models, Autoregressive Integrated Moving Average (ARIMA), Seasonal Autoregressive Integrated Moving-Average (SARIMA), Seasonal Autoregressive Integrated Moving-Average with Exogenous Regressors (SARIMAX), Vector Autoregression Moving-Average (SARIMA) VARMA), Vector Autoregression Moving-Average with Exogenous Regressors (VARMAX), Simple Exponential Smoothing (SES), Holt Winter's Exponential Smoothing (HWES) and others. However, the accuracy of these methods is still weak when extreme fluctuations occur and is unable to capture if there are observations that have a strong enough correlation with other observations even though the distance between observations is quite far (long memory). From several time series models, the Autoregressive Fractional Integrated Moving Average (ARFIMA) is a model that is able to capture extreme fluctuations and long memory. This happens if the case under study experiences a continual change over time. [13-20].

2. Methods

The method used in this research is a case study by applying theory to analyze data in concluding additional cases of dead patients caused by the COVID-19 pandemic, this is done to be able to determine preventive measures. To achieve this goal, the following steps are taken.

The first step is to describe the died characteristics of COVID-19 patient data from March 3, 2020 to June 1, 2020. Data is divided into two parts, in sample data and out sample data. In sample data starts from March 2, 2020 until June 1, 2020, while out sample data starts from June 2, 2020 to June 11, 2020. The second stage is to identify the existence of the long memory with $0 < d < 0,5$. This can be done by several estimation methods including Geweke and Porter Hudak (GPH), Nonlinear Least Square (NLS), Exact Maximum Likelihood (EML) and Modified Profile Likelihood (MPL) [21]. However, in this study using the GPH Estimator because the parameter d estimation in the GPH method can be done directly without knowing the values of the p and q parameters first [22].

$$\hat{d}_{GPH} = \frac{-0,5\sum_{j=1}^{m}\left(\tilde{X}_j - \bar{\tilde{X}}\right)\log I_j}{\sum_{j=1}^{m}\left(\tilde{X}_j - \bar{\tilde{X}}\right)^2} \qquad (1)$$

The next step is to create an ARFIMA model by making a time series plot, transforming data if the data does not meet the assumption of homogeneity in variance, making ACF and PACF plots of data that have been transformed, setting one or more ARFIMA models in accordance with the ACF and PACF plots of the results of the previous step, do the estimation of the model parameters and choose the best ARFIMA based model

$AICc = AIC + \frac{2M(M+1)}{n-M-1}$ for in sample data and $MSE = \frac{1}{L}\sum_{i=1}^{L} e_i^2$ for out sample data

After that carry out diagnostic tests assuming white noise and normal distribution using the Ljung-Box test and Kolmogorov Smirnov

Ljung-Box $Q = n(n+2)\sum_{k=1}^{K} \frac{\hat{\rho}_k^2}{n-k}$ and *"Kolmogorov Smirnov"*

$D = Sup\left|S(x) - F_0(x)\right|$ [23]

The last step is forecast the next 10 periods, then calculate the RMSE value from the forecast data obtained

$$RMSE = \sqrt{\frac{1}{L}\sum_{i=1}^{L} e_i^2} \qquad (2)$$

3. Results

3.1 Description of Data

The following is a description of data from COVID-19 patients who were declared dead from March 2, 2020 to June 11, 2020.

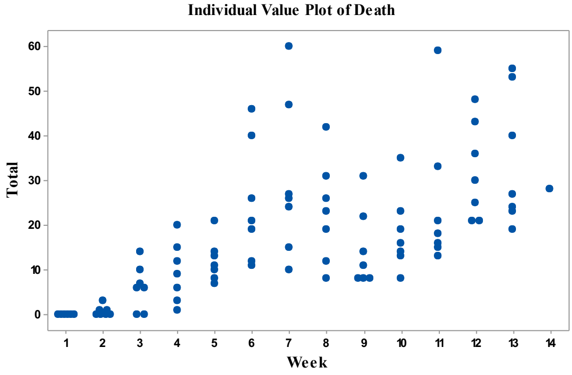


Fig. 1. Scatter Plot Description of Death Patient Data

Fig. 1. Shows the results from 1th week to 13 th week the number of patients dying from COVID-19 is increasing. On March 2, 2020, the highest total cases occurred in the 13th week, with an average of 34 patients per day and a minimum of 19 people. On April 14, 2020 which meant 60 people in one day. At the end of the 3 rd month or 9 th week there was a decrease but it did not need to be significant because positive confirmation was increasing. This phenomenon causes because in the 2 nd month the handling of this pademic did not go right, there was no enforcement of the rules of health protocols such as physical distension, social

distension and this was also the initial period of the pandemic entering Indonesia. While at the end of week 13 the easing of social boundaries began and coincided with the entry of Ramadan and Eid al-Fitr which is the culture of Indonesian people going home so that community mobilization is higher.

3.2 Testing Long Memory Data for Patients Died by COVID-19

Before create the ARFIMA model, the first thing to do is making a time series plot like in Fig. 1. to see the data patterns of patients dying from COVID 19 every day.

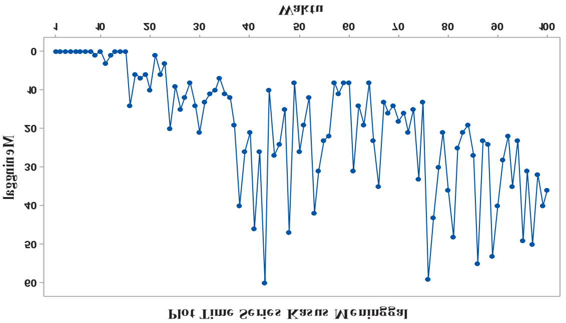


Fig. 2. Time Series plot for patients who have been declared dead due to COVID-19

Based on Fig. 2. it can be seen that data patterns are not stationary in variance, this causes the need for Box-Cox transformation as shown in Fig. 3.

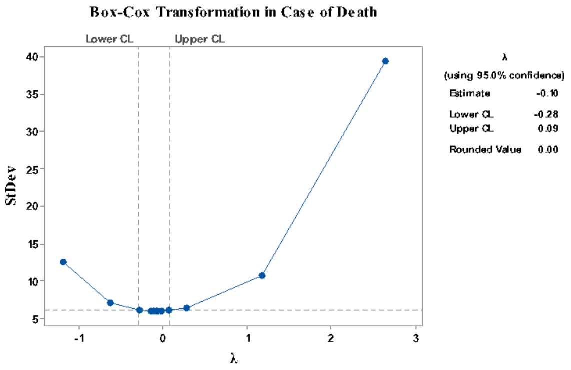


Fig. 3. Box Cox Transformation in Case of Death

Fig. 3. Shows the value of λ = 0, with each transformation in the patient's death data obtained a time series plot that resembles a straight line, the ACF and PACF plots also resemble the initial ACF and PACF data as shown in Fig. 4. and Fig. 5. for patients who died. Therefore, in this study ignoring data that was stationary on variants both data of patients recovering from COVID-19 disease and patients who died from COVID-19.

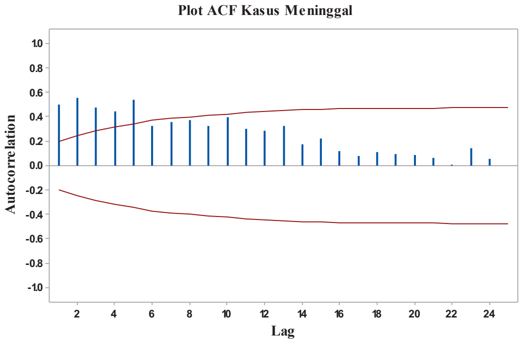


Fig. 4. ACF plot of patients who were declared dead due to COVID-19

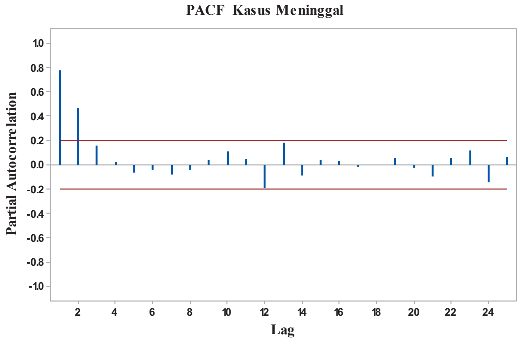


Fig. 5. PACF plot of data for patients who have been declared dead due to COVID-19

The next step is to identify the long memory data of patients who died from COVID-19, this is done to see whether there is a long memory effect (long-term dependency). The way to do this is to observe ACF in Fig. 4. The ACF plot indicates the alleged long memory in the recovered data of COVID-19 patients and the data of COVID-19 patients who die every day due to ACF plots moving down slowly.

Apart from the ACF plot, to prove that the data follows a long memory pattern by estimating the parameter d by using the GPH Estimator. The estimated parameter d in the data of patients who died due to COVID-19 disease was 0.488. This shows that the GPH Estimator value for data of patients who recovered and died had values between 0 and 1. This proves that the data is following a pattern of long memory. From the detection of long memory and long memory tests that have been done above, the data of patients who die from COVID-19 every day can be modeled using the ARFIMA model.

3.3 ARFIMA Model for data of patients who are declared dead every day due to COVID-19 disease

After the descriptive statistics and long memory identification phase, the next analysis is ARFIMA modeling of patients who died from COVID-19. If the normality assumptions of the model residuals are not met, then the analysis continues by showing the value of kurtosis.

Data of patients who die every day from COVID-19 disease are modeled with the ARFIMA model and get d = 0.431 with p-value = <2e-16 which means that the d value is significant in the model because it is less than α = 0.05. This modeling is based on Fig. 5 which is the absence of a significant lag (out of the $^2/_{\sqrt{n}}$ limit) after lag 2, whereas in Fig. 6 shows that ACF drops slowly. This shows that the AR model was used as an estimate in this case.

After conducting a number of trials that included significant lags (out of bounds), the estimation models obtained and the corresponding data for patients recovering daily from COVID-19 disease are presented in **Table 1.**

Table 1. Parameter Estimation of ARFIMA Model

	Estimate	Std. Error	z value	Pr(>|z|)
d	0,431	0,009	48,390	<2e-16
AR (1)	-0,236	0,102	-2,323	0,0202

Table 1. shows that the estimated parameter for AR (1) was -0.236 with an estimated value of d of 0.431. The overall coefficient of the parameter is significant for the model because the p-value < α, so the ARFIMA model formed is ARFIMA (1,0,431.0) with an AIC value of 712,699, so the model formed as follows.

$A_1\nabla^{0,431}Z_t \;= A_2 a_t$, with
A_1= $1 + 0{,}236B$
A_2= 1 ,
Or it can be written using the following model.

$$(1 + 0{,}236B)(1 - B)^{0{,}431}Y_t = a_t \qquad (3)$$

After getting the best ARFIMA model, the next step is checking the diagnosis, testing the residual to see whether the residual meets the assumption of white noise and has a normal distribution.

Table 2. ARFIMA Model Residual White Noise Testing (1,0.429.0)

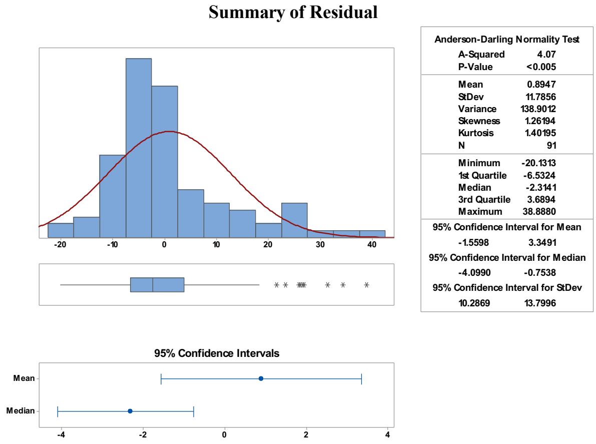
X-Squared	df	P-Value
55,32	48	0,218

Table 2 shows that the L-Jung Box test p-value statistic is more than 5%, which is equal to 0.218. This means that the ARFIMA residual model (1,0,431.0) meets the assumption of white noise residuals. The next step is to test the normality distribution using the Kolmogorov Smirnov test, where the results are presented in Table 3.

Table 3. Testing the Residual Normality Model of the ARFIMA (1,0,431.0)

D	P-Value
0,5023	1,332e-15

Based on the results of the normality test using the Kolmogorov Smirnov test in Table 3., the results of the residual assumptions are not normally distributed. After checking the kurtosis value, a value of 1.465 can be seen in Fig. 6., this causes the residuals not to have a normal distribution.



Anderson-Darling Normality Test	
A-Squared	4.07
P-Value	<0.005
Mean	0.8947
StDev	11.7856
Variance	138.9012
Skewness	1.26194
Kurtosis	1.40195
N	91
Minimum	-20.1313
1st Quartile	-6.5324
Median	-2.3141
3rd Quartile	3.6894
Maximum	38.8880
95% Confidence Interval for Mean	
-1.5598	3.3491
95% Confidence Interval for Median	
-4.0990	-0.7538
95% Confidence Interval for StDev	
10.2869	13.7996

Fig. 6. Graphycal Summary of Residual ARFIMA Model (1,0.431,0).

The results from ARFIMA modeling (1,0,431.0) yielded forecasting for the next 10 periods presented in Table 4.

Table 4. ARFIMA Model Forecasting Results (1,0.431.0)

Data	Forecast
92	31.2069
93	29.3407
…	…
101	25.8661
RMSE	**2.853**

Based on the forecast results in Table 4. it can be seen that the forecast results have a good value of the model for RMSE of 2.853, which means that the ARFIMA model (1.0.431.0) is very well used to predict patients who die from COVID-19 disease going forward.

4. Conclusion

The best Integrated Fractional Integrated Moving Average (ARFIMA) model obtained to predict cases of patients dying from a COVID-19 pandemic in Indonesia every day is $(1 + 0{,}236B)(1 - B)^{0{,}431}Y_t = a_t$ with Root Mean Square Error (RMSE) of 2,853. This is because the ARFIMA model is able to accommodate well the long memory effect, resulting in a small bias. Also in estimating model parameters, it is also simpler. By knowing the addition of patients who died from COVID-19, we can take anticipatory steps and decisions that need to be made.

Acknowledgement

This research was funded by LPPM Diponegoro University with the 2020 RPP scheme. The authors thank the LPPM Diponegoro Univeristy for its funding and support.

References

[1] C. Wang., P.W. Horby., F.G. Hayden., G.F. GaoA. Novel coronavirus outbreak of global health concern. Lancet. Vol. **395**. Issue 10223. Pp. 470-473. doi.org/10.1016/S0140-6736(20)30185-9.

[2] Q. Chen, M. Liang, Y. Li, J. Guo, D. Fei, L. Wang, L. He, C. Sheng, Y. Cai, X. Li, J. Wang, Z. Zhang. Mental health care for medical staff in China during the COVID-19 outbreak. Lancet Psychiatr. Vol. **7**. Issue. 4. Pp. PE15-E16. doi.10.1016/S2215-0366(20)30078-x.

[3] Yudong Shiab, Juan Wang, Yating Yanga, Zhiqiang Wang, Guoqing Wang, Kenji Hashimoto, Kai Zhang, Huanzhong Liuab. 2020. Knowledge and attitudes of medical staff in Chinese psychiatric hospitals regarding COVID-19. Brain, Behavior & Immunity-Health. Vol **4**. doi.org/10.1016/j.bbih.2020.100064.

[4] Kementerian Kesehatan Republik Indonesia. 2020. Pedoman Pencegahan dan Pengendalian Corona Virus Diaseases (COVID -19).

[5] Ramadhan Tosepua, Joko Gunawan, Devi Savitri Effendy, La Ode Ali Imran Ahmad, Hariati Lestari, Hartati Bahar, Pitrah Asfiang. Correlation between weather and COVID-19 pandemic in Jakarta, Indonesia. Science of The Total Environment. Available online 4 April 2020, 138436. In Press, Journal Pre-proof. doi.org/10.1016/j.scitotenv.2020.138436.

[6] Zheng F., Zhong S. 2011. Time series forecasting using a hybrid RBF neural network and AR model based on binomial smoothing. World Academy of Science. Eng Technol **75**. Pp. 1471- 1475.

[7] Pasaribu, Y. P., Fitrianti, H., Suryani, D. R. (2018). Rainfall forecast of merauke using autoregressive integrated moving average model. In The 3rd International Conference on Energy. Environmental and Information System, 73. Central Java, Indonesia.

[8] P. Arumugam., R. Saranya. (2018). Outlier Detection and Missing Value in Seasonal ARIMA Model Using Rainfall Data*. Materialstoday:proceedings. Vol. **5**. Issue 1. Part 1. Pp. 1791-1799. doi.org/10.1016/j.matpr.2017.11.277.

[9] Arya P., Paul R.K., Kumar A., Singh K.N., Sivaramne N, Chaudhary P. (2016). Predicting pest population using weather variables: An ARIMAX time series framework. International Journal of Agricultural and Statistical Sciences. Vol. **11**. No. 2. Pp. 381-386.

[10] Xiang C., Zhou Z. (2010). Application of ARIMA and SVM hybrid model in pest forecast. Acta Entomologica Sinica. Vol. **53** No. 9. Pp. 1055-1060.

[11] Al-Sakkaf A., Jones G. (2014). Comparison of time series models for predicting campylobacteriosis risk in New Zealand. Zoonoses Public Health, **61** (3). Pp. 167-174. doi.org/10.1111/zph.12046.

[12] Zhirui Heab., Hongbing Taoa. (2018). Epidemiology and ARIMA model of positive-rate of influenza viruses among children in Wuhan, China: A nine-year retrospective study. International Journal of Infectious Diseases. Vol. **74**. Pp. 61-70.doi.org/10.1016/j.ijid.2018.07.003

[13] Caporale, G. M., & Gil-Alana, L. A. (2010). Long Memory and Fractional Integration in High Frequency Financial Time Series. Economics and Finance Working Paper Series. Working Paper No. 10-10. Department of Economics and Finance.

[14] Oskar Vivero., William P. Heath. (2012). Regularised Estimators for ARFIMA Processes. IFAC Proceedings Volumes. Vol **45**. Issue 16. Pp. 298-303. doi.org/10.3182/20120711-3-BE-2027.00335.

[15] Baillie, R. T. & Morana, C. (2012). Adaptive ARFIMA Models With Apllications to Inflation. Economic Modelling. Vol. **29**. No.6. Pp. 2451-2459.

[16] Aye G. C, Balcilar M., Gupta R., Kilimani N., Nakumuryango A. & Redford S. (2014). Predicting BRICS stock returns using ARFIMA models. Applied Financial Economics. Pp. 1159-1166. doi: 10.1080/09603107.2014.924297.

[17] Kartikasari, P. 2015. Studi Simulasi Pengaruh Outlier Terhadap Pengujian Linieritas Dan Long Memory Beserta Aplikasinya Pada Data Return Saham. Masters thesis, Institut Teknologi Sepuluh Nopember.

[18] Krzysztof Burnecki., Grzegorz Sikora. (2017). Identification and validation of stable ARFIMA processes with application to UMTS data. Chaos, Solitons & Fractals. Vol. **102**. Pp. 456-466. doi.org/10.1016/j.chaos.2017.03.059.

[19] José M., Belbuteab., Alfredo M.Pereirac. (2015). An alternative reference scenario for global CO2 emissions from fuel consumption: An ARFIMA approach. Economics Letters. Vol. **136**. Pp. 108-111. doi.org/10.1016/j.econlet.2015.09.001.

[20] Kartikasari, P. (2020). Prediksi Harga Saham PT. Bank Negara Indonesia Dengan Menggunakan Model Autoregressive Fractional Integrated Moving Average (ARFIMA). Jurnal Statistika Universitas Muhammadiyah Semarang. Vol. **8**. No. 1. Pp. 1-7.

[21] Doornik, J. A., and Ooms, M. (1999). A Package for estimating, forecasting and Simulating ARFIMA Models: Arfima Package 1.0 for Ox. Nuffield College, Rotterdam.

[22] Geweke, J., & Hudak, S. P. (1983). The Estimation and Application of Long Memory Time Series Models. Journal of Time series Analysis 4, 221-237.

[23] Wei, W. W. S. (2006). Time Series Analysis Second Edition: Univariate and Multivariate Methods (2nd eds). New York, United States of America: Pearson Education.