Modeling and Forecasting of COVID-19 Confirmed Cases in Indonesia Using ARIMA and Exponential Smoothing

Hedi[1] M.V. Joyce Merawati BR[2,*]

[1] Energy Conversion Engineering Department, Politeknik Negeri Bandung, Indonesia
[2] English Department, Politeknik Negeri Bandung, Indonesia
[*]Corresponding author. E-mail: joyce.merawati@polban.ac.id

ABSTRACT
The number of confirmed COVID -19 cases in Indonesia is increasing rapidly. Therefore forecasting of the number of confirmed cases in the future needs to be predicted, so that the government can prepare to handle this pandemic case. The purpose of this study is to provide information on the estimated number of COVID -19 cases in the future in Indonesia. The data used are daily time series data, the number of confirmed cases of COVID-19 from March to August 2020. This study applies two mathematical models, namely: ARIMA and exponential smoothing. Based on ARIMA model, the parameter equations of the ARIMA model (0,2,1), (1,2,0), and (1,2,1) are obtained. The calculation results of Akaike's Information Criterion (AIC) and Schwartsz Bayesian Criterion (SBC), ARIMA (1,2,1) is the most suitable model. The exponential smoothing model, the model with the smallest root mean square error (RMSE) is obtained namely the exponential smoothing model involving trends. The results of the RMSE calculation of the two models, the ARIMA (1,2,1) model is the most suitable forecast the number of COVID -19 cases in Indonesia.

***Keywords**: Forecasting, COVID-19, ARIMA*

1. INTRODUCTION

COVID-19 confirmed in Indonesia was continuously increasing from March to August 2020, graphically there were no significant signs to decline (see figure 1). This pandemic case has been recognized as a global threat. Information to predict the number of people who will be infected needs to be conducted as accurately as possible. Statistical predictions can help the government to estimate the number of COVID-19 cases in the future so that they can anticipate prepare personnel, equipment, health services, and many others.

The number of COVID-19 cases recorded in daily time is time series data, making it possible to predict the future to model it with the model time series, namely: autoregressive integrated moving average (ARIMA) [1] and exponential smoothing model [2]. The ARIMA model is a very reliable model for predicting the number of confirmed COVID-19 cases. Several countries that apply ARIMA for forecasting COVID-19 [3], [4], Germany ARIMA(1, 4, 1), France ARIMA(0, 1, 3) and Turkey ARIMA(1, 4, 0) [5]. Furthermore, the exponential smoothing forecasting model, connects the new data with the previous data through exponential decreasing weighting. This process is very helpful if the parameters associated with the time series change gradually with time [2]. Some countries that are suitable for this model are Japan, Italy, and Canada [5]

This research attempts to determine the forecasting model of the number of COVID-19 pandemic cases in Indonesia using the ARIMA (p, d, q) estimation model and exponential smoothing. The purpose of this study is to provide information on the estimated number of covid-19 cases in Indonesia in the future.

1.1 Related Work

The COVID-19 cases analyzed were data that were confirmed in the period between March 2020 to June 2020. Then there were three stages, namely, ARIMA modeling, exponential modeling, and selection of the best model.

ARIMA modeling (p, d, q) was done through the Box-Jenkins method, namely identification, estimation, verification (diagnostic examination) and prediction [1]. The exponential model using data samples were determined in two models, namely the model with no trend assumptions (Brown method) and the model by incorporating trend elements (Holt method). Based on

RMSE calculations, the best model is the model with the smallest value.

1.2 Our Contribution

This paper compares two forecasting models, to predict the number of confirmed COVID-19 cases in Indonesia by applying the ARIMA model and exponential smoothing. The best model of the two models is expected to be used as a forecasting model for COVID-19 in Indonesia.

1.3. Paper Structure

This paper has four parts. The first part of this paper, introduces a model used to predict the number of confirmed cases of COVID-19, which includes ARIMA modeling and exponential smoothing modeling. The second part describes the application of the ARIMA model that was caried out by estimating the parameter d through a stationary of the COVID-19 data, followed by estimating the parameters p and q through the AIC and SBC criteria. The third part, explains the application of the exponential smoothing model

Through the Brown and Holts methods, by applying RMSE to determine the best method. The fourth part compares the ARIMA and exponential models by applying forecast data and observed data from August 8 to September 6.

2. BACKGROUND

2.1 ARIMA and Exponential Smoothing

ARIMA (p,d,q) is carried out using the Box-Jenkin method, namely identification, estimation, diagnostic check and prediction [1]. Parameter p, d, and q were estimated based on sample data. ARIMA (p,d,q) forecasting equation is:

$$\emptyset(B)(1-B)^d(y_t-\mu)=\theta(B)\varepsilon_t \qquad (1)$$

where:

t = 1,2,....T, T : number observed

B = backshift operator

d = differenced parameter

μ = mean

ε_t= residual

Single exponential smoothing modeling is in form of

$$\hat{y}_{t+1}=\alpha y_t+(1-\alpha)\,\hat{y}_t \qquad (2)$$

where

$\hat{y}_i$: predicting the number of cases of the COVID-19 pandemic in period i

y_t : number of cases of the COVID-19 pandemic in periods t, α : is weighted with $0<\alpha<1$

If the number of cases of the COVID-19 pandemic is influenced by the trend, the exponential smoothing model with the trend is called the Holt model as follows:

$$L_t=\alpha y_t+(1+\alpha)(L_{t+1}+T_{t-1}); \qquad (3)$$

$$T_t=\beta(L_t-L_{t-1})+(1-\beta)T_{t-1} \qquad (4)$$

and

$$\hat{y}_{t+p}=L_t+p\,T_t \qquad (5)$$

where

L_t= new smoothing

T_t= period t trend estimate

p = number of periods for future forecast

β = weight

2.2 Application of the ARIMA Model

The number of confirmed cases of COVID-19 every day from March 2, 2020 to August 7, 2020 for 160 days are depicted in figure 1.

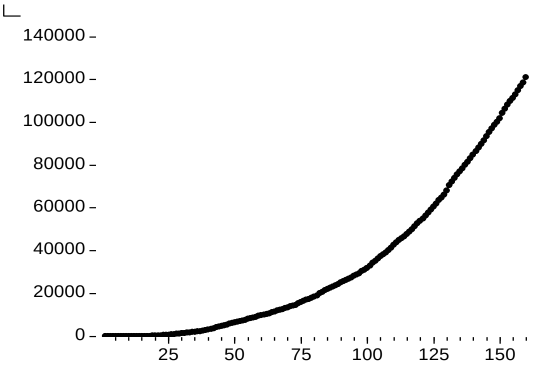


Figure 1 The number of COVID-19 cases from March to August 2020

In a graph of the time series data, the number of confirmed COVID-19 cases in figure 1 is not stationary because the number of confirmed cases is increasing exponentially. Furthermore, difference is made the time series data. While the data pattern resulting from the second difference in figure 2, shows that the trend element is not visible, and the data are stationary to the mean.

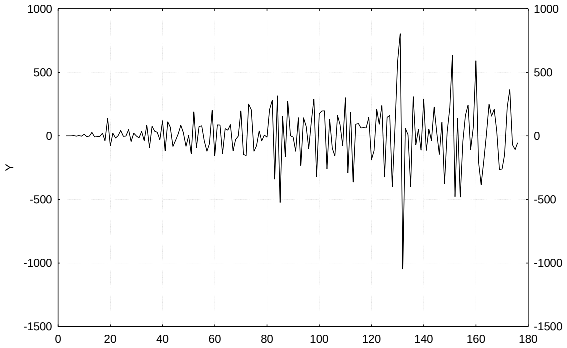


Figure 2 **S**econd difference the number cases of COVID-19

In table 1, stationary proved by applying the Augmented Dickey-Fuller test statistic on the second difference data. Through the null hypothesis (H_0) the data are not stationary, alternative hypothesis(H_1): the data are stationary. The results of the calculation of table 1, P-value = 0.0000 means that H_0 is rejected, which means that the second difference time series data are stationary. Through this stage, the model is ARMA (p,2,q).

Table 1. Stationary Test on Second difference Data with ADF

Null Hypothesis: D(Y,2) has a unit root Exogenous: Constant Lag Length: 5 (Automatic based on SIC, MAXLAG=13)			
		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-9.738703	0.0000
Test critical values:	1% level	-3.473672	
	5% level	-2.880463	
	10% level	-2.576939	

The next stage was the estimation of p and q which were determined based on the significant lag from autocorrelation function (ACF) plot, while the q parameter was determined from the significant lag from partial autocorrelation function (PACF) plot. The ACF plot result was exponential decay after lag 1(see figure 3), while PACF was similar (see figure 4)

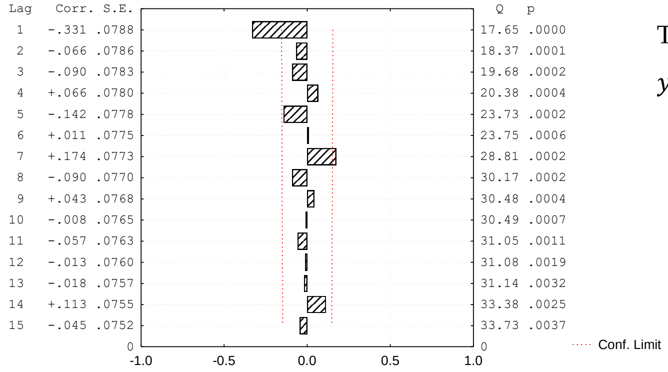


Figure 3 Second difference ACF

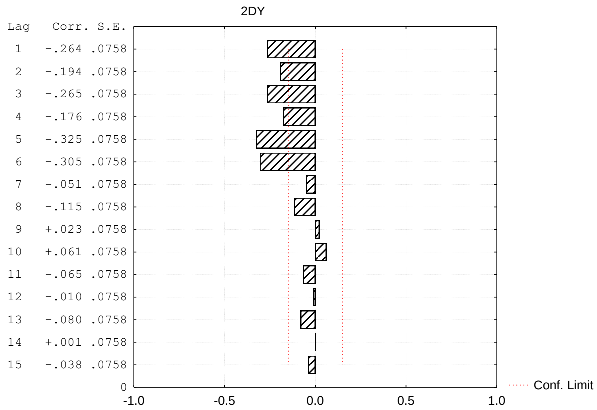


Figure 4 *Second difference* PACF

Therefore, there are three possible models for the number of confirmed COVID-19 cases, namely: ARIMA (1, 2, 0), ARIMA (0, 2, 1), and ARIMA (1, 2, 1). From the results of AIC and SBC calculation the smallest is ARIMA (1, 2, 1) as shown in Table 2.

Table 2. AIC and SBC values in the ARIMA model

MODEL	AIC	SBC
ARIMA(1,2,1)	13.21160	13.27000
ARIMA(0,2,1)	13,23856	13,27732
ARIMA(1,2,0)	13.41455	13.45349

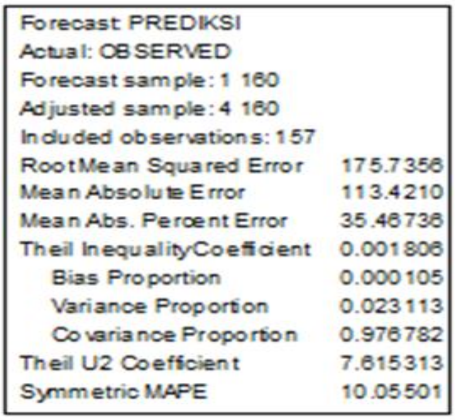
Thus, the model for the number of confirmed cases of COVID-19 that is suitable for Indonesia is ARIMA (1,2,1) and the results of the calculation of the parameter estimates are shown in table 3.

Table 3. Parameter Estimation of ARIMA (1,2,1)

	Param.	Asympt. - t(155)	p
Constant	12.77420	5.37433	0.000000
p(1)	0.24429	2.61679	0.009755
q(1)	0.87733	23.13852	0.000000

The model equation is

$$y_t = 12{,}77 + 0{,}24y_{t-1} + \varepsilon_t + 0{,}88\varepsilon_{t-1} \tag{6}$$



```
Forecast: PREDIKSI
Actual: OBSERVED
Forecast sample: 1 160
Adjusted sample: 4 160
Included observations: 157
Root Mean Squared Error        175.7356
Mean Absolute Error            113.4210
Mean Abs. Percent Error        35.46736
Theil Inequality Coefficient   0.001806
     Bias Proportion           0.000105
     Variance Proportion       0.023113
     Covariance Proportion     0.976782
Theil U2 Coefficient           7.615313
Symmetric MAPE                 10.05501
```

Figure 5 Predicting Error

Table 3 shows that, the parameter significance test uses the t distribution, with the null hypothesis H_0 of each parameter of the ARIMA model (1, 2, 1) which equal to zero, and the alternative hypothesis H_1 of each parameter are not equal to zero. For each parameter tested, the P-value is smaller the 5 %, this means that H_0 is rejected. Figure 5 indicates that the predicting error of the number of confirmed COVID-19 cases is quite small with RMSE = 175.7356

2.3 Application of the Exponential smoothing Model

In this section, exponential modeling is defined in two models, namely a model with no trend assumption (Brown's method) and a model that includes a trend element (Holts' method).

The exponential model without a trend element with the smallest RMSE = 1211,819 with a weight of α = 0.8, Table 4.

$$\hat{y}_{t+1}=0,8y_t+0,2\ \hat{y}_t \quad (7)$$

Table 4. RMSE Brown's method

α	RMSE
0.2	4708.374
0.3	3230.046
0.4	2468.593
0.5	2006.139
0.6	1696.662
0.7	1476.042
0.8	1211.819

The exponential model that involves the best trend element is the calculation result of the smallest RMSE = 198.4693, namely with α = 0.8 and β =0.8 see Table 5, the equation is

$$L_t = 0,8\,y_t + 1,8(L_{t+1} + T_{t-1}) \quad (8)$$

$$T_t = 0,8(L_t - L_{t-1}) + 0,2T_{t-1} \quad (9)$$

and

$$\hat{y}_{t+p}=L_t + p\,T_t \quad (10)$$

Table 5. RMSE Holt's method

α	β	RMSE
0.2	0.2	492.0285
0.3	0.3	317.3475
0.4	0.4	258.2318
0.5	0.5	236.1014
0.6	0.6	221.2391
0.7	0.7	207.6808
0.8	0.8	198.4693

2.4 The ARIMA and exponential models compares

The best comparison between exponential smoothing model and ARIMA (1, 2, 1) is ARIMA (1, 2, 1) with root mean square prediction error (RMSPE) = 2505.13. The results of the calculation of the RMSPE of the two smallest models are ARIMA (1, 2, 1) see Table 6.

Table 6. RMSPE ARIMA (1,2,1) and Exponential

MODEL	RMSPE
ARIMA(1, 2, 1)	2505.13
Exponential	2734.17

Figure 6, depicts the forecasted number of confirmed cases in the next thirty days with the ARIMA model (1,2,1) and exponential model from August 8 to September 6, 2020.

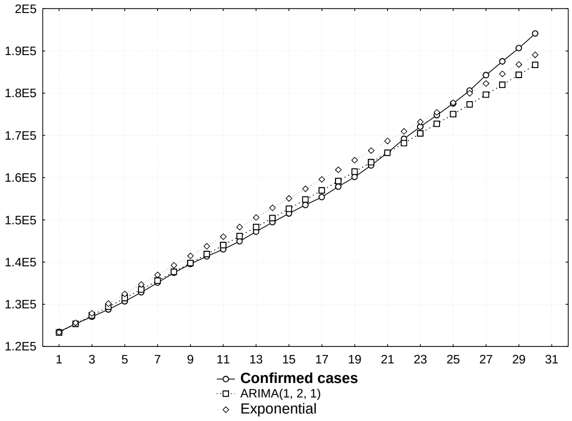


Figure 6 Forecasting ARIMA (1,2,1) and Exponential Models

Table 7. ARIMA (1,2,1) Forecast

Date	Confirmed cases	Forecast	Lower	Upper
8/8/20	123503	123339	122684	123993
8/9/20	125396	125373	124438	126308
8/10/20	127083	127398	126178	128618
8/11/20	128776	129430	127914	130946
8/12/20	130718	131474	129649	133298
8/13/20	132816	133530	131384	135676
8/14/20	135123	135599	133118	138080
8/15/20	137468	137680	134851	140509
8/16/20	139549	139774	136585	142964
8/17/20	141370	141882	138318	145445
8/18/20	143043	144001	140052	147951
8/19/20	144945	146134	141786	150482
8/20/20	147211	148279	143521	153038
8/21/20	149408	150438	145257	155618
8/22/20	151498	152608	146995	158222
8/23/20	153535	154792	148735	160849
8/24/20	155412	156989	150476	163501
8/25/20	157859	159198	152220	166176
8/26/20	160165	161420	153966	168873
8/27/20	162884	163655	155715	171594
8/28/20	165887	165902	157467	174338
8/29/20	169195	168163	159221	177104
8/30/20	172053	170436	160979	179892
8/31/20	174796	172722	162740	182703
9/1/20	177571	175020	164505	185536
9/2/20	180646	177332	166273	188390
9/3/20	184268	179656	168046	191267
9/4/20	187537	181993	169822	194164
9/5/20	190665	184343	171602	197084
9/6/20	194109	186706	173387	200024

The estimated future COVID-19 is the ARIMA model (1,2,1), and the forecasted number of COVID-19 cases in the next thirty days is provided in Table 7

Figure 7, is the forecasted number for next thirty days. This number graphically continues to increase, meaning that every day there is an increase in the number of cases

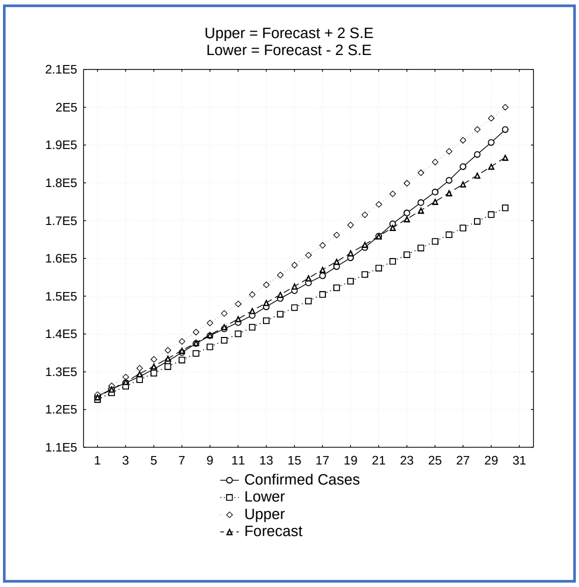


Figure 7 Forecasting ARIMA(1,2,1)

3. DISCUSSION

The ARIMA methodology is designed for short-term forecasting, for long-term forecasting, the daily data must be updated and of course it has implications for changing parameters with new data. Comparison of forecasting from 8 August to 28 August with actual data shows a good approach. While forecasting on September 29 to 6 shows a striking difference. The absolute average forecast bias from August 8 to August 28 was 0.51% while August 29 to September 6 was 18.6%. So for the correct prediction from 29 to 6 September, the data must be updated until August 28.

The use of interval forecasting in the ARIMA mode is much better than point forecasting, actual data lies within the forecasting interval. Interval forecasting of the number of confirmed cases in the future is also useful information. Interval forecasting the number of confirmed cases on August 8 to September 6 shows the actual data in that interval see figure 4.

4. CONCLUSION

Comparing the forecasting number of confirmed COVI-19 cases using the exponential smoothing model through the Brown's method and Holt's method, indicates that the best is the Holt's method.

The next forecasting method are ARIMA (1, 2, 0), ARIMA (0, 2, 1), and ARIMA (1, 2, 1). Comparing these three methods shows that is the best is ARIMA (1, 2, 1).

The best model from the ARIMA (1,2,1) and the exponential smoothing models is the ARMA (1,2,1) model. Thus the model for the number of confirmed COVID-19 cases in Indonesia is ARMA (1,2,1)

REFERENCES

[1] W. W. S. Wei, Time Series Analysis: Univariate and Multivariate Methods, Boston: Pearson Addison Wesley, 2006.

[2] S. Shastri, V. M. A. S. Amardeep Sharma and M. K. Arun SinghBhadwal, "A Study on Exponential Smoothing Method for Forecasting," International Journal of Computer Sciences and Engineering, vol. 6, no. 4, pp. 482-485, 2018.

[3] H. Yonar, A. Yonar, M. A. Tekindal and M. Tekindal, "Modeling and Forecasting for the number of cases of the COVID-19 pandemic with the Curve Estimation Models, the Box-Jenkins and Exponential Smoothing Methods," Eurasian Journal of Medicine and Oncology, vol. 4, no. 2, pp. 160-165, 2020.

[4] D. Benvenuto, M. Giovanetti, L. Vassallo, S. Angeletti and M. Ciccozzi, "Application of the Arima model on the COVID-2019 epidemic dataset," Data in Brief, vol. 29, pp. 1 - 4, April 2020.

[5] T. Dehesh and P. D. H.A. Mardani-Fard, "Forecasting of COVID-19 Confirmed Cases in Different Countries with Arima Models," MedRxiv, pp. 1 - 12, 2020.