



Application of ARIMA Model for Forecasting Additional Positive Cases of Covid-19 in Jember Regency

Wigid Hariadi^{a,1,*}, Sulantari^{b,2}

^a Mathematics Education of IKIP PGRI Jember, Jember, 68124, Indonesia

^b Mathematics Education of IKIP PGRI Jember, Jember, 68124, Indonesia

¹ wigid.hariadi@gmail.com*; ² sulantari89@gmail.com;

* corresponding author

ARTICLE INFO

Article history

Received February 10, 2021

Revised April 8, 2021

Accepted April 22, 2021

Keywords

ARIMA

Forecasting

Covid

Covid-19

Jember

ABSTRACT

The autoregressive integrated moving average (ARIMA) model is a popular method for forecasting univariate time series dataset. This method consists of four major stages, namely: identification, parameter assessment, diagnostic examination, and forecasting using the ARIMA model (p, d, q). ARIMA model can be applied in various fields, one of which is medical field. Currently, there had been a daily increase in the number of patients infected with Corona virus.. Jember is one of the regencies in East Java with a high number of confirmed patients. On February 5, 2021, it was recorded that 5,872 patients were confirmed positive for Corona, 5,241 patients had been declared cured, and 352 patients were declared dead. Given the high number of confirmed cases of Covid-19 in Jember, the authors would like to conduct a prediction research on the increasing number of confirmed cases of Covid-19 in Jember Regency for the upcoming period using the ARIMA model (p,d,q). The research was conducted in the Jember Regency, East Java. The data were collected from March 28, 2020 to January 30, 2021. The study showed that the ARIMA model (1,2,3) was the best model for predicting the additional positive cases of Covid-19 per week in Jember, with the sum squared resid of 7.9496. The data forecast for the additional positive cases of Covid-19 for the next 6 periods is: 224,56 patients, 247,84 patients, 273,53 patients, 301,89 patients, 333,18 patients, and 367,72 patients.

1. Introduction

The beginning of the year 2020 shocked the whole world with an unanticipated outbreak of a new virus, the coronavirus (SARS-CoV 2). The disease was named COVID-19 because it was first identified in Wuhan, China, at the end of December 2019. By the time of this research, the disease had spread to 65 countries. The illness severity of Covid-19 infection can range from mild, to

moderate, and severe symptoms. The main clinical symptoms of COVID-19 may include fever (temperature > 38°C), cough and shortness of breath. The condition of most patients with severe symptoms of the disease may worsen rapidly and progressively, as shown by the occurrence of ARDS, septic shock, metabolic acidosis, which is difficult to cure, and bleeding or dysfunction of the coagulation system within a few days. A small percentage of clinically depressed patients may even die [10]. According to [1] by March 31, 2020, the virus had infected more than 750.890 people worldwide. Thus, the World Health Organization (WHO) has declared a public health emergency relating to Covid-19. Since then, some citizens of many countries worldwide had contracted this strain of virus. At the global level, Covid-19 has infected at least 213 countries, territories, or areas. The high mortality rate due to the corona virus is caused by two factors: the internal factors of an individual (the existing congenital diseases, and the lack of individual awareness in applying strict health protocols to go against this virus) and the external factors (inadequate hospital facilities, ineffective government regulations, and so on)[4].

By March 28, 2020, the global confirmed cases of Covid-19 infection had reached 571,678 cases, with the death rate reaching 4-5%. In Indonesia, by March 29, 2020, the death rate of this virus had reached 9%, which was included as the highest mortality rate of the virus in the world[2]. Since then, there had been a rapid increase in the number of confirmed cases and escalating date toll. According to [9], on February 5, 2021, the number of infected patients in Indonesia was 1,134,854 patients, with 926,980 patients recovered and 31,202 patients died. This number indicates the alarming danger of the virus for humans. Jember is one of the regencies in East Java with a high number of confirmed patients. On February 5, 2021, it was recorded that 5,872 patients were confirmed positive for Corona, with 5,241 patients had been declared cured, and 352 patients were declared dead [6].

Forecasting is an activity to predict future events and is an applicable tool for effective and efficient planning [5]. Any epidemiological evolution is defined and influenced by different factors, in particular by randomness tendency. The abovementioned statistics tools cannot suffice the need for prediction research since they are difficult to generalize. This is why the Automatic Regressive Integrated Moving Average (ARIMA) model has been successfully applied at a much larger scale in various fields, mainly due to its easy-to-use concept and utility of algorithms. [3]. The autoregressive integrated moving average (ARIMA) model is a popular method for forecasting univariate time series dataset. According to [8], the autoregressive integrated moving average (ARIMA) models have been studied in depth by George Box and Gwilym Jenkins, 1976, and their names are often synonymous with the ARIMA process, which is applied to periodic series of analysis and forecasting. The Box-Jenskin method (method) is a standard procedure that is widely used to obtain an appropriate ARIMA model for time series dataset. This method consists of four major stages: identification, parameter assessment, diagnostic examination, and forecasting using the ARIMA model (p, d, q).

Seeing the high number of positive cases of Covid-19 in Jember Regency, the authors would like to conduct a prediction research on the increasing number of confirmed cases of Covid-19 in Jember Regency for the upcoming periods using the ARIMA model (p, d, q). The objectives of this study are: (1) to determine the estimation of the ARIMA model (p, d, q) for the additional positive cases of Covid-19 per week in Jember Regency, (2) to predict additional positive cases of Covid-19 in Jember Regency for the upcoming periods.

2. The Theoretical Framework

2.1. The ARIMA (p,d,q) Models

The autoregressive integrated moving average (ARIMA) model is a popular method for forecasting univariate time series dataset. According to [Rosadi], the AR (p) model is a p-order univariate autoregressive model for a stationary process, and is written as:

$$X_t = \phi_1 X_{t-1} + \phi_1 X_{t-1} + \phi_1 X_{t-2} + \dots + \phi_p X_{t-p} + \varepsilon_t \quad (1)$$

Model MA (q) is the q-order univariate moving average model for a stationary process, and is written as:

$$X_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q} \quad (2)$$

The Autoregressive integrated moving average (ARIMA (p, d, q)) model is a univariate time series of forecasting model that combines the autoregressive model (AR (p)) and the moving average model (MA (q)) for non-stationary data. The general form of ARIMA (p, d, q) can be written as:

$$\phi_p(B) (1-B)^d X_t = \theta_0 + \theta_q(B) \varepsilon_t \quad (3)$$

with, $\phi_p(B) = (1 - \phi_1 B^1 - \phi_2 B^2 - \dots - \phi_p B^p)$, and $\theta_q(B) = (1 - \theta_1 B^1 - \theta_2 B^2 - \dots - \theta_p B^p)$

with a value of $d > 0$. In order to make the stationary model, differencing is necessary. The concept of differencing is conducted by subtracting the X_t observations from the previous observations, namely X_{t-1} . Mathematically differencing can be written as:

$$\text{Differencing order 1 (d=1)} : \Delta X_t = X_t - X_{t-1} \quad (4)$$

$$\text{Differencing order 2 (d=2)} : \Delta X_t = X_t - X_{t-1} - X_{t-2} \quad (5)$$

Parameter estimation of ARIMA (p, d, q) model can be conducted using several methods, one of which is the maximum likelihood estimation method (MLE). The maximum likelihood estimation method is still considered as one of the efficient methods of estimating the ARIMA (p, d, q) model parameters.

3. Methodology

The research was conducted in the Jember Regency, East Java. The targets of this study were Covid-19 patients who had recently been tested positive. This study used the secondary data, which contains time series data for the additional positive cases of Covid-19 patients per week in the Jember Regency, which were obtained through the Jember Regency Government publication data (Instagram @pemkabjember) on the number of infected patients of Covid-19, the additional positive cases, the number of recovered patients, and the number of deaths.. The data were collected from March 28, 2020 to January 30, 2021. From the time series data, ARIMA model analysis will be carried out.

4. Results and Discussion

There had been a fluctuating trend of Covid-19 patients in Jember Regency, and there are no signs to indicate the end of this changing trend. This data can be seen from the high number of newly confirmed cases of Covid-19 in Jember Regency, of around 200 new patients per week. More details about the data on the increasing rate of confirmed cases of Covid-19 patients in the Jember Regency is presented in Figure.1. below in the form of weekly data.

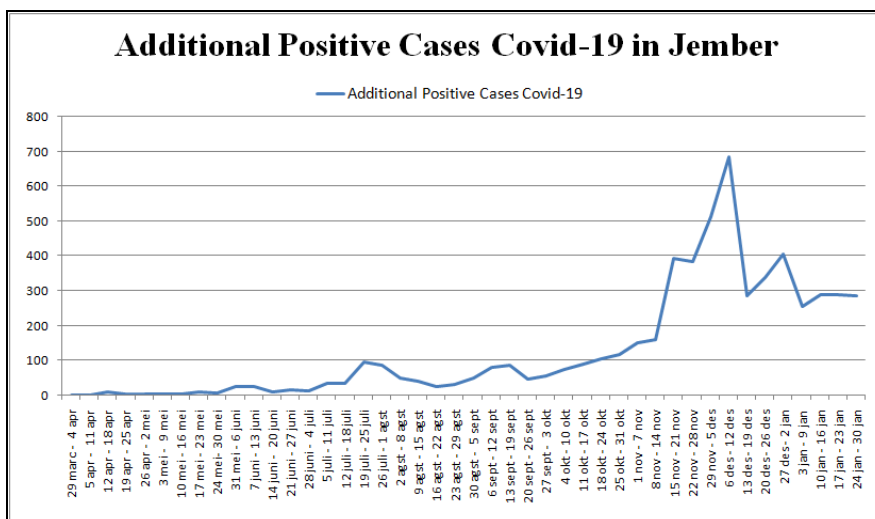


Figure. 1. Plot Data of Additional Positive Case Covid-19 Weekly in Jember

The covid-19 virus started to hit Jember Regency at the end of March 2020. Since then, there had been a gradual increase in the number of Covid-19 patients in this regency. The highest increase occurred in November to early December 2020, which was due to several factors, including the effect of joint holidays during collective leave, the massive demonstrations of the work copyright law, the large number of residents holding a celebration, the low awareness of the public in implementing the health protocols of Covid-19 as the prevention against the disease, and several other unknown factors. Given the high number of additional positive cases of Covid-19 at that time, the Jember Regency Government stipulated several measurements as an effort to reduce the spread of the Covid-19 virus in Jember, among which were the massive public campaigns to comply with 3M's health protocols, sanction for non-compliance, banning Christmas and New Year celebrations, and other measurements. Given this measurement, by the 2nd week of December 2020, there had been a declining rate of confirmed cases of Covid-19.

Before analyzing the ARIMA model (p, d, q), we need to check the stationarity of the data. For this purpose, the Augmented Dickey-Fuller (ADF) test will be conducted. The ADF test results are shown in Table.1. below:

Table. 1. Output of ADF Test from data

	t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic	-1.691493	0.4283
Test critical values:		
1% level	-3.592462	
5% level	-2.931404	
10% level	-2.603944	

*Mackinnon (1996) one-sided p-values.

Table.1 presents the P-value of 0.4283, and thus, we used the following hypothesis test:

$H_0: \phi = 0$ (the data are not stationary in mean or variance)

$H_0: \phi \neq 0$ (the data is stationary both in mean and variance)

By using $\alpha = 5\%$, we can conclude that the Failure to Reject H_0 is due to the P-value of $> \alpha$, so the data is not stationary in mean or variance. Therefore, the next step is data transformation and differencing. The data was transformed into a Natural Log (ln) and we performed a second order differencing, followed by the ADF test. The data graph of the results of transformation, second order differencing, and the ADF test are shown in Figure.2 below:

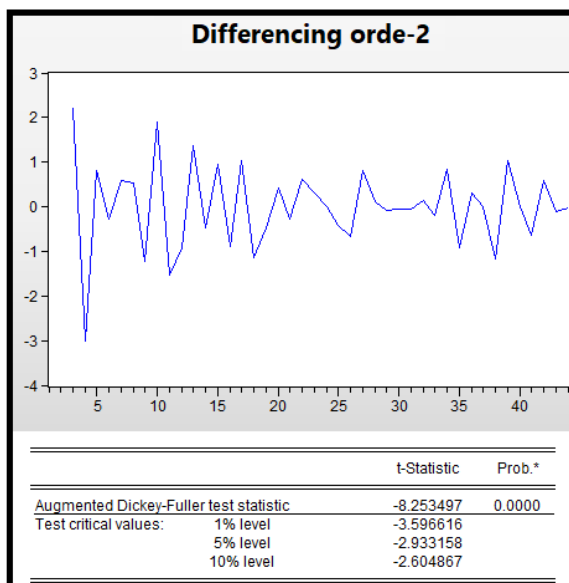


Figure.2. Plot and Output of ADF Test from Data after Transformation and Second Order Differencing

From Table. 2, it can be seen that the P-value is 0.000. Thus, by using $\alpha = 5\%$, it can be concluded that H_0 is rejected because the P-value is $< \alpha = 0,05$, so the data is stationary in both mean and variance. In addition, Figure.2 also indicates that the plot of data is already around the mean. This plot data shows that the data is stationary. Because the data is stationary in mean and variance, we will identify the ARIMA model (p,d,q). Identification of this model using the ACF and PACF plots is presented in Figure.3 below:

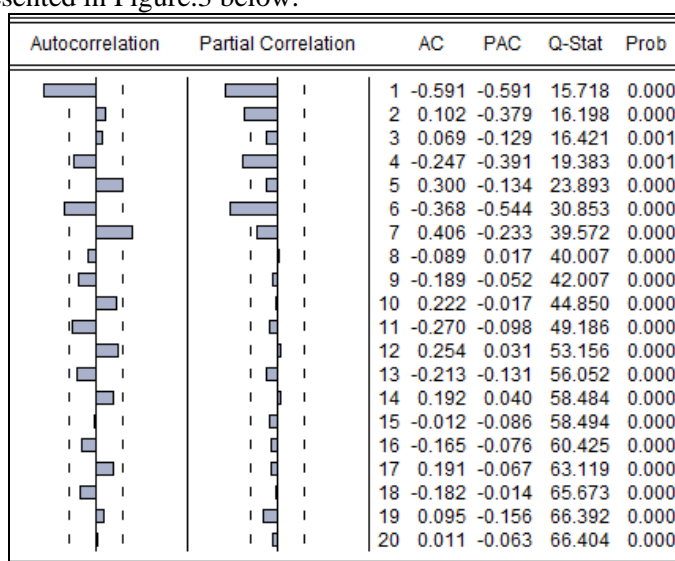


Figure. 3. Plot of ACF and PACF for Identifying ARIMA model

Figure 3 above indicates that there are up to 7 lags that exit the boundary line for the Autocorrelation plot. This data highlights that the possible model is MA (7), while the partial correlation plot shows up to 6 lags outside the boundary line, which insinuates that the possible model is AR (6). This description implies that the ARIMA model (p, d, q) that might occur is the ARIMA model (6,2,7). However, after analyzing and estimating using several models, we obtained the following results:

Table. 2. Summary of the Parameter Estimation from ARIMA models

		ARIMA (1,1,2)	ARIMA (1,2,3)	ARIMA (1,2,7)	ARIMA (1,2,1)	ARIMA (0,2,2)
a 1	p-value	0,0005	0,0002	0	0,0923	
b 1	p-value	0	0,0269	0,008	0	0
b 2	p-value	0	0	0		0,0038
b3	p-value		0,0062	0,0227		
b7	p-value			0		
Sum Squared		9,127	7,9496	8,6776	9,7231	13,6555
Akaike Info		1,4818	1,3926	1,5289	1,5452	1,8096
Schwarz criterion		1,6073	1,5597	1,7379	1,6705	1,8399
White Noise		Yes	Yes	Yes	Yes	No
Autocorrelation residual		Yes	No	No	No	No

From table.2. above, it is apparent that the best model is the ARIMA model (1,2,3). Apart from being a significant model, it has a smaller error value, fulfills the assumptions of white noise, and indicates no residual autocorrelation. For more details, the following Figure 4.4 presents the parameter estimation output of the ARIMA model (1,2,3).

Variable	Coefficient	Std. Error	t-Statistic	Prob.
AR(1)	-0.666911	0.162181	-4.112140	0.0002
MA(1)	-0.421543	0.182872	-2.305120	0.0269
MA(2)	-0.989436	0.129439	-7.644051	0.0000
MA(3)	0.431614	0.148771	2.901202	0.0062
R-squared	0.750929	Mean dependent var		-0.053930
Adjusted R-squared	0.730734	S.D. dependent var		0.893269
S.E. of regression	0.463525	Akaike info criterion		1.392555
Sum squared resid	7.949654	Schwarz criterion		1.559733
Log likelihood	-24.54739	Hannan-Quinn criter.		1.453432
Durbin-Watson stat	1.926433			

Figure.4. Output of the Parameter Estimation of ARIMA(1,2,3)

From Figure .4, we can see that the parameter estimation for the ARIMA (1,2,3) model is significant with the error value used of the sum squared resid of 7.9496. In subsequent, we will display the output of the residual autocorrelation assumption test as shown in Figure.5 below:

Correlogram of Residuals						
Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob	
		1	0.035	0.035	0.0541	
		2	0.087	0.086	0.3944	
		3	-0.156	-0.163	1.5173	
		4	-0.307	-0.313	6.0014	
		5	-0.059	-0.021	6.1724	0.013
		6	0.048	0.102	6.2908	0.043
		7	0.273	0.220	10.156	0.017
		8	-0.024	-0.173	10.187	0.037
		9	-0.080	-0.195	10.541	0.061
		10	-0.125	-0.005	11.425	0.076
		11	-0.245	-0.071	14.941	0.037
		12	-0.043	-0.098	15.054	0.058
		13	-0.015	-0.143	15.068	0.089
		14	0.041	-0.111	15.176	0.126
		15	0.019	-0.032	15.200	0.174
		16	-0.091	-0.131	15.785	0.201
		17	0.139	0.141	17.197	0.190
		18	-0.030	0.036	17.264	0.242
		19	0.177	0.106	19.787	0.180
		20	0.066	-0.000	20.154	0.213

Figure.5. the Output of ACF and PACF Plot for Residual Autocorrelation Test

In Figure.5 above, we can see that from the correlogram of residuals, the ACF and PACF plots no longer indicated any lag. Thus, it can be said that there is no autocorrelation in the residuals of the ARIMA model (1,2,3). Therefore, it can be said that the ARIMA model (1,2,3) is a good model

to be used for forecasting data on the additional positive cases of Covid-19 patients in Jember for weekly data, with the following mathematical model:

ARIMA Model (1,2,3)

$$\phi_1(B) (1-B)^2 X_t = \theta_0 + \theta_3(B) \varepsilon_t$$

$$(1 - \phi_1 B^1) (1-B)^2 X_t = \theta_0 + (1 - \theta_1 B^1 - \theta_2 B^2 - \theta_3 B^3) \varepsilon_t$$

$$(1 - (-0,666911) B^1) (1-B)^2 X_t = 0 + (1 - (-0,421543) B^1 - (-0,989436) B^2 - 0,431614 B^3) \varepsilon_t$$

$$(1 + 0,666911 B^1) (1-B)^2 X_t = (1 + 0,421543 B^1 + 0,989436 B^2 - 0,431614 B^3) \varepsilon_t$$

Based on this model, we will conduct data forecasting. In Figure 4.6, the plot of the data forecasting results using the ARIMA model (1,2,3) is shown below.

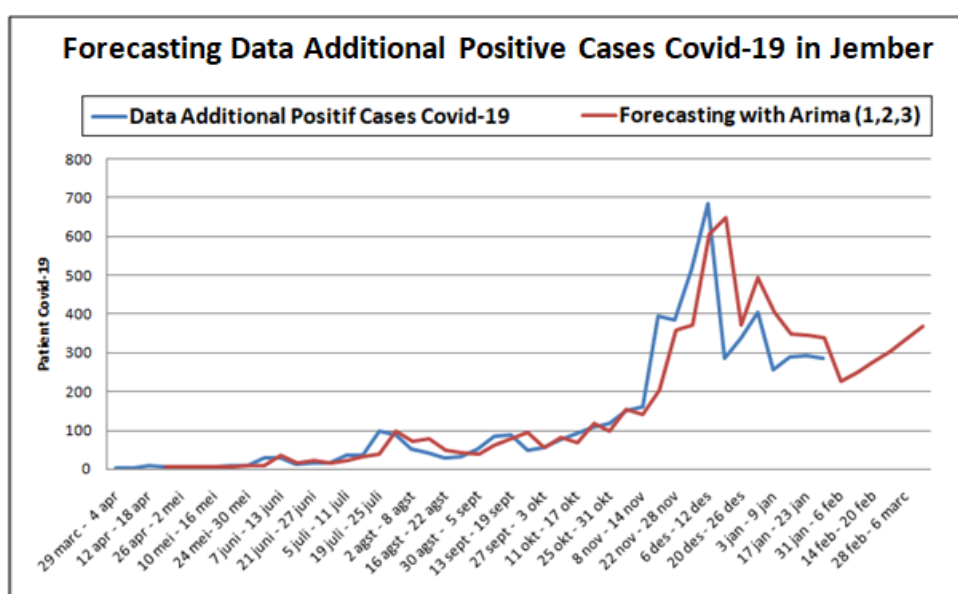


Figure.6. Plot forecasting Data of Additional Positive Cases Covid-19 in Jember

From Figure.6 above, we can see the plot of the results of forecasting data using the ARIMA model (1,2,3). It can be seen that the data forecast results move closer to the actual data, although they are not exactly the same. For more details, the results of the data forecast for the next 6 periods are presented in Table.3 below.

Table.3. Forecasting Data of Additional Positive Cases Covid-19 in Jember for the Next 6 Weeks

Date	Forecasting Data Additional Positive Cases Covid-19
31 jan - 6 feb	224,56
7 feb - 13 feb	247,84
14 feb - 20 feb	273,53
21 feb - 27 feb	301,89
28 feb - 6 marc	333,18
7 marc - 13 marc	367,72

Table.3 shows the results of the forecast data for the next 6 periods or the next 6 weeks with the high rate of additional positive cases of Covid-19. Therefore, it is necessary to handle and prevent the transmission of the Covid-19 virus with stricter health protocols. This attempt is done in order to reduce the number of positive cases of Covid-19 in the future.

5. Conclusion

The conclusions of this research are:

1. A good model for predicting additional data on positive cases of Covid-19 per week in Jember is the ARIMA (1,2,3) model, with the error value of the sum squared resid of 7.9496
2. Mathematically, the model is written as follows:
$$(1 + 0,666911 B^1) (1-B)^2 X_t = (1 + 0,421543 B^1 + 0,989436) B^2 - 0,431614 B^3) \varepsilon_t$$
3. Forecasting data for the additional positive cases of Covid-19 for the next 6 periods, namely:
 - From Jan 31 – Feb 6, 2021, the number of additional positive cases of covid-19 amounted to 224,56 patients.
 - From Feb 7 – Feb 13, 2021, the number of additional positive cases of Covid-19 was 247,84 patients.
 - From Feb 14 – Feb 20, 2021, the number of additional positive cases of Covid-19 was 273,53 patients.
 - From Feb 21 – Feb 27, 2021, the number of additional positive cases of Covid-19 was 301,89 patients.
 - From Feb 28 – March 6, 2021, the number of additional positive cases of Covid-19 was 333,18 patients.
 - From March 7 – March 13, 2021, the number of additional positive cases of Covid-19 was 367,72 patients.

References

- [1] P. Budholiya, A.W. Ali, D. Gunwan, S. Sahil, C.K. Tiyyagi, H. Sharma, "Covid-19: A Global Pandemic of 21st Century", *Journal of Drug Delivery and Therapeutics*. 2020; 10(3-s) 311-321. 2020.
- [2] D. Handayani, D.R. Hadi, F. Isbaniah, E. Burhan, H. Agustin, "*Penyakit virus corona 2019*", *Jurnal respirologi Indonesia*. Vol 40, nomor 2.119 -129.PDPI. Jakarta. 2020.
- [3] O.D. Ilie, R.O. Cojocariu, A. Ciobica, S.I. Timofte, I. Mavroudis, and B. Doroftei, "Forecasting the Spreading of COVID-19 across Nine Countries from Europe, Asia, and the American Continents Using the ARIMA Models", *Journal Microorganisms* 2020, 8, 1158, 1-18. 2020.
- [4] S.M. Ilpaj, N. Nurwati, "*Analisis Pengaruh Tingkat Kematian Akibat Covid-19 Terhadap Kesehatan Mental Masyarakat di Indonesia*", *Focus: Jurnal Pekerjaan Sosial*. Vol 3, no 1. 16-28. Jawa Barat. 2020.
- [5] D.C. Montgomery, C.L. Jennings, and M. Kulahci, "Introduction to Time Series Analysis and Forecasting", John Wiley & Son, Inc. Canada. 2008.
- [6] PemkabJember, "*Data Pantauan dan Sebaran Covid-19 di Kabupaten Jember*", *Instagram @pemkabjember*. 6 Februari 2021. 2021.
- [8] D. Rosadi, "*Analisis Ekonometrika dan Runtun Waktu Terapan dengan R*", *Andi Offset*. Yogyakarta. 2011.
- [9] Satgas Covid-19.2021, "*Peta Sebaran Covid-19*", Retrieved from <https://covid19.go.id/peta-sebaran-covid19> on February 6, 2021.
- [10] Yuliana, "*Corona virus disease (Covid-19); sebuah tinjauan literatur*", *Wellness and Healthy Magazine*, Vol 2, number 1, 187-192. *Fakultas Kedokteran Universitas Lampung*. Lampung. 2020.