talenta PUBLISHER

Journal of Mathematics Technology and Education

Implementation of Autoregresssive Integrated Moving Average (ARIMA) on Forecasting the Number of Positive Cases of Covid-19 in North Sumatera on September 2021

Herlina Fransiska Laia[1], Zahedi[2]

[1]*Student of Bachelor Degree, Department of Mathematics, Universitas Sumatera Utara, Medan, 20155, Indonesia*
[2]*Department of Mathematics, Universitas Sumatera Utara, Medan, 20155, Indonesia*

Abstract. North Sumatera is one of the provinces in Indonesia with a high number of positive cases of Covid-19 so North Sumatera is in the red zone and yellow zone for several months. The increasing number of positive cases can have an impact on various aspects of people's lives. Therefore, mitigation efforts are needed to control the spread of this virus. To be able to perform optimal mitigation, forecasting is needed. Forecasting is the activity of predicting a value in the future by considering current and past data. The forecasting method that will be used in this research is the ARIMA method. The ARIMA model obtained from this study is the ARIMA (1,1,0) model with the MAE, MAPE and MSE values of 453,175203, 3,312754324, and 6161032,938, respectively. Forecasting results for September show an increase every day so the government is expected to make efforts to prevent this.

Keyword: ARIMA, Covid-19, Forecasting

Abstrak. *Sumatera Utara merupakan salah satu provinsi di Indonesia dengan jumlah kasus positif Covid-19 yang cukup tinggi sehingga Sumatera Utara berada di zona merah dan zona kuning selama beberapa bulan. Pertambahan jumlah kasus positif yang terus meningkat dapat berdampak ke berbagai aspek kehidupan masyarakat. Oleh karena itu perlu dilakukan upaya mitigasi untuk mengendalikan penyebaran virus ini. Untuk dapat melakukan mitigasi yang optimal diperlukan peramalan. Peramalan adalah kegiatan memprediksi suatu nilai di masa yang akan datang dengan melakukan pertimbangan terhadap data masa kini maupun data masa lalu. Adapun metode peramalan yang akan digunakan pada penelitian ini adalah metode ARIMA. Model ARIMA yang diperoleh dari penelitian ini adalah model ARIMA(1,1,0) dengan nilai MAE, MAPE dan MSE masing-masing 453,175203, 3,312754324, dan 6161032,938. Hasil peramalan untuk bulan September menunjukkan kenaikan setiap harinya sehingga diharapkan pemerintah melakukan upaya untuk mencegah hal tersebut.*

Kata Kunci: *ARIMA, Covid-19, Peramalan.*

Received Day Month Year | Revised Day Month Year | Accepted Day Month Year

*Corresponding author at: Padang Bulan, Faculty of Mathematics and Natural Sciences, Department of Mathematics, Universitas Sumatera Utara, 20155, Medan, Indonesia

E-mail address: herlinalaia0@gmail.com,

Journal Homepage: ———————-

1. Introduction

Currently, the global is being hit by the pandemic Covid-19 or Coronavirus Disease 2019 or what is often called Covid-19 is caused by the SARS-CoV-two virus and was first discovered in 2019 in Wuhan, China. Common symptoms experienced by people infected with Covid-19 are fever, cough, fatigue, diarrhea, loss of sense of taste and smell, and experience shortness of breath. Since the first confirmed case was announced in Indonesia on March 2, 2020, in Indonesia there have been 4,043,736 total positive cases of Covid-19 until August 26, 2021 [1]. North Sumatra is a province that contributed 92,712 positive cases. The high spread of Covid-19 in North Sumatera has put North Sumatera in the red zone and yellow zone for several months increasing the number of cases that need attention. The high number of positive cases can have an impact on various aspects of community life so mitigation efforts need to be carried out to control of spread of this virus. One of the mitigation steps that need to be taken to control spread of Covid-19 is to forecast, both forecasting the number of positive cases, the number of recovered patients, and the number of deaths caused by Covid-19.

Forecasting the number of positive COVID-19 cases in various regions with different methods has been carried out, such as forecasting the number of active Covid-19 cases in West Java using the Multilayer Perceptron Feed Forward Neural Networks (Pangestu and Hidayat, 2021), forecasting positive cases of Covid-19 in Indonesia by using artificial neural networks [2]. In this study, to predict the number of positive cases of Covid-19, the time series forecasting method that can be used is ARIMA (Autoregressive Integrated Moving Average).

The ARIMA method is a model developed by George Box and Gwilyn Jenkins in 1976 so that ARIMA is often referred to as Box-Jenkins ARIMA or Box-Jenkins time series method. The ARIMA method is widely used in forecasting in various aspects. This because ARIMA is a forecasting technique that can be effective without requiring certain data patterns to emerge. ARIMA can therefore be applied to various forms of data. [3].

2. Related Work

2.1. Definition of Forecasting

Forecasting comes from the word forecast. Forecasting has the meaning of conjecture or prediction about how events will occur in the future. So Forecasting is a prediction of the value of a variable based on a known variable value or as a process of estimating how something will happen in the future [4].

2.2. Time Series Analysis Time

Time series is a collection of observational data that occurs based on a period of time intervals sequentially, carefully recorded according to the order in which they occur, and compiled as statistical data to see whether the observed events or phenomena occur regularly according to certain patterns [5].

2.3. Time Series Analysis Models

2.3.1. Autoregressive (AR) or ARIMA (p,0,0)

Autoregressive is a form of regression that does not relate the independent variable to the dependent variable. This model is useful for measuring the level of closeness (association) between , assuming that the impact of time lags 1,2,3,..., k-1 is separate. In general, the form of the autoregressive model is:

$$X_t = \mu' + \phi_1 X_1 + \phi_2 X_2 + ... + \phi_p X_{t-p} + e_t \tag{1}$$

Where: X_t = independent variable; μ' =constant; $X_{t-1}, X_{t-2}, ..., X_{t-p}$ = predictable variable; ϕ_p= AR parameter of order; e_t =error at time t.

2.3.2. Model Moving Average (MA) or ARIMA (0,0,q)

A relationship known as the moving average depicts the present value as the amount of white noise based on past time or past values. The general form of ARIMA (0,0,q) is:

$$X_t = \mu' + e_t - \theta_1 e_{t-1} - ... - \theta_q e_{t-q} \tag{2}$$

2.3.3. Model Autoregressive Moving Average(ARMA) or ARMA(p,q)

The Autoregressive Moving Average (ARMA) model combines the Moving Average (MA) and Autoregressive (AR) models. The ARMA model's general equation is shown below.

$$X_t = \mu' + \phi_1 X_{t-1} + ... + \phi_2 X_{t-p} + e_1 - \theta_1 e_{t-1} - ... - \theta_q e_{t-q} \tag{3}$$

2.3.4. Model Autoregressive Integrated Moving Average(ARIMA)

According to Wei (1990) in [6], the ARIMA model is a model that ignores independent variables to create a model that assumes the initial data is stationary. The number of independent variables and the residual value of the previous period determines the order of the ARIMA model. In general, the ARIMA model is written as follows.

$$X_t = \phi_1 \mu + e_1 \theta_1 e_{t-1} - \theta_2 e_{t-2} - ... - \theta_q e_{t-q} \tag{4}$$

2.4. Model Identification

2.4.1. Stationerity

A data is considered to be stationary, following Makridakis et al. (1999) [4], if there is no increase or decrease in the data, which indicates that data fluctuations are centered on a constant average value, regardless of the time and variance of fluctuations. The analysis of autocorrelation and partial autocorrelation can be used to verify stationarity. consists of two stationary elements, one each in the mean and variance. Differentiation and transformation will be performed if the stationarity conditions for the mean and variance are not met.

2.4.2. Differentiation (Differencing)

According to Box-Jenkins time series data that is not stationary can be transformed into a stationary data series by doing a process of differentiation (differentiating) on the actual data, namely:

$$X_t' = X_t - X_{t-1} \tag{5}$$

2.4.3. Transformation

For example $T(X_t)$ is transformation function of X_t and X_{t+k} to stabilize the variance is carried out using a power transformation, namely:

$$T(X_t) = \frac{X_t^{\lambda} - 1}{\lambda} \tag{6}$$

where λ is the transformation.

2.4.4. Autocorrelation Function (ACF)

The coefficient of autocorrelation is a function that displays the level of the linear correlation between two variables X_t with $X_{t-1}, X_{t-2}, X_{t-3}, ..., X_{t-k}$. Mathematically the following formula can be used to calculate the correlation coefficient.

$$\bar{X} = \frac{\sum_{t=1}^{n} X_t}{n} \tag{7}$$

with:

$$r_k = \frac{\sum_{t=1}^{n-k}(X - \bar{X})(X_{t+k} - \bar{X})}{\sum_{t=1}^{n}(X - \bar{X})^2} \tag{8}$$

2.4.5. Partial Autocorrelation Function (PACF)

Partial autocorrelation function is the set of partial autocorrelation for various lags. The purpose of the autocorrelation partial is used to calculate the degree of closeness between and if the influence of the time side is considered separate.

$$\phi_{kk} = \frac{r_k - \sum_{j=1}^{k-1} \phi_{k-1}\rho_{k-j}}{1 - \sum_{j=1}^{k-1} \phi_{k-1}\rho_j} \tag{9}$$

2.5. Parameter Estimation

Test whether the parameters in the model are equal to zero or not, parameter testing is necessary. The stages of the significant test can be carried out as follows. $H_0 : \theta = 0$ (there is at least one insignificant model parameter) $H_0 : \theta \neq 0$ (significant model parameter). With test statistics:

$$t = \frac{\hat{\theta}}{se(\hat{\theta})} \tag{10}$$

Decision : reject H_0 if $|t| > t_{\frac{\alpha}{2};df=n-np}$,np= sum of parameters

2.6. Diagnostic Check

2.6.1. Residual White Noise

A process X_t is called white noise , which is an independent and identical process if the successive variables are uncorrelated and follow a certain distribution. The average $E(X_t) = \mu_a$ of this process is assumed to be zero and has variance constant and the covariance $X_t = \sigma_a^2$ and value of the covariance for this process is $\gamma k = cov(X_t, X_{t-k}) = 0$ for $k \neq 0$.

2.6.2. Normal Distribution Assumption

This assumption test aims to see whether the data has met the normality assumption or not by looking at the histogram, which has a tendency to form abell shape. To test the normality of the residuals, the Kolmogorov-Smirnov test is used, namely if the $p < 0,05$ data values do not come from a normally distributed population and if the $p \geq 0,05$ data values come from a normally distributed population.

2.6.3. Selection of the Best Model

Model The best model chosen is the model that has the smallest error or error value. Forecasting results will be more precise if the resulting error rate is getting smaller. The measuring instrument used to calculate the prediction error is as follows.

1. Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i| \tag{11}$$

2. Mean Absolute Percentage Error (MAPE)

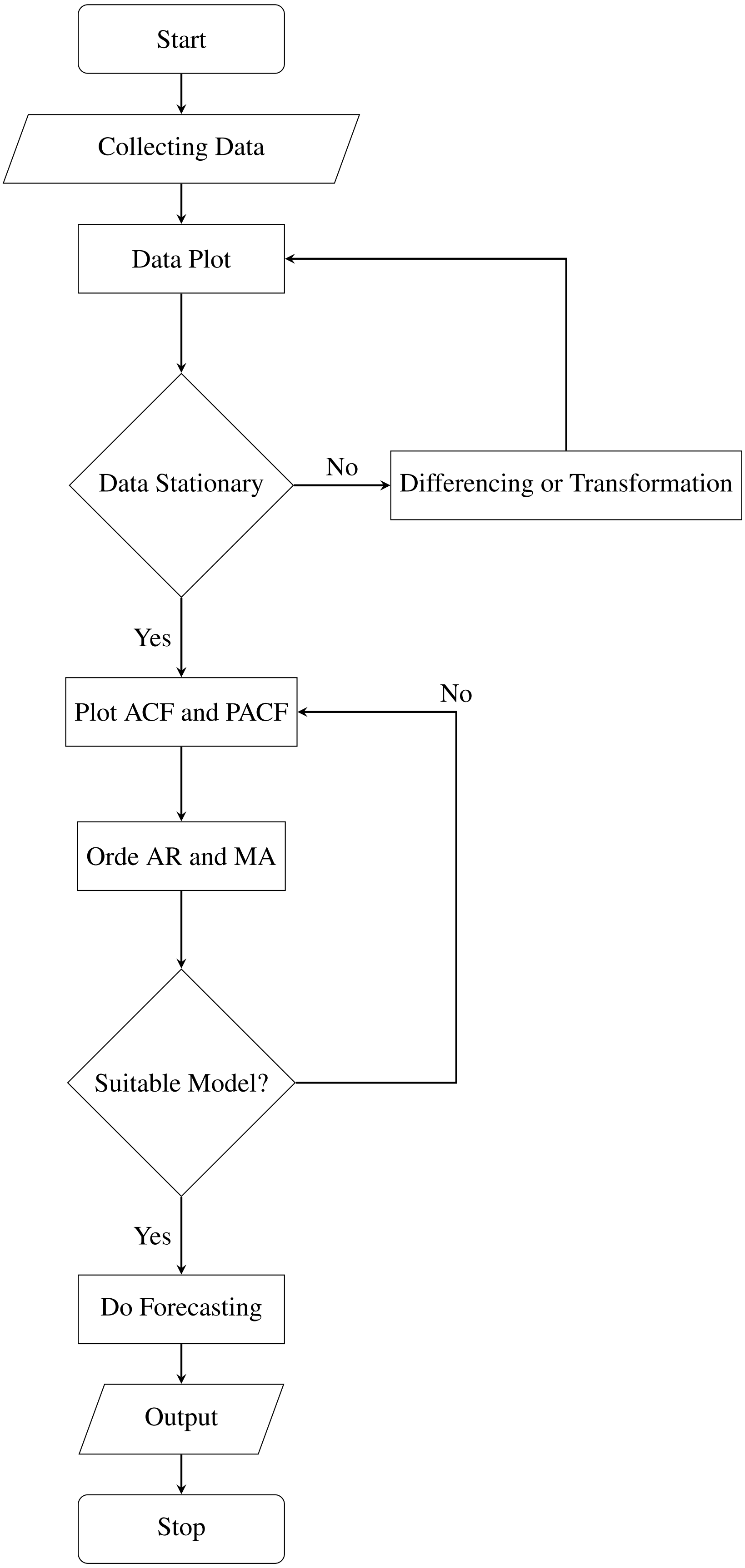
$$MAPE = \frac{1}{n}\sum_{t=1}^{n}(\frac{\hat{X}_t - X_t}{X_t})x100\% \tag{12}$$

3. Mean Square Error

$$MAE = \frac{1}{n}\sum_{t=1}^{n}(X_t - \hat{X}_t) \tag{13}$$

3. Methodology

This study relied on secondary data, which was gathered from other sources or was already available. This research uses data on the number of daily positive Covid-19 confirmation cases in North Sumatra starting in May 2021 to August 2021 obtained from the official website of the covid19.go.id Republic of Indonesia. After the data is obtained, the stage carried out is forecasting with the ARIMA method using the help of Minitab software. Stages forecasting with the ARIMA method is to create a data plot to see the data alignment. Once the data is stationary then it can be determined by some tentative model that is the model to be selected based on the established criteria. From these tentative models will The right model has a good level of accuracy. The next step is to estimate the parameters of the model itself by conducting hypothesis tests to find out whether the parameters are significant or not. The final result obtained will be used for the final model used in forecasting. To work on the stages in research, it is necessary to develop systematic research methods. The research methods in this study are as follows:


Start
Collecting Data
Data Plot
Data Stationary
No
Differencing or Transformation
Yes
Plot ACF and PACF
No
Orde AR and MA
Suitable Model?
Yes
Do Forecasting
Output
Stop

4. Results and Discussion

4.1. Data Collection

The data used in study is the data on the number of positive confirmed cases of Covid-19 in North Sumatra from May to August 2021 obtained from the official website covid19.go.id with a total of 123 data.

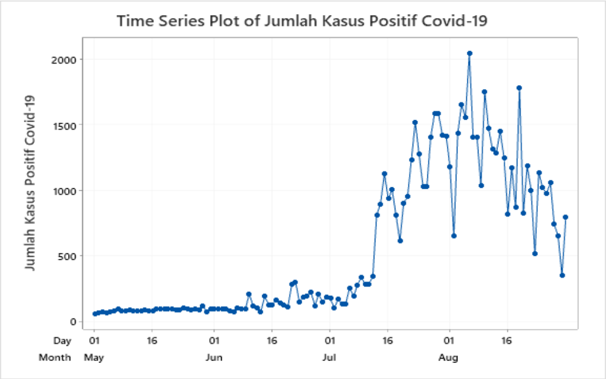


Figure 1. Time Series Plot Number of Positive Cases of Covid-19 North Sumatra

4.2. Stationary Examination

4.2.1. Stationary in Variance

To test the stationarity of the data on the variance, it can be done by transforming the data. To help the transformation to be used, you can see the rounded value λ on the Box – Cox transformation. A data will be said to be stationary if it has a value of equal to one.

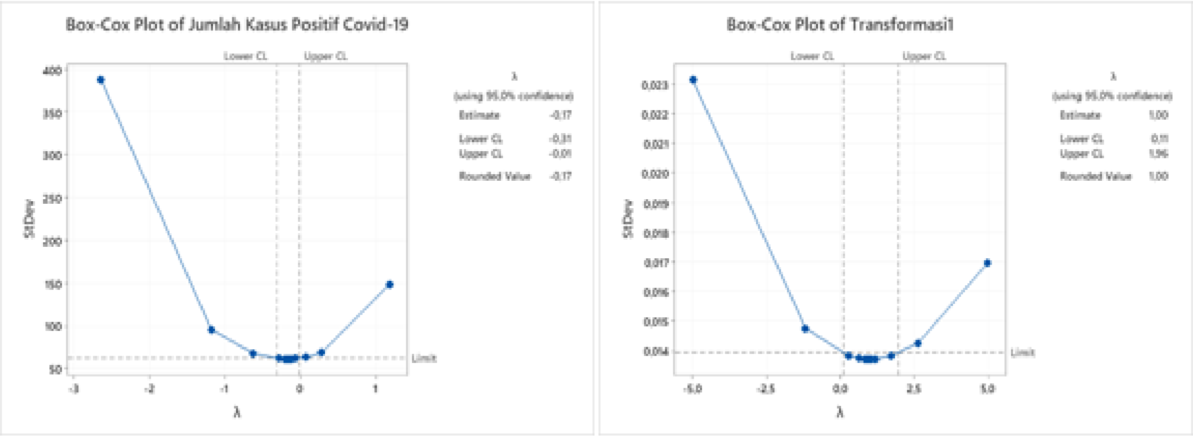


Figure 2. Box-Cox Plot Number of Positive Cases of Covid-19 North Sumatra and Box-Cox plot of the number of positive cases of Covid-19 North Sumatra From the First Transformation

Figure 2 shows that the Box-Cox transformation obtained a rounded value of $-0,17$ so the data cannot be said to be stationary in variance, so that must be done transformation. In figure Box-Cox plot of Transformasi1 it is obtained that the result of the first transformation is a rounded value of 1 so the data on the number of positive cases of Covid-19 is stationary in variance.

4.2.2. Stationary in the Mean

Time series plots or ACF plots can be used to observe the stationarity in the mean. It may be claimed that data is stationary in the mean if there is no trend component in the time series plot and the data on the ACF plot rapidly declines to close to zero.

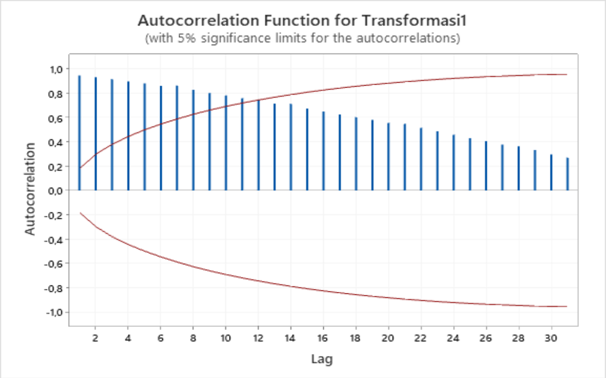


Figure 3. Plot ACF of data on the number of positive cases of Covid-19 North Sumatra From First Transformation

Figure 3 shows that 11 consecutive time lags are outside the significance limit, so the data cannot be said to be significant to the average, it is necessary to do this differencing. Based on the results differencing obtained, a data plot will be generated as shown in Figure 4 below.

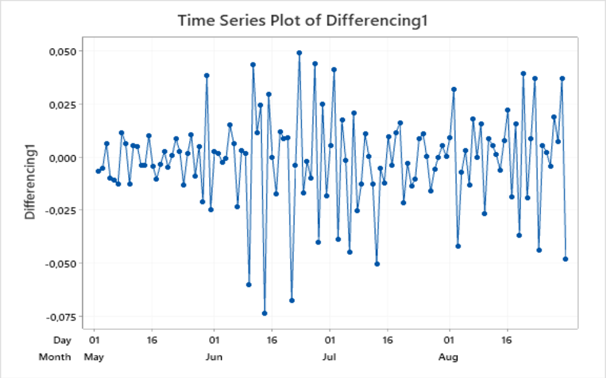


Figure 4. Plot First Differencing of Time Series the number of positive cases of Covid-19 in North Sumatra

Based on Figure 4, it can be seen that the pattern formed indicates that the data is stationary, i.e. the data moves around the average value of the data. So this stationary data can be directly used for the formation of the ARIMA model. The differencing process that has been carried out indicates that the value that can be used is the value of $d = 1$.

4.3. Analysis of ACF and PACF

According to the results of the first differentiation in Figure 4, stationary has been achieved. The following step is to estimate the values of ACF and PACF using Figure 5 below.

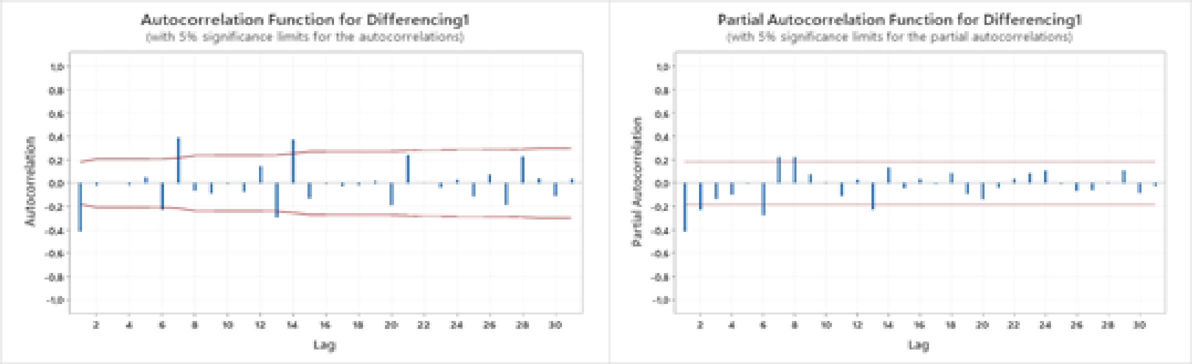


Figure 5. Plot ACF and PACF of data on the number of positive cases of Covid-19 in North Sumatra

From Figure 5, it is found that the ACF lag 1 plot crosses the red line, which is the lower significant limit. So it can be concluded that the differencing is significant at lag 1 so that the moving average (MA) pattern forms the MA(1) model. Plot PACF shows lag 1 and lag 2 crossing the red line, which is the lower significant limit. PACF plot differencing is significant at lag 1 and lag 2. From this, it is obtained that the autoregressive (AR) pattern forms the AR(1) and AR(2) models. Thus, the possible ARIMA models are ARIMA (0,1,1), ARIMA (1,1,0), ARIMA (1,1,1), ARIMA (2,1,1) and ARIMA (2,1,1) models. 1.0)

4.4. Estimation of Model Parameters

The estimation of model parameters is used to estimate the coefficients of the resulting model, which will then determine the significance of the parameters. The model to be chosen is a model that has a significant coefficient.

Table 1. Parameter Estimation of ARIMA Model

No.	**Model**	**Test Result**			**Decision**	
		Parameter	T.Value	P.Value	T.Value	P.Value
1	ARIMA(0,1,1)	MA(1)	7,87	0,000	Significant	Significant
2	ARIMA(2,1,0)	AR(1)	-5,13	0,000	Significant	Significant
		AR(2)	-0,84	0,401	Signifikan	Signifikan
3	ARIMA(1,1,1)	AR(1)	-0,70	0,484	Not Significant	Not Significant
		MA(1)	5,30	0,000	Significant	Significant
4	ARIMA(2,1,1)	AR(1)	-0,70	0,485	Not Significant	Not Significant
		AR(2)	-0,73	0,465	Not Significant	Not Significant
		MA(1)	3,55	0,001	Significant	Significant
5	ARIMA(1,1,0)	AR(1)	-0,28	0,000	Significant	Significant

From Table 1 above, it is found that the significant models are ARIMA(0,1,1) and ARIMA(1,1,0). These models have a significant test value $|t| > t_{\frac{\alpha}{2};df=n-np}$ thus it will be examined for residual values to select the best ARIMA model for projecting the number of positive Covid-19 North Sumatra cases in September 2021.

4.5. Diagnostic Testing

In the diagnostic test, a white noise assumption test and a normal distribution assumption test are performed. This assumption test is done to see if there are any residues from the residue. if there is a remnant then the model cannot be used. In this study, the ARIMA(0,1,1) and ARIMA(1,1,0) models met each assumption test so that to choose the best model to be used for forecasting was done by checking the error value of each model.

4.6. Selection the Best ARIMA Model

After the diagnostic test, the ideal model will be chosen based on the lowest MAE, MAPE, and MSE values. The greater the standard error of a statistic, the more unstable or less significant the statistic is [7] The results of the calculation of these values can be seen in the following table.

Table 2. Calculation Results of MAE, MAPE, and MSE Values on ARIMA

Model	**MAE**	**MAPE %**	**MSE**
ARIMA(0,1,1)	495,035	3,543	7351807,135
ARIMA(1,1,0)	453,175	3,312	6161032,938

In Table 2 the smallest MAE, MAPE, and MSE values are 453,175, 3,312, and 6161032,938, respectively, in the ARIMA (1,1,0) model. Therefore, the ARIMA (1,1,0) model will be used to forecast the number of Covid-19 positive cases in North Sumatera. The following is a model equation for the number of positive cases of Covid-19 in North Sumatera in the next period using the ARIMA (1,1,0) model.

$$X_t = 7,1 + (1 - 0,44)X_{t-1} - (-0,44X_{t-2}) + e_t \tag{14}$$

4.7. Forecasting Using Best Model

Forecasting will begin once the best model has been identified. Forecasting is done to find the values of forecasting the number of positive cases of Covid-19 in North Sumatera in September 2021, and the results are as follows.

Table 3. Results Forecasting the Number of Positive Cases of Covid-19 North Sumatera with ARIMA(1,1,0) in September 2021

Date	Result	Rounding	Date	Result	Rounding
01/09/2021	604,579	606	16/09/2021	737,059	737
02/09/2021	694,128	694	17/09/2021	741,971	742
03/09/2021	661,810	662	18/09/2021	764,885	747
04/09/2021	683,105	683	19/09/2021	751,789	752
05/09/2021	680,810	681	20/09/2021	756,711	757
06/09/2021	688,895	689	21/09/2021	761,626	762
07/09/2021	692,413	692	22/09/2021	771,451	767
08/09/2021	697,940	698	23/09/2021	771,451	771
09/09/2021	702,583	703	24/09/2021	776,364	776
10/09/2021	707,615	708	25/09/2021	781,127	781
11/09/2021	712,476	712	26/09/2021	786,191	786
12/09/2021	717,412	717	27/09/2021	791,104	791
13/09/2021	722,315	722	29/09/2021	796,017	796
14/09/2021	727,233	727	29/09/2021	800,931	801
15/09/2021	732,144	732	30/09/2021	805,844	806

Based on Table 3 above, it may be concluded that fewer Covid-19 positive cases will be reported in North Sumatera in September 2021 than in August 2021. However, in September itself it increases every day. In this case, it is possible that the number of positive cases of Covid-19 is not influenced by time but is also influenced by the presence of other external factors.

5. Conclusions

5.1. Conclusions

Forecasting using the time series shows that the number of positive cases of Covid-19 in North Sumatera in September 2021 has decreased compared to the previous month. The MAPE value of the ARIMA (1,1,0) model is 3.543649533 which means that the accuracy level of the ARIMA (1,1,0) the method is 96.456351%, which indicates that the ARIMA method is feasible to use for predicting the number of positive cases of Covid-19 in North Sumatera

REFERENCES

[1] W. Covid, “Coronavirus pandemic,” 19.

[2] W. W. Setialaksana, D. R. A. Sulaiman, S. S. Dewi, C. A. Lamasitudju, N. R. Ashadi, and M. M. Asriadi, “Model jaringan syaraf tiruan dalam peramalan kasus positif covid-19 di indonesia,” *Jurnal MediaTIK*, vol. 3, no. 2, pp. 53–56, 2020.

[3] H. Sugiarto, “Peramalan bisnis,” *PT. Gramedia Pustaka, Jakarta*, 2000.

[4] S. Makridakis, S. C. Wheelwright, V. E. McGee, U. Andriyanto, and A. Basith, “Metode dan aplikasi peramalan jilid 1 edisi kedua,” *Binarupa Aksara*, 1999.

[5] P. J. Brockwell and R. A. Davis, *Introduction to time series and forecasting*. Springer, 2001.

[6] W. W. Wei, “Time series analysis,” in *Univariate and Multivariate Methods*, 2006.

[7] L. Aritonang, *Peramalan Bisnis*. Ghalia Indonesia, 2009.