OPEN ACCESS
ISSN 2460-9056
http://socj.telkomuniversity.ac.id/ijoict/

Intl. Journal on ICT
Vol. 7, No. 1, Jun 2021. pp. 21-30
doi: 10.34818/ijoict.v7i1.554

Forecasting the Covid-19 Increment Rate in DKI Jakarta Using Non-Robust STL Decomposition and SARIMA Model

Rosmelina D. Satrisna [1*], Aniq A. Rohmawati [1], Siti Sa'adah [1]

[1]*School of Computing, Telkom University*
Jl. Telekomunikasi 1 Terusan Buah Batu, Bandung, Indonesia

[*] rosmelinads@student.telkomuniversity.ac.id,
aniqatiqi@telkomuniversity.ac.id,
sitisaadah@telkomuniversity.ac.id

Abstract

The Coronavirus, well known as Covid-19, was confirmed first in China. At this time, the world still works hard to tackle and control this pandemic concerning the increased levels of spread and severity. The daily data of new cases Covid-19 reported fluctuations during the early months of this outbreak, continued by the several trends and seasonal patterns also captured lately. As the massive severity of the virus, identifying the future increment rates become a major concern to support information and maintain essential health services. In this paper, we propose the Seasonal Trend Loess (STL) Decomposition to observe the trend and noise over the daily cases Covid-19 dataset, and forecast involving Seasonal Autoregressive Integrated (SARIMA) model. The observation involves the increment rate of Covid-19 cases of DKI Jakarta, as the largest capital and metropolitan district of Indonesia. The DKI Jakarta government has opted of imposing large-scale restrictions to curb a recent Covid-19 case. The numerical simulation demonstrates that STL-SARIMA model can be addressed to anticipate potentially overwhelming number of daily Covid-19 cases, with errors forecasting equal to 0.15.

Keywords: Autoregressive, Covid-19, Forecasting, Increment rate, Seasonal

I. INTRODUCTION

Covid-19 reported in China in December 2019, less than one month, the disease has spread in various other provinces in China, and spreading fast to other countries: Thailand, Japan, and South Korea. On March 2, 2019, the first Covid-19 confirmed in Indonesia uncovered two Covid-19 cases in Depok City. The Covid-19 easily spreads throughout Indonesia, especially DKI Jakarta. These days, DKI Jakarta has become the number one province with the massive Covid-19 cases in Indonesia [1]. The significant increase of Covid-19 in a short period, affects to stability system especially the economic sector in DKI Jakarta which has a strongly notion to Indonesia's economy [2]. The government executes policies to overcome the Covid-19 increment rate with enactment of large-scale social restrictions as stated in the regulation under the Indonesia ministry of national health [2].

Forecasting analysis for time series data has future importance on a wide variety of issues in many fields, as future events are an important input into many types of planning and decision-making processes. The most

Received on 15 May 2021, Revised at 14 Jun 2021, Accepted at 29 Jun 2021, Publish online on 30 Jun 2021

challenging tasks in time series forecasting analysis while the selection of the statistical models [3]. Several approaches are using various models of Covid-19 time series forecasting. In research [4], Feroze uses Bayesian Structural Time Series (BSTS) model to forecast the top five countries affected by Covid-19 and show effective results forecasting for the next thirty days. Moreover, Rumetna and Lina [5] explained uses the Moving Average and Exponential Smoothing models to help to forecast the Positive Covid-19 in Sorong City. The result shown that the performance of each model has advantages in different months, the Moving Average model is preferable for data in August and September, otherwise, data in July use the Exponential Smoothing model. Also, Riberio, et al. [6] uses several approaches to machine learning, namely cubist regression, random forest, ridge regression, SVR, stacking-ensemble learning, also ARIMA to forecast the cumulative positive cases of Covid-19 in several Brazilian states. They proposed the best model based on accuracy value were SVR, stacking-ensemble learning and ARIMA. Djakaria and Saleh [7] proposed the Holt-Winters model for their research to forecasting Covid-19 in Gorontalo, by 6 months also getting the smallest MAPE value of 6.14. From their successful research with various models to analyze Covid-19, we expected time series forecasting analysis can help improve this paper.

The essential feature of time series forecasting is inevitably involving the temporal dimension. In this research, the time series model revealing temporal historical dimension is combined by The STL Decomposition approach. The STL Decomposition was developed by Cleveland, et al. can be handled time series observation problems by determining the number of variations in seasonal, trend and noise components that are contained in each observation [8] [9]. In addition, A. Haritsah published a paper related to STL, entitled "Implementation of the Seasonal Trend Decomposition Procedure Based on Loess Model and ARIMA for Predicting Air Quality Concentrations". The result has shown that the combination of STL and ARIMA outperformed in most cases to predict the targeted value [9]. Then, the SARIMA model presents historical seasonal and autoregressive observations which are based on lag time series for dealing with forecasting. Buhl, et al. proposed to use Seasonally Decomposed Autoregressive and Exponential Smoothing Algorithms as a significant approach to predict the Material Footprint in Germany between 2015 and 2020. They aimed to reveal performance between two approaches: STL-ARIMA and STL-ETS. The result attracts our interest that combination model of STL-ARIMA reported good accuracy [10]. In this paper, we purposed to forecast the daily increment rate of Covid-19 involving STL Decomposition and SARIMA model. Our research is expected to develop the result of combination models [10], which explain in state-of-the-art. The observation involves the increment rate of Covid-19 cases of DKI Jakarta because DKI Jakarta the largest capital and metropolitan district of Indonesia, to anticipate potentially overwhelming Covid-19 cases.

II. LITERATURE REVIEW

A. Covid-19

Today, the disease Covid-19 that well-known for certain, which is a new type of Coronavirus that causes severe acute respiratory syndrome (SARS) and middle east respiratory syndrome (MERS CoV). The diagnosis is confirmed by the risk of travel or infection within 14 days accompanied by symptoms of upper or lower respiratory tract infection, accompanied by laboratory evidence of Covid-19 real-time polymerase chain reaction (RT-PCR) examination [11]. The World Health Organization divides Covid-19 into suspect, probable and confirmed cases. While the Ministry of Health of the Republic of Indonesia (Kemenkes RI) classifies, people under surveillance (ODP), patients under surveillance (PDP), people without symptoms (OTG) and the patient are confirmed if the RTPCR Covid-19 result is positive with any symptoms. Examination materials can be in the form of throat swabs, sputum and bronchoalveolar lavage (BAL). The diagnosis becomes critical, especially if accompanied by comorbid, elderly and have a history of the previous pulmonary disease [11].

B. Time Series Forecasting

Time series is an observation series on variables that will be observed sequentially overtime period and recorded based on the time sequence of occurrence [12]. Time series analysis is a forecasting method for the future based on past values or data from a variable and past errors. The purpose of this time series forecasting

method is to find time series data patterns and extrapolate these patterns to future periods. Every observation made can be expressed in the form of a random variable Z_t obtained based on a certain time index t_i with $i =$ 1,2,3… this is a sequence of observation times, so the writing of the time series data is $Zt_1, Zt_2, \ldots, Zt_n$.

1) Seasonal Trend Loess (STL) Decomposition Model: The Seasonal Trend Loess (STL) is an algorithm that is widely used to decompose an observation into three major components: trend, seasonality and remainder (residual) [8]. The STL has several parameters that can be analyzed. In many implementations, STL process can assist critical calculations quickly related to identifying data characteristics. According to Rob and George [13], the STL Decomposition is a time series decomposition model that has several advantages to other models such as Classical Decomposition, SEATS Decomposition and X11 Decomposition. The STL Decomposition has a local smotherer namely Loess. Loess (Local Regression) itself has high flexibility because the data will automatically form a curve estimate that is not influenced by subjective factors [9]. This Loess considers the STL Decomposition as the robust approach. The general equation of the STL Decomposition with the additive model can be described as follows,

$$Y_t = S_t + T_t + R_t, t = 1 \text{ until } t = n \quad (1)$$

Where, Y_t is time-series observation at time t, followed by S_t as a seasonal component at time t following SARIMA model, T_t as a trend-cyclical component at time t, and R_t as a component of the residual at time t.

2) Seasonal ARIMA (SARIMA) Model: SARIMA is an ARIMA model which contains autoregressive and seasonal elements [14]. According to K.E. ArunKumar, et al. used the SARIMA model in their research, they proposed that SARIMA is outperformed than ARIMA model because SARIMA has more realistic numbers. After all, they considered the variations that occurred of time series data [15]. The following is an Equation of the SARIMA model with order $(p, d, q)(P, D, Q)^S$.

$$\Phi_p(B)\Phi_p{}^*(B^S)(1-B)^{\mathrm{d}}(1-B^s)^D S_t = \theta_q(B)\theta_q{}^*(B^S)\epsilon_t \quad (2)$$

Where S_t is observed value at time t, Φ_p is the autoregressive (AR) parameter of the order p, θ_q as parameter(s) of the order moving average (MA) q, Φ_p as parameter(s) of the seasonal autoregressive (SAR) order P. Moreover, $\theta_q{}^*(B^S)$ as parameter(s) of the order seasonal moving average (SMA) Q, $(1-B)^{\mathrm{d}}$ as a non-seasonal differencing order d, $(1-B^s)^D$ as differencing seasonal orders D, s as the length of seasonal period, p, q as non-seasonal orders AR and MA. Also, we have P, Q as seasonal orders AR (SAR) and seasonal MA (SMA), and ϵ_t is white noise. To find the optimal order of time series model sequence, we consider the minimum value of Akaike Information Criterion (AIC).

$$AIC = 2\log(L) + 2(p + q + P + Q + k) \quad (3)$$

Since k = 1, and L is the Maximum Likelihood probability of the model used as described in Equation 3. However, it is not possible to fit every potential model and find the model with the lowest AIC. In the theory of Hyndman and Khandakar [16], they propose a stepwise selection algorithm to find the ordering model efficiently. In the second step, up to thirteen variations on the initially selected model will be considered for the model with the lowest AIC.

C. Accuracy

Accuracy is a ratio of the exactness between the forecast and targeted value in the data test. The logic behind accuracy that no forecast has an accuracy of 0% since every prediction may adversely contain an error. Therefore, to conclude the best approach to the particular phenomenon/observation, it is necessary to calculate the error rate in forecasting [17].

1) Root Mean Square Error: RMSE is valuable for evaluating forecasting techniques used to measure the level of accuracy of the forecast results according to a particular model. RMSE is the average value of the

squared error generated by a forecast model. A low RMSE value indicates that the variation in the value generated by a forecasting model is close to the variation in its observed value [18].

$$RMSE = \frac{\sqrt{\sum_{t=1}^{n}(A_t - F_t)^2}}{n} \quad (4)$$

where A_t, F_t, and n is real value, predictive value, and several prediction periods, respectively.

2) Mean Absolute Percent Error: MAPE is a measure of relative error. MAPE usually states the percentage error in forecasting the actual demand during a certain period which provides information on the percentage of errors that are too high or too low [19].

$$MAPE = \frac{\sum_{i=1}^{n} \frac{|A_t - F_t|}{A_t} \times 100}{n} \quad (5)$$

3) Mean Absolute Error: MAE is one of the methods used to measure the accuracy of forecasting models in time series analysis [20].

$$MAE = \frac{\sum_{i=1}^{n} |A_t - X_t|}{n} \quad (6)$$

III. RESEARCH METHOD

The flowchart design of STL-SARIMA model is described in Figure 1. Through the design of this model, which including data preparation, calculate the increment rate, splitting the training and testing data, stages to further carry out STL Decomposition and forecasting increment rate of the daily number of new Covid-19 cases based SARIMA using training dataset in DKI Jakarta.

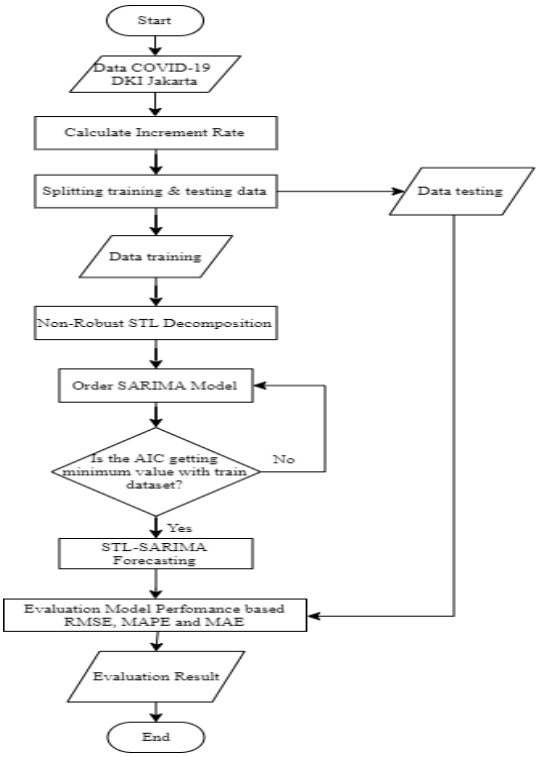

Fig 1. Flowchart System of Forecasting based STL-SARIMA Model

A. Data Covid-19

The Covid-19 dataset was gathered from the official website kawalcovid19.id. This website shows information regarding the level of spread and severity of Covid-19 in Indonesia, including the DKI Jakarta Province. The increment rate represents percentages of daily new cases of Covid-19, which depends on how much the number of positive cases in a particular period as well as the active total cases [1]. We consider the daily Covid-19 new cases and total cases in DKI Jakarta from May 1, 2020, to January 31, 2021, into STL-SARIMA model, the data shown in Table 1.

TABLE I
THE SAMPLE COVID-19 DATASET DKI JAKARTA

Date	Daily New Cases	Total Cases
2020-05-01	142	4317
2020-05-02	80	4397
2020-05-03	66	4463
…	…	…
2021-01-29	3448	262753
2021-01-30	3491	266244
2021-01-31	3474	269718

B. Calculate Increment Rate

The increment rate on this topic is the percentage rate of daily new cases of Covid-19 in DKI Jakarta. The Equation of increment rate covers several positive cases in a particular period and the active total cases as follows,

$$K = \frac{n}{m} * 100 \tag{7}$$

Where, K is the increment rate, followed by n as a number of new cases of Covid-19, m as a number of total cases of Covid-19. The plot of daily increment rate data for Covid-19 new cases in DKI Jakarta can be seen in Figure 2. This total observation is 276, and we evaluate into three scenarios of training and testing.

C. Training and Testing Data

We reveal three scenarios of training and testing data, as shown in Figure 2. The first scenario (70:30), the dataset splits into 192 training and 84 testing data. The second is scenario (80:20) for 220 training and 56 testing data, and the last scenario (90:10) splitting into 240 training and 36 testing data.

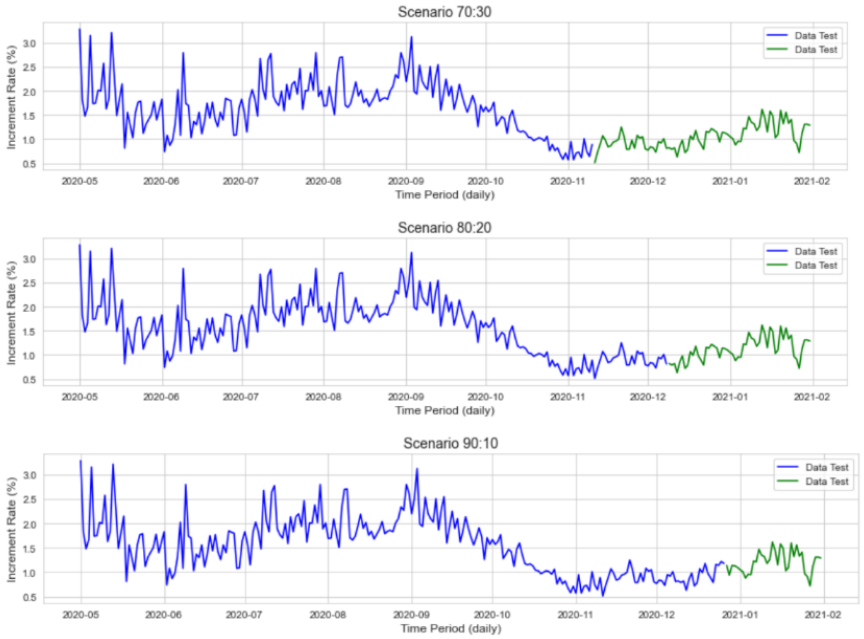

Fig 2. Scenario of Training and Testing Data

D. Training Model using STL Decomposition Model

In this paper, the STL model involves an additive approach to observe the seasonal, trend, and noise of train observation to compare non-robust and robust STL's which function to identify the performance of analyzed time series data that has smoothed or not smoothed, with calculation in Equation 1. Table II shows comparison results between robust and non-robust approaches.

TABLE II
RESIDUAL COMPARISON OF ROBUST AND NON-ROBUST METHODS

Model	Robust			Non-Robust		
	70:30	**80:20**	**90:10**	**70:30**	**80:20**	**90:10**
Seasonal	-0.96922	-0.93567	-0.93950	-0.70053	-0.70566	-0.70566
Trend	0.58854	0.86419	0.84436	0.58342	0.83518	0.83434
Residual	**-1.93441**	**-1.93080**	**-1.92686**	**-1.02134**	**-1.02020**	**-1.02020**

We evaluate STL Decomposition considering residual of robust and non-robust methods over three training and testing scenarios. The result has shown that non-robust STL well performed by the lowest residual compared to the robust-STL. The residual trend and the characteristics of observed value (trend and seasonal) are shown in Figure 3. According to that result, we construct SARIMA model-based non-robust.

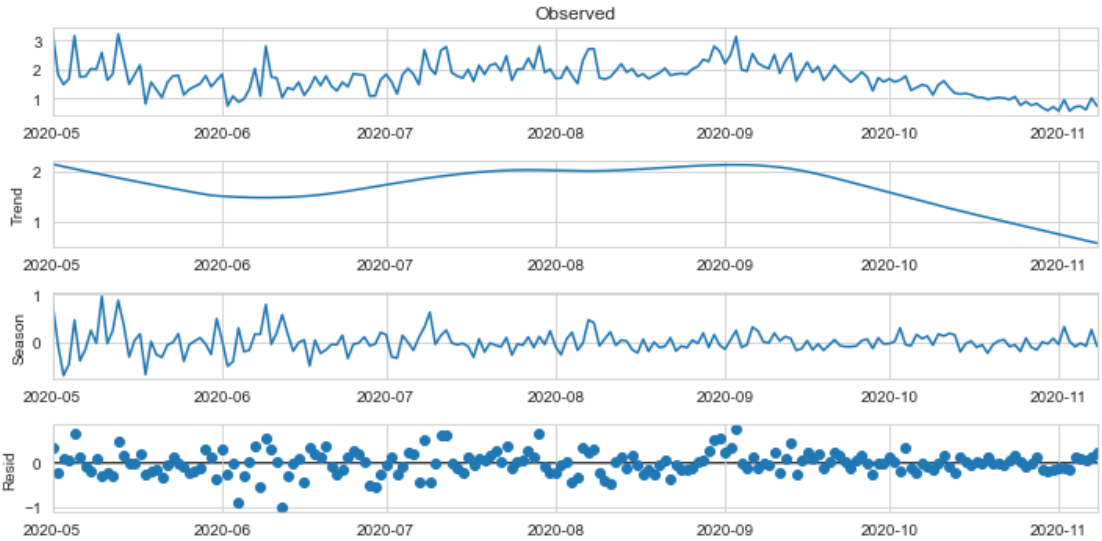

Fig 3. Non-Robust Plot of Data COVID-19 using STL Decomposition Model

E. Forecasting using STL-SARIMA

The STL-SARIMA model is a combination of the STL approach and Seasonal ARIMA to perform seasonal time series forecasting, with the calculations described in Equation 8,

$$X_t = T_t + S_t + R_t \tag{8}$$

where, X_t is the forecasting of STL-SARIMA model observation at a particular period t, T_t as trend decomposition at time t, S_t is seasonal decomposition following SARIMA model in Equation 2, and R_t is residual decomposition at time t.

IV. RESULTS AND DISCUSSION

According to the result of AIC, the optimal order of SARIMA model with the lowest AIC based on Equation 3, 257.794 is $(0,1,2)(2,1,0)^{12}$. Thus, S_t as seasonal of STL Decomposition can be written as follows,

$$(1-B)(1-B^{12})(1-\Phi_1 B^{12}-\Phi_2 B^{24})S_t = (1+\theta_1 B+\theta_2 B^2)\epsilon_t \quad (9)$$

Thus, we can construct combination of STL-SARIMA in Equation 8 considering Equation 9. We run three scenarios of training and testing dataset, to evaluate the error performance and analyze the parameter estimation in Table III, and plot of increment rate the forecasting scenarios shown in Figure 4, 5, 6.

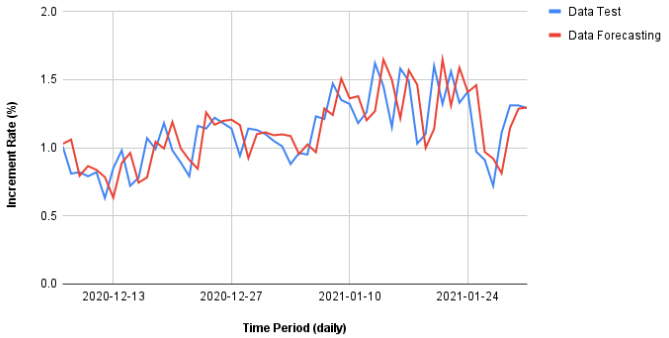

Fig 4. Plot of Increment Rate of Forecasting Scenario (70:30)

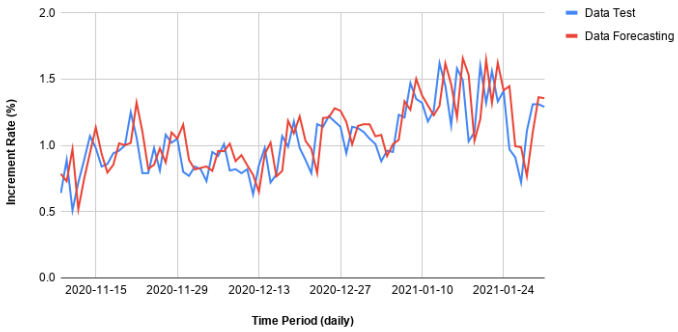

Fig 5. Plot of Increment Rate of Forecasting Scenario (80:20)

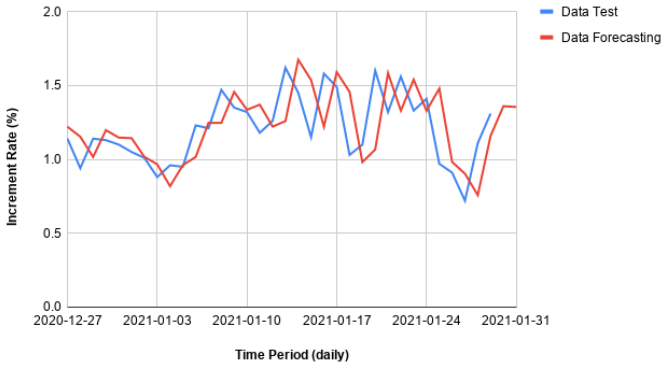

Fig 6. Plot of Increment Rate of Forecasting Scenario (90:10)

Based on Fig. 4, 5, and 6, for all scenarios, those figures show that the fluctuation trends of observed value well-captured by STL-SARIMA model. Therefore, can be seen from splitting training and testing data Increment Rate Covid-19, has increased drastically from December to January. In particular, the highest increment rate forecasting value was obtained on January 17, 2021, with an increment rate Covid-19 of 1.67%. Based on testing been carried out of increment rate Covid-19 provides the forecasting results is close to actual data with STL-SARIMA $(0,1,2)(2,1,0)^{12}$ model performance, shown in Table IV.

TABLE III
PARAMETERS ESTIMATION OF STL-SARIMA MODEL

Scenarios	Coefficient				z				$P > \|z\|$			
	x1	*ma.L1*	*ma.L2*	*sigma2*	*x1*	*ma.L1*	*ma.L2*	*sigma2*	*x1*	*ma.L1*	*ma.L2*	*sigma2*
70:30	-0.008	-0.657	-0.123	0.1411	-0.988	-10.839	-1.996	16.456	0.323	0.000	0.046	0.000
80:20	-0.007	-0.665	-0.112	0.1265	-0.935	-12.279	-2.302	19.009	0.350	0.000	0.042	0.000
90:10	**-0.005**	**-0.661**	**-0.113**	**-0.117**	**-0.832**	**-13.059**	**-2.217**	**20.342**	**0.405**	**0.000**	**0.027**	**0.000**

TABLE IV
THE STL-SARIMA ERROR PERFORMANCE

Scenarios	RMSE	MAPE	MAE
70:30	0.19×10^{-2}	0.16×10^{-2}	0.15×10^{-2}
80:20	0.21×10^{-2}	0.16×10^{-2}	0.17×10^{-2}
90:10	0.23×10^{-2}	$\mathbf{0.15 \times 10^{-2}}$	0.18×10^{-2}

Our state-of-the-art Model STL-SARIMA by forecasting method is proven in giving a better outcome than the previous research [10], which determines forecasting with the use of STL-ARIMA. Based on the comparison of the two models that have been built, our model generated from the combination of STL-SARIMA produces a lower error value according to what is listed in Table IV than the other one that has been done in previous research by using STL-ARIMA separately with the MAPE value of 2.6, RMSE 849, and MAE 686. Model of STL-SARIMA helps in overcoming time series data problem, which shows the best model with AIC for data training automatically.

From Table III, we have parameters estimation of STL-SARIMA $(0,1,2)(2,1,0)^{12}$ model considering Equation 9. The statistical test with scenario (90:10) outperforms than others, each parameter remarkably has $P > |z|$ less than 5% of the significance level. Thus, STL-SARIMA $(0,1,2)(2,1,0)^{12}$ model has prominent parameters contributing to the proposed model**.** Meanwhile, the result of the error forecasting can be seen in Table IV, STL-SARIMA $(0,1,2)(2,1,0)^{12}$ with scenario (90:10) shows the minimum error, since the largest numbers of training data performances the impressive model.

V. CONCLUSION

According to dataset of Covid-19 increment rate in DKI Jakarta from May 1, 2020, to January 31, 2021, STL Decomposition does not always require a local regression (Loess) to identify characteristics over this dataset. For overall scenarios of training and testing data, Seasonal Trend Loess (STL) Decomposition performs well to capture the trend and noise over the Covid-19 increment rate dataset. Furthermore, the forecast involving Seasonal Autoregressive Integrated (SARIMA) on STL Decomposition shows a significant performance with error value equal to 0.15. As the massive severity of the Coronavirus, identifying the future number of increment rates become a major concern to support information and maintain essential health services. The numerical result demonstrates that STL-SARIMA $(0,1,2)(2,1,0)^{12}$ model can be addressed to anticipate potentially overwhelming number of Covid-19 cases. Therefore, under the high volatility of increment rate of this outbreak, finding an appropriate model, such heteroscedastic time series model, is desirable and should be considered.

REFERENCES

[1] Indika Foundation, "Kawal Covid 19", 2020, Accessed: May 09, 2021. [Online]. Available: https://kawalcovid19.id/.

[2] F. R. Yamali and R. N. Putri, "Dampak Covid-19 Terhadap Ekonomi Indonesia", *Journal of Economics and Business*, vol. 4, no. 2, pp. 384-388, 2020.

[3] D. C. Montgomery, C. L. Jennings and M. Kulachi, "Introduction to Time Series Analysis and Forecasting", New Jersey, USA: John Willey & Sons, Inc., 2008.

[4] N. Feroze, "Forecasting the patterns of COVID-19 and causal impacts of lockdown in top five affected countries using Bayesian Structural Time Series Models", *Chaos, Solitons & Fractals*, vol. 140, 2020.

[5] M. S. Rumetna and T. N. Lina, "Forecasting Number of Covid-19 Positive Patients in Sorong City Using the Moving Average and Exponential Smoothing Methods", *International Journal of Informatics and Computer Science*, vol. 5, pp. 37-43, 2021.

[6] M. H. D. M. Riberio, R. G. da Silva, V. C. Mariani and L. d. S. Coelho, "Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil", *Chaos, Solitons & Fractals*, vol. 153, pp. 1-10, 2020.

[7] I. Djakaria and S. E. Saleh, "Covid-19 forecast using Holt-Winters exponential smoothing", *Journal of Physics: Conference Series*, vol. 1882, 2021.

[8] R. B. Cleveland, W. S. Cleveland, J. E. McRae and I. Terpenning, "STL: A Seasonal-Trend Decomposition Procedure Based on Loess", *Journal of Official Statistics*, vol. 6, no. 1, pp. 3-73, 1990.

[9] A. Haritsah, "Implementasi Model Stl Seasonal Trend Decomposition Based On Loess Dan Arima Untuk Prediksi Konsentrasi Kualitas Udara", 2015.

[10] J. Buhl, C. Liedtke, S. Schuster and K. Bienge, "Predicting the Material Footprint in Germany between 2015 and 2020 via Seasonally Decomposed Autoregressive and Exponential Smoothing Algorithms," *Resources*, vol. 9, no. 125, pp. 1-17, 2020.

[11] D. Handayani, D. R. Hadi, F. Isbaniyah, E. Burhan and H. Agustin, "Penyakit Virus Corona 2019", *Jurnal Respirologi Indonesia*, vol. 40, no. 2, pp. 119-129, 2020.

[12] W. W. Wei, "Time Series Analysis: Univariate and Multivariate Methods", Boston, USA: Pearson Addison Wesley, 2006.

[13] R. J. Hyndman and G. Athanasopoulos, "Forecasting: Principles and Practice (2nd ed)", Melbourne, Australia: OTexts, 2018.

[14] D. Mulya, Y. Asdi and F. Yanuar, "Penerapan Metode Holt Winter Dan Seasonal Arima Pada Peramalan Perkembangan Wisatawan Mancanegara Yang Datang Ke Indonesia", *Jurnal Matematika UNAND*, vol. VI, no. 4, pp. 29-36, 2017.

[15] K. E. Arun Kumar, D. V. Kalaga, C. M. Sai Kumar, G. Chilkoor, M. Kawaji and T. M. Brenza, "Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16

countries using statistical machine learning models: ARIMA and SARIMA", *Applied Soft Computing Journal*, vol. 103, pp. 1-26, 2021.

[16] J. R. Hyndman and Y. Khandakar, "Automatic Time Series Forecasting: The forecast Package for R", *Journal of Statistical Software*, vol. 27, no. 3, pp. 1-22, 2008.

[17] H. D. E. Sinaga and N. Irawati, "Perbandingan Double Moving Average Dengan Double Exponential Smoothing Pada Peramalan Bahan Medis Habis Pakai", *Jurnal Teknologi dan Sistem Informasi*, vol. IV, no. 2, pp. 197-204, 2018.

[18] A. I. Laksana, "Perbandingan Metode Single Moving Average dan Single Exponential Smoothing dalam Pengembangan Sistem Peramalan Penjualan Mobil Baru", 2017.

[19] T. S. Saputra and Terttiaavini, "Analisa Akurasi Penggunaan Metode Single Eksponential Smoothing untuk Perkiraan Penerimaan Mahasiswa Baru Pada Perguruan Tinggi XYZ", *Jurnal Ilmiah Informatika Global*, vol. 11, p. 66, 2020.

[20] A. A. Suryanto and A. Muqtadir, "Penerapan Metode Mean Absolute Error (Mea) Dalam Algoritma Regresi Linear Untuk Prediksi Produksi Padi", *Jurnal Sains dan Teknologi*, vol. 11, p. 79, 2019.