International Journal of Mathematics Trends and Technology
ISSN: 2231 – 5373 /doi:10.14445/22315373/IJMTT-V67I6P516
Volume 67 Issue 6, 136-144, June, 2021

Nonparametric Modeling Using Kernel Method for the Estimation of the Covid-19 Data in Indonesia During 2020

Subian Saidi[1], Netti Herawati[2], Khoirin Nisa[3], Eri Setiawan[4]

[1,2,3,4)]*Department of Mathematics, Fakulty of Mathematics and Natural Sciences, University Of Lampung, Indonesia*

Abstract *- One of the most research discussed topics is the prediction or forecasting of the COVID-19 data using the classical time series (such as exponential smoothing) and machine learning methods. In fact, the classical time series method often produces quite large error rates. In this study, the researchers try to use nonparametric modeling with the kernel method to get better results with the smallest error rate. Furthermore, the results of the kernel method are compared with the results of the classical time series method. As a comparison tool, the researchers use MAPE by paying attention to the smallest MAPE value. The data used in this study are the COVID-19 data in Indonesia in which its variable is the total of deaths per day. After comparing the classical time series method with the kernel method, the obtained better results are the results from the kernel method. In this study, the researchers use five kernel functions, namely the Gaussian, Epanechnikov, Triangular, Biweight, and Triweight. Then, these five kernel functions are compared to find the best function. After the comparison process is done, the triweight kernel function was determined as the best function with the smallest error rate with a MAPE value of 0,9%.*

Keywords — *Covid-19, Kernel Method, Mean Absolute Percentage Error, Time Series, Triweight Kernel Function.*

I. INTRODUCTION

In 2019 – 2020, an outbreak attacked the world called the COVID-19 (Coronavirus Disease-19). This 2019 coronavirus disease (Covid-19) is an infectious disease caused by severe acute respiratory syndrome from Coronavirus-2 (SARS-CoV-2). The virus emerged in China in December 2019 and since then it has spread rapidly throughout the world. As of 7 December 2020, WHO recorded that there were 65.8 million confirmed cases of the COVID-19, in which 1.5 million of them passed away and the COVID-19 has spread to 220 countries [1]. This COVID-19 has been declared as a pandemic by WHO on 11 March 2020.

This COVID-19 which has hit many countries in the world has made many scientists and academics conducting research on it. COVID-19 data that is calculated every day is time-series data. A study on time series analysis on the COVID-19 data has been conducted by kartis [2], Nikolopoulos et. al [3] and Gecil et. al [4] who used the classical time series method and machine learning approach to forecasting. COVID-19 forecasting has also been carried out by Kalantari [5] using the spectrum analysis method and Shi et. al [6] using weight kernel density estimation. Furthermore, [7], Triacca [8], Geng et. al [9] and Elson [10] has also conducted a similar study with a different analysis method. Mathematical modeling can also be used in estimating and predicting COVID-19 cases, as carried out by Zakary [11] in Morocco, Pande [12], Oyetunde [13], Oyetunde et. Al [14] and Hassan [15].

The methods commonly used in time series data studies are exponential smoothing. However, in fact, some data do not meet the time series assumptions so that if the analysis is still carried out using the classical time series method, it will result in large error rates and inaccurate forecasting. To overcome this problem, nonparametric modeling can be used, in which this modeling does not require time-series assumptions and is expected to produce smaller error rates. In several time series data studies, this nonparametric modeling is used and the results indicate that this modeling produces a smaller error rate than parametric modeling, such as studies conducted by [16] on the kernel regression model and [17] on a nonparametric regression approach with the kernel function, in which both studies discuss the composite stock price index. Studies on the kernel regression have been also carried out by [18] and [19] who analyzed inflation data in Indonesia. Furthermore, [20] also used the kernel method on time series data.

In Indonesia, COVID-19 cases are still increasing day by day. As of 10 December 2020, there were 598,933 confirmed cases with 491,975 recovered and 18,336 deaths [21]. For this reason, the researchers are interested in conducting a study on the COVID-19 data using nonparametric modeling to obtain better results than the parametric modeling. Therefore, the researchers conducted a study entitled "Nonparametric Modeling Using Kernel Method for the Estimation of the COVID-19 Data in Indonesia during 2020".

II. STATISTICAL MODEL

A. Classical Time Series Forecasting Methods

The classic time series method that is widely used is exponential smoothing which consists of

1) Single exponential smoothing **(SES).** It is a method that can reduce data storage problems so that all or part of historical data does not need to be stored. We just need to keep the last observations & forecasts and the constant value [22]. The following is the equation of this method.

$$F_t = \alpha\, Y_{t-1} + (1-\alpha) F_{t-1}$$

2) Double exponential smoothing **(DES).** It is a method whose prediction begins by determining the parameter size using trial and error. This method is more often used to predict trend-patterned data. According to [22], there are three equations concerning this method, namely as follows.

Exponential smoothing on overall data:

$$S_t = \alpha\, X_t + (1-\alpha)(\, S_{t-1} + b_{t-1})$$

Trend pattern smoothing:

$$b_t = \gamma\, (\, S_t - \, S_{t-1}) + (1-\gamma)\, b_{t-1}$$

Smoothing of the future *m* period:

$$F_{t+m} = \, S_t + \, b_t m$$

3) The Holt-Winters Method. It is also called triple exponential smoothing. This method is the development of exponential smoothing designed for trend and seasonal time series data. According to [22], there are four equations concerning this method, namely as follows.

Exponential smoothing on overall data:

$$S_t = \alpha \frac{X_t}{I_{t-L}} + (1+\alpha) + (\, S_{t-1} + b_{t-1})$$

Trend pattern smoothing:

$$b_t = \gamma\, (\, S_t - \, S_{t-1}) + (1-\gamma)\, b_{t-1}$$

Seasonal pattern smoothing:

$$I_t = \beta \frac{X_t}{S_t} + (1-\beta) I_{t-L}$$

Forecasting of the future *m* period:

$$F_{t+m} = (\, S_t + \, b_t m) I_{t-L+m}$$

B. Kernel estimator

The kernel estimator is an explicit modeling method based on a probability density function with a perpendicular observation distances that does not require any assumptions of the shape of the data distribution. kernel estimator is almost the same as the other linear estimators. However, the kernel method is more specific in the use of the bandwidth method [23]. According to [24], the kernel estimator is divided into three types, namely as follows.

Nadaraya-Watson Estimator:

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)}$$

Priestly-Chao Estimator:

$$\widehat{m}(x) = \frac{1}{h} \sum_{i=1}^{n} (x - x_{i-1})\, Y_i\, K\left(\frac{x - X_i}{h}\right)$$

Gasser-Muller Estimator:

$$\widehat{m}(x) = \frac{1}{h} \sum_{i=1}^{n} Y_i \int_{s_{i-1}}^{s_i} K\left(\frac{x - X_i}{h}\right) dx$$

According to [25], the general kernel function is as follows.

$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right)$$

with several kernel functions in the estimator kernel used for data estimation as shown in the following table.

Table I. Commonly Used Kernel Functions

o	Kernel	$K(x)$
	Uniform	$\frac{1}{2} I\ (\lvert x\rvert \le 1)$
	Triangle	$(1-\lvert x\rvert)\ I\ (\lvert x\rvert \le 1)$
	Epanechnikov	$\frac{3}{4}(1-x^2)\ I\ (\lvert x\rvert \le 1)$
	Biweight	$\frac{15}{16}(1-x^2)^2\ I\ (\lvert x\rvert \le 1)$
	Triweight	$\frac{35}{32}(1-x^2)^3 I(\lvert x\rvert \le 1)$
	Gaussian	$\frac{1}{\sqrt{2\pi}}\exp\left(\frac{1}{2}(-x^2)\right) -\infty < x < \infty$
	Cosine	$\frac{\pi}{4}\cos\left(\frac{\pi}{4}x\right) I(\lvert x\rvert \le 1)$
	Tricube	$\frac{70}{81}(1-\lvert x\rvert^3)^3 I(\lvert x\rvert \le 1)$
	Logistics	$\frac{1}{e^x+2+e^{-x}}$

where I is the indicator function with

$$I(x) = \begin{cases} 1\ if\ |x| \le 1 \\ 0\ if\ |x| > 1 \end{cases}$$

To get the best estimate, one of the most important things is to choose the optimal bandwidth with associated kernel functions. This can be done using the Generalized Cross-Validation (GCV) criterion with formula

$$GCV = \frac{MSE}{\left(\frac{1}{n}\ tr\big(I - H(h)\big)\right)^2}$$

where $H(h) = X(X'X + nhI)^{-1}X'$ and $MSE = \frac{1}{n}\sum_{i=1}^{n}\big(y_i - m_h(x_i)\big)^2$. The error rate measurement to compare the best estimator between classical time series forecasting methods and the kernel estimator is based the value of:

Root Mean Square Error (RMSE):

$$RMSE = \sqrt{MSE}$$

Mean Absolute Deviation (MAD):

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|e_i| = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Mean Absolute Percentage Error (MAPE):

$$MAPE = \sum_{i=1}^{n}\left|\frac{y-\hat{y}}{y}\right| \times 100\%$$

III. METHOD

In this study, researchers used mortality data of COVID-19 patients in Indonesia from April to November 2020 obtained from covid19.go.id. We estimated the data model using time series forecasting methods such as exponential smoothing. The results were compared with the kernel estimator method using several kernel functions (Gaussian, Epanechnikov, Triangular, Biweight, and Triweight) and the optimal bandwidth was selected based on the smallest GCV. The best estimator is determined using the smallest MAPE value.

IV. RESULTS AND DISCUSSION

A. *Data Distribution*

The following is a plot of the distribution of data on patients who died from COVID-19 in Indonesia

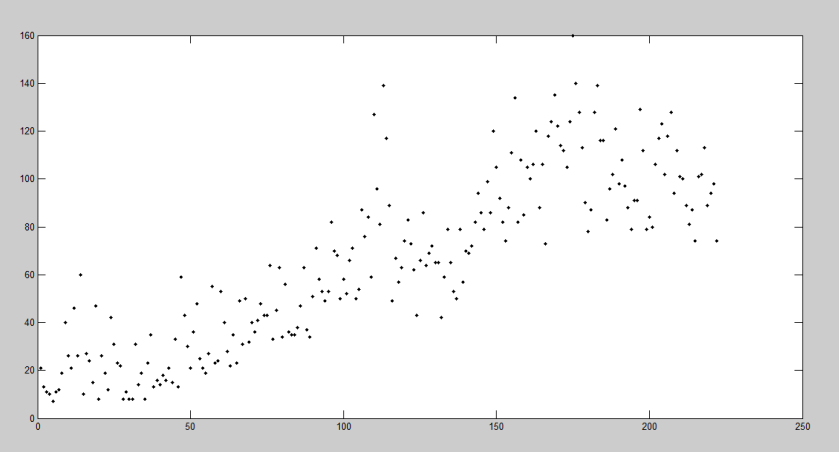

Figure 1. curve of the data distribution of patients who died due to covid-19 in Indonesia

From the figure 1 above, it can be concluded that the number of patients dying from COVID-19 in Indonesia tends to increase. But the more to the right the fluctuation in the number of deaths per day is getting bigger. And if we take a closer look, we can see that the end of the curve is showing signs of decline. However, a more in-depth analysis is needed to confirm this.

B. *Exponential Smoothing*

Here the author uses exponential smoothing as a comparison method because the data distribution follows the exponential smoothing pattern. The following is the plot of the results of the exponential smoothing (*Single exponential smoothing* (SES , double *exponential smoothing*, and Holt-Winters method) analysis using different value of parameters.

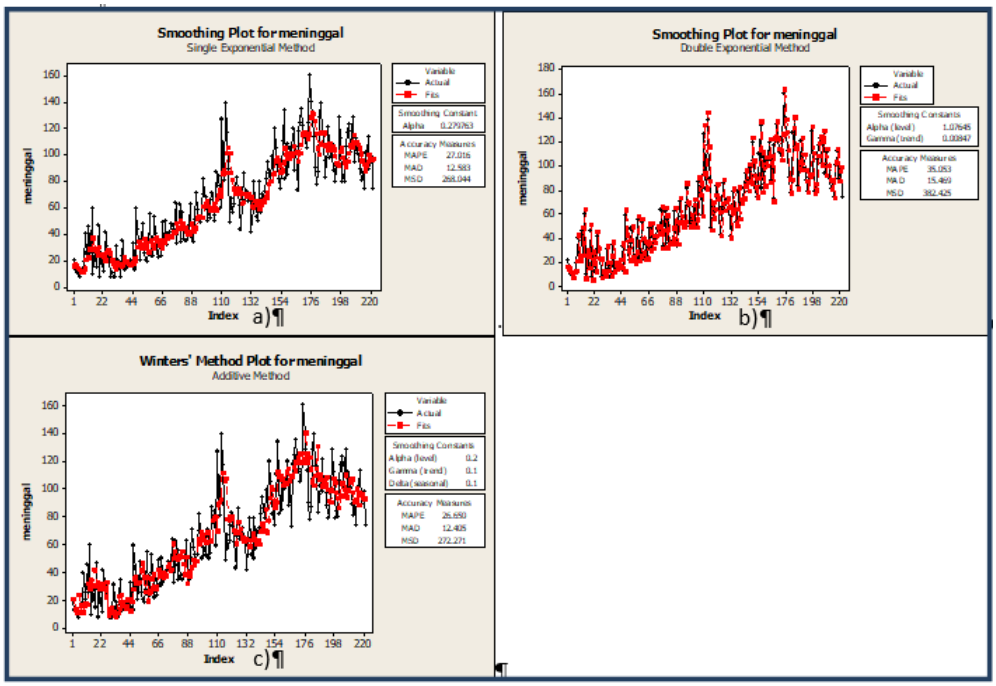

Figure 2 Plot of data using the exponential smoothing method

Based on Figure 2 above, in Single Exponential Method it can be seen that the parameter value of α= 0.279763 has MAPE= 27.016. in Double Exponential Method with α= 0.279763 and γ= 0.00847 has MAPE= 35.053. And in Holt-Winters Method, it can be seen that the parameter value of α is 0.2, the parameter value of γ is 0.1, and the parameter value

of δ is 0.1 with a MAPE value of 26.650.

C. Optimum Bandwidth Selection.

In this study, the method used to determine the optimum bandwidth is GCV. Using the range 0.1 to 2, the following results are obtained:

Table II. Optimum bandwidth selection

No	h	GCV	No	h	GCV
1	1.0	228.7283	11	1.3	224.4010
2	0.9	232.5138	12	1.4	224.6718
3	0.8	237.3243	13	1.2	224.8314
4	0.7	242.3303	14	1.5	225.4312
5	0.6	246.1749	15	1.1	226.2003
6	0.5	247.9758	16	1.6	226.5037
7	0.4	248.3623	17	1.7	227.7510
8	0.3	248.4032	18	1.0	228.7283
9	0.2	248.4063	19	1.8	229.0675
10	0.1	Inf	20	1.9	230.3755

Based on the output of the statistical program above, we can see that bandwidth 1.3 is the optimum bandwidth with the smallest GCV value of 224.4010. So that the bandwidth will be used for COVID-19 data modeling with kernel functions.

D. Kernel Method.

To perform analysis using the kernel method, the first thing to do is to determine the bandwidth. In this study, bandwidth 1.3 is used for those five functions. Furthermore, the results are presented in the following figure.

1) ***Gaussian Kernel***. Using the Gaussian kernel function, the smoothing curve is obtained as follows:

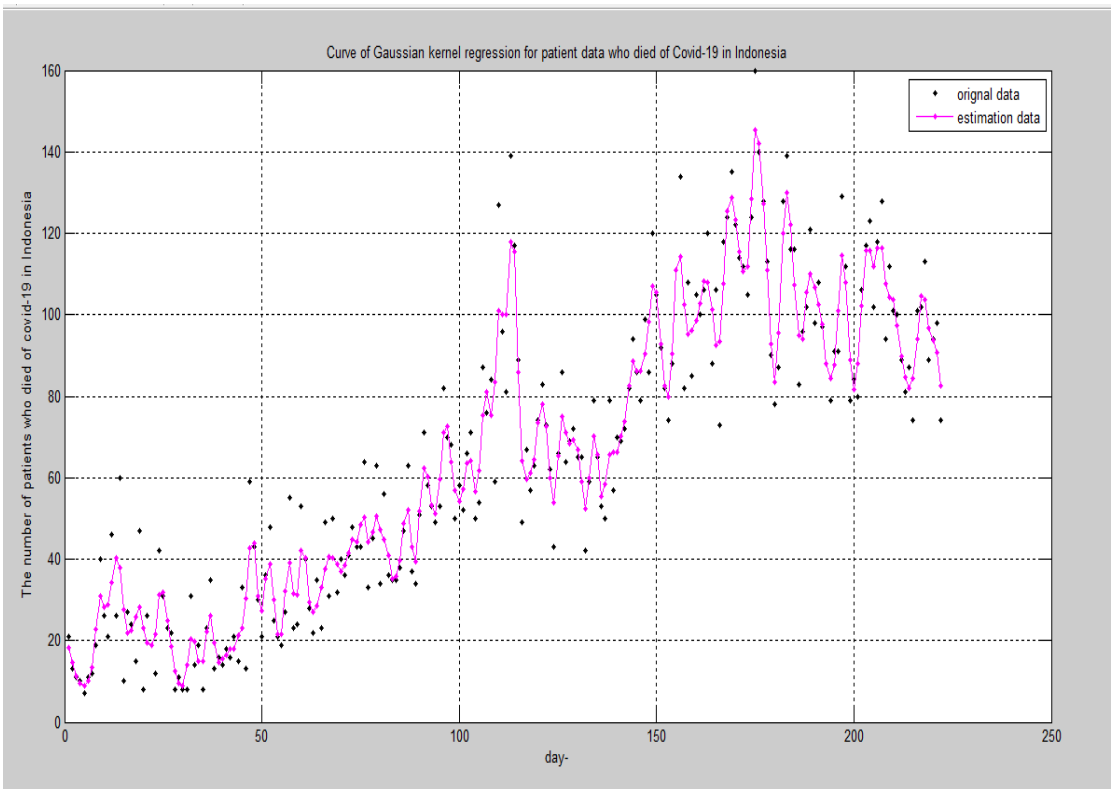

Figure 3. Curve of Gaussian Kernel

From the figure 3 above, it can be seen that the smoothing curve (purple line) is still very different from the original data plot. This shows that the function of the Gaussian kernel is not good enough for smoothing COVID-19 data in Indonesia.

2) ***Epanechnikov Kernel***. Using the epanechnikov kernel function, the smoothing curve is obtained as follows:

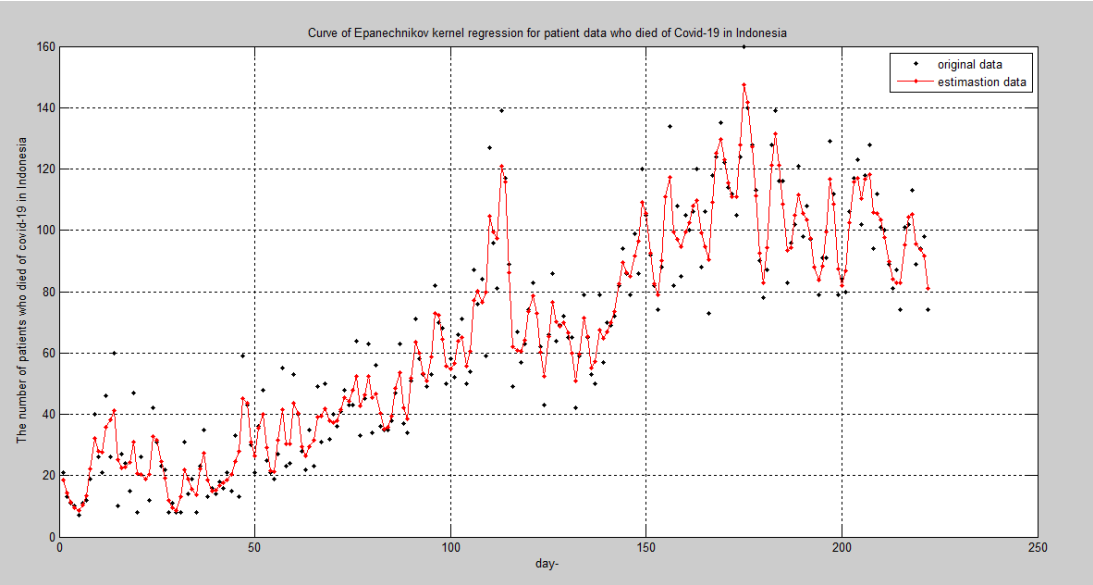

Figure 4. Curve of Epanechnikov Kernel

From Figure 4 above, it can be seen that the smoothing curve (red line) is still much different from the original data, but looks better when compared to the Gaussian kernel function. But we need to look at smoothing curves with other kernel functions.

3) ***Triangular Kernel.*** Using the triangular kernel function, the kernel smoothing curve is obtained as follows:

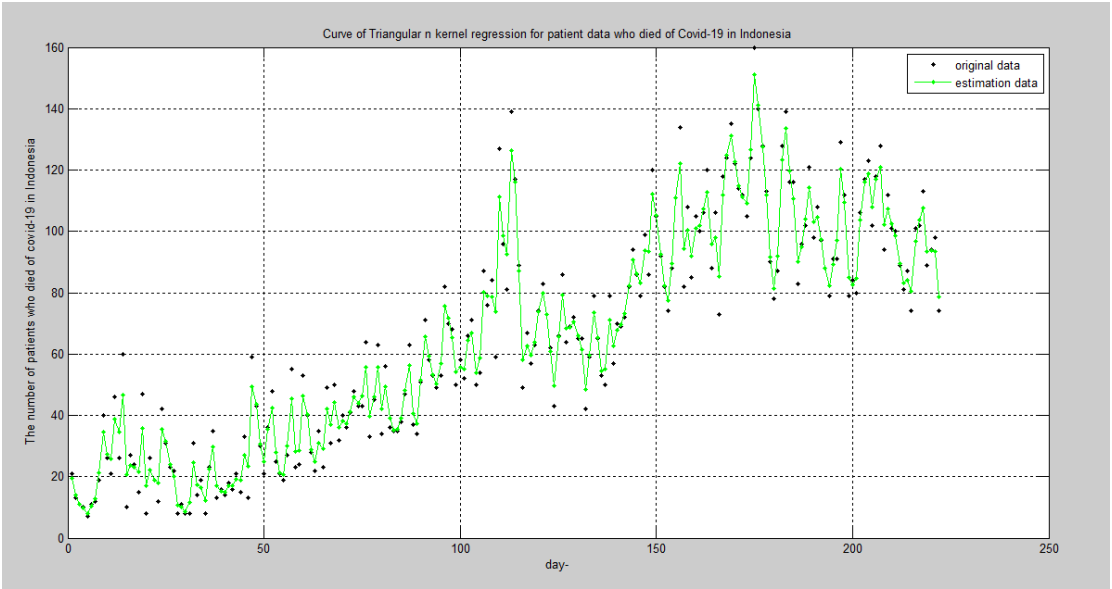

Figure 5. Curve of Triangular Kernel

From Figure 5 above, it can be seen that the smoothing curve with the triangular kernel function (green line) is getting closer to the original data plot. This shows that the triangular kernel is better than the Gaussian kernel and the epanechnikov kernel.

4) ***Biweight Kernel.*** Using the kernel biweight function, the smoothing curve is obtained as follows:

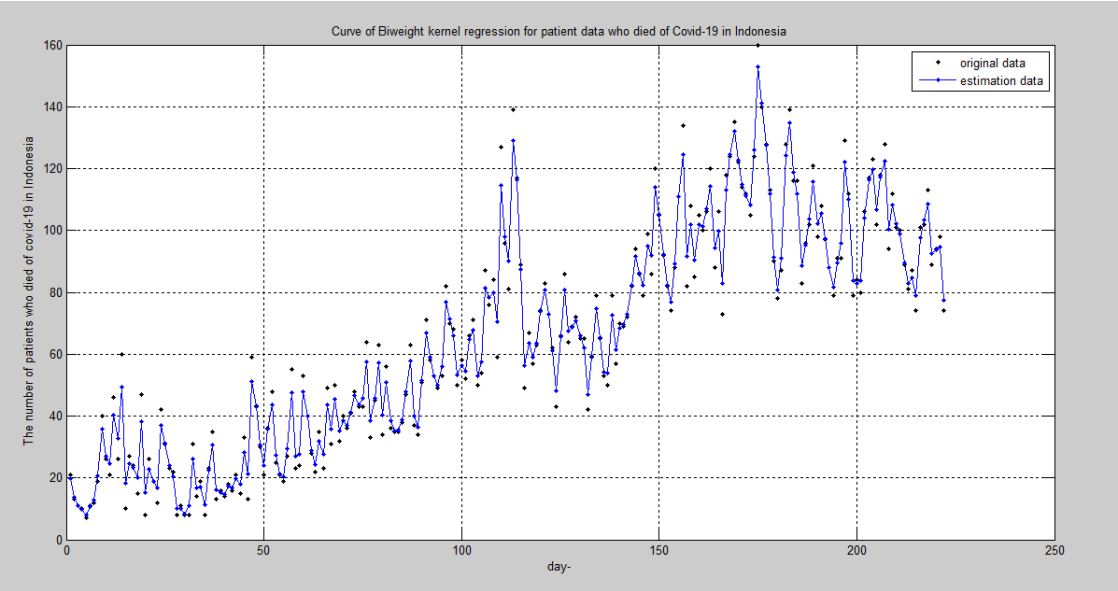

Figure 6. Curve of Biweight Kernel

From Figure 6 above, it can be seen that the kernel biweight curve (blue line) appears to be getting closer to the original data plot. Thus, it can be said that the biweight kernel function curve is better than the previous 3 kernel functions (Gaussian, Epanechnikov, and Triangular kernels).

5) ***Triweight Kernel.*** Using the kernel biweight function, the smoothing curve is obtained as follows:

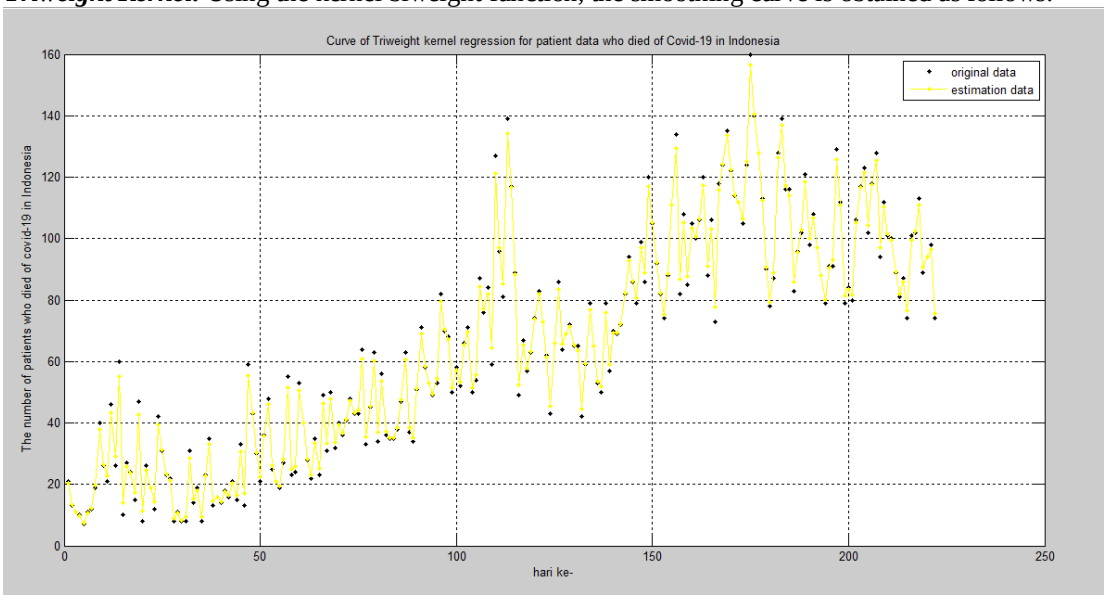

Figure 7. Curve of Triweight Kernel

From Figure 7 above, it can be seen that the kernel triweight curve (yellow line) is very close to the original data plot. The curve with the triwight kernel function is much better than the previous 4 kernel function curve. So by just looking at the curve, we can conclude that the best function is smoothing with a triweight kernel function. However, to determine the best kernel function, we should not only use curves because it only emphasizes the subjectivity of the researcher. Therefore, in determining the best kernel function, the author will compare the error values of 5 kernel functions and choose the smallest error value as the best kernel function.

E. Comparison of kernel methods

The following is a comparison curve between the five kernel functions.

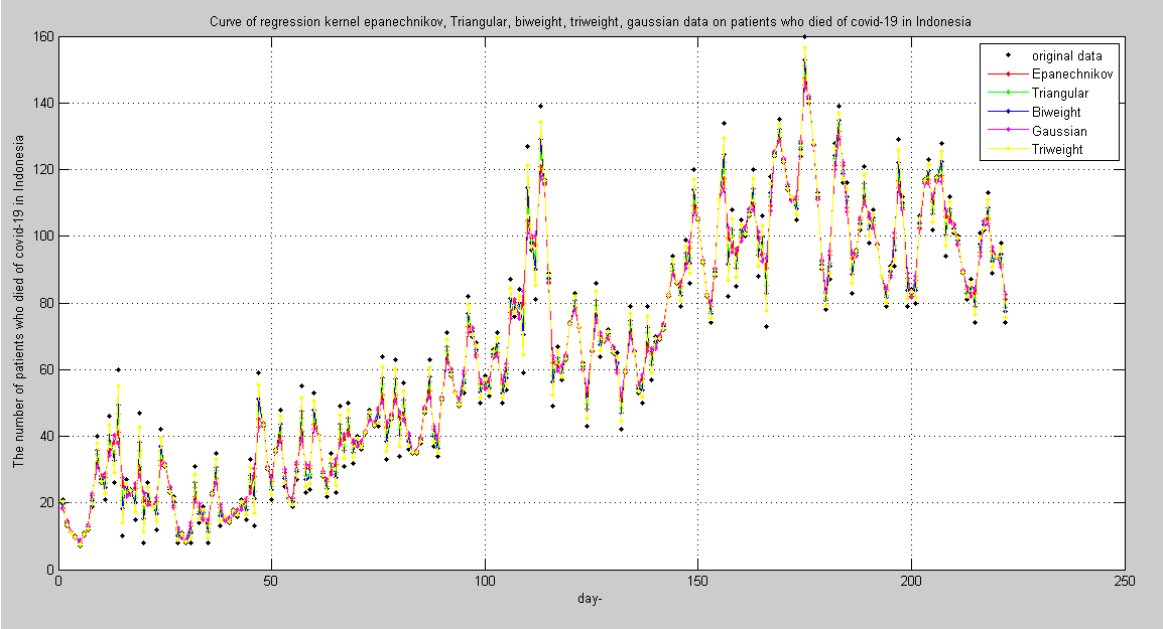

Figure 8. Comparison plot for five kernel functions

From Figure 4, it can be seen that, after those five functions are combined in one plot, there are differences between those five kernel functions. To find the best kernel function, we will look at the smallest MAPE value of the five kernel functions used. The following is a table of MSE, MAD, and MAPE values for the five kernel functions.

Table III. MSE, MAD, and MAPE values of the five kernel functions fungsi

No	Method		MSE	MAD	MAPE
1	Exponential Smoothing	1. SES	268.044	12.583	27.016
		2. DES	382.425	14.469	35.053
		3. Holt-Winters Method	272.271	12.405	25.650
2	Kernel Function	1. Gaussian	68.105	6.318	0.039
		2. Epanechnikov	49.839	5.404	0.034
		3. Triangular	25.578	3.794	0,024
		4. Biweight	15.404	3.003	0.019
		5. Triweight	3.535	1.438	0,009

From table 3 above, we can compare the nonparametric kernel method with the exponential smoothing method and it can be seen that the nonparametric kernel method is better than the exponential smoothing method with a much smaller MAPE value. Then in the kernel method, we will compare 5 kernel functions to find the best kernel function. And it can be seen that the triweight kernel function is the best kernel function because it has the smallest MAPE value of 0.009. So it can be concluded that the triweight kernel function is the kernel function that can most closely follow the original data pattern. After examining the data with five kernel functions, the smallest MAPE value is obtained using triweight kernel. Therefore, the best method to use is the triweight kernel method with the smallest error rate.

V. CONCLUSIONS

Based on the results of the discussion, it can be concluded as follows.

1. The nonparametric method is better than the parametric method, in which all nonparametric methods that have been examined get an error rate that is smaller than the parametric method.
2. From five examined kernel functions, the best method is triweight kernel because it produces the smallest error rate with a MAPE value of 0,9%.

ACKNOWLEDGMENT

Author thanks indonesia's covid-19 task force for allowing to retrieve covid-19 patient data

REFERENCES

[1] WHO, COVID-19 Weekly Epidemiological Update, In world health organization. (2020).

[2] C. Katris, A time series-based statistical approach for outbreak spread forecasting: Application of COVID-19 in Greece, Expert Systems with Applications, 166 (2021), 1-9.

[3] K. Nikolopoulos, S. Punia, A. Schafers, C. Tsinopoulos, & C. Vasilakis, Forecasting and Planing During a Pandemic: COVID-19 Growth Rates, Supply Chain Distuptions, and Governmental Decisions, Europan Journal of Operational Research. 290 (2020), 90-115.

[4] E. Geclll, A. Zlady, & R. D. Szczesniak, Forecasting COVID-19 confirmed cases, deaths and recoveries: Revisting estabilished time series modeling through novel applications for the USA and Italy, Plos One. 16 (1) 1-11.

[5] M. Kalantari, Forecasting COVID-19 pandemic using optimal singular spectrum analysis, Chaos, Solitons and Fractals, 142 (2021), 1-15.

[6] W. Shi, C. Tong, A. Zhang, B. Wang, Z. Shi, Y. Yao, & P. Jia, An extended Weight Kernel Density Estimation model forecasts COVID-19 onset risk and identifies spatiotemporal variations of lockdown effects in China, Communications Biology. 126 (4)(2021), 1-11.

[7] Petropoulos, F., & Makridakis, S. Forecasting the novel coronavirus COVID-19, PLoS ONE, 15(3) (2020), 1–8.

[8] M. Triacca & U. Triacca, Forecasting the number of cinfimed new cases of COVID-19 in italy for the period from 19 May to 2 June 2020, Infection Disease Modeling. 6 (2021), 362-369.

[9] X. Geng, G. G. Katul, F. Gerges, E. Bou-Zeid, H. Nassif, & M. C. Boufadel, A kernel-modulated SIR model for covid-19 contagious spread form country to continent, PNAS. 118(21), 1-9.

[10] R. Eslon, T. M. Davies, I. R. Lake, R. Vivancos, P. B. Blomquist, A. Charlett & G. Dabrera, The spatio-temporal distribution of COVID-19 infection in England between January and June 2020, Cambirdge University Press. 149 (2020) 1-6.

[11] O. Zakary, S. Bidah, M. Rachik, & H. Ferjouchia, Mathematical model to estimate and predict the COVID-19 infections in Morocco: Optimal control strategy, Journal of Applied Mathematics. (2020) 1-13.

[12] H. Pande, Mathematical Modelling of the Number of Transmitted Cases of COVID-19, International Journal of Mathematics Trends and Technology (IJMTT). 66 (4)(2020), 71-74.

[13] A. Oyutunde, Mathematical Modeling and Probability Distribution Function Analyses of Quarantine Control Strategies For Covid-19, International Journal of Mathematics Trends and Technology (IJMTT). 66 (6)(2020), 94-104

[14] A. Oyetunde, & U.Obiaderi, Mathematical Modeling Analysis For COVID-19 With Contact Tracing and Quarantine Control Measures, International Jouranl of Matematics and Technology (IJMTT). 66 (6)(2020), 311-315.

[15] N. Hassan, S. Mahmud, K. F. Nipa & Kamrujjaman, Mathematical Modeling and Covid-19 Forecast in Texas, USA: a prediction model analysis and the probability of disease outbreak, Disaster Med Public Health Prep. (2021), 1-27

[16] I. Puspitasari, Suparti, & Y. Wilandari, Analisis Indeks Harga Saham Gabungan (IHSG) dengan Menggunakan Model Regresi Kernel. Jurnal Gaussian, 1(1)(2012) 93–102.

[17] N. A. K.. Rifai, Pendekatan Regresi Nonparametrik Kernel pada Data Indeks Harga Saham Gabungan. STATISTIKA: Journal of Theoretical Statistics and Its Applications, 19 (1) (2019) 53–61.

[18] Suparti, D. Safitri, I. P. Sari, & Devi, A. R. Analisis Data Inflasi di Indonesia Menggunakan Model Regresi Kernel, Prosiding Seminar Nasional Statistika. (2013) 499-509.

[19] Suparti. (2013). Analisis Data Inflasi di Indonesia Pasca Kenaikan TDL dan BBM Tahun 2013 Menggunakan Model Regresi Kernel, Media Statistika, 6 (2) 103-112.

[20] G. Rubio, H. Pomares, L. J. Herrera, & I. Rojas, Kernel methods applied to time series forecasting. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer-Verlag. (2007) 782–789.

[21] Maliana, I. (2020). UPDATE Corona 10 Desember Pasien Positif Tambah 6. Tribunnews. https://www.tribunnews.com/corona/2020/12/10/update-corona-10-desember-pasien-positif-tambah-6033-sembuh-4530-meninggal-165

[22] S. Makridakis, S. C. Wheelwright, & V. E. McGee, Forecasting Methods and Applications (2 (ed.)). John wiley & Sons. (1983).

[23] R. L. Eubank, Nonparametric Regression and Spline Smoothing (2nd ed.). Marcel Dekker. (1999).

[24] S. Halim, & I. Bisono, Fungsi-fungsi kernel pada metode regresi nonparametrik dan aplikasinya pada, Jurnal Teknik Industri. 8(1) (2006) 73–81.

[25] Hardle, W. Applied Nonparametric Regression. Cambridge University Press. (1992).