Modeling determinant of COVID-19 mortality in Indonesia

Fajar, Muhammad and Magdhalena, Stephanie and Hartini, Sri and Nurfalah, Zelani

Badan Pusat Statistik-Statistics Indonesia

30 September 2020

Modeling Determinant of COVID-19 Mortality in Indonesia

M. Fajar[1*], S. Magdhalena[2], S. Hartini[3], Z. Nurfalah[4]

[1,2,3,4]BPS-Statistics Indonesia

[*]*Corresponding Author: mfajar@bps.go.id, Tel.: +62-89622459926*

Abstract — This study aims to examine the determinants of mortality-related to COVID-19 in Indonesia. Generalized additive models (GAM) was used for modeling the relationship between COVID-19-related deaths and predictor variables. Information used in this study was sourced from Badan Pusat Statistik (BPS Statistics Indonesia), Ministry of Health, and the Indonesian COVID-19 Task Force. The results obtained from GAM are statistically valid. Out of the eight predicting variables used in the analysis, six were significant and two were non-significant at 95 percent confidence interval. The significant variables are GRDP per capita, the proportion of population aged 60 years and over, life expectancy at birth, number of hospitals, number of people with tuberculosis, and number of diabetics. The model can explain the variation of COVID-19-related deaths by 98.5 percent, while the remaining 1.5 percent is attributed to other factors lying outside the model. In summary, this study suggests increasing the number of health facilities, carrying out health development programs, implementing health protocols, and mobility restrictions with prioritizing populations of vulnerable age or those with comorbidities can reduce mortality-related to COVID-19.

Keywords *— :* GAM, Covid-19, Determinant, Modeling Mortality, Indonesia

I. INTRODUCTION

COVID-19, as known as Novel Coronavirus, is a disease caused by coronavirus that can cause mild disorders of the respiratory system, severe lung infection, and even death. The virus, which was first discovered in Wuhan City, China at the end of December 2019, has spread to almost all countries in the world including Indonesia. As of June 12, 2020, the cumulative number of confirmed cases of COVID-19 in Indonesia reached 36,406 cases, with 12,213 patients recovered and 2,048 patients died. Figure 1 presents visually the trend of new confirmed cases and number of patients died due to COVID-19. However, the trend of COVID-19 patients died is far more volatile than the addition of new cases.

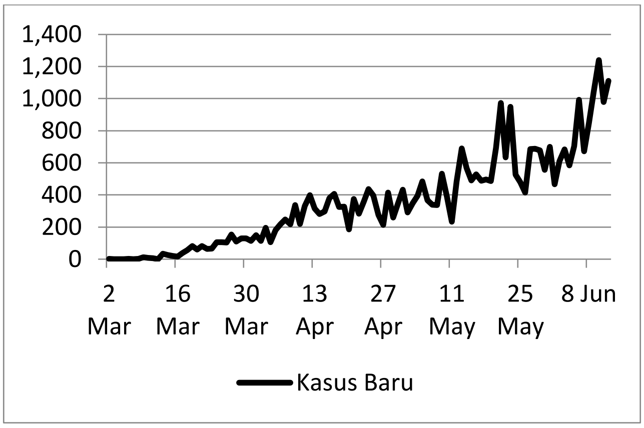

Figure 1. Trend: new cases of COVID-19 as of June 12, 2020

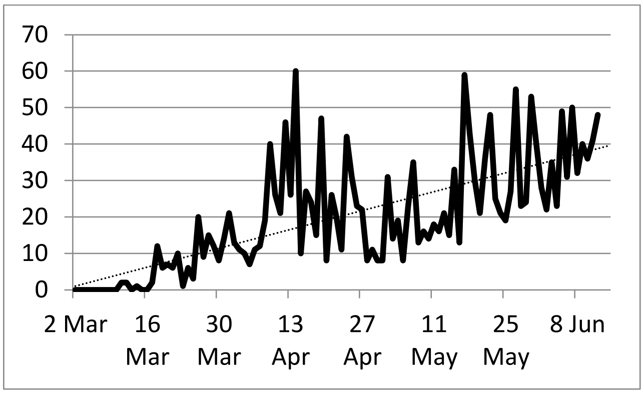

Figure 2. Trend: patients died due to COVID-19 as of June 12, 2020

Based on the trend of COVID-19 deaths in Indonesia, the highest percentage of deaths is in the age group of 60 years and over, reaching 43.5%. This is due to the immunity level of body decreases as people get older and the presence of other chronic diseases, such as hypertension, diabetes, heart, kidney, and lung diseases that suffered by the patient, aggravating the patient's condition[1] [3].

The application of health protocols to the public during this pandemic is very important. The availability of hospitals and adequate medical personnel is no less important. This is because hospitals and medical personnel play an important roles in providing health response to the COVID-19 pandemic and also become the backbone of a country's defense to restrain or overcome the spread of the virus (WHO). The hospital is used as a place to examine, treat, and isolate the COVID-19 patients, while medical personnel have the role of caring, treating, and monitoring patient's condition. Thus, the availability of hospital and medical personnel throughout Indonesia is expected to reduce the COVID-19 death rate, and increase the COVID-19 recovery rate [4].

The variables used in this study are health level, economic level, and education level which are suspected to influence the COVID-19 death rate. This is based on the fact that the level of public health is an inhibiting factor for the spread of COVID-19, implicating the public immunity level against the disease. Thus, if someone in the population is infected with COVID-19, the chances of recovery is very high and the risk of death is reduced. From economic point of view, high income may provide greater access for the population to create a healthy lives, conduct health care, and access health facilities compared to the low-income population. Furthermore, education level reflects public knowledge of the danger of COVID-19 and the application of health protocols. So that when people understand and are able to apply health protocols, it will have implications for the slowdown of the transmission that would affect the low COVID-19 death rate.

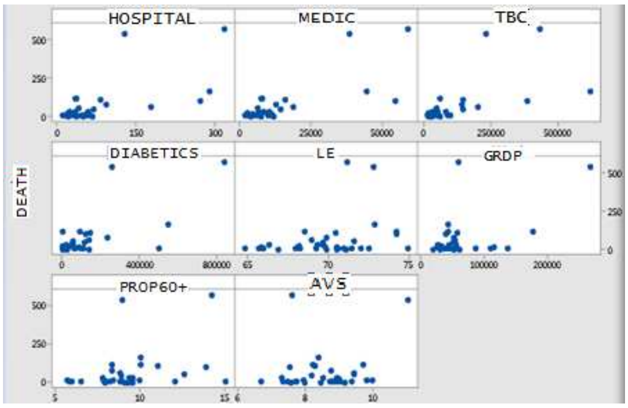

Figure 3. Scatter plot between response variable: number of COVID-19 death with each predictor variable (GRDP per capita, proportion of population aged 60 years and over, life expectancy, average length of school, number of hospitals, number of medical personnel, number of people with tuberculosis and number of diabetics)

Therefore, this study conducted a determinant model of COVID-19 death rate with number of patients died from COVID-19 as a response variable and predictor variables based on the factors mentioned earlier, including (1) demographic variables: proportion of population aged 60 years and over and life expectancy at birth which represents the level of public health, (2) economic variable: GRDP per capita that represents the economic level of the population, (3) health infrastructure variables: number of hospital, and number of medical personnel (doctors and nurses) to the population, (4) chronic disease variables: number of people with tuberculosis and diabetes mellitus (diabetics), and (5) education variables: average length of school that represents the level of population education. The method used in this study is the Generalized Additive Model (GAM) with the following considerations: (1) GAM can accommodate nonlinear or unrelated forms of parametric relationships between response variable and predictor variables, (2) GAM uses a flexible regression method based on data pattern [6], [7], [8], and (3) scatter plots between response variable with each of the predictor variable (mentioned earlier) do not have any specified form of relationship in a particular parametric model (Figure 2)

II. RELATED WORK

The study of found out the fact that the sole determinant strongly related to both coronavirus cases and coronavirus deaths is the level of economic development. That means wealthier countries display larger susceptibility to the disease. Besides that, the population size is also a credible predictor of the registered coronavirus cases per million population, with more populated economies showing greater resistance for being infected by the virus [9].

Another research about the determinant for fatality patients with COVID-19. Patients with comorbidities had a significantly high death risk. But, due to the unavailability of individual patient data, the influence of age could not be concluded because old patients were more likely to have the underlying comorbidities [2]. Reference [10] found a similar result that patients with any comorbidity yielded poorer clinical outcomes than those without. A greater number of comorbidities also correlated with poorer clinical outcomes.

Reference [7] was investigated the COVID-19 and its correlation between temperature and population density using generalized additive model (GAM). The results showed that temperature variation and humidity may not be the important factors affecting COVID-19 mortality.

III. METHODOLOGY

Data Source

Data used in this study were sourced from Badan Pusat Statistik (Statistics Indonesia), Ministry of Health, and Indonesian COVID-19 Task Force. GRDP per capita (GRDP), proportion of population aged 60 years and over (PROP60+), life expectancy at birth (LE), and average length of school (AVS) data were sourced from the Statistics Indonesia. Number of hospitals (HOSPITAL), number of medical personnel, limited to doctors and nurses, (MEDIC), number of people with tuberculosis (TB) and number of diabetics (DIABETICS) data were sourced from the Ministry of Health. Number of deaths (DEATH) by province was sourced from the Indonesian COVID-19 Task Force.

Generalized Additive Model (GAM)

Generalized Additive Model (GAM) is a statistical model in which the relationship between response variable and predictor variables is described as sum of the functions of predictor variables (where the function is a nonlinear function). GAM is formulated as follows:

$$y = \delta_0 + \sum_{k=1}^{p} f_k(x_k) \qquad (1)$$

where: y is the response variable (number of deaths due to COVID-19), δ_0 is intercept, x_k is the k-th predictor variable (in this case includes GRDP per capita, proportion of population aged 60 years and over, life expectancy at birth, average length of school, number of hospitals, number of medical personnel, number of people with tuberculosis and number of diabetics), p is number of predictor variable, and $f_k(.)$is a smoothing function of the k-th predictor variable, in this study, the authors use the identified B-Spline smoothing (P-Splines) as a smoothing function of the predictor variable.

In this study, the authors use the predetermined B-Spline smoothing (P-Splines) as a smoothing function of the predictor variables. B-splines are polynomial functions that have segmented properties at the x -interval formed by knot (piecewise polynomial) which are then estimated locally at these intervals to a certain degree of polynomial [1]. The j-th B-splines with degrees v based on a row of u knots $t_0, \dots, t_u$ for $j = 1, \dots, v + u$ (u express the number of knots) are denoted by the recursive formulation as follows:

$$B_j(x; v) = \frac{x - t_j}{t_{j+v-1} - t_j} B_j(x; v-1) - \frac{x - t_{j+v}}{t_{j+v} - t_{j+1}} B_{j+1}(x; v-1) \qquad (2)$$

with:

$$B_j(x; 0) = \begin{cases} 1 & \text{if } t_j \le x \le t_{j+1} \\ 0 & \end{cases} \qquad (3)$$

Based on equation (2) and (3) obtained that at interval$[t_v, t_{u+v+1}]$, then:

$$\forall\, x\colon \sum_{j=1}^{u+v} B_j(x; v) = 1$$

Based on equation (1) no functional form of $f(.)$ is known, then the function is approximated by the B-splines function, by formulation:

$$f(x) \approx \sum_{j=1}^{u+v=m} \alpha_j B_j(x; v) \qquad (4)$$

So the function f (.) is assumed to be a smooth function that is approximated by a linear combination of B-splines (B-splines function). Based on equation (4), then equation (1) becomes:

$$y_i = \sum_{j=1}^{m} \alpha_j B_j(x; v) + \varepsilon_i \qquad (5)$$

Equation (5) is denoted as a matrix equation:

$$\boldsymbol{Y} = \boldsymbol{B\alpha} + \boldsymbol{\varepsilon} \qquad (6)$$

with:

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \boldsymbol{B} = \begin{bmatrix} B_1(x_1; v) & B_2(x_1; v) & \cdots & B_m(x_1; v) \\ B_1(x_2; v) & B_2(x_2; v) & \cdots & B_m(x_2; v) \\ \vdots & \vdots & \ddots & \vdots \\ B_1(x_n; v) & B_2(x_n; v) & \cdots & B_m(x_n; v) \end{bmatrix}; \boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix}; \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

from equation (6) the least square is obtained:

$$\widehat{\boldsymbol{\alpha}} = (\boldsymbol{B}^t\boldsymbol{B})^{-1}\boldsymbol{B}^t\boldsymbol{Y} \qquad (7)$$

based on equation (7), then the estimator for the regression model in equation (5) is:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{B}(\boldsymbol{B}^t\boldsymbol{B})^{-1}\boldsymbol{B}^t\boldsymbol{Y} = \boldsymbol{AY} \qquad (8)$$

The use of too many knots makes the B-splines function curve tend to be over fit so penalties are needed on the adjacent coefficients of the B-splines [5]. In general, the objective functions of B-splines regression given a penalty or P-Splines are as follows:

$$\widehat{\boldsymbol{\alpha}} = \operatorname{argmin}_\alpha \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} \alpha_j B_j(x; v) \right)^2 + \lambda \int_{x_{min}}^{x_{max}} \left(\sum_{j=1}^{m} \alpha_j B_j''(x; v) \right)^2 dx$$

where $\lambda > 0$ is the smoothing parameter and $B_j''(x; v)$ is the second derivative of $B_j(x; v)$. Here are some derivatives of B-splines and B-splines functions that are useful in the estimation process:

- First derivative of B-splines

$$\frac{\partial B_j(x; v)}{\partial x} = \frac{v - t_j}{t_{j+v-1} - t_j} B_j(x; v-1) - \frac{v-1}{t_{j+v} - t_{j+1}} B_{j+1}(x; v-1), \text{ for } v > 1$$

$$\frac{\partial B_j(x; v)}{\partial x} = 0, \text{ for } v = 1$$

- Second derivative of B-splines

$$\begin{aligned} \frac{\partial^2 B_j(x; v)}{\partial x^2} = (v-1)(v-2)\Bigg(& \frac{1}{(t_{j+v-1} - t_j)(t_{j+v-2} - t_j)} B_j(x; v-2) \\ & - \left(\frac{1}{(t_{j+v-1} - t_j)} + \frac{1}{(t_{j+v} - t_{j+1})} \right) \frac{1}{(t_{j+v-1} - t_{j+1})} B_{j+1}(x; v-2) \\ & + \frac{1}{(t_{j+v} - t_{j+1})(t_{j+v} - t_{j+2})} B_{j+2}(x; v-2) \Bigg) \end{aligned}$$

- First derivative of B-splines function

$$\frac{\partial f(x)}{\partial x} = (v-1) \sum_{j=2}^{m} \frac{\alpha_j - \alpha_{j-1}}{t_{j+v-1} - t_j} B_j(x; v-1)$$

- Second derivative of B-splines function

$$\frac{\partial^2 f(x)}{\partial x^2} = (v-1)(v-2) \left(\sum_{j=3}^{m} \frac{\frac{\alpha_j - \alpha_{j-1}}{t_{j+v-1} - t_j} - \frac{\alpha_{j-1} - \alpha_{j-2}}{t_{j+v-2} - t_{j-1}}}{t_{j+v-2} - t_j} B_j(x; v-2) \right)$$

IV. RESULTS AND DISCUSSION

Based on Table 1, it is obtained that the variables that have a negative and significant relationship to COVID-19 deaths are life expectancy and number of hospitals, which means that when life expectancy and the number of hospitals are high, the number of COVID-19 deaths tends to be nonlinearly low. This indicates that with high level of public health, the risk of death due to COVID-19 tends to be low and as more hospitals treat COVID-19 patients, the risk of death due to COVID-19 is also low.

The predictor variables that have a positive and significant relationship to COVID-19 deaths are the proportions of the population aged 60 years and over, people with tuberculosis and diabetics. The more population aged 60 years and over in an area, the more COVID-19 patients tend to be there, consequently the risk of death is higher because body immunity at the age of 60 years and over is reduced compared to young human body, so it tends to increase the COVID-19 death. Likewise, when the number of people with tuberculosis and diabetics in an area is high, if the patients is also infected with COVID-19, then the risk of death due to COVID-19 is also high.

GRDP per capita variable, which is theoretically should have a negative and significant relationship to COVID-19 death, empirically related positively and significantly in this case. Authors suspect that the GRDP is a reflection of the economic transactions of all business actors in a certain area which also reflects the population's mobility indirectly. When the GRDP is high, then indirectly, mobility tends to be high so that the spread of COVID-19 is also increasingly widespread that implicates the risk of people getting affected by COVID-19 is also high so in the end COVID-19 deaths are also increasing.

Table 1. GAM Estimation Results

	estimation	**p-value**	**information**
δ_0	5.823×10^2	0.000	Significant at alpha level of 5 %
$f(\text{GRDP})$	1.382×10^{-3}	0.000	Significant at alpha level of 5 %
$f(\text{PROP60+})$	2.438×10	0.000	Significant at alpha level of 5 %
$f(\text{LE})$	-1.372×10	0.000	Significant at alpha level of 5 %
$f(\text{AVS})$	8.104	0.061	Not Significant at alpha level of 5 %
$f(\text{HOSPITAL})$	-7.141×10^{-1}	0.003	Significant at alpha level of 5 %
$f(\text{MEDIC})$	1.764×10^{-3}	0.081	Not Significant at alpha level of 5 %
$f(\text{TB})$	4.673×10^{-4}	0.000	Significant at alpha level of 5 %
$f(\text{DIABETICS})$	3.285×10^{-4}	0.000	Significant at alpha level of 5 %
$pseudo\ R^2$	98.472%		
AIC	320.701		

Then, the average length of school and medical personnel variables were not significant for COVID-19 deaths. The authors suspect that higher education is not enough to make population adhere to the COVID-19 health protocols, thus implicating in increasing COVID-19 death. From the medical point of view, the increase in the number of medical personnel cannot keep up with the growth of COVID-19 cases. Thus, when COVID-19 death increases on the other hand the number of medical personnel tends to be constant. The estimation results of this determinant model obtained a coefficient of determination equal to 98.472%, which means that this model is able to explain variations in the COVID-19 death variable by 98.472%, while the remaining 1.528% is explained by other factors outside the estimated model.

V. CONCLUSION & RECOMMENDATION

Based on the result, it can be concluded that the COVID-19 death determinant model estimated using GAM is valid. This is based on: (1) out of the eight predictor variables proposed by the authors, there are six significant predictor variables (GRDP per capita, proportion of population aged 60 years and over, life expectancy at birth, number of hospitals, number of people with tuberculosis, and number of diabetics) and two non-significant predictor variables at the alpha level of 5% (average length of school and limited to doctors and nurses), and (2) this model can explain the variation of COVID-19 death variable by 98.472 %, while the remaining 1.528% is explained by other factors outside the model.

GDRP contributed positively significant to COVID-19 death shows the high level of economic transactions that occur. So, health protocols have to be implemented in conducting economic transactions and optimizing online economic activities to avoid face-to-face transactions. Besides, mobility restrictions have to be implemented and prioritize for populations of vulnerable age or those with comorbidities. The government has to increase the number of health facilities to treat all patients with COVID-19 properly in health facilities so there is no overload of COVID-19 patients. Life expectancy at birth has always to be better by carrying out health development programs and other social programs contain environmental health, adequate nutrition and calories so it can reduce the death rate by COVID-19.

ACKNOWLEDGMENT

Give thanks to Allah, and thanks to my institutions Statistics-Indonesia.

REFERENCES

[1] C. de Boor, "A Practical Guide to Splines, Revised ed," *Springer*, *Berlin*, pp: 109-128, **2001**.
[2] G. Deng, M. Yin, X. Chen, and F. Zeng, "Clinical determinants for fatality of 44,672 patients with COVID-19," *Critical Care*, *24*(1), 1–3, **2020**.
[3] G. Onder, G. Rezza and S. Bruseferro, "Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy,",*JAMA*. Published online March 23, **2020**.
[4] H. Sahoo, C. Mandal, S. Mishra, and S. Banerjee, "Burden of COVID-19 pandemic in India: Perspective from Health Infrastructure," **2020**.
[5] P. Eilers and B. Marx, "Flexible Smoothing with B-Splines and Penalties,",*Statistical Science*, vol.11, no.2, pp. 89-102, **1996**.
[6] S.N. Wood, "Generalized Additive Models: An Introduction with R 2nd edition," *Chapman & Hall/CRC*, pp. 161-191, **2017**
[7] S. Dan, D. Sharma, M. Mandal, D. Sharma, "Incidence of COVID-19 and Its Correlation between Temperature and Population Density," *International Journal of Scientific Research in Biological Sciences*, Vol.7, Issue.2, pp.134-141, **2020**.
[8] T.J. Hastie and R.J. Tibshirani, "Generalized Additive Models," *Chapman & Hall*, pp: 136-166, **1990**.
[9] V. Stojkoski, Z. Utkovski, P. Jolakoski, D. Tevdovski, and L. Kocarev, "The Socio-Economic Determinants of the Coronavirus Disease (COVID-19) Pandemic," *SSRN Electronic Journal*, *April*, 1–22, **2020**.
[10] W.J. Guan, W.H. Liang, Y. Zhao, H.R. Liang, Z.S. Chen, Y.M. Li, X.Q. Liu, R.C. Chen, C.L. Tang, T. Wang, C.Q. Ou, L. Li, P.Y. Chen, L. Sang, W. Wang, J.F. Li, C.C. Li, L.M. Ou, B. Cheng, S. Xiong, Z.Y. Ni, J. Xiang, Y. Hu, L. Liu, H. Shan, C.L. Lei, Y.X. Peng, L. Wei, Y. Liu, Y.H. Hu, P. Peng, J.M. Wang, J.Y. Liu, Z. Chen, G. Li, Z.J. Zheng, S.Q. Qiu, J. Luo, C.J. Ye, S.Y. Zhu, L.L. Cheng, F. Ye, S.Y. Li, J.P. Zheng, N.F. Zhang, N.S. Zhong, and J.X. He, "Comorbidity and its impact on 1,590 patients with Covid-19 in China: A nationwide analysis. *European Respiratory Journal*, *55*(5), **2020**.

AUTHORS PROFILE

Muhammad Fajar is currently the Statistician of the BPS-Statistics Indonesia Office. In March 2017, he published paper with title; Spectral analysis and markov switching model of Indonesia business; in AIP Conference Proceedings, in November 2017 he published paper with title; PEMODELAN KURVA ENGEL SEDERHANA INDONESIA (Pendekatan Bayesian Quantile Regression); in National Statistics Seminar of Padjadjaran University. In February 2019 he published paper with title ; The Estimation of Production Function and Technical Efficiency Shallot Farming; in jurnal matematika Mantik of UIN Sunan Ampel. In October 2019 he published paper with title; An Application of Hybrid Forecasting Singular Spectrum Analysis – Extreme Learning Machine Method in Foreign Tourists Forecasting; in jurnal matematika Mantik of UIN Sunan Ampel.

Stephani Magdhalena is currently working as a Statistician at BPS-Statistics Indonesia of Kutai Timur Regency. She graduated from Institute of Statistics in 2015. In October 2019, she published a paper titled "Twitter Sentimen Gojek Indonesia dan Grab Indonesia" at National Seminar on Mathematics, Statistics, and Its Applications which held by Mulawarman University of Samarinda as a co-author.

Sri Hartini Rachmad is a researcher on Social-Economic Demography and Applied Statistics at BPS Statistics Indonesia; lecturer and research student adviser at STIS Polytechnique Statistics-Jakarta; and columnist at financial and entrepreneurship national magazine. She graduated from Statistics Institute- Jakarta and attained postgraduate degree from the Australian National University, Canberra-Australia. Her research interests are international and population development: climate change and vulnerability, disaster and water management, data assurance & quality, MDGs-SDGs, poverty, employment, migration, education, health, children, youth, gender, ageing and family planning. Sri Hartini is a prolific researcher with numerous publications in academic journals and in association with organizations such as Oxfam International, UNDP, WHO, UNICEF and others. Contact information: Email: shrachmat2@gmail.com

Zelani Nurfalah is currently a statistician at BPS-Statistics Indonesia Office. Previously, he was a statistician at BPS-Statistics Indonesia of Tidore Kepulauan Municipality. He graduated from the Institute of Statistics in 2016. His research interests are poverty, export-import, and forecasting method comparisons.