International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol. 7 Issue 11, November - 2023, Pages: 32-35

Modeling the Open Unemployment Rate in Indonesia Using Multivariate Adaptive Regression Spline

Andika Priyatama[1], Ardi Kurniawan[2], Elly Ana[3], M. Fariz Fadillah M[4], Sediono[5]

[1,2,3,4,5]Department of Mathematics, Faculty of Science and Technology, Campus C, Airlangga University, Surabaya, Indonesia
*Corresponding Author: [2]ardi-k@fst.unair.ac.id

Abstract: *The Open Unemployment Rate is the percentage of the number of unemployed to the total labor force. The open unemployment rate is an indicator used to measure labor that is not absorbed by the labor market and illustrates the underutilization of labor supply. The unemployment rate for the Indonesian workforce aged 15-24 years in 2021 is in second place after Brunei Darussalam in Southeast Asia. The aim of this research focuses on describing and modeling the level of open unemployment rate in Indonesia, and interpreting the best model results obtained. The method used is a method with a nonparametric regression approach, namely the Multivariate Adaptive Regression Spline. The results showed that the best model obtained was in a combination of 14 basis functions, 2 maximum interactions, and 1 minimum observation between knots. From this model, the predictor variables that have the most influence on the response variable in order based on the level of variable importance are education level, population density, Foreign Direct Investment (FDI), percentage of poor population, high school APK, economic growth rate, and gini ratio. In addition, there is also an interaction between two variables, namely education level with population density and FDI with population density. The best MARS model for open unemployment rate data in Indonesia produces a Generalized Cross Validation value of 1.915, R2 of 0.740, and Mean Square Error of 0.979.*

Keywords—Modeling, Open Unemployment Rate, Spline Regression, Multivariate Adaptive Regression Spline.

1. Introduction

Human resources bring the potential for economic development to a country and serve as an asset for the nation itself. When human resources are abundant and high in quality, the management of other resources possessed by a country will be optimized. A high population density should be accompanied by a high employment absorption rate. This is intended to ensure that the growing workforce each year can find employment due to a sufficient employment absorption. However, a high population density cannot guarantee proportional employment absorption, leading to a high level of unemployment. According to data from the International Labour Organization (ILO) collected by the World Bank, the unemployment rate for the Indonesian labor force aged 15-24 reached 16% in 2021, ranking second in Southeast Asia after Brunei Darussalam at 23,4%. Unemployment can be categorized based on its characteristics into open unemployment, hidden unemployment, seasonal unemployment, and underemployment. Open unemployment occurs due to the lower job opportunities compared to the growth of the labor force, resulting in many individuals not finding employment. The Open Unemployment Rate is the percentage of unemployed individuals in the labor force [1]. The higher this indicator, the more unused labor resources there are.

One method that can be used to analyze the open unemployment rate is the Multivariate Adaptive Regression Spline (MARS) method. This method is an approach to multivariate regression which was first introduced by Friedman in 1991. The advantage of this method is that it does not only see the effect of the predictor variable on the response variable, but also can see the interactions that occur between predictor variables [7]. In addition, the advantage of this method is that it can accommodate more than one predictor variable. The predictor variables commonly used are $3 \leq n \leq 20$ [5].

Based on the description above, the main issue discussed is regarding the open unemployment rate. These problems were analyzed using the MARS method. The results of this study are expected to be useful as material for consideration for taking innovative and progressive steps to support decreasing the open unemployment rate evenly in provinces in Indonesia so that global action known as the Sustainable Development Goals (SDGs) can be achieved and get the title as a developed country.

2. Literature Review

2.1 Open Unemployment Rate

Unemployment is a recurrent issue faced by developing countries worldwide, including Indonesia. Open unemployment refers to the working-age population who are not employed and are actively seeking employment, preparing to start a business, those who are not actively seeking employment due to perceived difficulty in finding a job, or individuals who have secured a job but have not started working [6]. High unemployment levels can diminish the welfare of the population as their income decreases. This declining welfare can lead to the emergence of a new issue, it is called poverty [3]. The open unemployment rate is calculated by comparing the number of unemployed individuals to the labor force [6]. In this case, the government must ensure the optimal utilization of the workforce if it intends to succeed in development, as the increasing number of underemployed labor force will become a burden and hindrance to the economy of a country.

2.2 Multivariate Adaptive Regression Spline (MARS)

The Multivariate Adaptive Regression Spline (MARS) method was first introduced by Friedman in 1991 [2]. The MARS method is a nonparametric multivariate regression approach capable of capturing the non-linearity effects of a model and examining the presence or absence of interactions among predictor variables [7]. The general MARS model, as outlined by Friedman, is as follows [2].

$$f(x) = a_0 + \sum_{m=1}^{M} a_m \prod_{k=1}^{K_m} \left[S_{km} \left(x_{v(k,m)} - t_{km} \right) \right] + \varepsilon$$

(1

)

With:

a_0 : The coefficient a_0 is a constant
a_m : The coefficient of the *m*th basis function
M : Maximum number of basis function
K_m : The number of interactions of the *m*th basis function
S_{km} : ±1
$x_{v(k,m)}$: Independent variable
t_{km} : Knots value of independent variable

The maximum number of basis functions is included in the model in the forward building phase [8]. After the forward phase, a typical overfitting model is produced, and so a backward deletion phase is engaged. In the backward phase, the model is simplified by deleting one least important basis function (i.e. deletion of which reduces training error the least) at a time. At the end of the backward phase, from those "best" models of each size, the one with the lowest Generalized Cross-Validation (GCV) is selected and outputted as the final one. Below is attached the GCV formulas [4].

$$GCV(M) = \frac{MSE}{\left[1 - \frac{C(\widehat{M})}{N}\right]^2}$$

(2

)

With:

MSE : Mean squared error
N : Amount of observations
$C(\widehat{M})$: $C(M) + dM$
$C(M)$: *Trace* $[\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T] + 1$; **B** is matrix of *M*th basis function
d : Value when each basic function reaches its optimal value (2≤d≤4)

3. Methodology

3.1 Data and Variable

The data used in this study is secondary data regarding the open unemployment rate and the factors that are thought to influence it. The data is taken from the official website of the Badan Pusat Statistik of the Republic of Indonesia. The units and observation data used in this study are 34 provinces in Indonesia in 2021. The research variables used in this study are the Percentage of Poor Population, High School APK, Economic Growth Rate, Gini Ratio, Education Level at high school level, Population Density, and Foreign Direct Investment (FDI). The data scale on all research variables is the ratio data scale.

4. Result

4.1 Descriptive Statistics

Descriptive statistics are the initial stage of data exploration used to describe research objects in general so as to produce useful information. Descriptive statistics can be presented in several forms. However, it needs to be adjusted to the needs of researchers by considering the usefulness of each form of data presentation.

Table 1. Descriptive Statistics

Research Variable	Minimum	Maximum
Open Unemployment Rate	3,01%	9,91%
Percentage of Poor Population	4,56%	27,38%
High School APK	75,05%	97,25%
Economic Growth Rate	-2,47%	16,4%
Gini Ratio	0,247	0,436
Education Level	32,95%	90,12%
Population Density	15.978 people/km^2	9 people/km^2
Foreign Direct Investmenr	5.217,7 million USD	5,9 million USD

To find out the pattern of distribution of research data as well as an early detection of the use of methods with a nonparametric approach, it can be seen through the scatterplot between the response variable and the predictor variable.

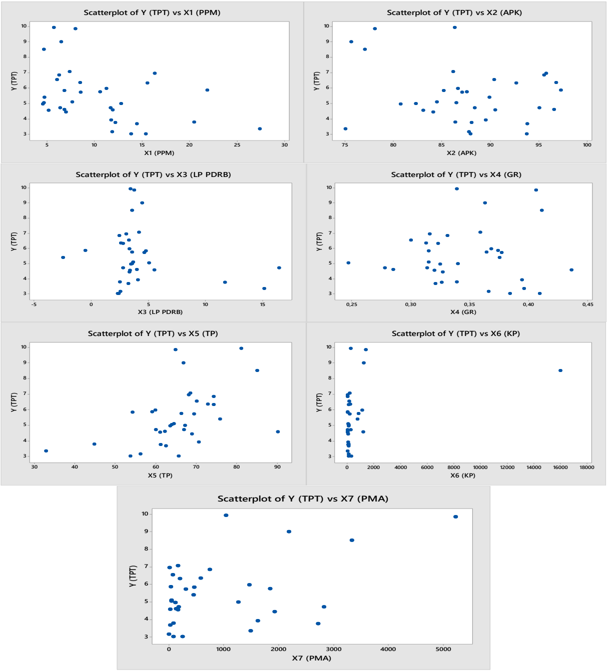

Fig 1. Scatterplot

Based on Figure 1, it can be seen that there is no particular pattern of data distribution (trend) for each predictor variable with the response variable. Therefore, a method with a nonparametric approach can be used.

4.2 Model Estimation

The basis functions used are 14 to 28, the maximum interactions are 1, 2, 3, and the minimum observations between knots are 0, 1, 2, and 3. This combination can determine the best model with minimum Generalized Cross Validation (GCV) criteria, maximum r-square, and minimum Mean Square Error (MSE). The best model is obtained from a combination of function base 14, maximum interaction 2, and minimum observation 1 with a GCV value is 1,915, r-square is 0,74, and MSE is 0.979. The model obtained is as follows.

$$\hat{Y} = 4{,}577 - 0{,}003BF_2 + 0{,}305375 \times 10^{-3}BF_3 + 0{,}166053 \times 10^{-5}BF_8 + 0{,}001BF_9$$

(3

)

4.3 Mars Model Significance Test

4.3.1 Simultaneous Test of the Coefficient of Function of the MARS Model Basis

Hypothesis:
$H_0: a_1 = a_4 = a_6 = a_8 = a_9 = 0$
H_1: There is at least one $a_m \neq 0; m = 2{,}3{,}8{,}9$

Based on the results of data processing, obtained F is 20,623 > $F_{(0{,}05{,}4{,}29)} = 2{,}701$. In addition, the resulting *p-value* is $0{,}388083 \times 10^{-7}$, which value is less than the level of significance ($\alpha = 0{,}05$), so that the decision is rejected H_0 and the conclusion was obtained that there is at least one $a_m \neq 0$ ($m = 2{,}3{,}8{,}9$).

This can be interpreted that the model obtained is appropriate and shows a relationship between the coefficients of the basis function and the response variable

4.3.2 Partial Coefficient Test of the MARS Model Base Function

Hypothesis:
$H_0: a_m = 0, m = 2{,}3{,}8{,}9$
$H_1: a_m \neq 0, m = 2{,}3{,}8{,}9$

Based on the results of data processing, obtained t dan *p-value* of each basis function as follows.

Table 2. Partial Coefficient Test of the MARS Model Base Function

Parameter	***T-Ratio***	***P-value***
BF_2	-5,781	$0{,}290742 \times 10^{-5}$
BF_3	6,306	$0{,}690954 \times 10^{-6}$
BF_8	4,673	$0{,}628411 \times 10^{-4}$
BF_9	4,827	$0{,}410028 \times 10^{-4}$

Based on Table 2, obtained t of each basis function in the model > $t_{(0{,}025{,}30)} = 2{,}042$. In addition, *p-value* of each basis function in the model is less than the significance level ($\alpha = 0{,}05$), so that the decision is rejected H_0 and concluded that $a_m \neq 0$ ($m = 2{,}3{,}8{,}9$).

This means that the model obtained shows a relationship between the coefficients of the basis function and the response variable.

4.3.3 Variable Importance Level

The level of importance of the variable is used to sort the predictor variables that affect the response variable. The level of importance of variables in modeling the open unemployment rate data is as follows.

Table 3. Variable Importance Level

Variable	**Variable Name**	**Importance Level**	**GCV Reduction**
X_5	Education Level	100%	3,636
X_6	Population Density	89,182%	3,284
X_7	Foreign Direct Investment	52,677%	2,392
X_1	Percentage of Poor People	0%	1,915
X_2	High School APK	0%	1,915
X_3	Economic Growth Rate	0%	1,915
X_4	Gini Ratio	0%	1,915

Based on Table 3, it can be seen that the predictor variable that has the most influence on the response variable is the education level variable with an interest level of 100%. In addition, the education level variable can reduce the Generalized Cross Validation (GCV) value is 3,636, if this variable is included in the model.

Furthermore, the predictor variable that influences the response variable based on the order of importance is the population density (X_6) and FDI (X_7). Meanwhile, the predictor variable that has an interest level of 0% is percentage of poor population (X_1), high school APK (X_2), economic growth rate (X_3), and gini ratio (X_4).

4.4 Discussion

After getting the best model and testing the significant variables, as well as the assumptions on the residuals, the response variable and the estimated results can be plotted to compare the two values.

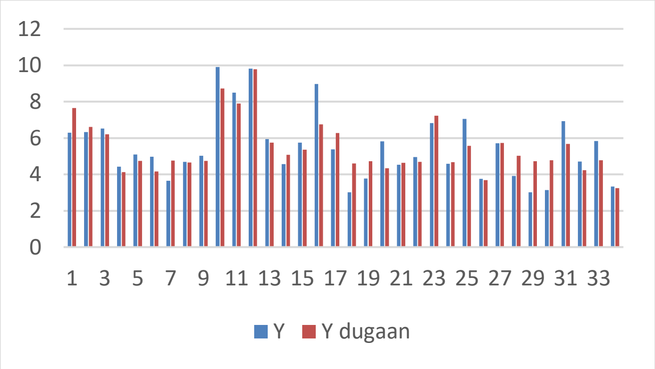

Fig 2. Plot Y with $\hat{Y}$

Based on Figure 2, it can be seen that the estimated value $(\hat{Y})$ is close to the value from factual data, namely the human capital index (Y). In addition, based on the best model that has been obtained, the results of the interpretation are as follows.

1. (BF_2); The interpretation of the value of the basis function two (BF_2) with a coefficient is -0,003 means that every one unit increase in BF_2 will decrease the open unemployment rate by 0,003 with the basis functions BF_3, BF_8, and BF_9 considered a constant. In addition, another meaning is that the population density variable will make a significant contribution to provinces that have a population density value of less than 1.248 people/km^2, namely in the form of a decrease in the open unemployment rate.
2. (BF_3); The interpretation of the value of the basis function three (BF_3) with a coefficient is $0{,}305375 \times 10^{-3}$means that every one unit increase in BF_3 will increase the open unemployment rate by $0{,}305375 \times 10^{-3}$ with the basis functions BF_2, BF_8, and BF_9 considered a constant. In addition, another meaning is that the education level and population density variable will make a significant contribution to provinces that have a education level more than 65,71% followed by the population density value of less than 1.248 people/km^2, namely in the form of an increase in the open unemployment rate.
3. (BF_8); The interpretation of the value of the basis function eight (BF_8) with a coefficient is $0{,}166053 \times 10^{-5}$ means that every one unit increase in BF_8 will increase the open unemployment rate by $0{,}166053 \times 10^{-5}$ with the basis functions BF_2, BF_3, and BF_9 considered a constant. In addition, another meaning is that the FDI and population density variable will make a significant contribution to provinces that have a FDI less than 1.921,4 million USD followed by the population density value of less than 1.248 people/km^2, namely in the form of an increase in the open unemployment rate.
4. (BF_9); The interpretation of the value of the basis function nine (BF_9) with a coefficient is 0,001 means that every one unit increase in BF_9 will increase the open unemployment rate by 0,001 with the basis functions BF_2, BF_3, and BF_8 considered a constant. In addition, another meaning is that the FDI variable will make a significant contribution to provinces that have a FDI value of more than 5,9 million USD, namely in the form of a increase in the open unemployment rate.

5. Conclusion

Based on the results of the analysis and discussion, the conclusions obtained from this study are as follows.

1. Descriptive statistics for the open unemployment rate variable with the highest value being 9,91% and the lowest being 3,01%. Furthermore, the scatterplot of each predictor variable on the response variable shows that there is no particular pattern of data distribution (trend).
2. The best model obtained with a trial-and-error system using the Multivariate Adaptive Regression Spline (MARS) method is a combination of 14 basis functions, 2 maximum interactions, and 1 minimum observations between knots. The model produces a Generalized Cross Validation (GCV) value of 1,915, r-square is 74%, and Mean Square Error (MSE) is 0,979. The following is the best model obtained by the MARS method.

   $$\hat{Y} = 4{,}577 - 0{,}003BF_2 + 0{,}305375 \times 10^{-3}BF_3 + 0{,}166053 \times 10^{-5}BF_8 + 0{,}001BF_9$$

3. Based on the best model that has been obtained, it is found that there are three predictor variables that have an importance level of more than 0%, namely the education level variable with an importance level is 100%, the population density variable with an importance level is 89,182%, and FDI variable with an importance level is 52,677%. In addition, it is known that there is an interaction between the two variables in the best model, namely BF_3 and BF_8.

6. References

[1] Badan Pusat Statistik. 2021. *Open Unemployment Rate 2021*. Jakarta: Badan Pusat Statistik.

[2] Friedman, J. H. 1991. Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1), 1-67.

[3] Mahmud, A., & Pasaribu, E. 2021. Spatial modeling to analyze factors influencing the open unemployment rate in Bangka Belitung 2018. *Engineering, Mathematics and Computer Science (EMACS) Journal*, 3(2), 47-58.

[4] Pintowati, W., Otok, B. W. 2012 Poverty Modeling in Jawa Timur Province Using a Multivariate Adaptive Approach, *Jurnal Sains dan Seni ITS*, 1(1), 283-288.

[5] Sita, E. D. A. A., & Otok, B. W. (2014). Multivariate Adaptive Regression Splines (MARS) Approach to Modeling the Poor Population in Indonesia 2008-2012. *Prosiding Seminar Nasional Matematika, Universitas Jember*, 175-191.

[6] Utama, S. S., Suparti, S., & Rahmawati, R. (2015). Modeling Open Unemployment Rates in Jawa Tengah Using Spline Regression. *Jurnal Gaussian*, *4*(1), 113-122.

[7] Wibowo, A. 2019. Multivariate Adaptive Regression Splines Modeling for Household Food Security in Central Borneo Province 2017, *Global Science Education Journal*, 1(1), 39-47.

[8] Zhang, W., Wu, C., Li, Y., Wang, L., & Samui, P. (2021). Assessment of pile drivability using random forest regression and multivariate adaptive regression splines. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, *15*(1), 27-40.