Vol. 09, No. 02: 34 — 39
August 2020
DOI: 10.22487/25411969.2019.v9.i2.15202

Natural Science
Journal of Science and Technology

## Original article

# Spline Regression Analysis to Modelling The Open Unemployment Rate in Sulawesi

Selvia Anggun Wahyuni, Ratnawati, Indriyani, Mohammad Fajri*

Department of Statistic, Faculty of Mathematics and Natural Sciences, Tadulako University, Jalan Raya Soekarno–Hatta, Palu 94117, Central Sulawesi, Indonesia.

**Keywords:** Spline Regression, Open Unemployment Rate, Optimum Knots

Article history:
Received 12 July 2020
Accepted 19 August 2020
Published 31 August 2020

* Coresponding Author :
m.fajri@untad.ac.id

**Abstract**

Unemployment is a very complex problem because it affects and influenced by several factors that interact with each other following a pattern that is not always easy to understand. When unemployment was not immediately addressed, it can cause social vulnerability and potentially lead to poverty. This research has used spline regression method in modeling the 2018 Sulawesi open unemployment rate. The results showed that the best spline model were from the optimum knots point with a combination of knots 3,3,1,1,3,3. This model has the minimum GCV value 1,97 with $R^2$ 77,67%. All variables were significantly affect the open unemployment rate.

## INTRODUCTION

Unemployment is a very complex problem because it affects and influenced by several factors that interact with each other following a pattern that is not always easy to understand. when unemployment is not immediately addressed then it can cause social vulnerability and potentially lead to poverty (BPS, 2018). Unemployment is a waste of resources. Due to the high unemployment, the burden of living becomes complex (Mankiw, 2006). The unemployment rate is also the key to economic performance. The unemployment rate shows the percentage of the labor force that is not working. A decrease in the unemployment rate is a good indicator of the economy. This is because companies that add workers are considered successful in increasing production and sales. Nevertheless, the unemployment rate and the number of people working can rise at the same time (Hotchkiss & Kaufman, 2006).

An example in 2017, TPT in Central Sulawesi was 3.81 percent and decreased to 3.43 percent in 2018. The number of Central Sulawesi Workforce in August 2018 was 1,502,972 people, an increase of 74.4 thousand people compared to August 2017. The explanation is the growth of new workers is greater than the provision of new jobs. This means, although the TPT trend is declining, the unemployment problem is still important in Central Sulawesi.

Regression analysis is a measuring tool that used to measure the presence or absence of correlation between variables. A regression term that means forecast or estimation. In a regression analysis not all cases of parametric patterned data such as linear, quadratic or cubic patterns, so the need for other regression approaches such as the nonparametric and semiparametric regression approaches According to Hardle (1990), if the data pattern tends to follow a linear, quadratic or cubic model then the regression approach that fits the data pattern is the parametric regression approach, whereas if the data pattern is not the regression curve form, the appropriate regression approach is the nonparametric regression approach (Budiantara, 2010).

The problems in nonparametric regression can be overcome by several methods that have been developed at this time, the kernel method or commonly called the normal polynomial method and the method that uses the spline function. The spline function has high flexibility and able to handle data that changes in certain subintervals because the spline function is a segmented form of polynomial. This spline function is better than the kernel method as shown by Aydin (2007). Spline has the advantage in overcoming data patterns that show sharp up/down with the help of knots and the resulting curve is relatively smooth (Hardle, 1990) and can describe changes (piecewise) behavior patterns of certain sub-interval functions (Eubank, 1999).

A regression model that uses a spline function is commonly known as a spline regression. The form of the spline estimator is strongly influenced by the value of the smoothing parameter (Budiantara, 2000). The shape of the spline estimator is also influenced by the location and the number of knots. Eubank (1988) reported that the optimal selection in spline regression is essentially the choice of the location of the knot point. Determination of optimal knot points to choose the best spline regression model based on GCV values. The problem that arises is how to determine the number of knots and the location of the knots and also how to choose the best spline regression model using GCV criteria. Basically, choosing the best spline criteria or model can be done in 2 ways, by looking at the MSE value and its GCV value, the best spline model is the smallest MSE and GCV values compared to the spline model with other knots.

Research on spline regression has been carried out, for example Anwar (2014) use spline regression for Modeling Open Unemployment Rates in West and Sari (2012) modelling Unemployment case in East java Using Multivariable Spline Regression. In this research, Open Unemployment Rate Modeling in Central Sulawesi will be used using spline regression approach. The spline regression model is considered suitable for modeling the open unemployment rate in Central Sulawesi because the spline regression model can be used to explain the relationship between the response variable and the predictor variable whose regression curve is unknown or if the shape of the pattern changes at each particular sub interval. With the hope that the model obtained can be taken into consideration in developing intervention steps and decisions in monetary policies that are more effective and efficient.

## MATERIALS AND METHODS

The data used in this study are secondary data obtained from the Central Statistics Agency (BPS) in the Central Sulawesi publication in 2019. The objects in this study consist of 13 districts / cities in Central Sulawesi Province. The variables used are:

a. Open unemployment rate $(Y)$
b. Total population $(x_1)$
c. Population density $(x_2)$
d. Labor force participation rate $(x_3)$
e. Population growth rate $(x_4)$
f. Average length of school $(x_5)$
g. Life expectancy $(x_6)$

The method used in this study is spline regression analysis. The stages of analysis in this study are:

1. Make descriptive statistics of each variable.
2. Make a scatter plot between the dependent variable (Y) and each independent variable $x_1$ until $x_6$.
3. Determine the optimum knot value based on minimum GCV.
4. Make the best model using spline with optimum knots.
5. Perform parameter tests and test residual assumptions.
6. Interpreting the obtained spline regression model.
7. Make conclusions.

## RESULT AND DISCUSSION

Descriptive statistics of each research variable can be seen in Table 1.

**Table 1.** Descriptive statistics.

| Variabel | Average | Varians | Min | Max |
|---|---|---|---|---|
| $\boldsymbol{Y}$ | 4,1449 | 5,248 | 0,70 | 12,19 |
| $\boldsymbol{X_1}$ | 25,4419 | 306,567 | 4,29 | 4,29 |
| $\boldsymbol{X_2}$ | 400,3388 | 1142422,419 | 12,10 | 12,10 |
| $\boldsymbol{X_3}$ | 66,9998 | 31,892 | 56,22 | 56,22 |
| $\boldsymbol{X_4}$ | 1,4375 | 0,603 | 0,13 | 0,13 |
| $\boldsymbol{X_5}$ | 8,2162 | 1,429 | 6,21 | 6,21 |
| $\boldsymbol{X_6}$ | 68,5667 | 6,202 | 61,05 | 61,05 |

The pattern of relationships formed between the Open Unemployment Rate which is the response variable with x_1,x_2,x_3,x_4,x_5 and x_6 is visualized in Figure 1, which shows the pattern of relationships that do not form a particular pattern, This indicates that there is a nonparametric component where the function of The regression curve is unknown so the estimation of the model uses nonparametric regression.

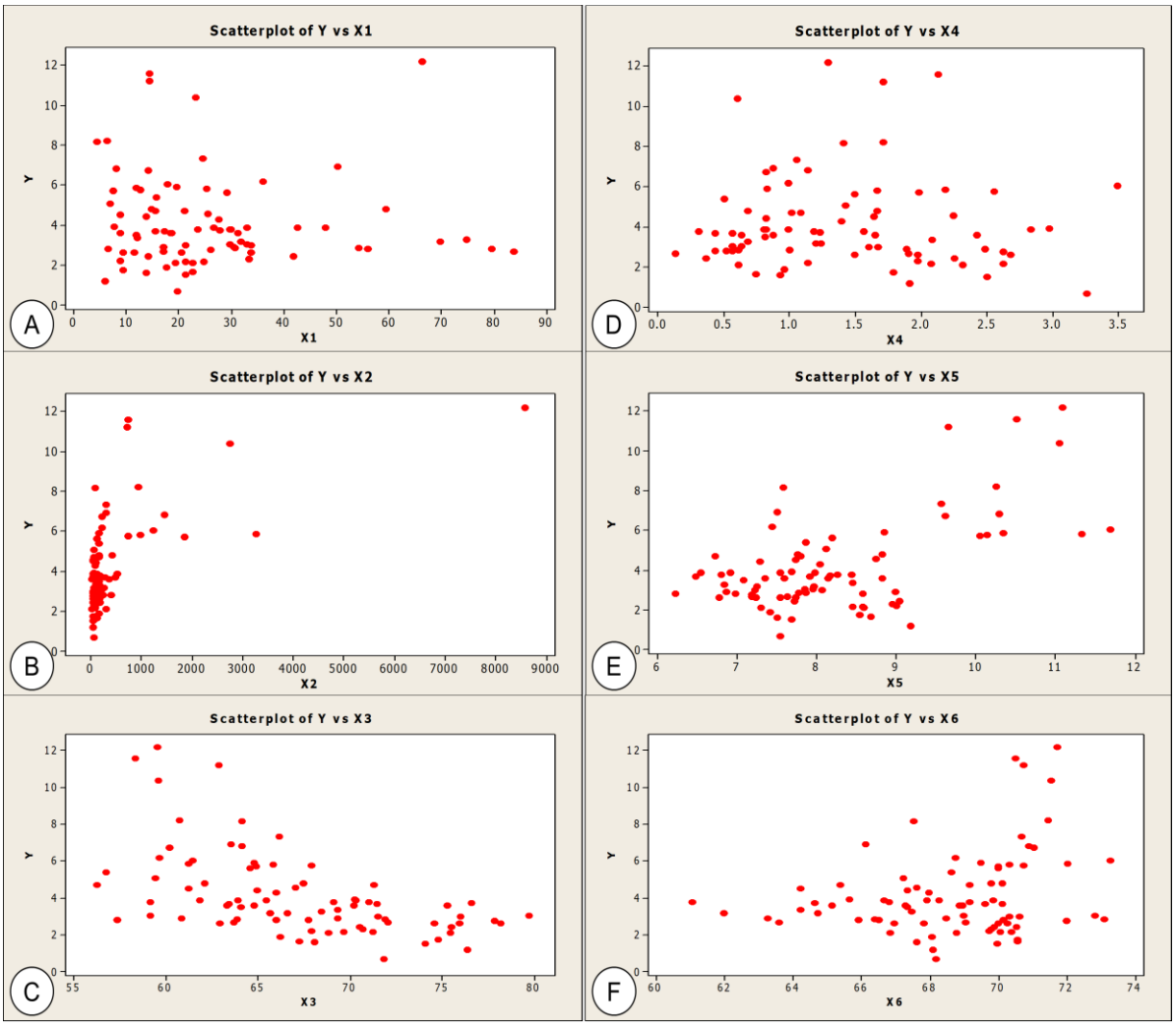


**Fig 1.** *Scatterplot* between Open Unemployment Rate (Y): with Total population $(x_1)$ (A), Population density $(x_2)$ (B), Labor force participation rate $(x_3)$ (C), Population growth rate $(x_4)$ (D), Average length of school $(x_5)$ (E), Life expectancy $(x_6)$ (F).

Based on the results obtained, it is known that the smallest GCV value is 2,54, with the optimum knots point being:

$K_1 = 37,13$ $K_4 = 1,52$
$K_2 = 3557,43$ $K_5 = 8,47$
$K_3 = 65,94$ $K_6 = 66,1$

This GCV value will be compared with the GCV value with 2 knots and 3 knots, then the minimum is chosen.

After getting the optimum knots from one knot point, the next step is to find the optimum knot point with two knot points. The experiments were carried out in the same way, and the minimum GCV was chosen. Based on the results obtained, it is known that the smallest GCV value is 2,37, with the optimum knots point being:

($K_1 = 48,08$; $K_2 = 50,81$) ($K_7 = 1,98$; $K_8 = 2,09$)
($K_3 = 4739,21$; $K_4 = 5034,66$) ($K_9 = 9,23$; $K_{10} = 9,42$)
($K_5 = 69,18$; $K_6 = 69,99$) ($K_{11} = 67,78$; $K_{12} = 68,20$)

The GCV value is smaller than the optimum knots point selection results with one knot point.

After getting the optimum knots with one and the knots then the next is to find the optimum knots with three knots. The experiments were carried out in the same way, and the minimum GCV was chosen.

Based on the results obtained, it is known that the smallest GCV value is 2,25, with the optimum knots point being:

($K_1 = 7,02$; $K_2 = 64,5$; $K_3 = 67,23$)
($K_4 = 307,54$; $K_5 = 6511,88$; $K_6 = 6807,33$)
($K_7 = 57,03$; $K_8 = 74,04$, $K_9 = 74,85$)
($K_{10} = 0,24$; $K_{11} = 2,67$, $K_{12} = 2,79$)
($K_{13} = 6,39$; $K_{14} = 10,36$, $K_{15} = 10,55$)
($K_{16} = 61,47$; $K_{17} = 70,31$, $K_{18} = 70,73$)

When compared with previous experiments using one knot point and two knot points, this model has a smaller GCV value. So it can be said that this spline nonparametric regression model is better.

Knots combination is a combination of one knot point, two knot points, and three knot points. This combination is used to choose the optimum

knot point. In choosing the optimum knots point in the nonparametric spline regression model with this knot combination, a minimum GCV value of 1,97 (Table 2) with a combination of knots of 3,3,1,1,3,3 was chosen.

**Table 2.** GCV with 1,2,3 and knots combination.

| Knot | GCV |
|---|---|
| 1 Knot | 2,54 |
| 2 Knot | 2,37 |
| 3 Knot | 2,25 |
| Knot Combine | 1,97 |

The selection of the optimum knot point is done by finding the lowest GCV value produced. The lowest GCV produced is when using a combination of knots that is equal to 1,97.

The results of the optimum selection of knots, the regression model using three knots is the best result of estimating parameters using a combination of knots is as follows:

$$\hat{y} = -0{,}0215 - 1{,}0507x_1 + 1{,}0552\,(x_1 - 7{,}0268) - 0{,}6163(x_1 - 64{,}5017) + 0{,}6462\,(x_1 - 67{,}2386) + 0{,}0062x_2 - 0{,}0063(x_2 - 307{,}5448) + 0{,}0008(x_2 - 6511{,}886) + 0{,}0007(x_2 - 6807{,}331) - 0{,}1793x_3 + 0{,}0637(x_3 - 65{,}94) + 1{,}4739x_4 - 2{,}1513(x_4 - 1{,}5203) + 5{,}6865x_5 - 5{,}6674(x_5 - 6{,}3989) + 31{,}2944(x_5 - 10{,}3672) - 36{,}6403(x_5 - 10{,}55621) - 0{,}2505x_6 + 0{,}1582(x_6 - 61{,}4710) + 5{,}1693(x_6 - 70{,}31276) - 5{,}5262(x_6 - 70{,}7337)$$

The model formed has a coefficient of determination 77,67%. To test the significance of the parameters carried out by simultaneous and individual tests. Simultaneous test uses the following hypothesis.

H_0: γ_1=γ_2=γ_3= 〖…=γ〗 _20=0

H_1: there is at least one β_k≠0 where k=1,2,…,20.

The ANOVA results for the nonparametric regression model are given in Table 3 as follows. Based on the ANOVA results in Table 3 it can be seen that the p-value is 6.94415929697796e-13, this value is less than α (0,05) so that it can be decided that $H_0$ is rejected, then there is at least one significant parameter to the variable the response. To be able to find out which parameters affect the response it is necessary to test individually. The hypothesis for individual testing is as follows:

$H_0$: $\gamma_1 = 0$

$H_1$: $\gamma_1 \neq 0$ where k=1,2,…,20.

Individual test results are presented in Table 4. which shows that there are 11 parameters that produce a p-value less than the 0,05 significance level, namely the parameter Population $(x_1)$, Population density $(x_2)$ Labor force participation rate $(x_3)$, Population growth rate $(x_4)$, Average length of school $(x_5)$ and Life expectancy $(x_6)$. So it can be concluded that all of these variables significantly influence the model.

Based on table, we get the final significant model

$$\hat{y} = -1{,}0507x_1 + 1{,}0552\,(x_1 - 7{,}0268) + 0{,}0062x_2 - 0{,}0063(x_2 - 307{,}5448) - 0{,}1793x_3 + 1{,}4739x_4 - 2{,}1513(x_4 - 1{,}5203) + 31{,}2944(x_5 - 10{,}3672) - 36{,}6403(x_5 - 10{,}55621) + 5{,}1693(x_6 - 70{,}31276) - 5{,}5262(x_6 - 70{,}7337)$$

Residual assumption testing is carried out to determine whether the residuals generated from the model meet identical assumptions and are normally distributed. Identical test results can be seen in Table 5.

The value of the F test statistic is 0,4146. The P-value generated in the Glejser test is 0,9841, which is greater than α (0,05) so that it can be decided that it failed to reject H_0. So it can be interpreted that heteroscedasticity does not occur. This shows that the residuals have fulfilled identical assumptions.

Normal testing is carried out by the Kolmogorov Smirnov test presented in Table 6. where the results of the Kolmogorov Smirnov test are 0,082 and the p-value is 0,6176 where the p-value obtained is greater than α (0,05) so it can be decided that it failed reject H_0. So it can be concluded that the residuals are normally distributed.

The interpretation of the model that has been obtained is if x_2,x_3,x_4,x_5 and x_6are considered constant then if the Population Amount is less than or equal to 7,0268%, then if the Population Amount increases by 1% the Open Unemployment Rate decreases by 1,0507% . If at the time the Population is more than or equal to 7,0268% then if the Population Increase 1%, the Open Unemployment Rate will tend to increase by 0,0045%.

x_1,x_3,x_4,x_5 and x_6 are considered constant then when Population Density is less than or equal to 307,5448%, then if Population Density increases by 1% the Open Unemployment Rate tends to increase by 0,0062%. If at the time the Population Density is is more than or equal to 307,5448%, then if the Population Density rises 1%, the Open Unemployment Rate will tend to decrease by 0,0001%.

x_1,x_2,x_4,x_5 and x_6 are considered constant then if the Labor Force Participation Rate rises by 1% the Open Unemployment Rate tends to decrease by 0,1793%. When x_1,x_2,x_3,x_5 and x_6are considered constant then when the Population Growth Rate is less than or equal to 1,5203%, then if the Population Growth Rate rises by 1% the Open Unemployment Rate tends to increase by 1,14739%. If when the Population Growth Rate is more than 1,5203%, then if the Population Growth Rate goes up by 1%, the Open Unemployment Rate will tend to fall by 0,6774%.

**Table 3.** ANOVA concurrent test

| SV | Df | SS | MS | F | P-*value* |
|---|---|---|---|---|---|
| Regresion | 20 | 326,079 | 16,304 | 10,4358 | 6,9441e-13 |

**Table 4.** Individual test

| Variable | Parameter | Estimator | $t_{hit}$ | P-*value* |
|---|---|---|---|---|
| Constant | $\beta_0$ | -0,0215 | -0,1748 | 0,8616 |
| $X_1$ | $\beta_1$ | -1,0507 | -2,3882 | 0,0192* |
| | $\beta_2$ | 1,0552 | 2,3747 | 0,0199* |
| | $\beta_3$ | -0,6163 | -1,1301 | 0,2617 |
| | $\beta_4$ | 0,6462 | 1,0059 | 0,3174 |
| $X_2$ | $\beta_5$ | 0,0062 | 3,0331 | 0,0032* |
| | $\beta_6$ | -0,0063 | -2,8698 | 0,0052* |
| | $\beta_7$ | 0,0008 | 0,9107 | 0,3651 |
| | $\beta_8$ | 0,0007 | 0,9108 | 0,3651 |
| $X_3$ | $\beta_9$ | -0,1793 | -2,5737 | 0,0119* |
| | $\beta_{10}$ | 0,0637 | 0,5775 | 0,5651 |
| $X_4$ | $\beta_{11}$ | 1,4739 | 2,8000 | 0,0064* |
| | $\beta_{12}$ | -2,1513 | -2,5976 | 0,0111* |
| $X_5$ | $\beta_{13}$ | 5,6865 | 0,7962 | 0,4282 |
| | $\beta_{14}$ | -5,6674 | -0,7898 | 0,4319 |
| | $\beta_{15}$ | 31,2944 | 3,7724 | 0,0003* |
| | $\beta_{16}$ | -36,6403 | -3,6644 | 0,0004* |
| $X_6$ | $\beta_{17}$ | -0,2505 | -0,3348 | 0,7386 |
| | $\beta_{18}$ | 0,1582 | 0,2103 | 0,8339 |
| | $\beta_{19}$ | 5,1693 | 3,0716 | 0,0029* |
| | $\beta_{20}$ | -5,5262 | -2,7527 | 0,0073* |

*variables have a significant effect

x_1,x_2,x_3,x_4 and x_6 are considered constant then when the Average Length of School is less than or equal to 10,3672%, then if the Average Length of School rises by 1%, the Unemployment Rate will tend to increase by 31,2944%. If the average length of school is between 10,3672% and 10,55621%, if the average length of school increases by 1% the Unemployment Rate will tend to decrease by 5,3459%. If the Average Length of School is equal or more than 10,55621%, if the average length of school increases by 1% the Unemployment Rate will tend to fall by 36,6403%.

**Table 5.** ANOVA Glejser Test

| SV | df | SS | MS | F | P-*value* |
|---|---|---|---|---|---|
| Regresi | 20 | 4,11 | 0,2055 | 0,4146 | 0,984162 |
| Error | 60 | 29,736 | 0,4956 | | |
| Total | 80 | 33,864 | | | |

**Table 6.** Kolmogorov Smirnov

| D | P-*value* |
|---|---|
| 0,082 | 0,6176 |

x_1,x_2,x_3,x_4 and x_5 are considered constant then when the Life Expectancy is less than or equal to 70,31727%, then if the Life Expectancy rises by 1% the Open Unemployment Rate will tend to rise by 5,1693%, and if the Life Expectancy rises between 70,31727% to 70,7337%, then if Life Expectancy Rises by 1% the Open Unemployment Rate will tend to fall by 0,3569%. If the Life Expectancy is equal or more than 70,7337%, then if Life Expectancy Rises by 1% the Open Unemployment Rate will tend to decrease by 5,5262%.

## CONCLUSION

Based on the analysis results above, the following conclusions are the best spline model is obtained from the optimum knots point with a combination of knots 3,3,1,1,3,3. This model has the minimum GCV value of 1,97 with $R^2$ of 77,67%. So the spline model formed as follows:

$$\hat{y} = -1{,}0507x_1 + 1{,}0552\,(x_1 - 7{,}0268) + 0{,}0062x_2 - 0{,}0063(x_2 - 307{,}5448) - 0{,}1793x_3 + 1{,}4739x_4 - 2{,}1513(x_4 - 1{,}5203) + 31{,}2944(x_5 - 10{,}3672) - 36{,}6403(x_5 - 10{,}55621) + 5{,}1693(x_6 - 70{,}31276) - 5{,}5262(x_6 - 70{,}7337)$$

Based on the parameter test it can be seen that all variables significantly influence the Open Unemployment Rate. These variables are the population, population density, labor force participation rate, population growth rate, average length of schooling, and life expectancy. By knowing the influencing factors, it can be taken into consideration in matters relating to unemployment issues.

## ACKNOWLEDGEMENTS

The author would like to thank profusely to the BPS in Sulawesi for allowing the authors to retrieve unemployment data for the completion of this study.

## REFERENCES

Anwar, S. 2014. Regresi nonparametrik spline untuk pemodelan tingkat pengangguran terbuka di Jawa Barat. Doctoral dissertation, Institut Teknologi Sepuluh Nopember, Surabaya.

Aydin, D. 2007. A comparison of the nonparametric regression models using smoothing spline and kernel regression. *World Academy of Science, Engineering and Technology*. *36*: 253-257.

Badan Pusat Statistik [BPS]. 2018. Definisi Pengangguran [online]. https://www.bps.go.id/subject/6/tenaga-kerja.html (accesed on 7 June 2020).

Budiantara, IN. 2000. Metode U, GML, CV dan GCV Dalam Regresi Nonparametrik Spline. *Majalah Ilmiah Himpunan Matematika Indonesia (MIHMI)*. *6*: 285-290.

Budiantara, IN., Lestari, B., & Islamiyati, A. 2010. Estimator Spline Terbobot dalam Regresi Nonparametrik dan Semiparametrik Heteroskedastisitas untuk Data Longitudinal. Hibah Penelitian Kompetensi. LPPM Institut Teknologi Sepuluh Nopember, Surabaya.

Eubank, RL. 1999. Nonparametric Regression and Spline Smoothing, 1st Edition. CRC press, USA.

Eubank, RL. 1999. Nonparametric Regression and Spline Smoothing, 2nd Edition. CRC press, USA.

Hardle, W. 1990. Applied nonparametric regression No.19. Cambridge University press, USA.

Hotchkiss, J., Kaufman, B. 2006. *The economics of labor markets*. Thomson/South Western, New York.

Mankiw, NG. 2006. Principles of microeconomics Vol.10. Cengage Learning,

Sari, RS., Budiantara, IN. 2012. Pemodelan Pengangguran Terbuka di Jawa Timur dengan Menggunakan Pendekatan Regresi Spline Multivariabel. *Jurnal Sains dan Seni ITS*. 1(1): 236-241.