MODELING OF OPEN UNEMPLOYMENT RATE AND PERCENTAGE OF POOR POPULATION WITH NONPARAMETRIC REGRESSION

Ahid Nur Istinah[1], Pardomuan Robinson Sihombing[2]
[1]BPS Kota Bandung, [2]BPS-Statistics Indonesia
Email : robinson@bps.go.id

ABSTRACT: Poverty is often a topic discussed and debated in various national and international forums. The problem of poverty does not seem to be finished being discussed every day. Many factors have contributed to the high poverty rate in Indonesia, and one of them is the high unemployment rate. This research is aimed at modeling of open unemployment rate and percentage of poor population with nonparametric regression. The methods used are Nadaraya Watson Estimator (NEW), Local Polynomial Estimator (LPE) and Smoothing Spline regression. In choosing the best model using the smallest Mean Square Error (MSE) value. The pattern of the relationship between unemployment and the percentage of poor population based on the scatter plot shows an unclear pattern and does not follow a parametric regression pattern. The results of the comparison of the MSE value, smoothing spline has the smaller value than NWE.

Keyword: MSE, NEW, Nonparametric, Poverty, Spline, Unemployment

INTRODUCTION

In 2015, the member states of the United Nations (UN) consisting of 194 countries have agreed Sustainable Development Goals (SDGs). The SDGs are an international agenda that is a continuation of the Millennium Development Goals (MDGs). The SDGs consist of 17 (seventeen) global goals with 169 (one hundred and sixty-nine) targets that will serve as policy and funding guidelines for the next 15 years and are expected to be achieved by 2030. One of the objectives of the SDGs is to answer the demands of world leadership. in overcoming poverty. Eradication of all forms of poverty everywhere is the first goal of the SDGs.

Poverty is often a topic discussed and debated in various national and international forums. The problem of poverty does not seem to be finished being discussed every day. The problem of poverty is a complex and multidimensional problem. Poverty that occurs in a country is seen as a serious problem, because nowadays poverty makes people unable to make ends meet.

Poverty is a condition that is often associated with needs, difficulties and deficiencies in various life circumstances. The development of poverty conditions in a country economically is one indicator to see the development of the level of community welfare. Therefore, with the decreasing level of poverty, it can be concluded that the welfare of the people in a country will increase.

Poverty continues to be a major global problem. Poverty is a big and fundamental classic problem for most developing countries, including Indonesia. Poverty in Indonesia is still quite high. Based on data from Statistics Indonesia, the poverty rate in Indonesia as of September 2018 touched 25.67 million people or 9.66% of the total population of Indonesia.

In realizing the country's goals, the Indonesian government has continuously carried out national development programs. One of the main targets that always receive attention in national development programs is poverty alleviation. Quality improvement in various sectors has been improved by the government to reduce poverty. However, in reality, poverty is difficult to eradicate completely.

Many factors have contributed to the high poverty rate in Indonesia (Wulandari & Budiantara, 2014), such as HDI (Mulyani, 2017), enviromental (Laswinia & Chamid, 2016), education (Hudoyo, 2017) and unemployment rate (Amins, 2017), etc. Theoretically, the poverty rate moves with the unemployment rate (Sukirno, 2006). The high unemployment rate of a country will lead to a high level of poverty in that country. In theory, if people are not unemployed, it means that they have jobs and income, and with the income they have from work, they are expected to meet the needs of life. If the necessities of life are met, it will not be poor. So, the ideal condition is that when the unemployment rate decreases, the poverty rate will also decrease.

The relationship between the unemployment rate and the poverty rate can be shown by a regression analysis model (Yacoub, 2013). Regression analysis is one of the statistical methods used to analyze the relationship between response variables and predictor variables. The form of the relationship pattern of the response variable with the predictor variable is known, but there is also an unknown form of the relationship pattern. If the form of the relationship pattern between the response variables is unknown, the nonparametric regression approach is the appropriate approach for the case.

There are two problems in this study, namely how is the relationship between the unemployment rate and the percentage of the poor population using nonparametric regression methods and how is the comparison between the Nadaraya Watson (NWE), Local Polynomial (LPE) and Smoothing Spline estimators for the model of the relationship between the open unemployment rate.

LITERATURE REVIEW

Nonparametric Regression

Regression analysis is a statistical tool that is widely used to determine the relationship between two or more random variables. Relationship betweenx_i and y_i if suppose X is the independent variable and Y is the dependent variable with $\{(x_i, y_i), i = 1,2, \dots, n\}, x_i \in X, y_i \in Y$ can be assumed to follow the regression model as follows:

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1,2, \dots, n \tag{1}$$

Where ε_i is a random error and $f(x_i)$is a regression function. If function$f(x_i)$ the shape of the curve is not known, the approach used is nonparametric regression.

Nonparametric regression is not bound by certain assumptions as in parametric regression. Nonparametric regression is a very flexible approach (Hermawan, 2018). The estimation of nonparametric regression function is based on observational data using smoothing technique. The nonparametric regression curve is only assumed to be smooth, where the data will find its own estimation form of the estimation function $f(x_i)$ without being influenced by the subjectivity of the researcher (Eubank, 1999). There are several smoothing techniques in nonparametric regression models, including kernel estimator and spline estimator.

Kernel Estimator

One approach to estimating nonparametric regression curves is through kernel functions. The kernel estimator was introduced by Rosenblatt (1956) and Parzen (1962) so it is called kernel density estimator Rosenblatt-Parzen (Hardle, 1994). The kernel estimator or the Rosenblatt-Parzen kernel density estimator is an extension of the histogram estimator. The kernel density estimator gives a probability distribution for each observation that is not always evenly distributed over a fixed interval, but is smooth around several observations, usually symmetrically.

A Kernel K is a finite continuous function in that satisfies:

$$\int_{-\infty}^{\infty} K(u)\,du = 1 \qquad (2)$$

(In particular can also $K(u) \geq 0$, meaning that K can be a density function). Rescaled Kernel, Kh with bandwidth h is as follows:

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

meaning,

$$\int_{-\infty}^{\infty} K(u)\,du = 1$$

The general functions of the kernel estimator are:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) \qquad (3)$$

The bandwidth h, also called the window-width or smoothing parameter, is a function that depends on the sample size n and goes to zero when $n \to \infty$.

Some types of kernel functions include the following (Hardle, 1994):

Table 1. Types Kernel Function

Name	k	μ	Kernels
Uniform	2	0	$\frac{1}{2} 1_{[-1,1]}(u)$
triangle	2	0	$(1-\lvert u\rvert) 1_{[-1,1]}(u)$
Normal	2	∞	$\frac{1}{\sqrt{(2\pi)}} \exp\left(-\frac{1}{2}u^2\right)$
Epanechnikov	2	1	$\frac{3}{4}(1-u^2) 1_{[-1,1]}(u)$
Bisquare	2	2	$\frac{15}{16}(1-2u^2+u^4) 1_{[-1,1]}(u)$
Triweight	2	3	$\frac{35}{32}(1-3u^2+4u^4-u^6) 1_{[-1,1]}(u)$
Cauchy	2	0	$[\pi(1-u^2)]^{-1}$
Picard	2	∞	$\frac{1}{2}\exp(-\lvert u\rvert)$

μ is called the degree of smoothness of the kernel function and $k = 0, \dots, \mu - 1$ is the order of the kernel functions.

Nadaraya-Watson Estimator (NWE)

One of the non-parametric regression techniques to estimate the regression function (m(.)) is to use the Nadaraya-Watson estimator. Separately in the same year

1964 Nadaraya and Watson published the method of estimating m(.), hereinafter referred to as the Nadaraya – Watson Estimator (NWE) method, as follows:

$$\hat{m}(X) = \frac{\sum_{i=1}^{n} K_h(X_i - x)Y_i}{\sum_{i=1}^{n} K_h(X_i - x)} \tag{4}$$

For X fixed, for estimator θ drink

$$\sum_{i=1}^{n} (Y_i - \theta)^2 K_h(X_i - x) \tag{5}$$

have shape $\sum_{i=1}^{n} a_i Y_i$, NWE is the minimizer of the above equation, where

$$a_i = \frac{K_h(X_i - x)}{\sum_{i=1}^{n} K_h(X_i - x)} \tag{6}$$

Local Polynomial Estimator (LPE)

Approximation $m(X)$ locally with a constant $m(X) \approx m_0$ through local least squares,

$$\sum_{i=1}^{n} (Y_i - \beta)^2 K_h(X_i - x) = \min_{\beta}!$$

$$\frac{\partial}{\partial \beta} \sum_{i=1}^{n} (Y_i - \beta)^2 K_h(X_i - x) \overset{!}{=} 0$$

so that it is obtained,

$$\hat{\beta} = m_0 = \hat{m}(X) = \frac{\frac{1}{n} \sum_{i=1}^{n} K_h(X_j - x)Y_i}{\frac{1}{n} \sum_{i=1}^{n} K_h(X_i - x)} \tag{7}$$

Regression with local polynomial estimator is a nonparametric regression method, where the regression function $m(X)$estimated using the polynomial form. The form of local polynomial weights is determined by the kernel function K(.) while the size of the weights is determined by a parameter h called bandwidth.

The polynomial regression model is as follows:

$$Y_i = \beta_0 + \beta_1(X_i - x) + \beta_2(X_i - x)^2 + \cdots + \beta_p(X_i - x)^p + \varepsilon_i \tag{8}$$

The least squares estimator with polynomial form is obtained as follows:

$$\sum_{i=1}^{n} (Y_i - \sum_{k=0}^{p} \beta_k (X_i - x)^k)^2 K_h(X_i - x) = \min_{\beta_0,\dots,\beta_p} !$$

Then the estimator of the local polynomial regression coefficient is

$$\hat{\beta} = (X^T W X)^{-1} X^T W Y \tag{9}$$

Where $W = diag(W_{ii})$ with $W_{ii} = K_h(X_i - x), Y = (Y_1, \dots, Y_n)^T$, and

$$X = \begin{pmatrix} 1 & (X_1 - x) & (X_1 - x)^2 & \cdots & (X_1 - x)^p \\ 1 & (X_2 - x) & (X_2 - x)^2 & \cdots & (X_2 - x)^p \\ \vdots & \vdots & \vdots & & \vdots \; \vdots \\ 1 & (X_n - x) & (X_n - x)^2 & \cdots & (X_n - x)^p \end{pmatrix}$$

with the note that $X^T W X$ invertible.

Then the estimate of Y at point x is β_0 and can be obtained through the following equation:

$$\hat{Y}(x) = \hat{\mu}(x; p, h) = \beta_0 = e_1^T (X^T W X)^{-1} X^T W Y$$

(10)

Where $e_1^T = (1,0,0,\dots,0)$ with $p + 1$element. So, to estimate$Y(x)$, there are three steps to do:

- Build an X matrix for each $x \in (X_1, \dots, X_n)$
- Build a matrix W for each $x \in (X_1, \dots, X_n)$
- Estimate $\hat{\mu}(x; p, h)$ with the above equation

Optimal Bandwidth Selection

The main problem with kernel estimation is bandwidth selection, not kernel selection. Bandwidth is a form of kernel weight specified by K, where the size of the weight is determined by h. Bandwidth h is a smoothing parameter that functions to control the smoothness of the estimated curve (Loader, 1999). A bandwidth that is too small will result in an under-smoothing curve that is very rough and very volatile, and conversely a bandwidth that is too wide will result in an over-smoothing curve that is very smooth, but does not match the data pattern. The bandwidth of the kernel is an independent parameter which shows a strong influence on the resulting estimates.

Optimum bandwidth selection is done by minimizing the error rate (Adisantoso, 2010). This method is a method used to predict prediction errors. The smaller the error rate, the better the estimate. To find out the size of the error rate of an estimator, one way is to look at the MSE (Mean Square Error).

Smoothing Spline Regression

Spline regression is an approach to plotting the data while taking into account the smoothness of the curve. Spline is a segmented or divided polynomial model where the nature of this segment provides better flexibility than ordinary polynomial models.

Smoothing function estimator is a function estimator that is able to map data well and has a small error range. Therefore, by using n observed data, then$f(x_i)$ obtained by minimizing the Penalized Least Square (PLS) function, namely:

$$PLS = \underbrace{\sum_{i=1}^{n} (y_i - f(x_i))^2}_{a} + \underbrace{\lambda \int [f''(x)]^2 dx}_{b} \qquad (10)$$

where part (a) is the sum of the residual squares or a function of the distance between the data and the estimate, part (b) is the roughness penalty, which is a measure of the smoothness of the curve in mapping the data, and $0 < \lambda < 1$is the smoothing parameter, which controls the balance between the fit to the data

(Goodness of Fit) and the smoothness of the curve (penalty). If the value ofλ large is close to 1, it will give a large penalty weight and have a small variance.

METHODOLOGY

The data used in this study is secondary data obtained from BPS-Statistics Indonesia, which was downloaded from the BPS website, www.bps.go.id. The secondary data that will be used refers to 2018 with an observation unit of 34 provinces in Indonesia.

The variables to be studied in this study consist of one independent variable (X) and one dependent variable (Y), namely:

X: Unemployment Rate (TPT)

Y: Percentage of Poor Population

The steps of the research carried out are as follows:

1. *Plotting* predictor variable and response variable with Scatter plot.
2. Modeling with a nonparametric approach to Kernel Regression, with the following steps:
   a. Choose the optimum bandwidth by minimizing cross validation (CV).
   b. Determine the estimation of the regression curve function using the Nadaraya-Watson (NWE) and local polynomial (LPE) estimator approaches.
3. Performing modeling with Smoothing Spline nonparametric approach.
4. Comparing the results of fitting curves resulting from the kernel method approach and Smoothing Spline

RESULTS AND DISCUSSION

Descriptive statistics

The following is a descriptive of the Open Unemployment Rate (TPT) and Percentage of Poor Population. From table 1 above, information is obtained that the response variable (Percentage of Poor Population) has the highest value of 27.43 percent. The predictor variable (Open Unemployment Rate) has the highest value of 8.52 percent. From the table above, it can be seen that the percentage of poor people has a data distribution that is larger than the Open Unemployment Rate (TPT) and also has a larger mean than the TPT.

Table 2. Descriptive Data

Information	Unemployment Rate	Percentage of Poor Population
Min.	1.37	3.55
median	4.38	8.91
mean	4.86	10.61
Max.	8.52	27.43
Stdev	1.64	5.70

Unemployment Rate Plot with Percentage of Poor Population

The relationship pattern between TPT as a predictor variable on the Percentage of Poor Population as a response variable is shown in Figure 1. The relationship pattern formed on the scatter diagram can be used to determine the appropriate approach in estimating the regression function, namely the parametric or nonparametric approach.

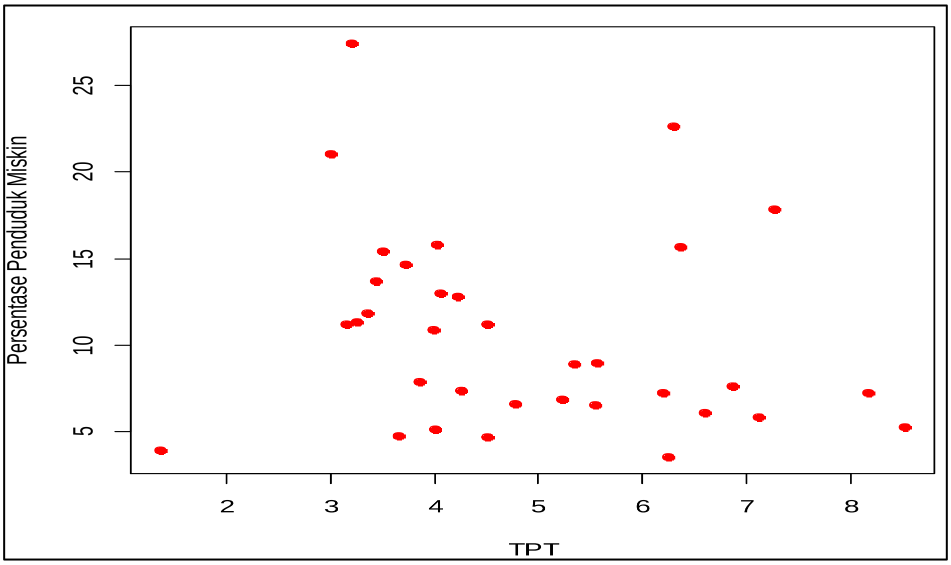

Figure 1. Pattern of relationship between TPT and Percentage of Poor Population

From the picture above, it can be seen that the relationship between TPT and Percentage of Poor Population spreads in a pattern that does not follow the parametric regression pattern. The pattern of data distribution is not clear and it is difficult to determine the exact form of the relationship between the two variables. Therefore, the relationship between TPT and Percentage of Poor Population will be analyzed using a nonparametric approach.

Modeling with Kernel Regression

The most important problem associated with using kernel density estimates is the selection of bandwidth (h). The selection of the optimum bandwidth (h) is done by minimizing the error rate. The smaller the error rate, the better the estimate. The method used to determine the optimum bandwidth (h) is the Cross Validation (CV) method. The optimum bandwidth value is obtained when the CV is minimum. The optimum bandwidth (h) selected in the study is 0.5621203.

After determining the optimal bandwidth and the kernel function used, the regression function estimation will then be carried out. The kernel estimators used in this study are Nadaraya-Watson (NWE) and local polynomials (LPE). The estimation results of the regression curve for the TPT variable and the Percentage of Poor Population are shown in Figure 2 below.

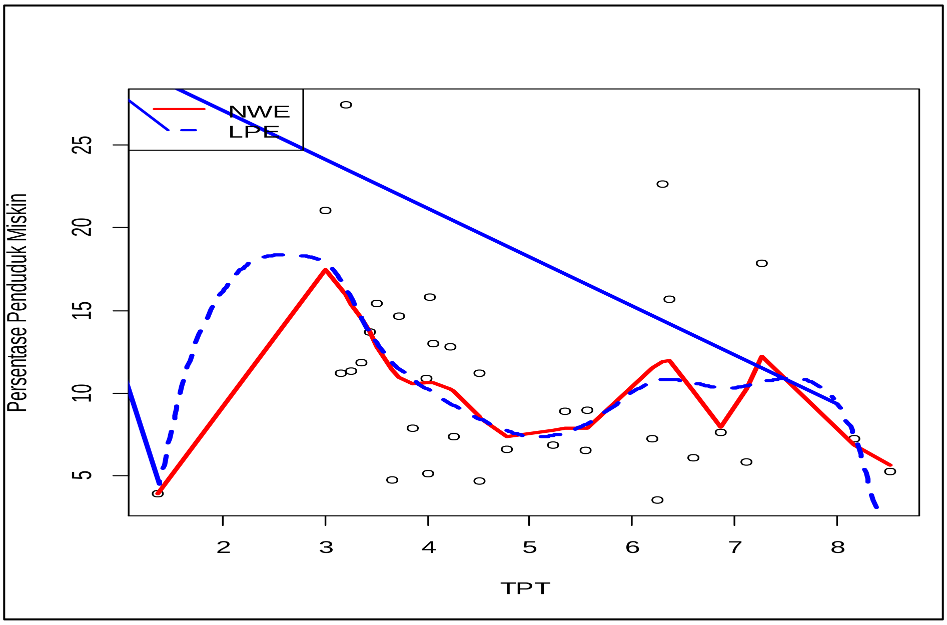

Figure 2. Estimation Result of Kernel Regression Curve with Nadaraya-Watson Estimator (NWE) and Local Polynomial Estimator (LPE)

From the picture above, it can be seen that there is a difference in the fitting curve between NWE and LPE. Although different, it can be seen that the fitting curve between NWE and LPE is almost close. The regression curve formed with LPE looks smoother than NWE. The LPE kernel regression model looks more like the trend of the data, while the NWE looks closer to the data points. But unfortunately, the comparison can only be seen from the fitting curve.

Comparison of the model by looking at the MSE cannot be done because the LPE kernel regression cannot calculate the MSE value. However, from the existing graph, the LPE kernel regression model looks more able to describe the pattern of the relationship between TPT and the Percentage of Poor Population

Modeling with Smoothing Spline Regression

Non-parametric regression approach other than the kernel is a nonparametric spline regression approach. In this study, non-parametric smoothing spline regression will be discussed. The results of the estimated regression curve for the TPT variable and the Percentage of Poor Population with Smoothing Spline are as follows.

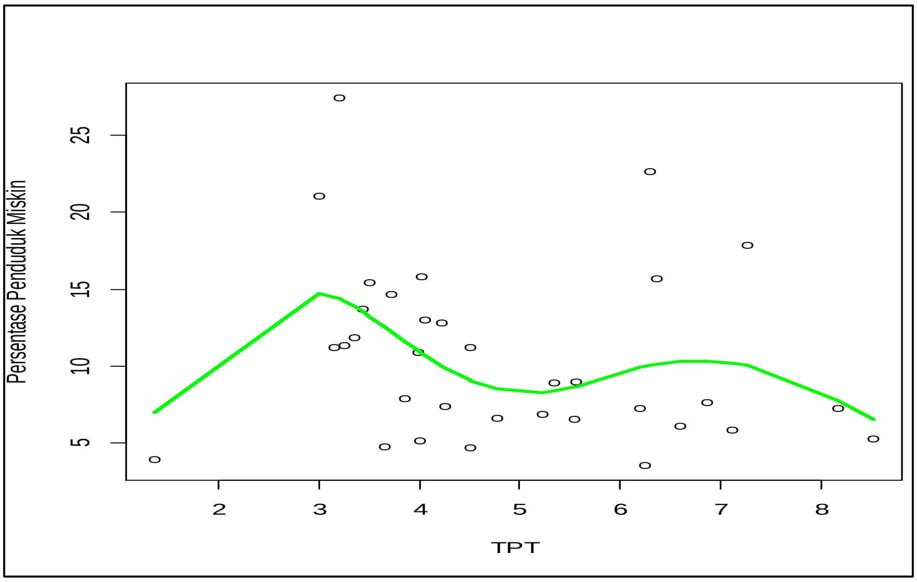

Figure 3. Estimation Results of Smoothing Spline Regression Curve

From the picture above, it can be seen that by using the optimal smoothing parameter (λ), the regression curve formed follows the distribution of the data. Estimating the regression curve with smoothing spline has a shape similar to kernel regression with a local polynomial estimator.

Comparison of NWE Kernel, LPE Kernel and Smoothing Spline

The comparison of the regression curves obtained from the three nonparametric regression methods used can be seen in the following figure:

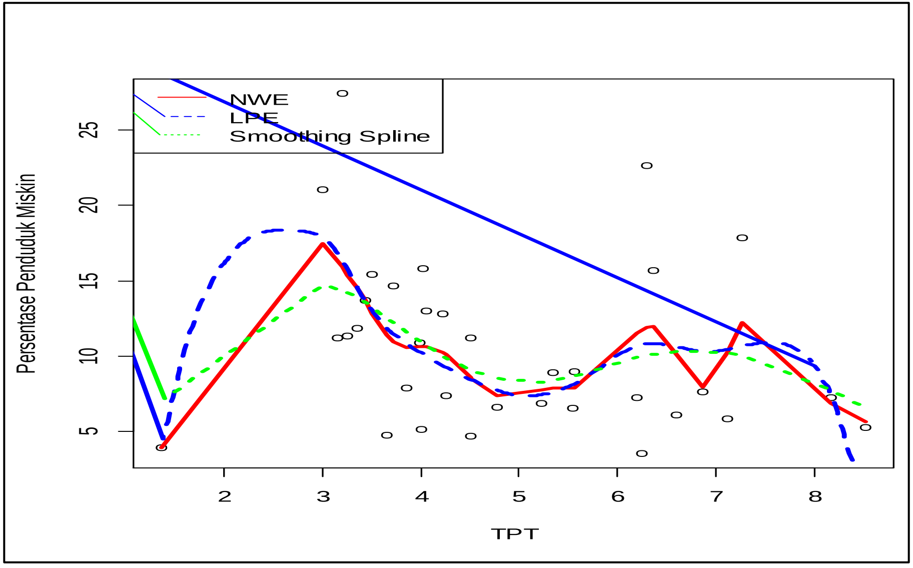

Figure 4. Comparison of NWE Kernel Regression Curve Estimation Results, LPE Kernel, and Smoothing Spline

From Figure 4, the curve shape of the LPE kernel regression model with smoothing spline regression looks smoother than the curve from the NWE kernel regression model. For the NWE kernel regression model looks more crude. Subjectively, it can be said from the three non-parametric regression models, the LPE kernel regression and smoothing spline are relatively better in modeling compared to the kernel regression model with NWE. Because the curve looks smoother and looks closer to the distribution of the data.

To obtain the best approach for estimating the regression function for the variable TPT and Percentage of Poor Population, a comparison of the MSE values between the NWE kernel regression and Smoothing Spline methods was carried out. The LPE kernel regression method was not included because the MSE value could not be found. The smaller the MSE value, the better the model formed. The following table presents a comparison of model accuracy. In table 3 it can be seen that between the Nadaraya Watson Estimator (NWE) and Smoothing Spline models when viewed from the MSE value, the one that produces better accuracy is the Smoothing Spline model because the MSE value is smaller.

Table 3. Model Accuracy Comparison

Model	**MSE**
Nadaraya-Watson Estimator	7.1576
Smoothing Spline	6.6415

CONCLUSION

Based on the results of research regarding the relationship between the variables of the unemployment rate and the percentage of the Poor Population, it can be concluded as follows: The pattern of the relationship between TPT and the Percentage of Poor Population based on the scatter plot shows an unclear pattern and does not follow a parametric regression pattern. Estimating the kernel regression curve between the TPT variable and the Percentage of Poor Population with the local polynomial estimator (LPE) shows a smoother curve than the Nadaraya-Watson (NWE) estimator. The curve shape of the LPE kernel regression model with smoothing spline regression is smoother than the curve of the NWE kernel regression model. Subjectively, it can be said that from the three nonparametric regression models used, the LPE kernel regression and smoothing spline are relatively better in modeling compared to the kernel regression model with NWE. The results of the comparison of the MSE value, smoothing spline has the smaller value than NWE.

REFERENCES

Adisantoso, J. (2010). *Penentuan Parameter Pemulus pada Regresi Smoothing Spline.* Bogor: Thesis: IPB.

Amins, D. B. (2017). Pengaruh Pengangguran Terhadap Tingkat Kemiskinan di Kabupaten Berau. *ECOBUILD : Economy Bring Ultimate Information All About Development Journal, 1*(2), 112-124.

Eubank, R. (1999). *Spline Smoothing and Nonparametric Regression.* New York: Marcel Dekke.

Hardle, W. (1994). *Applied Nonparametric Regression.* Cambridge: University Press.

Hermawan, D. (2018). *Pemodelan Pengeluaran Rumah Tangga di Papua dengan Regresi Nonparametrik Aditif Kuantil.* Bandung: Tesis:Universitas Padjadjaran.

Hudoyo, L. P. (2017). *Pemodelan Hubungan Antara Rata-Rata Lama Sekolah dan Pengeluaran Rumah Tangga Menggunakan Constrained B-Splines (COBS) pada Regresi Kuantil.* Bandung: Tesis:Universitas Padjadjaran.

Laswinia, V. D., & Chamid, M. S. (2016). Analisis Pola Hubungan Persentase Penduduk Miskin dengan Faktor Lingkungan, Ekonomi, dan Sosial di Indonesia Menggunakan Regresi Spasial.,ITS, Vol.5 No.2. *Jurnal Sains dan Seni,ITS, 5*(2).

Loader, C. R. (1999). Bandwidth Selection: Classical or Plug-In. *The Annals of Statistics, 27*(2), 415-438.

Mulyani, S. (2017). *Pemodelan Hubungan Indeks Pembangunan Manusia dan Persentase Penduduk Miskin Menggunakan Regresi Kuantil Smoothing Splines.* Bandung: Tesis:Universitas Padjadjaran.

Sukirno, S. (2006). *Ekonomi Pembangunan (proses, masalah dan dasar kebijakan), Edisi Kedua.* Jakarta : Prenada Media Group.

Wulandari, I., & Budiantara, I. (2014). nalisis Faktor-faktor yang Mempengaruhi Persentase Penduduk Miskin dan Pengeluaran Perkapita Makanan di Jawa Timur menggunakan Regresi Nonparametrik Birespon Spline. *Jurnal Sains dan Seni,POMITS*.

Yacoub, Y. (2013). Pengaruh Tingkat Pengangguran terhadap Tingkat Kemiskinan Kabupaten/Kota di Provinsi Kalimantan Barat. *EKSOS, 8*(3), 176-185.