USE OF GCV METHOD IN FORMING ESTIMATORS FOURIER SERIES AND ITS APPLICATIONS

Agustini Tripena*, Niken Larasati, Agus Sugandha, Princess Nurrifkadillah
Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Jenderal Soedirman, Purwokerto, Indonesia

*Email: agustini.brsurbakti@unsoed.ac.id

Abstract. Linear regression model with function curve regression and error are normally distributed with a mean of zero and a deviation of standard sigma square. The problems that arise is how form estimate from function curve regression. For estimate function curve regression, there are two approaches that can be taken used that is approach parametric and approximation nonparametric. Approach nonparametric done if there is no assumption form curve regression. Function curve regression only assumed loaded in a room dimensionless function not up to. Research this aiming for to study series estimator form fourier in regression model nonparametric. Next under review selection of optimal lambda *smoothing* parameters with CV and GCV methods. Form of series estimator fourier in regression model nonparametric bulk data rain Cilacap Central Java in the month January 2010 – Dec 2022 is $\hat{g}_\lambda(t) = \hat{b}(\lambda)t + \frac{1}{2}\hat{a}_0(\lambda) + \sum_{k=1}^{K}\hat{a}_k\,(\lambda)\,cos\cos k\,t$. Next under review selection of optimal lambda smoothing parameters if k = 5 then lambda value = 0.001584 and CV value = 0.0052237 and GCV = 0.0005400.

Keywords: regression nonparametric, Fourier series, CV, GCV, rainfall

A. Introduction

Fourier series is a regression model nonparametric used in estimate a data pattern in the form of trigonometry [3]. Data patterns that can used in fourier series is repeating data patterns because nature periodic [2]. Periodic has the meaning namely a condition happen with hose fixed time [5],[10]. The condition can called with pattern seasonal or trend where lots research that uses periodic data and determines method his research [7],[8]. Rainfall and elements other climates have a regular pattern [28],[29]. The pattern caused by the presence of condition climate in a particular area. Variables used in this study is rainfall (Y) and time (X). Rainfall data pattern that forms a series periodic influenced by time so that time is variables that greatly influence rainfall.

In statistics, to find out the model of the relationship pattern between the predictor variable t_j and the response variable y_j, regression analysis can be used. Assume that paired data (t_j , y_j) follows the regression model.

$$y_j = f\,(t_j) + \varepsilon_j\,, \qquad (1)$$

j = 1, 2, …, n. The function $f\,(t_j)$ is a regression curve and ε_j random *error* assumed to be normally distributed independently with a mean of zero and a variance of σ^2. If in the regression analysis the shape of the regression curve is known, then the regression model approach is called a parametric regression model [4],[5]. If the data pattern tends to follow a

linear/quadratic/cubic model, then the regression approach that is appropriate for the data is linear/quadratic/cubic parametric regression [11],[13].

In some cases, the response variable can have a linear relationship with one of the predictor variables, but the relationship pattern is unknown with the other predictor variables. In such circumstances, [25],[26] suggest the use of a semiparametric regression approach. Among the nonparametric and semiparametric regression models, the Fourier Series is one of the models that has a very special and excellent statistical and visual interpretation [12]. The spline estimator is obtained from a *penalized least square* (PLS) optimization and has high flexibility [8],[9]. The Fourier Series Estimator, Kernel Estimator, Spline Estimator in nonparametric regression developed by the researchers above, are only for regression models with one response variable. In some cases in the real world, cases are often encountered where variable measurements are carried out at the same time, so that it will involve a regression model with more than one response variable and the response variables are correlated with each other [14]. Therefore, in this study, a Fourier series estimator was derived to estimate the nonparametric regression curve [11]. The main problem in regression analysis, whether parametric, nonparametric or semiparametric regression, is finding an estimate for the regression curve. The Fourier series estimator is a very good and popular approximation model for this purpose [15].

Curve estimators regression This have background background and motivation alone, as a approach for the data model. If function $f \in C\,(0,\pi) = (f, f\ continuous\ at\ (0,\pi)$ so size conformity curve to the data is $n^{-1}\sum_{j=1}^{n}(y_j - f(t_j))^2$ and size rudeness curve is $\int_0^{\pi}\frac{2}{\pi}(f^{(2)}(t)^2\,dt$. The estimator f is obtained with minimize *Penalized Least Square* [14],[15]

$$n^{-1}\sum_{j=1}^{n}(y_j - f(t_j))^2 + \int_0^{\pi}\frac{2}{\pi}(f^{(2)}(t)^2\,dt \qquad (2)$$

λ are the smoothing parameters and $\lambda > 0$.

To obtain a good regression curve estimate, an optimal selection is required λ and is very important. If the Fourier Series is used to estimate the regression curve in (1), then an λ optimal value must be selected. Several methods for selecting λ are *Unbiased Risk (UBR)* [27]. Nonparametric regression curve estimation is highly dependent on the smoothing parameters λ [28].

This study aims to determine the form of the Fourier Series estimator in a nonparametric regression model. The properties of the Fourier Series estimator are also investigated. Furthermore, the CV and GCV methods will be compared to select the optimal smoothing parameters in the Fourier Series estimator λ using rainfall data in Cilacap Central Java in the month January 2010 – Dec 2022.

1. Nonparametric Regression

According to [12] nonparametric regression is an approach to data patterns where the regression curve shape is unknown or there is no complete past information regarding the shape of the data pattern, for example, wind speed data in certain areas where the curve shape cannot be determined. Nonparametric regression is used to determine the relationship between response variables and unknown predictors. the form of the function and is only assumed to be smooth. Nonparametric regression has high flexibility because it is not bound by assumptions about the shape of the curve as in parametric regression [13]. [23] suggests the use of nonparametric regression because it has good flexibility. Nonparametric regression models can be formed in general regression, namely

$$y_i = f(x_i) + \varepsilon_i,\ i = 1,2,3,\ldots,n \qquad (3)$$

with y_i is the *i*-th response variable, x_i is the *i*-th predictor variable, $f(x_i)$ is the *i*-th nonparametric regression function, and ε_i is the *i*-th *error assumed* to be normally distributed with zero *mean* and variance σ^2.

2. Fourier Series Estimators in Nonparametric (One Response) Regression

Research concerning estimator Univariate (one response) Fourier series, in regression in recent years, it has received a lot of attention from several nonparametric regression researchers. The Fourier Series estimator in univariate nonparametric regression is generally used when the data being investigated has an unknown pattern and there is a tendency for a seasonal pattern [21],[22]. Meanwhile, the partial Fourier Series estimator in univariate semiparametric regression is used when some of the data tends to have a certain pattern and some of the pattern is unknown and tends to be seasonal [12],[16].

Given a regression model nonparametric $y_j = f(t_j) + \varepsilon_j$, $j = 1, 2, \ldots, n$. Form curve regression f assumed no known and contained in room function continuous $f \in C(0, \pi)$. Random *error* ε_j assumed independent normal distribution with zero mean and variance σ^2. Since $f(t)$ is continuous on the interval $(0, \pi)$ so can approached by function $F(t)$, with:

$$F(t) = \gamma t + \frac{1}{2}\alpha_0 + \sum_{i=1}^{K} \alpha_i \cos it\,, \qquad (4)$$

where γ, α_0, α_1, $I = 1, 2, \ldots, K$ are model parameters.

3. Fourier Series Estimator in Nonparametric (Multi Response) Regression

According to [20], [21], estimate to $f(x)$ is $f_\lambda(x)$ a smooth estimator. The general form of the th-order spline regression $m$is as follows:

$$y = \beta_0 + \sum_{j=1}^{m} \beta_j x^j + \sum_{k=1}^{N} \beta_{j+k} (x - K_k)_+^m + \varepsilon \qquad (5)$$

With using observation data as much as n, then the matrix form of equation (5) is

$$y = X_1\delta_1 + (X - K)\delta_2 + \varepsilon \qquad (6)$$

For reason simplicity, then matrix (6) can written return become

$$y = X\beta + \varepsilon \qquad (7)$$

In relation to the estimation curve smooth $f(x)$, that has the optimal smoothing parameter value (λ), then to choose the $f(x)$best estimator among the estimator class $C(\Lambda) = \{f_\lambda : \lambda \in \Lambda, \Lambda = \text{index set}\}$. The index set is a set that contains indices. With using the spline regression model as estimate curve smooth f_λ, the equation is adjusted to become$b_\lambda = \hat{\beta}_\lambda$

$$b_\lambda = (X'_\lambda X_\lambda)^{-1} X'_\lambda\, y \qquad (8)$$

with X_λ is the design matrix of the model that forms the estimation model f_λ with λ the optimal one. In this case,

$$f_\lambda = X_\lambda b_\lambda = X_\lambda (X'_\lambda X_\lambda)^{-1} X'_\lambda\, y = H_\lambda\, y$$
$$\lambda \in \Lambda \qquad (9)$$

It should be noted H_λ that $H_\lambda = X_\lambda (X'_\lambda X_\lambda)^{-1} X'_\lambda$ .it is symmetric, positive definite, and idempotent. For get curve smooth which has optimally using λ as much observation n data as possible, a universally acceptable performance measure of the estimator is required.

4. Quadratic Spline Regression

According to [23] the quadratic *spline* function ik is a second-order *spline function*. The quadratic *spline function* ik with one knot point *K* can be presented in the form

$$f_1(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 (x-K)_+^2 \quad (10)$$

This function can also be presented as

$$f_1(x) = \begin{cases} \beta_0 + \beta_1 x + \beta_2 x^2 & , x < K \\ \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 (x-K)_+^2 & , X \le K \end{cases} \quad (11)$$

Spline function with four knot points on ($x = K_1$, $x = K_2$, $x = K_3$, $x = K_4$)

$$f_4(x) = \beta_0 + \beta_1 x + \beta_2 (x-K_1)_+^1 + \beta_3 (x-K_2)_+^1 + \beta_4 (x-K_3)_+^1 + \beta_5 (x-K_4)_+^1$$

5. Selection of the Best Estimation Model

Knot points can affect the performance of the model produced in spline regression. Therefore, the selection of the knot point location is very important in forming a spline regression model. Choosing the right knot point location will give good results and can affect the error value in the spline regression model obtained. One method that can be used to determine the optimal knot point is the method cross validation (CV) and generalized cross validation (GCV). [1],[6]

a. Generalized Cross Validation

According to [12], *Generalized Cross Validation* (GCV) is a modification of *cross validation* (CV). *Cross validation* (CV) is a method for selecting a model based on the predictive ability of the model.

Furthermore, to select the λ optimal smoothing parameters using the GCV method, the equation used is as follows:

$$GCV(\lambda) = n^{-1} \sum_{i=1}^{n} \frac{\left(y_i - \hat{g}_\lambda(t_i)\right)^2}{\left(1 - n^{-1} \sum_{i=1}^{n} a_{ii}(\lambda)\right)^2} \quad (12)$$

The model form of CV is as follows:

$$CV = n^{-1} \sum_{i=1}^{n} \left(\frac{y_i - f(x_i)}{1 - g_{ii}} \right)^2 \quad (13)$$

with g_{ii} being the *i*-th diagonal element *of* the matrix *G*.

[24],[25],[26], Equations *GCV* is obtained by replacing 1-g_{ii} in equation (13) with $\sum_{i=1}^{n} g_{ii} = n^{-1} Tr(I - G)$. The value of *Tr (I – G)* is the sum of the diagonal elements of the matrix (*I – G*). The *GCV function* is defined as:

$$CV = \sum_{i=1}^{n} \left(\frac{y_i - f(x_i)}{n^{-1} Tr(I-G)} \right)^2$$

$$CV = \sum_{i=1}^{n} \frac{MSE}{(n^{-1} Tr(I-G)^2} \quad (14)$$

with n -1 Tr (I –G) < n and G = W(W T W) -1 W T.

B. Method

Methodology study in the form of studies library and study case. In the study library done assessment about Fourier Series estimators and regression nonparametric which has been conducted by researchers. Furthermore, it will Fourier series estimator formation model was built and its application to rainfall data rain**.** Most of the study conducted in the Department Mathematics, BMKG Cilacap Central Java.

C. Results and Discussion

1. Application Fourier series on rainfall data rain in the city Cilacap with Method, CV and GCV

In this study, an application was conducted to provide an overview of the Fourier Series regression model. The application in this study was conducted to evaluate the goodness of the GCV and CV methods. [17],[18],[19] The reliability of the measurement is based on the smallest value obtained in the GCV and CV methods using Veiten Software. The form of the series estimator fourier in regression model nonparametric bulk data rain Cilacap Central Java in the month January 2010 – Dec 2022 is $\hat{g}_\lambda(t) = \hat{b}(\lambda)t + \frac{1}{2}\hat{a}_0(\lambda) + \sum_{k=1}^{K}\hat{a}_k(\lambda)\, cos\, \cos k\, t.$

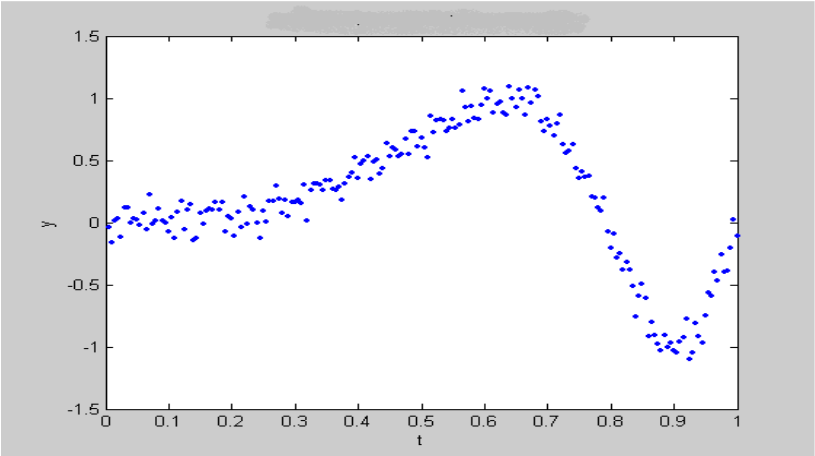

Figure 1. Plot (t_i, y_i) with n = 30, σ^2 = 0,1

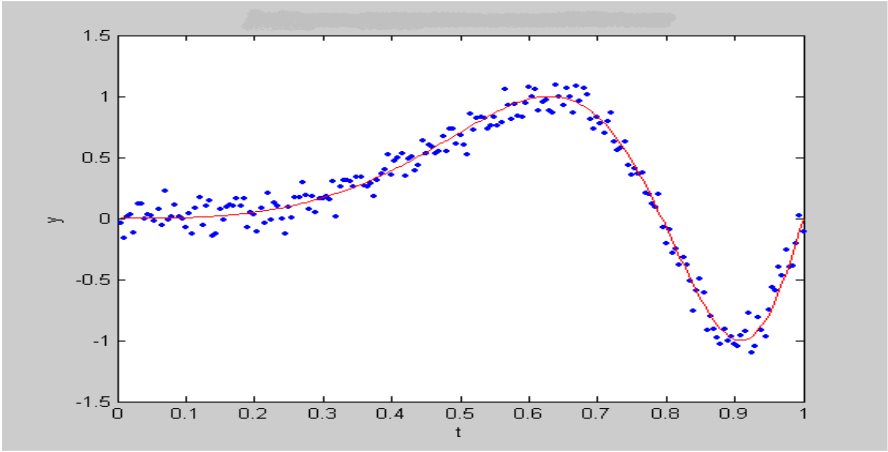

Figure 2. Plot (t_i, y_i) with n = 100, σ^2 = 0,1

2. Selection of the Best Estimation Model

[1],[27],[28], Next, the selection of optimal smoothing parameters λ is selected using the GCV and CV methods which have previously been studied in sub-chapter 1.5. The plot (t_i, y_i) of the Fourier Series estimator for trigonometric functions $g(t)$ using the CV method with n = 100, σ^2 = 0,1 and K = 5 is presented in Figure 3 as follows:

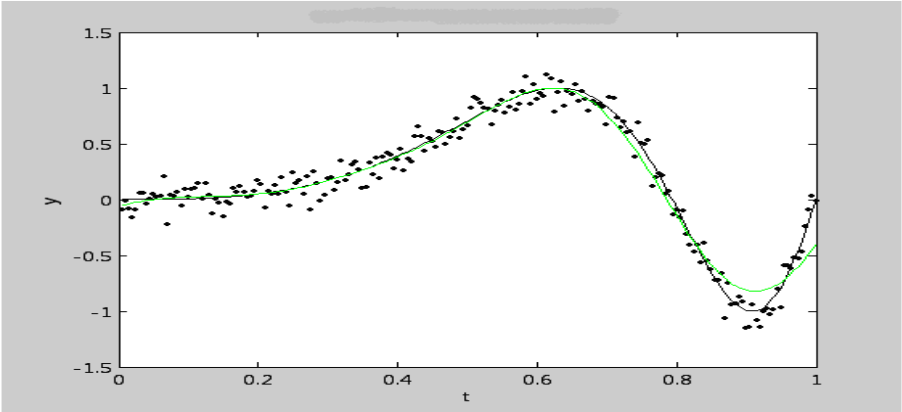

Figure 3. Plot (t_i, y_i), and Fourier series estimator with CV method for n = 100, σ^2 =0,1 and K = 5

Furthermore, to select the λ optimal smoothing parameters using the GCV method, the equation used is as follows:

$$GCV(\lambda) = n^{-1}\sum_{i=1}^{n}\frac{\left(y_i - \hat{g}_\lambda(t_i)\right)^2}{\left(1 - n^{-1}\sum_{i=1}^{n} a_{ii}(\lambda)\right)^2}$$

Given the plot (t_i, y_i) of the Fourier Series estimator for trigonometric functions $g(t)$ using the GCV method with n = 100, σ^2 = 0,1 and K = 5 is presented in Figure 4 as follows:

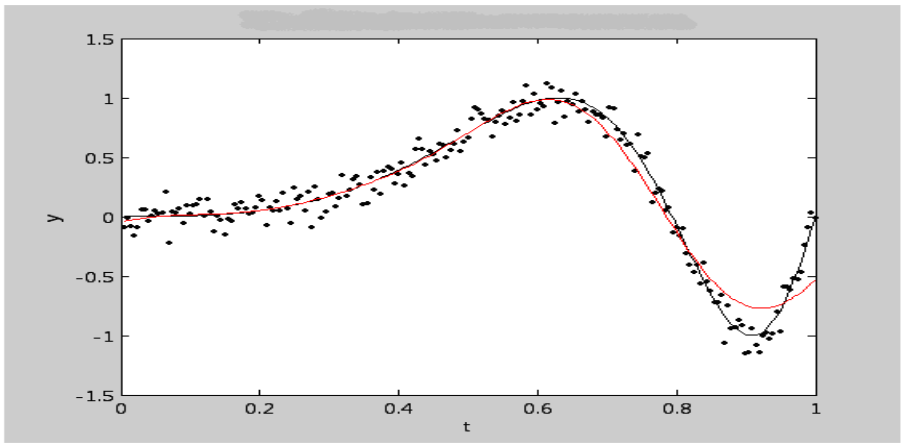

Figure 4. Plot (t_i, y_i), and Fourier series estimate with GCV method for n = 100, σ^2 = 0,1 and K = 5

Evaluation of the goodness of CV and GCV methods in selecting smoothing parameters λ is reviewed based on the smallest value produced by each method. The following data results, for the λ optimal values of CV and GCV are presented in Table 1 as follows:

Table 1. λ optimal values of CV and GCV in the Fourier series estimator where n = 30, n = 100, σ^2 = 0,1 and K = 5, K = 20

n	var	k	CV Method		GCV Method	
			λ Optimal	CV	λ Optimal	GCV
30		5	0.077352	0.0040737	0.010822	0.01583
0.1		20	0.043624	0.0219820	0.003099	0.0336500
100		5	0.079471	0.0052237	0.001584	0.0005400
0.1		20	0.043365	0.0203590	0.003076	0.0334470

Overall, if the value of K is greater, it will give a greater value for both methods, CV and GCV for sample sizes n = 30, n = 100, σ^2 = 0,1 and K = 5, K = 20. So from these two methods,

the CV method and the GCV method on the Fourier Series estimator in nonparametric regression can be seen in Table 1.

D. Conclusion

Based on the application results for $n = 30$, $n = 100$, $\sigma^2 = 0{,}1$ and $K = 5$, $K = 20$, it is obtained that the value of the CV method and the GCV value in each model. For the GCV value is smaller than the CV value in each model. The greater the value of K, the greater the CV and GCV values. It can be concluded that the selection of λ optimal smoothing parameters of the GCV method is better than and CV. Fourier Series estimator in nonparametric regression $\hat{g}_\lambda(t) = \hat{b}(\lambda)t + \frac{1}{2}\hat{a}_0(\lambda) + \sum_{k=1}^{K}\hat{a}_k\,(\lambda)\,cos\,\cos k\,t$.

E. Reference

[1]. Budiantara, I N., and Subanar, 2017, Selection of Smoothing Parameters in Spline Regression, *Scientific Journal of Mathematics and Natural Sciences*, 7.37 – 49.

[2]. Bilodeau, M. 2012. Fourier Smoother and Additive Models, *The Canadian of Statistics*, 3, p. 257- 259

[3]. Bahtiyar et al. 2014. Ordinary Kriging in Rainfall Estimates in Semarang City. *Journal Gaussian FMIPA University Diponegoro Semarang.* Volume 3 No.2 pp. 151-159

[4]. Budiantara , IN, Subanar , and Soejoeti , Z., 2017, Weighted Spline Estimator, *Bulletin of the International Statistical Institute* , 51, 333-334.

[5]. Budiantara, I N., 2010 *Spline Estimators in Nonparametric and Semiparametric Regression*, Doctoral Dissertation at Gadjah Mada University, Yogyakarta.

[6]. Budiantara, IN, and Subanar, 2017, Asymptotic Properties of Weighted Spline Estimators, *Magazine Periodic Mathematics and Science Knowledge Nature (BMIPA), Gadjah Mada University,* 2, 23-36.

[7]. Budiantara, IN, 2011b, Parametric and Nonparametric Estimation for Curve Approach Regression, Keynote Speaker Paper at the 5th National Statistics Seminar, Department of Statistics, Faculty of Mathematics and Natural Sciences, Sepuluh Nopember Institute of Technology (ITS), Surabaya.

[8]. Budiantara, IN, 2002 *a*, Penalized Type Estimator Likelihood, *Natural Journal of FMIPA Unibraw*, Edition Special, 231-235.

[9]. Budiantara, I. N, 2004, Spline History, Motivation, and Role In Regression Nonparametric, Paper Speaker Main at the Conference National Mathematics XII, Department Mathematics, Faculty Mathematics and Science Natural Sciences, University University of Indonesia, Denpasar.

[10]. Chamidah, N. and Lestari, B., 2016. Spline Estimator in Homoscedastic Multi-Response Nonparametric Regression Model in Case of Unbalanced Number of Observations. *Far East Journal of Mathematical Sciences* (*FJMS*), 100(9), 1433- 1453.

[11]. DRS Saputro , A. Sukmayanti , and P. Widyaningsih , "The Nonparametric Regression Model Using Fourier Series Approximation and Penalized Least Squares (Case on Data Proverty in East Java)," in *The Sixth National Seminar on Mathematics Education , Ahmad Dahlan University , IOP Conference Series: Journal Physics* 1188, 201.

[12]. Eubank, R.L., 1988, *Spline Smoothing and Nonparametric Regression*, Mercel Decker, New York.

[13]. Hardle, W., 1990, *Applied Nonparametric Regression*, Cambridge University Press, New

York.

[14]. Hidayati, L., Chamidah, N., and Budiantara, IN, 2019 Spline Truncated Estimator in Multiresponse Semiparametric Regression Model For Computer Based National Exam In West Nusa Tenggara. *Proceedings* 9th Annual Basic Science International Conference. IOP Conf. Series: Materials Science and Engineering 546 (2019) 052029 doi: 10.1088/1757-899X/546/5/052029

[15]. Hidayati, L., Chamidah, N., and Budiantara, IN, 2020 Confidence Interval of Multiresponse Semiparametric Regression Model Parameters Using Truncated Spline. *International Journal of Academic Applied Research (IJAAR)* ISSN: 2643-9603, Published in vol. 4, Issue 1 January 2020, Pages 14 -18.

[16]. Octavanny, MAD, Budiantara, IN, Kuswanto, H., and Rahmawati, DP (2021). "Modelling Children Eve Born in Indonesia Using Fourier Series Nonparametric Regression", Journal of Physics: Conference Series, Vol. 1752, No. 1, pp. 1-7.

[17]. Prahutama, A. 2013. Regression Model Nonparametric with Approach Fourier Series in the Case of Open Unemployment Rate in East Java. *Proceedings of the National Seminar Diponegoro University Statistics 2013.*

[18]. Pane, R. S., and Ampa, A. T. (2020). "Estimation of Heteroskedasticity Semiparametric Regression Curve Using Fourier Series Approach", Journal of Research in Mathematics Trends and Technology (JoRMTT), Vol. 2, no. 1, p. 14-20.

[19]. Sahidah, Kuzairi, Mardianto, MFF Fourier Series Estimator in Nonparametric with Penalty For Planning Sale Product Seasonal. *Zeta-Math Journal*. 2020; 7(2), 69-78.

[20]. Tripena, A, Agung Prabowo, Yosita Lianawati, Abdul Talib Bon (2021). *Estimated Splines in Nonparametric Regression with a Generalized Cross Validation and Unbiased Risk Approach*, Proceedings of the 11th Annual International Conference on Industrial Engineering and Operations Management Singapore, 3788-3798.

[21]. Tripena, A. and Budiantara, IN., 2006, *Fourier Estimator in Nonparametric Regression*, International Conference on Natural Sciences and Applied Natural Sciences, Ahmad Dahlan University, Yogyakarta.

[22]. Tripena, A. 2005. *Linear Spline Regression Model Approach*. Department of Mathematics and Natural Sciences, Faculty of Science and Engineering, UNSOED.

[23]. Tripena, A., 2011. *Quadratic Spline Regression Analysis*. Proceedings of the National Seminar on Mathematics and Mathematics Education, Department of Mathematics Education, FMIPA UNY. Yogyakarta. ISBN: 978-979 -16353-6-3

[24]. V. Ratnasari, IN Budiantara, M. Ratna, and I. Zain, "Estimation of Nonparametric Regression Curve using Mixed Estimator of Multivariable

[25]. Wahba, G., 1990. *Spline Models For Observation Data*, SIAM Pennsylvania.

[26]. Wahba, G. (2000). "A Comparison of GCV and GML for Choosing the Smoothing Parameter in Generalized Spline Smoothing Problems", The Annal of Statistics, Vol. 13, no. 4, p. 1378-1402.

[27]. _____ (1990). Spline Models for Observation Data. Philadelphia: SIAM, CBMS-NSF Regional Conference Series in Applied Mathematics

[28]. Wisisono, IRN, Nurwahidah, AI, Andriyana, Y., & Sunengsih, N. Regression Nonparametric with Approach Fourier Series on Citarum River Water Discharge Data.

Journal Mathematics "Logic". 2018; 04 (02), 76-78.

[29]. Yatcchew, A., 2003. *Semiparametric Regression for the Applied Econometrician*. Cambridge University Press, Cambridge.