Modeling with Mixed Kernel, Spline Truncated and Fourier Series on Human Development Index in East Java

N Y Adrianingsih, I N Budiantara* and J D T Purnomo
Department of Statistics, Institut Teknologi Sepuluh November, Surabaya 60111, Indonesia

*Corresponding author: nyomanbudiantara65@gmail.com

Abstract. The pattern of the relationship between the response variable and the unknown predictor can be determined using a nonparametric regression approach. Nonparametric regression allows it to be used for response variables following different curves between one variable and another. In this regression, there are several types of approaches including the kernel, spline, and Fourier series. In its use, there is not only one type of approach, but can be in the form of a mixture, such as a mixture of a spline and Fourier series, a kernel and a Fourier series, and so on. In this study, in modeling the HDI cases in East Java Province, nonparametric regression mixed kernels, Spline Truncated, and Fourier series were used. The results of research that have been applied to HDI in East Java with the predictor variables for APM SMA, Morbidity, and GRDP per Capita are using the mixed kernel, Spline Truncated, and Fourier series nonparametric regression approach with three-knot points and three oscillations. Good because the coefficient of determination of the estimator is 75.1041%.

1. Introduction
One of the successes in regional development can be seen from the resolution or failure of basic problems in the community, for example, poverty, illiteracy, unemployment, and so on. Recently, the Indonesian government has been very interested in achieving the success of human development. The United Nations (United Nations) has established a measure for the achievement of human development standards, namely the Human Development Index (IPM). HDI is a summary measure of the average achievement in the main dimensions of human development. This human development can be done by increasing the aspects which are very important for life, namely the age of life, education, and a decent standard of living. There are 4 indicators formed in this index, namely life expectancy, literacy rate, the average length of schooling, and purchasing power. In human development, there are important factors that a region needs to have, including social and economic factors. From BPS data in East Java, HDI reached a value of 70.27 in 2017. This value is influenced by several factors, including the open unemployment rate, education level, and per capita GRDP.

In modeling HDI with the factors that influence it, there are several ways, one of which is by using nonparametric regression. Nonparametric regression is a regression analysis used to determine the relationship between the response variable and the predictor variable, for which no information is available from before about the relationship between these two types of variables. This regression is very good to use for unknown data patterns because it is flexible [1]. There are many types of
estimators in this regression, namely kernel, wavelet, Fourier series, and Spline Truncated. There are many nonparametric regression analyzes used in a study can be seen in ref [2]-[7].

The kernel estimator is quite in demand by researchers because it can handle data that has a good ability to model data that does not have a certain pattern [8], its mathematical form is easy, and is more flexible, and can achieve relatively fast convergence [9], and from a computational point of view, it is easier to do. The spline truncated estimator can handle data characteristics that have variable behavior at certain sub-intervals [10]. The Fourier series estimator has the flexibility to adapt effectively to local data properties [11], and there is a recurring trend [12]. Based on the advantages possessed by the three estimators, the researchers are interested in modeling the human development index in East Java by using a nonparametric regression additive mixed model between the spline, kernel, and Fourier series, because the pattern between the response variables and each predictor, in this case, has The pattern is unknown, follows a repeating pattern at certain intervals, and has a changing pattern at certain sub-intervals.

2. Methodology

2.1. Kernel Nonparametric Regression

Initially, the most widely used nonparametric regression research is kernel estimator. The first researchers were Watson [13] and Nadaraya [14]. The regression curve is approximated by the kernel function, the regression curve estimate can be written in equation (1).

\[ g_\alpha(t) = n^{-1} \sum_{i=1}^{n} \frac{K_\alpha(t-t_i)}{n^{-1} \sum_{i=1}^{n} K_\alpha(t-t_i)} y_i = n^{-1} \sum_{i=1}^{n} W_{\alpha i}(t) y_i \]  

(1)

If the form of the addition in the equation is described completely, it can be written in equation (2).

\[ g_\alpha(t) = n^{-1} W_{\alpha 1}(t) y_1 + n^{-1} W_{\alpha 2}(t) y_2 + \ldots + n^{-1} W_{\alpha n}(t) y_n \]  

(2)

If notated in the form of a matrix, it will be like equation (3).

\[ g_\alpha(t) = V(\alpha) y \]  

(3)

2.2. Spline Truncated Nonparametric Regression

Spline Truncated is a very popular estimator that is often used, because it has a good visual interpretation, can handle smooth functions, and is flexible [1] [15]. In general, the Spline Truncated function with degrees and knots is a function that can be written in equation (4).

\[ h(z_i) = \sum_{j=0}^{M} \alpha_j z_i^j + \sum_{k=1}^{r} \beta_k (z_i - \xi_k)^M \]  

(4)

with truncated function

\[ (z_i - \xi_k)_+^M = \begin{cases} (z_i - \xi_k)^M, & z_i \geq \xi_k \\ 0, & z_i \leq \xi_k \end{cases} \]

In general, the Spline Truncated model can be written in equation (5).

\[ y_i = \sum_{j=0}^{M} \alpha_j z_i^j + \sum_{k=0}^{r} \beta_k (z - \xi_k)_+^M + \varepsilon_i \]  

(5)

Equation (5) can be denoted in the form of a matrix written in equation (6).
\[ h(z_i) = G(\xi)\theta \]  

(6)

### 2.3. Fourier Series

Fourier series is an estimator whose regression curve shows sine and cosine waves and has high flexibility. If the data being investigated has an unknown pattern and there is a seasonal trend, the Fourier series estimator can be used [12] [16]. Recently, many alternatives have been developed and studied by researchers. The first researcher of the Fourier series was De Jong [17]. The Fourier series equation is in equation (7).

\[
f(x_i) = bx_i + \frac{1}{2} \alpha_0 + \sum_{k=1}^{K} \alpha_k \cos k x_i \tag{7}
\]

where \(b, \alpha_0, \alpha_k, k = 1,2,\ldots,K\) are the model parameters.

Equation (7) can be written as equation (8).

\[
f(x_i) = bx_i + \frac{1}{2} \alpha_0 + \alpha_1 \cos x_i + \alpha_2 \cos 2x_i + \ldots + \alpha_K \cos Kx_i \tag{8}
\]

If equation (8) is written in matrix form in equation (9).

\[
f(x) = N(K)\rho \tag{9}
\]

### 2.4. Kernel Mix Estimator, Spline Truncated and and Fourier Series

In this mixed estimator, there are three predictor components, the first predictor component, the regression curve is approached using the Spline Truncated, the second uses the kernel, the third uses the Fourier series. Paired data \((z_{i1}, z_{i2}, \ldots, z_{ip}, t_{i1}, t_{i2}, \ldots, t_{iq}, x_{i1}, x_{i2}, \ldots, x_{ir}, y_i)\) with \(i = 1, 2, \ldots, n\) which has a relationship following a nonparametric regression model. The variable \(z_{i1}, z_{i2}, \ldots, z_{ip}, t_{i1}, t_{i2}, \ldots, t_{iq}, x_{i1}, x_{i2}, \ldots, x_{ir}\) is the predictor variable and \(y_i\) is the response variable. In general, the nonparametric regression model for the mixture of the kernel, spline, and Fourier series is in equation (10).

\[
y_i = \mu(z_{i1}, z_{i2}, \ldots, z_{ip}, t_{i1}, t_{i2}, \ldots, t_{iq}, x_{i1}, x_{i2}, \ldots, x_{ir}) + \varepsilon_i
\]

(10)

where \(z_i = (z_{i1}, z_{i2}, \ldots, z_{ip})', t_i = (t_{i1}, t_{i2}, \ldots, t_{iq})', \) and \(x_i = (x_{i1}, x_{i2}, \ldots, x_{ir})'\). Random error \(\varepsilon_i\) is normally distributed with \(\mu(\varepsilon_i) = 0\) and \(\text{Var}(\varepsilon_i) = \sigma^2\). The regression curve of \(\mu(z_i, t_i, x_i)\) is assumed to be additive, so it can be written as in equation (11).

\[
\mu(z_{i1}, t_{i1}, x_i) = h(z_{i1}) + h(z_{i2}) + \ldots + h(z_{ip}) + g(t_{i1}) + g(t_{i2}) + \ldots + g(t_{iq}) + f(x_{i1}) + f(x_{i2}) + \ldots + f(x_{ir}) \tag{11}
\]

The regression curve \(\mu(z_{i1}, t_{i1}, x_i)\) is called a nonparametric mixed regression curve which is grouped into three components of the regression curve, namely the Spline Truncated, kernel, and Fourier series regression curve components. The equation can be written like equation (12).

\[
\mu(z_{i1}, t_{i1}, x_i) = \sum_{i=1}^{K} h(z_{ip}) + \sum_{i=1}^{q} g(t_{iq}) + \sum_{i=1}^{r} f(x_{ir}) \tag{12}
\]
3. Result and Discussion

This section discusses the results of the modeling process using a mixed nonparametric regression of the kernel, Spline Truncated, and Fourier series which will be applied to the human development index case in East Java Province in 2017.

Table 1 shows that each Regency / City in East Java Province has different characteristics for all variables, namely Human Development Index, morbidity, APM SMA, GRDP Per Capita.

| Variable | Number of Observations | Min    | Max    | Range  | Mean   | Standard Deviation |
|----------|------------------------|--------|--------|--------|--------|--------------------|
| (1)      | (2)                    | (3)    | (4)    | (5)    | (6)    | (7)                |
| Y        | 38                     | 59.90  | 81.07  | 21.17  | 70.35  | 5.31               |
| X1       | 38                     | 8.48   | 27.04  | 18.56  | 14.33  | 3.72               |
| X2       | 38                     | 34.22  | 84.33  | 50.11  | 63.75  | 11.89              |
| X3       | 38                     | 16.96  | 408.66 | 391.70 | 51.84  | 66.33              |

Identification using a scatter plot will determine the pattern of the relationship between the response variable and each predictor variable in Figure 1.

**Figure 1.** Scatter plot for each response variable with predictor variables

In Figure 1a, it can be seen that the relationship pattern of $Y$ and $X_1$ shows a relationship pattern that does not follow a certain pattern so that it can use the Fourier series nonparametric regression approach. Figure 1b shows that the relationship pattern of $Y$ and $X_2$ shows a relationship pattern that does not follow a certain pattern so that it can use the nonparametric spline regression approach. Figure 1c shows the relationship pattern of $Y$ and $X_3$ shows a relationship pattern that does not follow a certain pattern so that it can use the nonparametric kernel regression approach.
Judging from the scatterplot above, the mixed estimation model of the Kernel, Spline Truncated, and Fourier series is an appropriate method, because the pattern of the relationship between predictor variables and response variables is unknown, changes in certain sub-intervals, and there is a loop with a downward trend. The mixed estimation model of the Kernel, Spline, and Fourier series are selected by looking at the smallest GCV. This study uses knot points 1 to 3 and oscillations 1 to 3 so that the smallest GCV and MSE values, and maximum $R^2$, optimal knot points, bandwidth, and oscillations, as well as the parameter values of the best models.

The results obtained from the HDI modeling process on the factors that are thought to influence it, namely morbidity, APM SMA, and GRDP per capita using a mixed estimator of the kernel, spline, and Fourier series with one oscillation and one-knot point are shown in Table 2. Table 2 shows that the smallest GCV and MSE values are 6.6273 and 28.3273, respectively. In Table 3, the GCV and MSE values are generated using two knots and two oscillations. Table 3 shows that the smallest GCV and MSE values are 14.9144 and 7.1923, respectively.

Table 2. GCV and MSE values with one-knot point and one oscillation

| No | $\xi_1$ | $\phi$ | $k$ | MSE  | GCV  |
|----|---------|--------|-----|------|------|
| 1  | 35.5049 | 7.9939 | 1   | 7.1773 | 16.2666 |
| 2  | 36.7897 | 7.9939 | 1   | 7.1773 | 17.6161 |
| 3  | 38.0746 | 7.9939 | 1   | 7.1773 | 18.4572 |
| 4  | 35.5049 | 7.9939 | 1   | 7.1773 | 19.2121 |
| 5  | 39.3595 | 7.9939 | 1   | 7.1773 | 19.8570 |

Table 3. GCV and MSE values with two knots and two oscillations

| No | $\xi_1$ | $\xi_2$ | $\phi$ | $k$ | MSE  | GCV  |
|----|---------|---------|--------|-----|------|------|
| 1  | 44.4990 | 48.3536 | 7.9939 | 2   | 7.1923 | 14.9144 |
| 2  | 50.9233 | 59.9174 | 7.9939 | 2   | 6.9389 | 14.9214 |
| 3  | 53.4931 | 61.2023 | 7.9939 | 2   | 7.0015 | 14.9541 |
| 4  | 48.3536 | 63.7721 | 7.9939 | 2   | 7.0298 | 14.9702 |
| 5  | 40.6444 | 44.499  | 7.9939 | 2   | 7.0561 | 14.9786 |

Table 4. GCV and MSE values with three knots and three oscillations

| No | $\xi_1$ | $\xi_2$ | $\xi_3$ | $\phi$ | $k$ | MSE  | GCV  |
|----|---------|---------|---------|--------|-----|------|------|
| 1  | 53.4931 | 58.6326 | 53.4931 | 7.9939 | 3   | 6.5369 | 10.2322 |
| 2  | 47.0687 | 59.9174 | 65.0569 | 7.9939 | 3   | 6.3407 | 10.2618 |
| 3  | 41.9292 | 63.7720 | 70.1964 | 7.9939 | 3   | 6.2501 | 10.2800 |
| 4  | 47.0687 | 53.4931 | 70.1964 | 7.9939 | 3   | 6.5296 | 10.2803 |
| 5  | 53.4931 | 66.3418 | 70.1964 | 7.9939 | 3   | 6.8384 | 10.2947 |

In Table 4, the GCV and MSE values are generated using three knots and three oscillations. Table 4 shows that the smallest GCV and MSE values are 10.2322 and 6.3569, respectively. The comparison of the minimum MSE and GCV values obtained from one-knot point and one oscillation, two knot points and two oscillations, and three knot points and three oscillations are shown in Table 5.
Table 5. GCV, MSE, and R² values for each model

| Model                        | MSE   | GCV   | R²   |
|------------------------------|-------|-------|------|
| 1 Knot Point and 1 Oscillation | 6.6273 | 28.3273 | 75.8877 |
| 2 Knot Point and 2 Oscillation | 7.1923 | 14.9144 | 73.8321 |
| 3 Knot Point and 3 Oscillation | 6.3570 | 10.2328 | 75.1041 |

Based on Table 5, it can be seen that the smallest MSE and GCV values, as well as the largest R², are the three-point knot and three oscillation models with an MSE value of 6.3570, the GCV value is 10.2328, and R² is 75.1041 so it shows that the best model used in the case of HDI in East Java Province is with using a mixed regression model of the spline, kernel, and Fourier series with three knots and three oscillations. The parameter estimates are in Table 6.

The model is formed in equation (13).

$$
\hat{y} = \bar{a}_{0} + \alpha_{11}z_{i1} + \beta_{11}(z_{i1} - \xi_{11})_{+} + \beta_{21}(z_{i1} - \xi_{21})_{+} + \beta_{31}(z_{i1} - \xi_{31})_{+} + n^{-1} \sum_{i=1}^{n} \frac{K_{\phi}(t - t_{j})}{n^{-1} \sum_{j=1}^{n} K_{\phi}(t - t_{j})} y_{i} + \hat{b}x_{ii} + \frac{1}{2} \hat{a}_{0} \hat{w} + \sum_{k=1}^{3} \alpha_{i} \cos kx_{ii}
$$

(13)

Table 6. Estimated parameters

| Parameter | Estimate | Parameter | Estimate |
|-----------|----------|-----------|----------|
| a₀        | -5.280   | a₀        | -0.143   |
| a₁        | 0.070    | a₁        | -0.327   |
| β₁₁       | 0.855    | Φ         | 7.994    |
| β₂₁       | -1.035   | ξ₁₁       | 53.493   |
| β₃₁       | 0.295    | ξ₂₁       | 58.633   |
| b         | 0.000    | ξ₃₁       | 66.342   |

Thus, the mixed nonparametric regression estimator model of Spline Truncated, Kernel, and Fourier series in equation (14).

$$
\hat{y} = -5.280 + 0.070z_{i1} + 0.855(z_{i1} - 53.493)_{+} + 1.035(z_{i1} - 58.633)_{+} + 0.295(z_{i1} - 66.342)_{+} + \frac{1}{7.994} \sum_{i=1}^{3} \frac{K(t - t_{i})}{7.994} y_{i} + 0.072 + \sum_{i=1}^{3} \cos kx_{ii}
$$

(14)

where K is the Gaussian kernel function. With R² is 75.1041. Visualization of y and \( \hat{y} \) values for each possible knot point and the tested oscillations are as follows:
Figure 2. Visualization of values from 1 knots and 1 oscillations

Figure 3. Visualization of values from 2 knots and 2 oscillations

Figure 4. Visualization of values from 3 knots and 3 oscillations

It can be seen that Figure 4, Y the prediction is more similar to the actual Y, compared to Figure 2 and Figure 3.
4. Conclusions

Based on the results and discussion, it can be concluded that the case modeling of District / City Human Development Index in East Java Province in 2017 using a nonparametric regression approach of a mixture of the kernel, Spline Truncated, and Fourier series with three knots and three oscillations gave quite good results because the coefficient of determination of the estimator is 75.1041%, with the estimator model is

\[ \hat{y} = -5.280 + 0.070z_{it} + 0.855(z_{it} - 53.493) - 1.035(z_{it} - 58.633) + 0.295(z_{it} - 66.342) + \sum_{i=1}^{38} \frac{1}{7.994} K(t - t_i) \sum_{j=1}^{38} \frac{1}{7.994} K(t - t_j) y_{it} - 0.072 + \sum_{i=1}^{3} (-0.327) \cos kx_{it} \]

References

[1] Eubank R L 1999 Nonparametric Regression and Spline Smoothing Second Edition New York: Marcel Dekker
[2] Saputro D R S, Sukmayanti A, Widyaningsih P 2019 The nonparametric regression model using Fourier series approximation and penalized least squares (PLS) (case on data poverty in East Java) *Journal of Physics: Conf. Series*, **1188** 012019
[3] Saputro D R S, Demu K R, and Widyaningsih, P 2018 Nonparametric Spline Truncated Regression on the Data of Human Development Index (HDI) in Indonesia *Journal of Physics: Conference Series* **1028** (1) 012219
[4] Budiantara I N, Mariati N P A M, Ratnasari V, Zain I, Ratna M, Sudiarsa I W, Mardianto M F F and Hendayanti N P N 2015 Comparison Spline Truncated and Fourier Series in Multivariable Nonparametric Regression Models (Application: Data of Poverty in Papua, Indonesia) *International Journal of Basic and Applied Sciences* **15** 9-12
[5] Mariati N P A M, Budiantara I N and Ratnasari V 2019 Modeling Poverty Percentages in the Papua Islands using Fourier Series in Nonparametric Regression Multivariable *Journal of Physics: Conference Series* **1397** (1) 012071
[6] Sudiarsa I W, Budiantara I N and Purnami S W 2015 Combined estimator Fourier series and spline truncated in multivariable nonparametric regression *Applied Mathematical Sciences* **9** (100) 4997-5010
[7] 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
[8] Hardle W 1994 *Applied Nonparametric Regression* Berlin: Springer-Verlag
[9] Amini M and Roozbeh M 2015 Optimal partial ridge estimation in restricted semiparametric regression models *Journal of Multivariate Analysis* **136** 26-40
[10] Budiantara I N 2009 *Spline in Nonparametric and Semiparametric Regression: A Present and Future Model of Statistics*, Inagural Speech fot Professor’s Position at the Department of Statistics Surabaya: ITS.
[11] Asrini L J and Budiantara I N 2014 Fourier Series Semiparametric Regression Models (Case Study: The Production of Law Land Rice Irrigation in Central Java). *ARPN Journal of Engineering and Applied Sciences* **9** 1501-1506
[12] Bilodeau M 1992 Fourier Smoother and Additive Models *The Canadian Journal of Statistics* **3** 257-259
[13] Watson G S 1964 Smooth regression analysis *Sankhya: The Indian Journal of Statistic, Series A* **26** (4) 359-372
[14] Nadaraya E A 1964 On Estimating regression. *Theory of Probability and its Applications* 9 (1) 1412

[15] Budiantara I N 2007 Inferensi Statistik untuk Model Spline *Jurnal Mat-Stat* 7 1-14

[16] Tripena A and Budiantara I N 2006 Fourier Estimator in Nonparametric Regression *International Conference On Natural and Applied Natural Sciences. Ahmad Dahlan University*

[17] De Jong P 1977 The Fast Fourier Transform Spectral Estimator *Journal of Statistical Soc.* 39 (3) 327-330