Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/130318

An Extended Model of Penalized Spline with The Addition of Kernel Functions in Nonparametric Regression Model

Rahmat Hidayat[1,2], *I. Nyoman Budiantara*[2,*], *Bambang W. Otok*[2] *and Vita Ratnasari*[2]

[1] Department of Mathematics, Faculty of Science, Cokroaminoto Palopo University, South Sulawesi/ 91913, Indonesia
[2] Department of Statistics, Faculty of Mathematics, Computation and Data Science, Institut Teknologi Sepuluh Nopember, Surabaya/ 60111, Indonesia

Received: 2 Feb. 2019, Revised: 12 Apr. 2019, Accepted: 18 Apr. 2019
Published online: 1 May 2019

Abstract: The decline in the unemployment rate is an indication of the success of economic development in a country, so it needs to be managed to a stability point. The unemployment rate is influenced by various factors (predictor variables). One of the most widely used models if there are many predictor variables is regression model, which is how to know the pattern of functional relationships between one response variable with one or more predictor variables. In this study, a nonparametric regression model is developed by expanding the smoothing Spline model. Based on the results obtained, the model can capture cases of unemployment rate well and can make good predictions based on the data obtained.

Keywords: Economic, unemployment, regression, smoothing Spline, prediction

1 Introduction

In general, the national and regional development puts more emphasis on the economy in order to improve the standard of living and prosperity among the community. Specific understanding of the characteristics and potential differences of a region is important to be considered in implementing the economic development at both regional and national levels. Every region is demanded to be able to carefully identify its characteristics and potentials so that the objective of economic development can be achieved and right on the target. East Java is the third province that becomes the national development barometer, the first is Jakarta and followed by West Java. High economic growth is the main target of every region in implementing national development. The rate of Gross Domestic Product (GDP) is one of the indicators used to measure the magnitude of economic growth at the national level, while the rate of Regional Domestic Product (GRDP) is used for the regional level. In addition, another important thing that becomes a benchmark for economic development in a country is the unemployment rate.

East Java Economy in 2015, measured using Gross Regional Domestic Product (GDP) on the basis of current prices, reached IDR. 1,689.88 trillion while the GDP at constant prices reached IDR. 1,331.42 trillion. There was a decrease in the growth of East Java Economy from 5.86% in 2014 to 5.44 in 2015. From all production categories, only the supply of electricity and gas experienced a decrease by 3.00%. The highest growth rate was found in mining and quarrying sectors at 7.92%; followed by provision of accommodation food and beverage at 7.91%. In terms of expenditure, the highest growth rate was achieved by the inter-regional export net at 13.39%. The unemployment problem also causes the level of national income and the level of prosperity of the people not to reach maximum potential. The high unemployment rate in Indonesia is caused by the increasing population which is not balanced with the growth of existing business fields. Based on Statistics Indonesia (BPS) publication, it was noted that the unemployment rate in East Java had increased by 0.28% compared to the previous year.

Econometrics is a branch of science integrating economics, mathematics, and statistics. Qualitative

* Corresponding author e-mail: nyomanbudiantara65@gmail.com

economic phenomenon is mathematically modeled without verification of empirical theory. Statistics play a role in collecting data, turning mathematical models into econometric models, then statistical methods are used to estimate the parameters and suitability of the model formed.

One of the most frequently used models in econometrics is a regression model. The model is applied to figure out the pattern of functional relationships between one variable to another or between one variable and multiple variables. The use of multiple regression analysis to determine the factors that influence the unemployment rate makes an important contribution because it can be used as a consideration in government policy making, see [1,2,3,4]. Based on these studies it was obtained that factors that influence the unemployment rate include investment, number of workers, level of education, Gross Domestic Product (GDP) and economic growth. In addition to multiple regression models, the use of other statistical models such as Weibull regression models and log-linear models to find out the factors that influence the unemployment rate have been widely developed by researchers, see, among others, [5,6,7]. Based on the model, in addition to the previously mentioned factors, the Human Development Index (HDI) and income distribution also influence the unemployment rate.

Regression analysis can be used to predict and to figure out the pattern of the relationship. There are three approaches used in regression analysis in estimating the regression curve, namely parametric regression, nonparametric regression, and semi-parametric regression. Most researchers used parametric regression in modeling. There are very strong and rigid assumptions in the parametric regression approach. The form of the regression curve known includes linear, quadratic, cubic, polynomial, exponents and so on. Besides that, previous knowledge about the characteristics of the data is needed in order to obtain good modeling. So that in this study, unemployment rate is modeled by nonparametric regression. Meanwhile, in the nonparametric regression model, the form of the regression curve is assumed to be unknown. The regression curve is assumed to be smooth, which means that it is in a certain function space function (Hilbert space, Sobolev space, Banach space, Entropi space, etc) [8]. There is a difference between parametric and nonparametric approach in which the parametric approach tends to follow certain patterns, while the latter has more freedom to find its regression curve patterns which makes the nonparametric approach very flexible and objective. The regression approach that is widely used by researchers is Spline and Kernel. The benefit of using Spline is its ability to handle data patterns that change at certain intervals. Whereas if the data plot is unclear and the standard deviation is large, the Kernel approach is used, see [9].

In some real cases, there is often a different data pattern between predictors. Thus, this study develops a new method in nonparametric regression that combine smoothing Spline and Kernel approach, which we then call the Mix Spline-Kernel (M S-K). This approach is developed to be able to handle different data patterns between each predictor in which there are predictors that follow the Spline data pattern and some other predictors follow the Kernel data pattern.

2 Materials

Spline Function

Spline function is a function that is widely used in numerical methods for interpolation purposes. In statistics, this function is also used for the purposes of modeling relationships between variables. The Spline function is built based on the polynomial form. In general, the form of order polynomials m is given by:

$$\wp_m = \left\{ p(x) : p(x) = \sum_{i=1}^{m} c_i x^{i-1}, x, c_1, c_2, \ldots, c_m \in \mathbb{R} \right\} \quad (1)$$

At relatively small data intervals, polynomial functions work well. However, if the data interval is getting bigger, the polynomial function is not always able to approach the data properly [10]. One important type of polynomial cut is the polynomial Spline. Spline as a data approach was introduced by Whittaker in 1923. Whereas Spline which is based on an optimization problem, was developed by Reinch in 1967 [11,12]. Smoothing Spline is obtained if the estimation of the regression curve is resulted by minimizing the penalized least square

$$S(f) = n^{-1} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_a^b \left(f^{(m)} \right)^2 dx \quad (2)$$

with $a \leq x_1 \leq x_2 \leq \ldots \leq x_n \leq b$

The first term is the sum of the residual squares, and the second term is a penalty for function f with λ as the smoothing parameter.

Solution $S(f)$ is a natural Spline polynomial with degree $2m-1$ that no longer interpolates $y_1, y_2, \ldots, y_n$, but passes it smoothly with the smoothing parameters λ [13]. The smoothing parameter λ acts to regulate the smoothness of the function f. If $\lambda \to 0$, then the form of the function f is getting more rough by interpolating the points of observation. Conversely, if $\lambda \to \infty$, then the form of the f function approaches a straight line, as in linear regression.

Kernel Function

A function $K : R \to R$ is called a kernel function if the function is continuous, symmetrical, limited and

$$\int_{-\infty}^{\infty} K(t)dt = 1$$

From this definition, if K is a nonnegative function then K is also interpreted as a function of chance density (density function). Generally, Kernel K with bandwidth α are defined by:

$$K_\alpha(t_i) = \frac{1}{\alpha} K\left(\frac{t}{\alpha}\right); \ -\infty < t < \infty, \ \alpha > 0 \qquad (3)$$

Paired data is given (t_i, y_i) in which the relationship pattern can be expressed in the regression model $y_i = h(t_i) + \varepsilon_i$. The regression curve h is approached by the Kernel regression curve. One of the estimators that can be used to approach the h curve in nonparametric regression is to use the Kernel estimator, see [14].

$$\hat{h}_\alpha(t) = n^{-1} \sum_{i=1}^{n} W_{\alpha i}(t) y_i \qquad (4)$$

with $W_{\alpha i}(t) = \frac{\alpha^{-1} K\left(\frac{t-t_i}{\alpha}\right)}{n^{-1} \sum_{j=1}^{n} \alpha^{-1} K\left(\frac{t-t_i}{\alpha}\right)}$.

K is the Kernel function. Several types of kernel functions are commonly used: uniform, triangle, Epanechnikov, quartic, tricube, triweight, Gaussian, quadratic and cosine, but in this study the Kernel function to be used is the Gaussian Kernel

$$K(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}.$$

$\hat{h}_\alpha(t)$ is a Kernel regression estimation function, t is a predictor variable whose value is not observed but is used to estimate, t_i is the predictor variable, and α is the width of the bandwidth.

The Kernel approach depends on bandwidth α, which functions to control the smoothness of the estimation curve. Choosing the right bandwidth is very important in Kernel function [15,16,17,18]. The too large bandwidth produces a very smooth estimation curve and it leads to the mean of the variable response, whereas, if the bandwidth is too small, it produces a less smooth estimation curve in which the estimation results reach the data.

3 Method

Penalized Maximum Likelihood Estimation (PMLE)
Multiple regression analysis is used to find out the pattern of the relationship between the response variable and the predictor variable. Given a set of data pairs $(x_{1i}, ..., x_{pi}, t_{1i}, ..., t_{si}, y_i)$ then the general form of a nonparametric regression model is given by:

$$y_i = m(x_{1i}, ..., x_{pi}, t_{1i}, ..., t_{qi}) + \varepsilon_i \qquad (5)$$

The regression curve m in equation (5) is unknown and additive, so that it can be written in the form of:

$$m(x_{1i}, ..., x_{pi}, t_{1i}, ..., t_{si}) = \sum_{j=1}^{p} f_j(x_{ji}) + \sum_{k=1}^{s} h_k(t_{ki}) \qquad (6)$$

Function f and h are respectively approached using Spline dan Kernel functions. Function f_j is assumed to be smooth and in the Sobolev space. Estimator $\hat{m}$ by using M S-K model is obtained through PMLE optimization by completing likelihood function

$$L(a, b, \sigma^2 | \lambda, \theta, \alpha) = \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right) \right] \qquad (7)$$

with a constraint function

$$\int_{a_j}^{b_j} \left(f_j^{(m)}(x_j)\right)^2 dx_j \le \Upsilon_j, \Upsilon_j \ge 0 \qquad (8)$$

The estimation of the nonparametric regression model with the M S-K approach is obtained by the following steps:

- Step 1. Forming an additive regression curve m has an equation model (6)
- Step 2. Assume nonparametric regression curve f is in the Sobolev space $f_j \in W_2^m(a_j, b_j)$
- Step 3. Determining the f curve with the Spline function
- Step 4. Determining the h curve using the Kernel function
- Step 5. Obtaining the Penalty component:

$$\sum_{j=1}^{p} \lambda_j \int_{a_j}^{b_j} \left(f_j^{(m)}(x_j)\right)^2$$

  where λ_j is a smoothing parameter
- Step 6. Obtaining maximum likelihood function based on the results in steps 3,4 and 5
- Step 7. Completing the penalized maximum likelihood function based on the results in step 6

4 Results and Discussion

This section discusses the completion of the M S-K model. Firstly, the regression curve f is approached using Spline function, then the regression curve h is approached using Kernel function. After the estimator for each curve f and h, then the M S-K model estimator is completed.

Estimated Spline Function
The form of the regression curve f in (6) is unknown and assumed to be in the Sobolev space. Suppose Hilbert

space H is decomposed $H = H_0 \oplus H_1$, where $H_0 \perp H_1$. H_0 is finite-dimensional space on the basis $\varphi_1, \varphi_2, \ldots, \varphi_m$ and H_1 is with reproducing Kernel $\vartheta(x, x_i)$. Smoothing Spline equation is obtained by forming curve $f(x)$ in the form $L_i f$ where f is a member of H and $L_i f$ is a linear function limited to H. If v_i is a functional representation L_i then based on Riesz representation theory, it is obtained:

$L_i f = f(x_i) = \langle v_i, f \rangle$

$L_i \varphi_v = \varphi_v(x_i) = \langle v_i, \varphi_v \rangle$ and $L_i \vartheta_\ell = \vartheta_\ell(x_i) = \langle v_i, \vartheta_\ell \rangle = \langle \vartheta_\ell, \vartheta_\ell \rangle$

So that the regression curve f has the basis $\varphi_1, \varphi_2, \ldots, \varphi_n$ and $\vartheta_1, \vartheta_2, \ldots, \vartheta_n$ stated as:

$$f = \sum_{v=1}^{m} a_v \varphi_v + \sum_{\ell=1}^{n} b_\ell \vartheta_\ell \tag{9}$$

to get the smoothing Spline estimator in the multipredictors cases, an extension is made to the equation (9)

$$\mathbf{f}(\mathbf{x}) = \sum_{j=1}^{p} \sum_{v=1}^{m} a_{jv} \varphi_{jv} + \sum_{i=1,\ell=1}^{n} b_i \sum_{j=1}^{p} \theta_j \vartheta_j \tag{10}$$

or can be written in the form of vectors and matrices:

$$\mathbf{f}(\mathbf{x}) = \mathbf{T}\mathbf{a} + \mathbf{S}_\theta \mathbf{b} \tag{11}$$

with

$\mathbf{a} = \begin{pmatrix} a_1 & a_2 & \cdots & a_p \end{pmatrix}'$

$\mathbf{b} = \begin{pmatrix} b_1 & b_2 & \ldots & b_n \end{pmatrix}'$

$\mathbf{T} = \begin{pmatrix} T_1 & T_2 & \ldots & T_p \end{pmatrix}$

$\mathbf{S}_\theta = \theta_1 S_1 + \theta_2 S_2 + \ldots + \theta_p S_p$

$S_k = \left\{ \vartheta_j(x_i, x_\ell) \right\}_{i=1,\ell=1}^{n,n}$ $T_j = \left\{ \varphi_{jv}(x_{ji}) \right\}_{i=1,v=1}^{n,m}$

Estimated Kernel Function

From equation (4), Kernel functions are given in the form:

$$h_{\alpha_k}(t_{ki}) = n^{-1} \sum_{i=1}^{n} W_{\alpha_k i}(t_i) y_i \tag{12}$$

for each $k = 1, 2, \ldots, s$, for $k = 1$ we obtain $h_{\alpha_1}(t_i) = n^{-1} \sum_{i=1}^{n} W_{\alpha_1 i}(t_i) y_i$ because it applies for $i = 1, 2, ..., n$ then:

$$\begin{aligned}
h_{\alpha_1}(t_1) &= W_{\alpha_1 1}(t_1) y_1 + W_{\alpha_1 2}(t_1) y_2 + \ldots + W_{\alpha_1 n}(t_1) y_n \\
h_{\alpha_1}(t_2) &= W_{\alpha_1 1}(t_2) y_1 + W_{\alpha_1 2}(t_2) y_2 + \ldots + W_{\alpha_1 n}(t_2) y_n \\
&\vdots \\
h_{\alpha_1}(t_1) &= W_{\alpha_1 1}(t_n) y_1 + W_{\alpha_1 2}(t_n) y_2 + \ldots + W_{\alpha_1 n}(t_n) y_n
\end{aligned}$$

in the form of a matrix written:

$$\begin{pmatrix} h_{\alpha_1}(t_1) \\ h_{\alpha_2}(t_2) \\ \vdots \\ g_{\alpha_1}(t_n) \end{pmatrix} = \begin{pmatrix} n^{-1} W_{\alpha_1 1}(t_1) & \cdots & n^{-1} W_{\alpha_1 n}(t_1) \\ n^{-1} W_{\alpha_1 1}(t_2) & \cdots & n^{-1} W_{\alpha_1 n}(t_2) \\ \vdots & \ddots & \vdots \\ n^{-1} W_{\alpha_1 1}(t_n) & \cdots & n^{-1} W_{\alpha_1 n}(t_n) \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$h_1 = Z_1(\alpha_1) y$$

for $k = 2, 3, ..., s$ then we obtain the following function:

$$\begin{aligned}
\sum_{k=1}^{s} h_k &= Z_1(\alpha_1) y + Z_2(\alpha_2) y + ... + Z_s(\alpha_s) y \\
&= (Z_1(\alpha_1) + Z_2(\alpha_2) + ... + Z_s(\alpha_s)) \mathbf{y} \\
&= \mathbf{Z}(\alpha) \mathbf{y}
\end{aligned}$$

Then we obtain the following equation

$$\sum_{k=1}^{s} h_k = \mathbf{Z}(\alpha) \mathbf{y} \tag{13}$$

Then the constraint function (8) can be written as :

$$\begin{aligned}
\sum_{j=1}^{p} \lambda_j \int_{a_j}^{b_j} \left(f_j^{(m)}(x_j) \right)^2 dx_j &= \lambda \sum_{j=1}^{p} \left\| P_j f_j \right\|^2 \\
&= \lambda \sum_{j=1}^{p} \left(\mathbf{b}' \theta_j \mathbf{S}_j \mathbf{b} \right) \\
&= \lambda \mathbf{b}' \mathbf{S}_\theta \mathbf{b}
\end{aligned} \tag{14}$$

Theorem 4.1. If the sum of the squared errors from nonparametric regression model is given by equation (5), the error model is multivariate normal distribution with zero mean and $E(\varepsilon\varepsilon') = \sigma^2 I$, where $L(\mathbf{a}, \mathbf{b}, \sigma^2 | \lambda, \theta, \alpha)$ is a likelihood function, then estimator MLE for the parameter vector $\mathbf{a}$ and $\mathbf{b}$ is obtained from optimization of : $\max_{\substack{\mathbf{a} \in \mathbb{R}^{pm} \\ \mathbf{b} \in \mathbb{R}^n}} \left\{ L(\mathbf{a}, \mathbf{b}, \sigma^2 | \lambda, \theta, \alpha) \right\} =$

$\min_{\substack{\mathbf{a} \in \mathbb{R}^{pm} \\ \mathbf{b} \in \mathbb{R}^n}} \left\{ \| \mathbf{y} - \mathbf{T}\mathbf{a} - \mathbf{S}_\theta \mathbf{b} - \mathbf{Z}(\alpha) \mathbf{y} \|^2 \right\}$

Proof. Given nonparametric regression model (5), ε normal multivariate distribution with $E(\varepsilon) = 0$ and $E(\varepsilon\varepsilon') = \sigma^2 I$, then likelihood function $L(\mathbf{a}, \mathbf{b}, \sigma^2 | \lambda, \theta, \alpha)$ is given by:

$$\begin{aligned}
L(\mathbf{a}, \mathbf{b}, \sigma^2 | \lambda, \theta, \alpha) &= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\varepsilon_i^2}{2\sigma^2} \right) \right] \\
&= \left(2\pi\sigma^2 \right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \| \varepsilon \|^2 \right)
\end{aligned}$$

Based on equation (11) and (13), we obtain the likelihood function:

$$= \left(2\pi\sigma^2 \right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \| \mathbf{y} - \mathbf{T}\mathbf{a} - \mathbf{S}_\theta \mathbf{b} - \mathbf{Z}(\alpha) y \|^2 \right)$$

Using MLE method, estimator for parameter $\mathbf{a},\mathbf{b}$ is obtained from optimization:

$$\max_{\substack{\mathbf{a}\in\mathbb{R}^{pm}\\ \mathbf{b}\in\mathbb{R}^{n}}}\left\{\left(2\pi\sigma^2\right)^{-\frac{n}{2}}\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{T}\mathbf{a}-\mathbf{S}_\theta\mathbf{b}-\mathbf{Z}(\alpha)\mathbf{y}\|^2\right)\right\}$$

If likelihood function $L\left(\mathbf{a},\mathbf{b},\sigma^2|\lambda,\theta,\alpha\right)$ is transformed into logarithm:

$$l\left(\mathbf{a},\mathbf{b},\sigma^2|\lambda,\theta,\alpha\right)=\ln L\left(\mathbf{a},\mathbf{b},\sigma^2|\lambda,\theta,\alpha\right)$$

$$=-\frac{n}{2}\ln\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{T}\mathbf{a}-\mathbf{S}_\theta\mathbf{b}-\mathbf{Z}(\alpha)\mathbf{y}\|^2$$

Maximum optimization occurs when the component $\|\mathbf{y}-\mathbf{T}\mathbf{a}-\mathbf{S}_\theta\mathbf{b}-\mathbf{Z}(\alpha)\mathbf{y}\|^2$ has a minimum value, so the equation applies:

$$\max_{\substack{\mathbf{a}\in\mathbb{R}^{pm}\\ \mathbf{b}\in\mathbb{R}^{n}}}\left\{L\left(\mathbf{a},\mathbf{b},\sigma^2|\lambda,\theta,\alpha\right)\right\}=$$

$$\min_{\substack{\mathbf{a}\in\mathbb{R}^{pm}\\ \mathbf{b}\in\mathbb{R}^{n}}}\left\{\|\mathbf{y}-\mathbf{T}\mathbf{a}-\mathbf{S}_\theta\mathbf{b}-\mathbf{Z}(\alpha)\mathbf{y}\|^2\right\}\tag{15}$$

Theorem 4.2. If a regression model is given (5), the error model is multivariate normal distribution with zero mean and $E(\varepsilon\varepsilon')=\sigma^2\mathbf{I}$, and estimator PMLE for parameter $\mathbf{a},\mathbf{b}$ is obtained from optimization Theorem 1, then the MLE estimator for the regression curve mixture $\hat{\mathbf{m}}$ is given by:

$$\hat{\mathbf{m}}=\mathbf{A}_{\lambda,\alpha}\mathbf{y}$$

with

$$\mathbf{A}_{\lambda,\alpha}=\mathbf{T}\left(\mathbf{T}'\mathbf{M}^{-1}\mathbf{T}\right)^{-1}\mathbf{T}'\mathbf{M}^{-1}\left(\mathbf{I}-\mathbf{Z}(\alpha)\right)$$
$$+\mathbf{S}_\theta\mathbf{M}^{-1}\left(\mathbf{I}-\mathbf{T}\left(\mathbf{T}'\mathbf{M}^{-1}\mathbf{T}\right)^{-1}\mathbf{T}'\mathbf{M}^{-1}\right)\left(\mathbf{I}-\mathbf{Z}(\alpha)\right)+\mathbf{Z}(\alpha)$$

Proof. Based on Theorem 1 and the solution of the constraint function in equation (14), the PMLE estimator for the mixed regression $\hat{\mathbf{m}}$ is given by:

$$\max_{\substack{\mathbf{a}\in\mathbb{R}^{pm}\\ \mathbf{b}\in\mathbb{R}^{n}}}\left\{n^{-1}(\mathbf{y}-\mathbf{T}\mathbf{a}-\mathbf{S}_\theta\mathbf{b}-\mathbf{Z}(\alpha)\mathbf{y})'\right.$$
$$\left.(\mathbf{y}-\mathbf{T}\mathbf{a}-\mathbf{S}_\theta\mathbf{b}-\mathbf{Z}(\alpha)\mathbf{y})+\lambda\mathbf{b}'\mathbf{S}_\theta\mathbf{b}\right\}$$

Deriving the partial in equation above to $\mathbf{a}$ and $\mathbf{b}$, we obtain:

$$\hat{\mathbf{a}}=\left(\mathbf{T}'\mathbf{M}^{-1}\mathbf{T}\right)^{-1}\mathbf{T}'\mathbf{M}^{-1}\left(\mathbf{I}-\mathbf{Z}(\alpha)\right)\mathbf{y}\tag{16}$$

$$\hat{\mathbf{b}}=\mathbf{M}^{-1}\left(\mathbf{I}-\mathbf{T}\left(\mathbf{T}'\mathbf{M}^{-1}\mathbf{T}\right)^{-1}\mathbf{X}'\mathbf{M}^{-1}\right)\left(\mathbf{I}-\mathbf{Z}(\alpha)\right)\mathbf{y}\tag{17}$$

Based on equation (16),(17) and (13), the estimator for Spline component and Kernel component are obtained:

$$\hat{\mathbf{f}}=\mathbf{T}\hat{\mathbf{a}}+\mathbf{S}_\theta\hat{\mathbf{b}}\tag{18}$$

$$\hat{\mathbf{h}}=\mathbf{Z}(\alpha)\mathbf{y}\tag{19}$$

Based on equation (18) and (19), we obtain following M S-K model:

$$\hat{\mathbf{m}}=\hat{\mathbf{f}}+\hat{\mathbf{h}}$$
$$=\mathbf{T}\hat{\mathbf{a}}+\mathbf{S}_\theta\hat{\mathbf{b}}+\mathbf{Z}(\alpha)$$

$$\hat{\mathbf{m}}=\left\{\mathbf{T}\left(\mathbf{T}'\mathbf{M}^{-1}\mathbf{T}\right)^{-1}\mathbf{T}'\mathbf{M}^{-1}\left(\mathbf{I}-\mathbf{Z}(\alpha)\right)+\right.$$
$$\left.\mathbf{S}_\theta\mathbf{M}^{-1}\left(\mathbf{I}-\mathbf{T}\left(\mathbf{T}'\mathbf{M}^{-1}\mathbf{T}\right)^{-1}\left(\mathbf{I}-\mathbf{Z}(\alpha)\right)\right)+\mathbf{Z}(\alpha)\right\}\mathbf{y}$$

$$\hat{\mathbf{m}}=\mathbf{A}_{\lambda,\alpha}\mathbf{y}\tag{20}$$

5 Smoothing Parameter and Bandwidth Selection

In nonparametric regression, one of the important things is to find the estimator $\hat{\mathbf{m}}$ that is most suitable for a set of data. This is related to the smoothing parameters and appropriate bandwidth parameters. If the smoothing parameter value and bandwidth parameters are very small, then it provide a very rough estimator [11]. Conversely, if the smoothing parameter value and bandwidth parameters are very large, it will produce a very smooth estimator. As a result, smoothing parameters and optimal bandwidth parameters are needed to obtain the most suitable estimator for the data. One method in selecting the optimal smoothing parameters in the spline estimator is Cross Validation (CV) method, see [19]. The Unbiased Risk method (UBR) method can also be used to select the optimal smoothing parameters on the spline estimator [20]. An excellent method for selecting optimal smoothing parameters in the Spline-Kernel estimator, namely the general validation cross validation (GCV) method [21]. Theoretically, the GCV has asymptotic optimal properties, which do not belong to other methods, see [11]. The strength possessed by GCV makes it very well known in nonparametric and semiparametric. The GCV formula is often generalized and adjusted by researchers in other estimators to select the optimal smoothing parameters. The following is given a method for selecting smoothing parameters and optimal bandwidth in the M S-K model in multiple nonparametric regression. The goodness of fit in the estimator model M S-K is given by:

$$MSE=n^{-1}\left(\mathbf{y}-\mathbf{A}_{\lambda,\alpha}\right)'\left(\mathbf{y}-\mathbf{A}_{\lambda,\alpha}\right)$$
$$=n^{-1}\left\|\left(I-\mathbf{A}_{\lambda,\alpha}\right)\mathbf{y}\right\|^2$$

then:

$$G(\lambda,\alpha)=\frac{n^{-1}\left\|\left(I-\mathbf{A}_{\lambda,\alpha}\right)\mathbf{y}\right\|^2}{\left[n^{-1}tr\left(I-\mathbf{A}_{\lambda,\alpha}\right)\right]^2}\tag{21}$$

Smoothing parameter and optimal bandwidth are obtained by completing optimization of function $G(\lambda,\alpha)$ as presented in equation (21).

6 Application

In this study, the model obtained is applied to the unemployment rate data in East Java province in 2015, as the response variable with 38 districts/cities unit of observation. While predictor variables used are Gross Domestic regional Product (GDRP), Literacy Rate (LR), Population Growth (PG), and Rough Participation Rate (RPR). Several alternative models are used to ensure the best model is based on predictor variables.

$$\text{Model A} : y = h(GDRP) + h(LR) + g(PG) + g(RPR)$$
$$\text{Model B} : y = h(GDRP) + h(PG) + g(LR) + g(RPR)$$
$$\text{Model C} : y = h(LR) + h(PG) + g(GDRP) + g(RPR)$$
$$\text{Model D} : y = h(LR) + h(RPR) + g(GDRP) + g(PG)$$
$$\text{Model E} : y = h(PG) + h(RPR) + g(GDRP) + g(LR)$$
$$\text{Model F} : y = h(GDRP) + h(RPR) + g(LR) + g(PG)$$

The function estimation model M S-K in multiple nonparametric regression is more effective and can handle the local nature of functions or data on different domains in an integrated manner. So that the unemployment rate can be formed with multiple nonparametric regression models using several alternatives. To compare the six models, GCV criteria is used to get the lowest GCV. Comparison of GCV values is presented in the following table:

Model	GCV
Model A	$3.867 \times 10^{-4}*$
Model B	2.508×10^{-3}
Model C	3.552×10^{-4}
Model D	6.133×10^{-3}
Model E	2.836×10^{-3}
Model F	1.475×10^{-3}

Based on the goodness criteria for GCV, each model shows that the A model is the best model to estimate the unemployment rate of East Java. The model gives value of $R^2 = 87.01\%$. For comparison, the data set is analyzed using multiple linear regression. Based on the results of analysis with multiple linear regression on the unemployment rate data, the value of R^2 = 23.98% is obtained. So based on the value of R^2 it can be concluded that the M S-K model is better than multiple linear regression.

Model validation is the next stage to do after the best model formation obtained to check the accuracy of the model in predictions. Cross validation method is applied in this stage. The cross validation process applied to the multiple nonparametric regression models M S-K formed from the unemployment rate data in 2015 is used to predict the East Java unemployment rate data in 2016 using predictor variables in 2016. The estimation of unemployment rate in 2016 using the multiple nonparametric regression model M S-K can be seen in Figure 1. Based on the bar diagram, it can be seen that the results of the estimated unemployment rate in 2016 have closed to the real data.

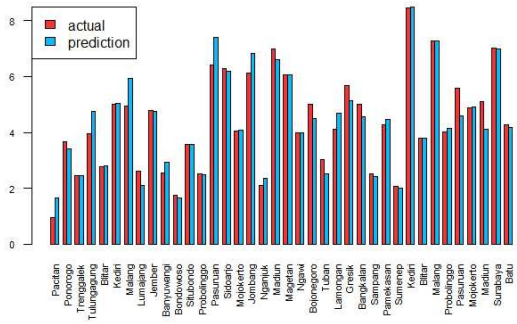

Fig. 1: Comparison between observation and prediction data

To examine the reliability of the model, statistical explorations is carried out using the Kolmogorov-Smirnov test. The test is to find the pattern similarity between the predictive data and the observation data. Based on the results of the test on prediction and observation data, it can be seen that the pattern of predictive data with observation data has the same pattern. This means that the multiple nonparametric regression model with the M S-K model approach is still valid to be applied in the unemployment rate data in 2016.

7 Conclusion

This paper introduces a new method in nonparametric regression. This model is a combination of Spline and Kernel functions. We have given explicit mathematical expressions for some basic functions such as the Spline multipredictor function, the Kernel function multipredictors, and the new model itself. Also, the method for determining optimal smoothing parameter and bandwidth parameter. Estimation of the model was approached through the method of penalized maximum likelihood estimation. The application of this model was illustrated with a real data set and the results obtained show that the model is quite good at modeling data

Acknowledgement

This research was supported by Lembaga Pengelola Dana Pendidikan (LPDP). We thank our colleagues who provided insight and expertise that greatly assisted the research.

References

[1] N. Sitompul, Analysis of The Effect of Investment and Labor on Gross Domestic Product. University of North Sumatra, Sumatra: Master Thesis (2007).

[2] Yunan, Analysis of Factors Affecting Economic Growth in Indonesia. University of North Sumatra, Sumatra: Master Thesis (2009).

[3] M. Maitah, D. Toth, and E. Kuzmenko, Exploring The Relationship between Economic Growth and Employment in The Czech Republic and Belgium, Review of European Studies, Vol. 7, No. 11, pp. 115-124 (2015).

[4] K. Malec, S. Gouda, E. Kuzmenko, D. Soleimani, H. Rezbova, and P. Sanova, Gross Domestic Product Development and Employment in Egypt, International J of Eco and Fin, Vol. 6, No. 1, pp. 199-206 (2016).

[5] J. M. Kirigia, D. Oluwole, G. M. Mwabu, D. Gatwiri, and L.H. Kainyu, Effects of Maternal Mortality on Gross Domestic Product in the WHO African Region, African J of Health Sci, Vol. 13, No. 1, pp. 86-95 (2006).

[6] G. Ranis and F. Stewart, Economic Growth and Human, Development World Dev, Vol. 28, No. 2, pp. 197-219 (2000).

[7] P. C. Sutton, C. Elvidge and T. Ghosh, Estimation of Gross Domestic Product and Unemployment at Sub-National Scales using Nighttime Satellite Imagery, International J of Ecol Eco and Stat, Vol. 8, No. 7, pp. 5-21 (2007).

[8] I. N. Budiantara, M. Ratna, I. Zain, and W. Wibowo, Modeling the Percentage of Poor People in Indonesia using Spline Nonparametric Regression Approach, International J Basic and App Sci, Vol. 12, No. 6, pp. 199-124 (2012).

[9] D. Hurley, J. Hussey, R. McKeown, and C. Addy, An evaluation of Splines in linear regression, SAS Conference Proceedings: SAS Users Group International, Vol. 147, pp. 1-11 (2006).

[10] W. Wibowo, S. Haryatmi, and I. N. Budiantara, Modeling of Regional Banking Activities using Spline Multiresponse Semiparametric Regression, International J of App Math and Stat, Vol. 44, No. 14, pp. 391-398 (2013).

[11] G. Wahba, Spline Models for Observation Data, Pennsylvania: SIAM (1990).

[12] R. L. Eubank, Spline Smoothing and Nonparametric Regression, New York: CRC Press (1999).

[13] H. Becher, G. Kauermann, P. Khomski, and B. Kouyate, Using Penalized Splines to Model Age and Season of Birth Dependent Effects of Childhood Mortality Risk Factors in Rural Burkina Faso, Biometrical Journal, Vol. 51, No. 1, pp. 110-122 (2009).

[14] W. Hardle, Applied Nonparametric Regression, New York: Cambridge University Press (1990).

[15] M. Kayri, and G. Zirhhoglu, Kernel Smoothing Function and Choosing Bandwidth for Nonparametric Regression Methods, Ozean J of App Sci, Vol. 2, No. 1, pp. 49-54 (2009).

[16] W. J. Braun and L. S. Huang, Kernel Spline Regression, Canadian J of Stat, Vol. 33, No. 2, pp. 259-278 (2005).

[17] M. Y. Cheng, R. L. Paige, S. Sun, and K. Yan, Variance Reduction for Kernel Estimators in Clustered/longitudinal Data Analysis, Journal of Stat Plan and Inf, Vol. 140, No. 6, pp. 1389-1397 (2010).

[18] Y. P. Chaubey, N. Laib, and J. Li, Generalized Kernel Regression Estimator for Dependent Size-Blazed Data, Journal of Stat Plan and Inf, Vol. 142, No. 3, pp. 708-727 (2011).

[19] P. Craven, and G. Wahba, Smoothing Noisy Data with Spline Functions, Numerische Mathematik, Vol. 31, No. 4, pp. 377-403 (1978).

[20] J. Opsomer, Y. Wang, and Y. Yang, Nonparametric Regressin with Correlated Errors, Statistical Sci, Vol. 16, No. 2, pp. 134-153 (2001).

[21] R. Hidayat, I. N. Budiantara, B.W. Otok and V. Ratnasari, A Reproducing Kernel Hilbert Space Approach and Smoothing Parameters Selection in Spline-kernel Regression, Journal of Theo and App Inf Tech, Vol. 97, No. 2, pp. 465-475 (2019).

Rahmat Hidayat is a Lecturer at Department of Mathematics, Faculty of Sciences, Cokroaminoto Palopo University, Indonesia. He received the Magister degree in Applied Mathematics at Bogor Agricultural University, Indonesia. His research interest are in the areas of Applied Mathematics and Statistics more specifically in the field of nonparametric regression and survival analysis. His research has been published in National and International Journals.

I Nyoman Budiantara is Professor of Statistics at Department of Statistics, Faculty of Mathematics, Computation and Data Sciences, Institut Teknologi Sepuluh Nopember, Indonesia. He received the PhD degree in Mathematics and Natural Sciences at Gadjah Mada University, Indonesia. His research interest is Nonparametric-Semiparametric Regression and its applications. The result of his research are published in many national and international journals.

Bambang W. Otok is Associate Professor at Department of Statistics, Faculty of Mathematics, Computation and Data Science, Institut Teknologi Sepuluh Nopember Surabaya, Indonesia. He received the PhD degree in Mathematics and Natural Sciences at Gadjah Mada University, Indonesia. His research interests are in the areas of multivariate statistics, structural equation model, and re-sampling analysis.

Vita Ratnasari is Senior Lecturer at Department of Statistics, Faculty of Mathematics, Computation and Data Science, Institut Teknologi Sepuluh Nopember Surabaya, Indonesia. He received the PhD degree in Statistics at Institut Teknologi Sepuluh Nopember Surabaya, Indonesia. Her research interests are in the areas of regression analysis, and categorical data analysis.