THE PARAMETRIC AND NONPARAMETRIC ESTIMATOR IN SEMIPARAMETRIC REGRESSION FOR LONGITUDINAL DATA WITH SPLINE APPROACH

[a]Kuzairi*, [b]Noer Azizah, [c]M. Fariz Fadillah Mardianto, [d]Ira Yuditira, [e]Faisol, [f]Tony Yulianto, [g]Rica Amalia

[a,b,d,e,f,g] Departement of Mathematics, Universitas Islam Madura, Pamekasan
[c]Departement of Mathematics, University of Airlangga, Surabaya
E-mail: kuziri@uim.ac.id, noerazizahone@gmail.com, m.fariz.fadillah.m@fst.unair.ac.id,
irayudistira91@gmail.com, faisol.munif@gmail.com, toniyulianto65@gmail.com,
ricaamalia5@gmail.com

Abstract

Regression analysis aims to determine the relationship between response variables and predictor variables. There are three approaches to estimate regression curves, there are parametric, nonparametric, and semiparametric regression. In this study, the form of spline semiparametric regression curve estimator for longitudinal data assessed. Based on the estimator that be obtained by using Weighted Least Square (WLS) optimization applied to model electricity consumption in Madura by choosing a model for longitudinal data based on linear spline estimator with two knot. The good criterion of the model is using the GCV value, the coefficient of determination and the value of MSE. The best model is a model that has a high coefficient of determination and a small MSE value. This spline model has a determination coefficient value of 99,72911% and MSE 32,50458.

Key words: Semiparametric Regression, Longitudinal Data, Spline, GCV (Generalized Cross Validation), Electricity Consumption in Madura.

INTRODUCTION

Regression analysis is a study used to determine the relationship pattern between response variables and predictor variables [1][2]. There are three approaches to determine the shape of the regression curve, there are parametric, nonparametric and semiparametric. In the parametric regression model it is assumed that the function patterns are known as linear, quadratic, cubic, polynomial, exponential, and many more. While the nonparametric regression model is assumed

that the function pattern is not known like spline, neural network, kernel, polynomial, wavelet, histogram, MARS, Fourier series and others. If a regression model contains parametric components and other nonparametric estimated components then a semiparametric model [3][4].

In this research, we use semiparametric regression approach with linier estimator for parametric component and spline estimator for nonparametric component. Spline estimator has a high flexibility and ability to estimate data behavior that tends to differ at different

intervals and spline is a model providing superior and very respectable visually statistical interpretation [5]. Therefore, the spline method developed in the last decade. Pratiwi[6] introduce a semiparametric regression to estimate the average of age at first marriage in East Java with linier spline estimator. Budiantara [7] developed a spline estimator in nonparametric regression by using a base spline function family. Truncated spline estimator gives easier and simpler mathematical calculations than other estimator.

Along with the development of data analysis in big data era, study about regression analysis requires not only cross section data that often be used, but also longitudinal data. Longitudinal data consists of cross section and time series data. Research with longitudinal data is more reliable in finding answers about the dynamics of change. Longitudinal data potentially provides more complete information. Another advantage of using longitudinal data, we can know the changes that occur in a subject, because the observations are repeated for each subject [8]. Some semiparametric regression studies that have used longitudinal data include spline [9], mixed-effects modeling [10] and kernel logistic [11].

However, a study for longitudinal data with spline estimator is needed to accommodate repetitive data patterns, in this case seasonal, periodic, and seasonal trend combinations.

This manuscript consists of four parts. This part is the first part which give explanation how important to determine spline estimator in semiparametric regression for longitudinal data. The first part also gives some motivation based on previous study. The second part presents study about longitudinal data structure, the form of spline estimator in semiparametric regression for longitudinal data generally and matrices equation. Based on matrices equation, Weighted Least Square (WLS) optimization is done to determine parameter vector estimator. This section is ended by present estimator form from spline in semiparametric regression for longitudinal data. An application case based on simulation data with spline estimator in semiparametric regression for longitudinal data is result and discussion in the third part. The fourth part gives conclusion based on the result.

MATERIAL AND METHODS

Consider a longitudinal data structure that be presented in Table 1 as follows:

Table 1 Longitudinal data structure in semiparametric regression

Subject	Response	Predictors					
		Parametric			Nonparametric		
	y_{ij}	x_{ij1}	...	x_{ijp}	t_{ij1}	...	t_{ijq}
1st Subject	y_{11}	x_{111}	...	x_{11p}	t_{111}	...	t_{11q}
	y_{12}	x_{121}	...	x_{12p}	t_{121}	...	t_{12q}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$y_{1.n}$	$x_{1.n.1}$	...	$x_{1.n.p}$	$t_{1.n.1}$	...	$t_{1.n.q}$
2nd Subject	y_{21}	x_{211}	...	x_{21p}	t_{211}	...	t_{21q}
	y_{22}	x_{221}	...	x_{22p}	t_{221}	...	t_{22q}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$y_{2.n}$	$x_{2.n.1}$	...	$x_{2.n.p}$	$t_{2.n.1}$	...	$t_{2.n.q}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
nnd Subject	y_{m1}	x_{m11}	...	x_{m1p}	t_{m11}	...	t_{m1q}
	y_{m2}	x_{m21}	...	x_{m2p}	t_{m21}	...	t_{m2q}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$y_{m.n}$	$x_{m.n.1}$	...	$x_{m.n.p}$	$t_{m.n.1}$	...	$t_{m.n.q}$

Based on Table 1, there are pairs of data with form $(y_{ij}, x_{ijp}, t_{ijr})$, x_{ijp} denotes p^{th} parametric predictor variable and t_{ijr} denotes r^{th} nonparametric predictor variable for j^{th} observation in i^{th} subject. Here, $i = 1,2, \dots, n$ denote the number of subjects, $j = 1,2, \dots, n_i$ denote the number of observations for each subject, and p represents the number of predictor for parametric component, r represent the number of predictor for nonparametric component. Response variable for j^{th} observation in i^{th} subject is denoted by y_{ij}. The pairs of data that be presented in Table 1, follows semiparametric regression equation for longitudinal data.

$$y_{ij} = \sum_{k=1}^{p} \beta_{kj} x_{kij} + \sum_{r=1}^{R} f_r(t_{rij}) \quad (1)$$
$$+ \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma^2)$$

$f_r(t_{rij})$ represents a regression curve. Random error for j^{th} observation in i^{th} subject is denoted by ε_{ij} that independent, identically normal distributed with mean 0, and variance σ^2. In this case, $f_r(t_{rij})$ approached by spline function as follows:

$$f(t_{rij}) = \sum_{s=1}^{m} \alpha_{rs} t_{rij}^s$$
$$+ \sum_{l=1}^{q} \gamma_l (t_{rij} \quad (2)$$
$$- K_{rl})_+^m$$

Equation (2) is substituted to equation (1), the result is a semiparametric regression equation for longitudinal data that be approached by spline function as follows:

$$y_{ij} =$$
$$\sum_{k=1}^{p} \beta_{kj} x_{kij} + \sum_{r=1}^{R} \left\{ \sum_{s=1}^{m} \alpha_{rs} z_{rij}^s \right.$$
$$+ \sum_{l=1}^{q} \gamma_l (z_{rij} \quad (3)$$
$$\left. - K_{rl})_+^m \right\}$$

$$y_{ij} =$$
$$\sum_{k=1}^{p} \beta_{kj} x_{kij} + \sum_{r=1}^{R} \left\{ \sum_{s=1}^{m} \alpha_{rs} z_{rij}^s \right.$$
$$+ \sum_{l=1}^{q} \gamma_l (z_{rij} \quad (4)$$
$$\left. - K_{rl})_+^m \right\}$$

with s is the order of the spline, l as the representation of the knot point, q the number of knots. K_{rl} value of point knot to-l, m is the number of spline order. For spline function with $m = 1$ is called linear spline function, while $m = 2$ is called quadratic spline function, and so on. In this case $\alpha_{11}, \alpha_{12}, \dots, \alpha_{1m}, \alpha_{21}, \alpha_{22}, \dots, \alpha_{Rm}, \gamma_1, \gamma_2, \dots, \gamma_q$ are unknown parameters and their value is estimated, and truncated function $(z_{rij} - K_{rl})_+^m$ is defined as.

$$(z_{rij} - K_{rl})_+^m = \begin{cases} (z_{rij} - K_{rl})^m, z_{rij} \geq K_{rl} \\ 0, z_{rij} < K_{rl} \end{cases}$$

Equation (3) can be presented to matrices form as follows:

$$y = X\beta + T\eta + \varepsilon; \varepsilon \sim N(0, \sigma^2 I) \quad (4)$$

with $y = [y_1 \quad y_2 \quad \cdots \quad y_n]^T$, where $y_i = [y_{i1} \quad y_{i2} \quad \cdots \quad y_{in_i}]^T$ as vectors that include response data. Vectors that include random error denoted by $\varepsilon = [\varepsilon_1 \quad \varepsilon_2 \quad \cdots \quad \varepsilon_n]^T$, where $\varepsilon_i = [\varepsilon_{i1} \quad \varepsilon_{i2} \quad \cdots \quad \varepsilon_{in_i}]^T$. ε is independent, identically normal distributed with mean 0, and variance $\sigma^2 I$, I is an identity matrix. Vectors that include regression parameters denoted by $\beta = [\beta_0 \quad \beta_1 \quad \cdots \quad \beta_n]^T$, Vectors that include regression predictor denoted by

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{31} & \cdots & x_{np} \end{bmatrix}$$

In this case, $T[K]$ is a matrix with structure as follows:

$$T[K] = \begin{bmatrix} T_1[K] & 0 & \cdots & 0 \\ 0 & T_2[K] & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & T_p[K] \end{bmatrix}$$

Where

$$T_i[K] = \begin{bmatrix} t_{i11}^1 & \cdots & t_{in_i1}^m & (t_{i11} - K_{1i1})_+^m & \cdots & t_{i1p}^1 & \cdots & t_{i1p}^m & (t_{i1p} - K_{1ip})_+^m \\ t_{i21}^1 & \cdots & t_{i21}^m & (t_{i21} - K_{1i1})_+^m & \cdots & t_{i2p}^1 & \cdots & t_{i2p}^m & (t_{i2p} - K_{1ip})_+^m \\ \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ t_{in_i1}^1 & \cdots & t_{in_i1}^m & (t_{in_i1} - K_{1i1})^m & \cdots & t_{in_ip}^1 & \cdots & t_{in_ip}^m & (t_{in_ip} - K_{1ip})_+^m \end{bmatrix}$$

$$\eta = [\eta_1 \quad \eta_2 \quad \cdots \quad \eta_n]^T$$

where

$$\eta_i = (\alpha_{1i1} \quad \cdots \quad \alpha_{mi1} \quad \eta_{1i1} \quad \cdots \quad \eta_{qi1} \quad \alpha_{1ir} \quad \cdots \quad \alpha_{mir} \quad \eta_{1ir} \quad \cdots \quad \eta_{qir})$$

Weighted Least Square (WLS) optimization is done by using matrices form that be presented in equation (4). The form of WLS optimization is presented as follows:

$$\min_{\beta,\eta}\left[R(\beta,\eta)\right]=\min_{\beta,\eta}\left\{(y-X\beta-T\eta)^{T}W(y-f(x,t))\right\} \quad (5)$$

with

$$W=\begin{bmatrix}\frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_p^2}\end{bmatrix} \quad (6)$$

as a weighted matrix

WLS optimization to get estimator for β, is called $\widehat{\beta}$, can be done by expanding right side in equation (5), so that
$R(\beta,\eta) = \{(y - X\beta - T\eta)^{T}W(y - f(x,t))\} = 0$ that can be expanded as follows:

$$R(\beta,\eta) = y^{T}Wy - 2y^{T}W \\ - 2\eta^{T}T^{T}Wy \\ + 2\beta^{T}X^{T}WT\eta \quad (7) \\ + \beta^{T}X^{T}WX\beta \\ + \eta^{T}T^{T}WT\eta$$

For determining $\hat{\beta}$ derivation process is done with $R(\beta,\eta)$ respect to β with requirement right side $R(\beta,\eta)$ equals to zero vector. The result is an estimator for parameter vector as follows:

$$\hat{\beta} = (X^{T}WX)^{-1}\{X^{T}Wy - X^{T}WT\hat{\eta}\} \quad (8)$$

with similar procedure to get an estimator from parameter η so that it is obtained

$$\hat{\eta} = (T^{T}WT)^{-1}\{T^{T}Wy - T^{T}WX\hat{\beta}\} \quad (9)$$

Estimator in the equation (8) and equation (9) not yet parameter free, so an estimator must be searched that is free from parameters with mutual substitution. If equation (9) is substituted to equation (8), the result is

$$\hat{\beta} = M\ [\![(X^{\wedge}T\ WX)]\!] \\ {}^{\wedge}(-1)\ \{X^{\wedge}T - \\ X^{\wedge}T\ WT(T^{\wedge}T\ WT)^{\wedge}(-1)\ T^{\wedge}T\ \}Wy \quad (10) \\ = B(K)y$$

where

$$B(K) = M(X^{T}WX)^{-1}\{X^{T} \\ - X^{T}WT(T^{T}WT)^{-1}T^{T}\}W$$

with

$$M \\ = (I - (X^{T}WX)^{-1}X^{T}WT(TWT)^{-1}T^{T}WX)^{-1}$$

if equation (9) is substituted to equation (8), the result is

$$\hat{\eta} \\ = N(T^{T}WT)^{-1}\{T^{T} \\ - T^{T}WX(X^{T}WX)^{-1}X^{T}\}Wy \quad (11) \\ = C(K)y$$

where

$$C(K) = N(T^{T}WT)^{-1}\{T^{T} \\ - T^{T}WX(X^{T}WX)^{-1}X^{T}\}W$$

with

$$N = (I - (T^{T}WT)^{-1}T^{T}WX(X^{T}WX)^{-1}X^{T}WT)^{-1}$$

Based on equation (10), a weighted matrix gives affect for determining $\widehat{\beta}$. Similar with Ordinary Least Square (OLS) optimization method, WLS is not bound with an error distribution. A weighted matrix plays a role in the smoothing process for nonparametric regression with longitudinal data (G, 1990).

The spline estimator in semiparametric regression for longitudinal data is presented by equation (10) as follows:

$$\widehat{y}_{ij} = \sum_{k=1}^{p} \hat{\beta}_{kj}x_{kij} \\ + \sum_{r=1}^{R}\left\{\sum_{s=1}^{m} \hat{\alpha}_{rs}z_{rij}^{s}\right. \quad (12) \\ + \sum_{l=1}^{q} \hat{\gamma}_{l}\left(z_{rij}\right. \\ \left.\left. - K_{rl}\right)_{+}^{m}\right\} + \varepsilon_{ij}$$

Equation (13) can be presented to matrix equation as follows:

$$\hat{y} = X\hat{\beta} + T\hat{\eta} \quad (13)$$

Equation (10) and equation (11) is substituted to equation (13), the result is as follows:

$$\hat{y} = X\hat{\beta} + T\hat{\eta} = XB(K)y + TC(K)y \quad (14) \\ = D(K)y$$

$B(K)$ is a hat matrix for parametric components. $C(K)$ hat matrix for nonparametric components. Matrix hat for semiparametric regression model with spline approach is denoted by $D(K)$. $K = (K_1, \ldots, K_q)$ shows the knot point contained in the matrix $D(K)$.

The goodness indicators that often be used in semiparametric regression is Mean Square Error (MSE), Generalized Cross Validation (GCV), and determination coefficient (R^2). All of goodness indicators can be applied to splne

estimator in semiparametric regression for longitudinal data.

In spline estimator in semiparametric regression for longitudinal data, an optimal knots point (K) is determined. In determining optimal knots point can be used GCV formula. GCV often be used because have asymptotically optimal properties [9]. For determining an optimal knots point can be seen based on the smallest GCV value. The formula of GCV given as follows:

$$\text{GCV}(K) = \frac{\text{MSE}(K)}{(N^{-1}trace(I - D(K)))^2}$$

with

$$\text{MSE}(K) = N^{-1}y^T(I - D(K))^T W(I - D(K))y.$$

By choosing of an optimal knots point will give impact to produce a determination coefficient with high value, or approximate to 100%. The determination coefficient formula given as follows:

$$\text{R}^2 = \frac{(\hat{y} - \bar{y})^T(\hat{y} - \bar{y})}{(y - \bar{y})^T(y - \bar{y})} \; ; \; 0 \leq \text{R}^2 \leq 1$$

with $\hat{y}$ is a vector that include of estimation result for all of subjects, and $\bar{y}$ is a vector that include mean value for each subject. The best model that can be used for prediction met he goodness of criteria. The goodness of criteria is the smallest GCV value for an optimal knots point, the smallest mean square error (MSE) value, and the big of determination coefficient value.

In this part, study about spline estimator in semiparametric regression for longitudinal data is applied in electricity consumption data. The electricity consumption data consists of one response, and two predictors. Response variable represents monthly peak load denoted by y. The first predictor represents monthly current denoted by x, the second predictor represents observation time denoted by t. There are three subjects in this application, each subject is observed for 12 months. The data analysis procedure uses spline estimator in semiparametric regression for longitudinal data is given as follows:

1. Input pairs of data (y_{ij}, x_{ij}, t_{ij}), $i = 1,2,...,n$, and $j = 1,2,...,n_i$.
2. Presenting scatter plot for every pairs of data (y_{ij}, x_{ij}) and (y_{ij}, t_{ij}).

3. Determining an optimal knots point based on the smallest GCV.
4. Determining goodness of indicators for selectedknots point.
5. Determining the elements values from $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\eta}}$ based on equation (10) and equation (11) for selected knots point.
6. Determining spline estimator in semiparametric regression for longitudinal data based on equation (12).
7. Presenting plot between response data, and estimator data.

RESULT AND DISCUSSION

Open Source Software (OSS) R is used to analysis based on spline estimator in semiparametric regression for longitudinal data. The scatter plot that shows how the pattern of data for the Sampang Regency is presented in Figure 1. The scatter plot that shows how the pattern of data for the Pamekasan Regency is presented in Figure 2. The scatter plot that shows how the pattern of data for the Sumenep Regency is presented in Figure 3.
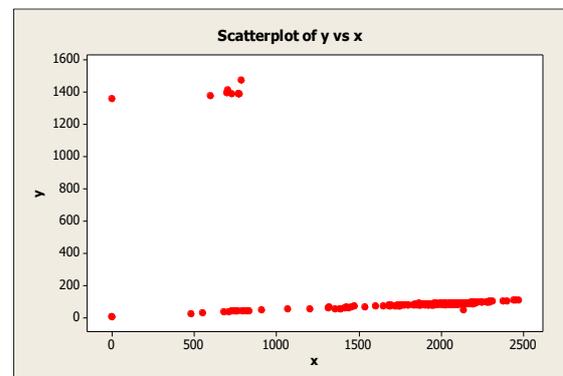

(a)
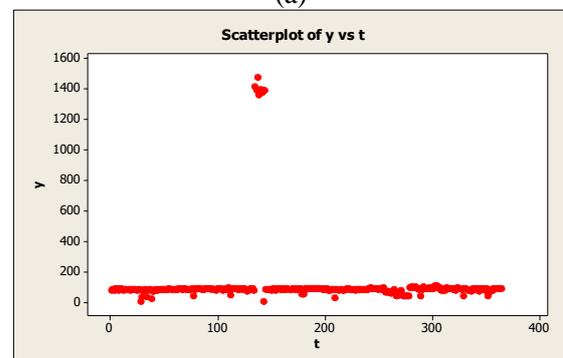
(b)

Fig 1. Scatter plot for the Sampang Regency, (a) Scatter plot between response and the first predictor, (b) Scatter plot

between response and the second predictor.
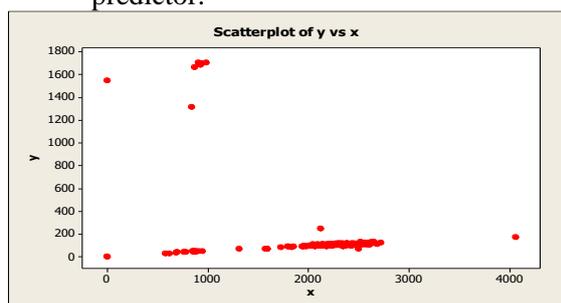

(a)
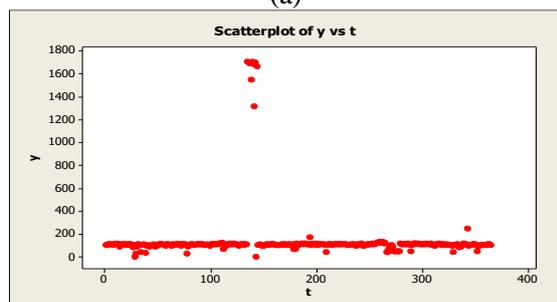
(b)

Fig 2. Scatter plot for the Pamekasan Regency, (a) Scatter plot between response and the first predictor, (b) Scatter plot between response and the second predictor
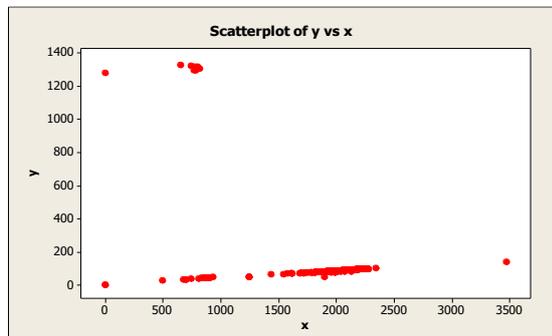

(a)
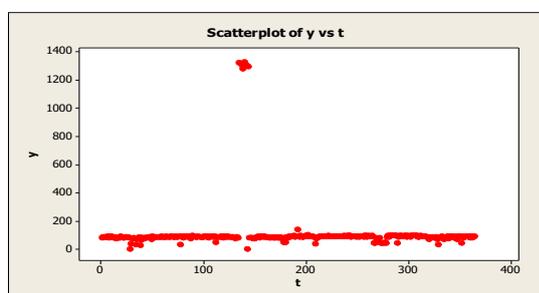
(b)

Fig 3. Scatter Plot for the Sumenep Regency, (a) Scatter plot between response and the first predictor, (b) Scatter plot between response and the second predictor

Based on Figure 1, Figure 2, and Figure 3 part (a) the pattern shape is known, so Figure 1, Figure 2, and Figure 3 part (a) includes the parametric component. Based on Figure 1, Figure 2, and Figure 3 part (b) there are unspecifict trend pattern between response and predictor, so Figure 1, Figure 2, and Figure 3 part (b) includes a nonparametric component. Time series plot for each variable can present data pattern more clearly. However, we use the data to apply spline estimator in semiparametric regression for longitudinal data.

Parameter estimates in this study depend on the optimum knot value obtained by considering the minimum GCV value. An overview of the selection of optimum knot points obtained from the minimum GCV will be shown in Table 2

Table 2. Comparison of GCV value and determination coefficient

Knots	GCV Value Minimum	Determination Coefficient
1	2733,23	89,87586
2	73,13531	99,72911

Based on Table 2, the minimum GCV value was obtained from 2 knots with a minimum GCV value of 73.13531 with knots for the 1st, 2nd, and 3rd subjects respectively were 1616,263; 1643,622; 1964,132; 1993,104; 1590,769; 1621,570. Then the parameter estimation is obtained for each subject as well as the estimated model which will be shown in Table 3

Table 3. Parameter estimation

Subjek	Parameter	Parameter Estimation
Sampang	β_0	0,033
	β_1	0,147
	β_2	0,288
	α	-16,222
	γ	16,078
Pamekasan	β_0	0,025
	β_1	0,581
	β_2	0,281
	α	-16,094
	γ	15,838
Sumenep	β_0	0,024
	β_1	-0,055
	β_2	0,279
	α	-12,482
	γ	12,248

Based on the optimal knot points and parameter estimates obtained for each subject, the truncated spline semiparametric regression model for longitudinal data can be written as follows:

Model estimates for the 1st subject are:

$$\hat{y}_{1j} = 0,033 + 0,147x_{1j1} \\ + 0,288t_{1j1} \\ - 16,222(t_{1j1} \\ - 1616,263)_+ \\ + 16,078(t_{1j1} \\ - 1643,132)_+ \tag{15}$$

Model estimates for the 2 nd subject are:

$$\hat{y}_{2j} = 0,025 + 0,581x_{2j1} \\ + 0,281t_{2j1} \\ - 16,094(t_{2j1} \\ - 1964,132)_+ \\ + 15,838(t_{2j1} \\ - 1993,104)_+ \tag{16}$$

Model estimates for the 3 rd subject are:

$$\hat{y}_{3j} = 0,024 - 0,055x_{3j1} \\ + 0,279t_{3j1} \\ - 12,482(t_{3j1} \\ - 1590,769)_+ \\ + 12,248(t_{3j1} \\ - 1621,57)_+ \tag{17}$$

The estimation of this model produces a coefficient of determination of R^2 of 99.72911 and MSE of 32.50458.

After obtaining the spline semiparametric regression model for longitudinal data in each subject. Next is the interpretation of the model. Interpretation is done by making the translation into a truncated form on each subject to facilitate predictions of electrecity consumption. Equation (15) for the first subject, Sampang can be truncated as follows:

$$\hat{y}_{1j} = \\ \begin{cases} 0,288t_{1j1} & t_{1j1} < 1616,263 \\ -15,934t_{1j1} + 26219,018 & 1616,263 \le t_{1j1} < 1643,132 \\ 0,144t_{1j1} - 26418,276 & t_{1j1} > 1643,132 \end{cases} \tag{18}$$

Based on the truncated equation (18) model for Sampang Regency, several interpretations can be determined, namely when the average monthly electric current in Sampang is below 1616,263 and if there is an increase in the average monthly flow of 1 unit then the peak load tends to increase 0.288 knots. The same pattern occurred when the monthly average flow of Sampang Regency between 1616,263 to 1643,132, then the peak load tended to decrease by 15,934 knots times with the sum of values of 26219,018. If the current in Sampang Regency is 1643.132 or more, then the peak load tends to change by 0.144 knots by reducing the value of 26418.276. A similar interpretation method applies to the other two subjects. Following are given truncated models of equations (16) and (17), respectively for second and third subjects.

The truncated model of equation (16) for Pamekasan Regency is

$$\hat{y}_{2j} = \\ \begin{cases} 0,281t_{2j1} & t_{2j1} < 1964,132 \\ -15,813t_{2j1} + 31610,74 & 1964,132 \le t_{2j1} < 1993,104 \\ 0,025t_{2j1} - 31566,781 & t_{2j1} > 1993,104 \end{cases}$$

The truncated model of equation (17) for Sumenep Regency is

$$\hat{y}_{3j} = \\ \begin{cases} 0,279t_{3j1} & t_{3j1} < 1590,769 \\ -12,203t_{3j1} + 19855,979 & 1590,769 \le t_{3j1} < 1621,57 \\ 0,045t_{3j1} - 19860,989 & t_{3j1} > 1621,57 \end{cases}$$

CONCLUSION

The model the relationship between indicators in the field of electrecity consumption that is packaged according to longitudinal data structures can be done by using a spline semiparametric regression estimator for longitudinal data. Based on the result of the semiparametric spline regression estimator for longitudinal data and the concept of the goodness of the model, the model for the best electricity consumption indicator is a linier spline model with two knots. The result based on data analysis gives satisfied goodness of indicator, like the small MSE, and the high value of determination coefficient. This spline model has a determinition value of 99,72911% and MSE 32,50458.

REFERENCES

[1] Fan, J. and Gijbels, I. Local Polynomial Modelling and Its Applications. London: Chapman and Hall. 1997

[2] Mardianto, M. F. F., Kartiko, S. H., and Utami, H. Forecasting Trend-Seasonal Data Using Nonparametric Regression with Kernel and Fourier Series Approach. Proceedings of the Third International Conference on Computing, Mathematics and Statistics (ICMS2017)

[3] Yani, N. W., Srinadi, I. G., & Sumarjaya, I. W. Aplikasi Model Semiparametrik Spline Truncated (Study Kasus: Pasien Demam Berdarah Dengue di Rumah Sakit Puru Raharja). *E-Journal Matematika*.2017

[4] Kuzairi, Miswanto, Budiantara, I., N., Three form fourier series estimator semiparametric regression for longitudinal data. Journal of Physics: Conference Series, 2020

[5] Eubank, R. L. *Nonparametric Regression and Spline Smoothing (Second ed.)*. New York: Marcel Dekker.1999

[6] Pratiwi, D. A. *Linier Spline Semiparametric Approach for The Average of Age at First Marriage in East Java.* Surabaya: Undergraduate Programe Departemen of Statistic Faculty of Mathematics and Natural Science Institut Teknologi Sepuluh Nopember.2016

[7] Budiantara, I. N. Model of Truncated Polinomial Spline Family in Semiparametric Regression. *Berkala MIPA, 15*(3).2005

[8] Weiss, R. E. *Modelling Longitudinal Data.* New York: Spinger Texts in Statistic.2005

[9] Fitriana, D., Budiantara, I. N., & Ratnasari, V. Semiparametric Spline Truncated Regression on Modelling AHH in Indonesia. *The 3rd International Seminar on Science and Technology* (p. 26). Surabaya: Postgraduate Program Institut Teknologi Sepuluh Nopember.2017

[10] Crainiceanu, C. M. Nonparametric Regression Methods for Longitudinal Data Analysis Mixed-effects Modelling Approaches. *Journal of the American Statistical Association, 102*(476).2007

[11] Shim, J., & Seok, K. H. Semiparametric Kernel Logistic Regression with Longitudinal Data. *Journal of the Korean Data & Information Science Society, 23*(2).2012

[12] Wahba, G. Spline Model for Observational Data. *SIAM XII.* USA: Philadelphia. 1990