

PAPER • OPEN ACCESS

Comparing between mGCV and aGCV Methods to Choose The Optimal Knot Points in Semiparametric Regression with Spline Truncated Using Longitudinal Data

To cite this article: Aviolla Terza Damaliana *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **546** 032003

View the [article online](#) for updates and enhancements.



IOP | ebooks™

Bringing you innovative digital publishing with leading voices to create your essential collection of books in STEM research.

Start exploring the collection - download the first chapter of every title for free.

Comparing between mGCV and aGCV Methods to Choose The Optimal Knot Points in Semiparametric Regression with Spline Truncated Using Longitudinal Data

Aviolla Terza Damaliana¹, I Nyoman Budiantara^{1*}, Vita Ratnasari

¹Department of Statistics, Faculty of Mathematics, Computation, and Data Science, Institut Teknologi Sepuluh Nopember (ITS), Surabaya 60111, Indonesia

*Corresponding author: nyomanbudiantara65@gmail.com

Abstract. Semiparametric regression is a combination of parametric and nonparametric components. The estimation of the semiparametric regression function uses a parametric approach and a nonparametric approach. This study uses longitudinal data. The estimation technique in this study uses Spline truncated which has very special, excellent statistical interpretation and visual interpretation. The estimation technique in longitudinal semiparametric regression uses the weighted least square (WLS). The choice of knots in semiparametric spline truncated regression is very important because the number of knot points and locations of each knot will affect the regression estimation form. The method of selecting knots in this study uses a Modification of Generalized Cross-Validation (mGCV) and aGCV. This study uses cases of life expectancy in East Java Province 2001-2015. Comparison of the two methods based on the R-square value and the value of Mean Square Error (MSE). The results show that the R-square value of mGCV is greater than aGCV and the MSE value of mGCV is smaller than aGCV. So, it can be concluded that the mGCV method is better than the aGCV method for optimal selection of knot points in the case of life expectancy in East Java Province.

1. Introduction

Regression analysis is a statistical method used to estimate the relationship between predictor variables and the response variable. The main purpose of regression analysis is to the estimated form of the regression curve. In this study, the regression approach used is semiparametric regression. The estimation of the non-parametric component in semiparametric regression uses smoothing techniques. The smoothing estimation technique used is spline estimation [1]. The spline is one of estimation approach that has very specific and very good statistical and visual interpretations. Spline has a very good ability to accommodate data that has behavior changing at certain sub-intervals [2]. Budiantara developed a spline method based on spline truncated functions that provide easier and simpler mathematical calculations and least square optimization without involving penalties [3]. The previous study on semiparametric truncated regression is still limited to cross-section data while in this study has used longitudinal data.

In the semiparametric spline truncated model, we must know about the points of knots and its location [4]. The point of knots is integration points where its functions change into patterns at different sub-intervals. The number of knots and the location of the suitable knots will get the best truncated semiparametric spline model. An example of the conventional knot point selection method is



Generalized Cross Validation (GCV) [5]. Wahba uses the minimum MSE criteria to get optimal GCV [5]. The advantages of the GCV Method are having asymptotic optimal properties, being invariant against the population, and variance of the population need not be known in its calculation [5]. The disadvantages of GCV are providing parameter values that are too small, producing very rough estimates for small or medium samples because it has significant variability [6], and estimating GCV into under smooth (biased and changes vary in point estimation) [7]. Then, Cummins *et al* [8], Kim and Gu [9], and Lukas *et al* [6] made the modification of GCV to make more stable for small or medium samples. It is called mGCV. Their results of the study showed that the performance of mGCV is better than GCV. Han *et al* made new weights in the residual sum of square and generalized degrees of freedom [7]. The new method is called aGCV. Their results of the study showed that the performance of aGCV is better than GCV.

Previously, Lukas *et al* has compared the optimal knot point selection method using the RGCV method and the mGCV method [6]. Han *et al* has compared the optimal knot point selection method using the aGCV method and the GCV method [7]. Previously studies used only one variable and cross-sectional data. Furthermore, there are no studies comparing the mGCV method and the aGCV method using longitudinal data.

The data used in this study is the life expectancy in East Java Province. The population of East Java Province in 2017 was 38.9 million. It becomes one of the role models of the Indonesian economy. However, high economic development does not guarantee the welfare of its people. Another benchmark in the welfare of society is the level of quality of life based on the degree of public health. In general, the increase in the value of life expectancy is influenced by 3 main factors, namely health factors, educational factors, and economic factors [10]. Based on [10], there is a life cycle model that explains how life expectancy can be determined through the influence of income, health, and consumption. Study about life expectancy in East Java was conducted by Lilliard *et al* [11], but this study uses cross-section data.

Here, the study is comparing between aGCV method and the mGCV method to select the optimal knot points with truncated semiparametric regression in data of East Java's life expectancy. This study uses longitudinal data from 2001 to 2015 in 37 districts/cities. Based on some studies that have been done, the variables used in this study are infant mortality, literacy rates, and labor force participation rates.

2. Theoretical Review

In this section we will review some theory about mGCV and aGCV methods

2.1. Estimation of Semiparametric Regression with Spline Truncated for Longitudinal Data

Let (x_{ij}, z_{ij}, y_{ij}) , $i = 1, 2, \dots, n$, $j = 1, 2, \dots, t$ be paired data which is assumed to follow a semiparametric regression model for longitudinal data. The model can be expressed as below :

$$y_{ij} = f(x_{ij}) + g(z_{ij}) + \varepsilon_{ij} = f(x_{1ij}, x_{2ij}, \dots, x_{pij}) + g(z_{1ij}, z_{2ij}, \dots, z_{rij}) + \varepsilon_{ij} \quad (1)$$

Where, $f(x_{1ij}, x_{2ij}, \dots, x_{pij})$ is parametric component functions and $g(z_{1ij}, z_{2ij}, \dots, z_{rij})$ is nonparametric component functions.

The characteristic of the longitudinal semiparametric regression model is additive, then equation (1) can be expressed as below:

$$y_{ij} = f(x_{1ij}) + \dots + f(x_{pij}) + g(z_{1ij}) + \dots + g(z_{rij}) + \varepsilon_{ij} = \sum_{u=1}^p f(x_{uij}) + \sum_{s=1}^q g(z_{sij}) \quad (2)$$

Where, $u = 1, 2, \dots, p$ $s = 1, 2, \dots, q$

This study uses the 1st degree polynomial regression curve or ($m = 1$), then the model of semiparametric regression can be expressed as below :

$$y_{ij} = \beta_{oi} + \sum_{u=1}^p \beta_{ui} x_{uij} + \sum_{s=1}^q \left(\alpha_{1i} z_{sij} + \sum_{k=1}^r \alpha_{s(1+k)i} (z_{sij} - K_{ski})_+^1 \right) + \varepsilon_{ij}$$

(3)

with

$$(z_{sij} - K_{ski})_+^1 = \begin{cases} (z_{sij} - K_{ski})^1, & z_{sij} \geq K_{ski} \\ 0, & z_{sij} < K_{ski} \end{cases}$$

Equation (3) can also be written in vector as,

The vector of variable response :

$$\mathbf{y}_1 = (y_{11} \ y_{12} \ \dots \ y_{1t})^T$$

$$\vdots$$

$$\mathbf{y}_n = (y_{n1} \ y_{n2} \ \dots \ y_{nt})^T$$

The matrix of variables predictor :

$$\mathbf{Z}_1[\mathbf{K}] = \begin{pmatrix} 1 & x_{111} & \dots & x_{p11} & z_{111} & (z_{111} - K_{111})_+^1 & \dots & (z_{111} - K_{1r1})_+^1 & \dots & z_{q11} & (z_{q11} - K_{111})_+^1 & \dots & (z_{q11} - K_{qr1})_+^1 \\ 1 & x_{112} & \dots & x_{p12} & z_{112} & (z_{112} - K_{111})_+^1 & \dots & (z_{112} - K_{1r1})_+^1 & \dots & z_{q12} & (z_{q12} - K_{111})_+^1 & \dots & (z_{q12} - K_{qr1})_+^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{11t} & \dots & x_{p1t} & z_{11t} & (z_{11t} - K_{111})_+^1 & \dots & (z_{11t} - K_{1r1})_+^1 & \dots & z_{q1t} & (z_{q1t} - K_{111})_+^1 & \dots & (z_{q1t} - K_{qr1})_+^1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

$$\mathbf{Z}_n[\mathbf{K}] = \begin{pmatrix} 1 & x_{1n1} & \dots & x_{pn1} & z_{1n1} & (z_{1n1} - K_{11n})_+^1 & \dots & (z_{1n1} - K_{1rn})_+^1 & \dots & z_{qn1} & (z_{qn1} - K_{11n})_+^1 & \dots & (z_{qn1} - K_{qrn})_+^1 \\ 1 & x_{1n2} & \dots & x_{pn2} & z_{1n2} & (z_{1n2} - K_{11n})_+^1 & \dots & (z_{1n2} - K_{1rn})_+^1 & \dots & z_{qn2} & (z_{qn2} - K_{11n})_+^1 & \dots & (z_{qn2} - K_{qrn})_+^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1nt} & \dots & x_{pnt} & z_{1nt} & (z_{1nt} - K_{11n})_+^1 & \dots & (z_{1nt} - K_{1rn})_+^1 & \dots & z_{qnt} & (z_{qnt} - K_{11n})_+^1 & \dots & (z_{qnt} - K_{qrn})_+^1 \end{pmatrix}$$

The vector of parameters :

$$\boldsymbol{\delta}_1 = (\beta_{01} \ \beta_{11} \ \dots \ \beta_{p1} | \alpha_{11} \ \alpha_{1(1+1)1} \ \dots \ \alpha_{1(1+r)1} \ \dots \ \alpha_{q1} \ \alpha_{q(1+1)1} \ \dots \ \alpha_{q(1+r)1})^T$$

$$\vdots$$

$$\boldsymbol{\delta}_n = (\beta_{0n} \ \beta_{1n} \ \dots \ \beta_{pn} | \alpha_{1n} \ \alpha_{1(1+1)n} \ \dots \ \alpha_{1(1+r)n} \ \dots \ \alpha_{qn} \ \alpha_{q(1+1)n} \ \dots \ \alpha_{q(1+r)n})^T$$

The vector of errors :

$$\boldsymbol{\varepsilon}_1 = (\varepsilon_{11} \ \varepsilon_{12} \ \dots \ \varepsilon_{1t})^T$$

$$\boldsymbol{\varepsilon}_2 = (\varepsilon_{21} \ \varepsilon_{22} \ \dots \ \varepsilon_{2t})^T$$

$$\vdots$$

$$\boldsymbol{\varepsilon}_n = (\varepsilon_{n1} \ \varepsilon_{n2} \ \dots \ \varepsilon_{nt})^T$$

The Model at i^{th} subject can be expressed as below:

$$\mathbf{y}_i = \mathbf{Z}_i[\mathbf{K}] \boldsymbol{\delta} + \boldsymbol{\varepsilon}_i \quad (4)$$

The model for overall data set can be written in matrix as :

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_n \end{pmatrix} = \begin{pmatrix} \mathbf{Z}_1[\mathbf{K}] & 0 & 0 & 0 \\ 0 & \mathbf{Z}_2[\mathbf{K}] & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{Z}_n[\mathbf{K}] \end{pmatrix} \begin{pmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \\ \vdots \\ \boldsymbol{\delta}_n \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{pmatrix} \quad (5)$$

Based on equation (4) and (5), The model of semiparametric regression with spline truncated can be expressed as :

$$\mathbf{y} = \mathbf{Z}[\mathbf{K}]\boldsymbol{\delta} + \boldsymbol{\varepsilon} \quad (6)$$

Where :

The size of $\mathbf{y}$ is $nt \times 1$

The size of $\mathbf{Z}[\mathbf{K}]$ is $nt \times n(1 + p + q(1 + r))$

The size of $\boldsymbol{\delta}$ is $n(1 + p + q(1 + r)) \times 1$

The size of $\boldsymbol{\varepsilon}$ is $nt \times 1$

The estimation of $\boldsymbol{\delta}$ uses optimisation of Weighted Least Square (WLS) with weighted matrix $\mathbf{W}$. Matrix $\mathbf{W}$ can be written as :

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_1 & 0 & \dots & 0 \\ 0 & \mathbf{W}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{W}_n \end{pmatrix} \quad (7)$$

There are 3 weighted matrix $\mathbf{W}$, it can be written as [12] :

1. If the diagonal element of matrix $\mathbf{W}$ is $\mathbf{W}_i$ with $i = 1, 2, \dots, n$, then $\mathbf{W}_i = (nt)^{-1}\mathbf{I}_{nt}$ is a weighted matrix that gives the same treatment at each observation.
2. If the diagonal element of matrix $\mathbf{W}$ is $\mathbf{W}_i$ with $i = 1, 2, \dots, n$, then $\mathbf{W}_i = (nt)^{-1}\mathbf{I}_{nt}$ is a weighted matrix that gives the same treatment for each observation in the same subject.
3. If the diagonal element of matrix $\mathbf{W}$ is $\mathbf{W}_i$ with $i = 1, 2, \dots, n$, then $\mathbf{W}_i = \mathbf{V}_i^{-1}$, where $\mathbf{V}_i = \text{Cov}(\mathbf{y}) = \text{diag}[\mathbf{V}_1, \mathbf{V}_2, \dots, \mathbf{V}_n]$ is a weighted matrix from correlation in each subject of observation.

Technique estimation of WLS is to minimize $(\boldsymbol{\varepsilon}^T \mathbf{W} \boldsymbol{\varepsilon})$, so we get the equation as follows :

$$(\mathbf{y} - \mathbf{Z}[\mathbf{K}]\boldsymbol{\delta})^T \mathbf{W} (\mathbf{y} - \mathbf{Z}[\mathbf{K}]\boldsymbol{\delta}) = \mathbf{y}^T \mathbf{W} \mathbf{y} - 2\boldsymbol{\delta}^T \mathbf{Z}[\mathbf{K}]^T \mathbf{W} \mathbf{y} + \boldsymbol{\delta}^T \mathbf{Z}[\mathbf{K}]^T \mathbf{W}^{-1} \mathbf{Z}[\mathbf{K}]\boldsymbol{\delta}$$

let Q is $\mathbf{y}^T \mathbf{W} \mathbf{y} - 2\boldsymbol{\delta}^T \mathbf{Z}[\mathbf{K}]^T \mathbf{W} \mathbf{y} + \boldsymbol{\delta}^T \mathbf{Z}[\mathbf{K}]^T \mathbf{W}^{-1} \mathbf{Z}[\mathbf{K}]\boldsymbol{\delta}$, then the estimation of $\boldsymbol{\delta}$ is minimizing partial derivatif equation Q against $\hat{\boldsymbol{\delta}}^T$ which is equal to 0 as follows :

$$\frac{\partial Q}{\partial \hat{\boldsymbol{\delta}}^T} = \frac{\partial (\mathbf{y}^T \mathbf{W} \mathbf{y} - 2\boldsymbol{\delta}^T \mathbf{Z}[\mathbf{K}]^T \mathbf{W} \mathbf{y} + \boldsymbol{\delta}^T \mathbf{Z}[\mathbf{K}]^T \mathbf{W}^{-1} \mathbf{Z}[\mathbf{K}]\boldsymbol{\delta})}{\partial \hat{\boldsymbol{\delta}}^T} = 0 \quad (8)$$

The result of partial equation (8) can be written as :

$$-\mathbf{Z}[\mathbf{K}]^T \mathbf{W} \mathbf{y} + \mathbf{Z}[\mathbf{K}]^T \mathbf{W}^{-1} \mathbf{Z}[\mathbf{K}]\hat{\boldsymbol{\delta}} = 0$$

$$\mathbf{Z}[\mathbf{K}]^T \mathbf{W}^{-1} \mathbf{Z}[\mathbf{K}]\hat{\boldsymbol{\delta}} = \mathbf{Z}[\mathbf{K}]^T \mathbf{W} \mathbf{y}$$

So, we get $\hat{\boldsymbol{\delta}}$ estimator which can be written as :

$$\hat{\boldsymbol{\delta}} = (\mathbf{Z}[\mathbf{K}]^T \mathbf{W} \mathbf{Z}[\mathbf{K}])^{-1} \mathbf{Z}[\mathbf{K}]^T \mathbf{W} \mathbf{y} \quad (9)$$

Based on equation (9), we get $\mathbf{y}$ estimation which can be written as :

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{Z}[\mathbf{K}]\hat{\boldsymbol{\delta}} \\ &= \mathbf{Z}[\mathbf{K}](\mathbf{Z}[\mathbf{K}]^T \mathbf{W}^{-1} \mathbf{Z}[\mathbf{K}])^{-1} \mathbf{Z}[\mathbf{K}]^T \mathbf{W} \mathbf{y} \\ &= \mathbf{A}[\mathbf{K}]\mathbf{y} \end{aligned}$$

Where :

$$\mathbf{A}[\mathbf{K}] = \mathbf{Z}[\mathbf{K}](\mathbf{Z}[\mathbf{K}]^T \mathbf{W}^{-1} \mathbf{Z}[\mathbf{K}])^{-1} \mathbf{Z}[\mathbf{K}]^T \mathbf{W}$$

$\mathbf{A}[\mathbf{K}]$ is a positive semi-definite matrix and the size of $\mathbf{Z}[\mathbf{K}]$ matrix is $nt \times n(1 + p + q(1 + r))$. The estimation of $\mathbf{y}_i$ for i^{th} subject can be written as :

$$\hat{\mathbf{y}}_i = \mathbf{Z}_i[\mathbf{K}]\hat{\boldsymbol{\delta}} = \mathbf{Z}_i[\mathbf{K}](\mathbf{Z}[\mathbf{K}]^T \mathbf{W}^{-1} \mathbf{Z}[\mathbf{K}])^{-1} \mathbf{Z}[\mathbf{K}]^T \mathbf{W} \mathbf{y}$$

2.2. mGCV and aGCV Methods

In this section we discuss about method for selecting point knots. These methods are mGCV and aGCV. The mGCV and aGCV methods are derived from the GCV function given the following equation [12] :

$$GCV(\mathbf{K}) = \sum_{i=1}^n \left(\frac{(\mathbf{y}_i - \hat{\mathbf{y}}_i)}{1 - \frac{\text{Trace}(\mathbf{A}[\mathbf{K}])}{nt}} \right)^T \mathbf{W}_i \left(\frac{(\mathbf{y}_i - \hat{\mathbf{y}}_i)}{1 - \frac{\text{Trace}(\mathbf{A}[\mathbf{K}])}{nt}} \right) = \frac{\sum_{i=1}^n (\mathbf{y}_i - \hat{\mathbf{y}}_i)^T \mathbf{W}_i (\mathbf{y}_i - \hat{\mathbf{y}}_i)}{\left(1 - \frac{\text{Trace}(\mathbf{A}[\mathbf{K}])}{nt} \right)^2} \quad (10)$$

Based on [8], [9], and [6], mGCV method is changing $\left(1 - \frac{\text{Trace}(\mathbf{A}[\mathbf{K}])}{nt}\right)$ with $\left(1 - \frac{\rho \text{Trace}(\mathbf{A}[\mathbf{K}])}{nt}\right)$. So, the function of mGCV is

$$mGCV(\mathbf{K}) = \frac{\sum_{i=1}^n (\mathbf{y}_i - \hat{\mathbf{y}}_i)^T \mathbf{W}_i (\mathbf{y}_i - \hat{\mathbf{y}}_i)}{\left(1 - \frac{\rho \text{Trace}(\mathbf{A}[\mathbf{K}])}{nt}\right)^2} \quad (11)$$

where $1 > \rho > 0$

The aGCV method was developed by [7]. aGCV method has a new weighted which can be written as :

$$aGCV(\mathbf{K}) = \frac{\sum_{i=1}^n (\mathbf{y}_i - \hat{\mathbf{y}}_i)^T \mathbf{W}_i (\mathbf{y}_i - \hat{\mathbf{y}}_i)}{\left(1 - \frac{\text{Trace}(\mathbf{A}[\mathbf{K}])}{nt}\right)^l} \quad (12)$$

where $2 \leq l < 10$

2.3. MSE and R-Square Criterion

One of the purposes of regression analysis is to get the best model that is able to explain the relationship between predictor variables and response variables based on certain criteria. The criteria used in selecting the best model is using MSE and R-Square. In MSE criteria, the best model is the model that has the smallest MSE value and In R-Square criteria, the best model is the model that has the largest R-Square value. The equation of MSE and R-Square can be expressed as :

$$\text{MSE} = \frac{\|\mathbf{I} - \mathbf{A}(\mathbf{K})\mathbf{Y}\|^2}{nt} \quad (13)$$

Where $\mathbf{A}(\mathbf{K})$ is obtained from equation (9)

$$R^2 = \frac{SS \text{ Regression}}{SS \text{ Total}} \quad (14)$$

3. The Data

In this study, we use secondary data. we get this data from the publication of BPS (Central Bureau of Statistics) East Java Province. This data is related to the value of life expectancy in East Java Province. The subjects of this study are 37 districts/cities without Batu city. The variables that will be used in this study are

Table 1. List of Variables

Variables		Variables name	Measurement Scale
Response	y_{ij}	The value of life expectancy (AHH)	Ratio
	x_{ij1}	The value of infant morality (AKB)	Ratio
Predictors	x_{ij2}	Literacy rates (AHH)	Ratio
	x_{ij3}	Labor force participation rates (TPAK)	Ratio

4. Main Result

Before analyzing, we check the relationship between the response variable and the predictor using scatter plots.

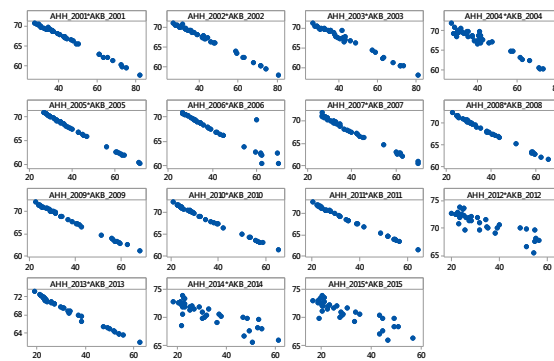


Figure 1. The relationship between AHH and AKB

Figure 1 shows that there is a relationship between AHH and AKB. The pattern of these relationships is the higher value of AHH, the lower value of AKB in East Java Province. Based on this condition, the pattern of AHH and AKB has been known, so the variable of AKB is a parametric component.

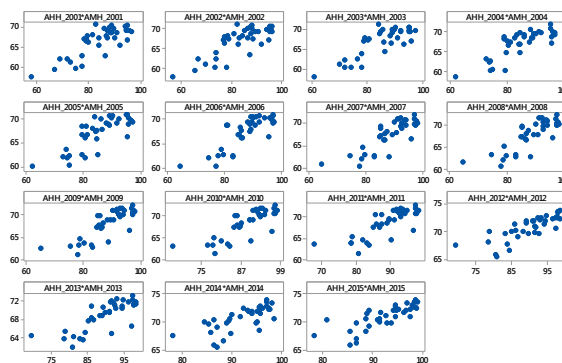


Figure 2. The relationship between AHH and AMH

Figure 2 shows that there is a relationship between AHH and AMH. The pattern of these relationships is the higher value of AHH, the higher value of AMH in East Java Province. Based on this condition, the pattern of AHH and AMH has been known, so the variable of AMH is a parametric component.

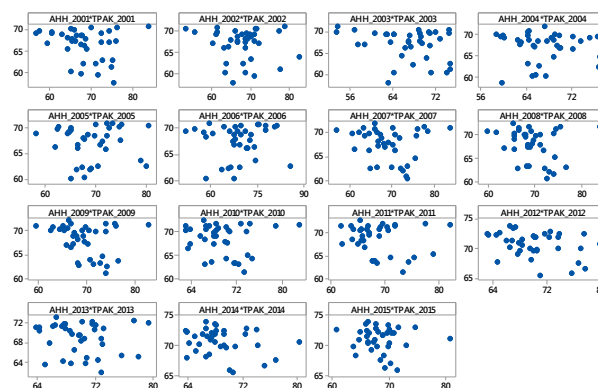


Figure 3. The relationship between AHH and TPAK

Figure 3 shows that there is no pattern of relations between AHH and TPAK. In the theory, the relationship between AHH and TPAK is the higher value of AHH, the higher value of TPAK. Based on

this condition, the pattern of AHH and TPAK is unknown, so the variable of TPAK is a nonparametric component.

4.1. Estimation of mGCV method

Firstly, we estimate mGCV used $\mathbf{W}_i = (nt)^{-1}\mathbf{I}_{nt}$ weighted matrix. The smallest estimated value of mGCV is the criteria for choosing the optimal knot point. This study use one knot point and 14 increments. Based on equation (11), ρ in this study are 0,1, 0,2, 0,3, and $\geq 1,1$. In longitudinal data, we get a knot point in each district/cities. The result of optimal mGCV estimation with $\rho < 1$ can be seen in Table 2 :

Table 2. List of minimum mGCV in each ρ with $\mathbf{W}_i = (nt)^{-1}\mathbf{I}_{nt}$

ρ	Minimum mGCV	ρ	minimum mGCV	ρ	minimum mGCV
0.1	0.513	0.5	0.690	0.9	0.970
0.2	0.550	0.6	0.749		
0.3	0.592	0.7	0.815		
0.4	0.638	0.8	0.891		

Table 2 shows that the minimum mGCV is located at $\rho = 0.1$. The minimum mGCV in this study is 0.513. Knots point in each district/city based on table 2 can be seen in table 3.

Table 3. List of Knot Points in each district/city of East Java Province with $\mathbf{W}_i = (nt)^{-1}\mathbf{I}_{nt}$ Use mGCV Method

Districts/Cities	Knot Points	Districts/Cities	Knot Points	Districts/Cities	Knot Points	Districts/Cities	Knot Points
Pacitan	68.625	Bondowoso	67.430	Ngawi	69.235	Blitar	64.085
Ponorogo	67.055	Situbondo	65.605	Bojonegoro	68.565	Malang	63.015
Trenggalek	65.435	Probolinggo	70.420	Tuban	67.330	Probolinggo	63.965
Tulungagung	73.020	Pasuruan	69.090	Lamongan	69.780	Pasuruan	67.425
Blitar	70.915	Sidoarjo	60.925	Gresik	66.225	Mojokerto	66.310
Kediri	67.540	Mojokerto	67.750	Bangkalan	69.610	Madiun	62.235
Malang	68.415	Jombang	66.730	Sampang	68.110	Surabaya	61.220
Lumajang	66.640	Nganjuk	68.545	Pemekasan	77.740		
Jember	65.915	Madiun	64.225	Sumenep	74.515		
Banyuwangi	69.385	Magetan	73.090	Kediri	68.640		

Secondly, we estimate mGCV used $\mathbf{W}_i = (nt_i)^{-1}\mathbf{I}_{nt}$ weighted matrix. The result of optimal mGCV estimation with $\rho < 1$ can be seen in Table 4 :

Table 4. List of minimum mGCV in each ρ with $\mathbf{W}_i = (nt_i)^{-1}\mathbf{I}_{nt}$

ρ	Minimum mGCV	ρ	minimum mGCV	ρ	minimum mGCV
0.1	7.692	0.5	10.350	0.9	14.548
0.2	8.251	0.6	11.230		
0.3	8.873	0.7	12.228		
0.4	9.569	0.8	13.365		

Table 4 shows that the minimum mGCV is located at $\rho = 0.1$. The minimum mGCV in this study is 7.692. Knots point in each district/city based on Table 4 as same as Knots point in first weighted matrix which can seen in Table 3.

Thirdly, we estimate mGCV used $\mathbf{W}_i = \mathbf{V}_i^{-1}$, where $\mathbf{V}_i = Cov(\mathbf{y}) = diag[\mathbf{V}_1, \mathbf{V}_2, \dots, \mathbf{V}_n]$. The result of optimal mGCV estimation with $\rho < 1$ can be seen in Table 5. Table 5 shows that the minimum mGCV

is located at $\rho = 0.1$. The minimum mGCV in this study with $\mathbf{W}_i = \mathbf{V}_i^{-1}$ is 1474.129. Knots point in each district/city based on table 5 can be seen in table 6.

Table 5. List of minimum mGCV in each ρ with $\mathbf{W}_i = \mathbf{V}_i^{-1}$

ρ	Minimum mGCV	ρ	minimum mGCV	ρ	minimum mGCV
0.1	1474.129	0.5	1972.988	0.9	2775.637
0.2	1579.412	0.6	2137.373		
0.3	1696.390	0.7	2323.191		
0.4	1826.863	0.8	2534.339		

Table 6. List of Knot Points in each district/city of East Java Province with $\mathbf{W}_i = \mathbf{V}_i^{-1}$ Use mGCV Method

Districts/Cities	Knot Points	Districts/Cities	Knot Points	Districts/Cities	Knot Points	Districts/Cities	Knot Points
Pacitan	77.314	Bondowoso	69.801	Ngawi	71.484	Blitar	70.311
Ponorogo	71.995	Situbondo	69.734	Bojonegoro	71.094	Malang	66.275
Trenggalek	72.415	Probolinggo	73.277	Tuban	70.273	Probolinggo	67.716
Tulungagung	75.563	Pasuruan	71.159	Lamongan	73.266	Pasuruan	71.119
Blitar	72.826	Sidoarjo	66.225	Gresik	68.319	Mojokerto	71.899
Kediri	68.786	Mojokerto	69.601	Bangkalan	72.387	Madiun	66.049
Malang	70.109	Jombang	68.187	Sampang	73.104	Surabaya	65.014
Lumajang	69.354	Nganjuk	70.868	Pemekasan	82.134		
Jember	67.375	Madiun	67.554	Sumenep	77.101		
Banyuwangi	71.691	Magetan	77.296	Kediri	73.697		

After we know the minimum mGCV in each weighted matrix, then we compare the result based on R-Square and MSE in each weighted matrix. The results can be seen in Table 7. Table 7 shows that the biggest value of R-Square is 95.907 and the smallest value of MSE is 0.570. It is located in $\mathbf{W}_i = (nt)^{-1}\mathbf{I}_{nt}$ weighted matrix. So, the optimal knot points in mGCV method use weighted matrix with $\mathbf{W}_i = (nt)^{-1}\mathbf{I}_{nt}$.

Table 7. Comparing mGCV method based on weighted matrix Use mGCV Method

Pembobot	mGCV minimum	R-Square	MSE
$\mathbf{W}_i = (nt)^{-1}\mathbf{I}_{nt}$	0.570	95.907	0.719
$\mathbf{W}_i = (nt_i)^{-1}\mathbf{I}_{nt}$	7.692	95.907	0.719
$\mathbf{W}_i = \mathbf{V}_i^{-1}$	1474.129	92.979	1.346

4.2. Estimation of aGCV method

Same with previously methods, we used $\mathbf{W}_i = (nt)^{-1}\mathbf{I}_{nt}$ weighted matrix in first analysis and we use the smallest estimated value of aGCV is the criteria for choosing the optimal knot point. Based on equation (12), l in this study is $2 \leq l < 10$. The result of optimal aGCV estimation can be seen in Table 8 :

Table 8. List of minimum aGCV in each l with $\mathbf{W}_i = \mathbf{V}_i^{-1}$

l	Minimum aGCV	l	minimum aGCV	l	minimum aGCV
2	1.042	6	3.602	10	12.454
3	1.420	7	4.912		
4	1.937	8	6.698		
5	2.641	9	9.133		

Table 8 shows that the minimum aGCV is located at $l = 2$. The minimum aGCV in this study is 1.042. Knots point in each district/city based on Table 8 can be seen in table 9. After that, we estimate aGCV with $\mathbf{W}_i = (nt_i)^{-1}\mathbf{I}_{nt}$. The result of optimal aGCV estimation can be seen in Table 10. Table 10 shows that the minimum aGCV is located at $l = 2$. The minimum aGCV in this study is 15.625. Knots point in each district/city based on Table 10 as same as Knots point in first weighted matrix which can be seen in Table 9.

Table 9. List of Knot Points in each district/city of East Java Province with $\mathbf{W}_i = (nt)^{-1}\mathbf{I}_{nt}$

Districts/Cities	Knot Points	Districts/Cities	Knot Points	Districts/Cities	Knot Points	Districts/Cities	Knot Points
Pacitan	53.420	Bondowoso	63.280	Ngawi	65.300	Blitar	53.190
Ponorogo	58.410	Situbondo	58.380	Bojonegoro	64.140	Malang	57.310
Trenggalek	53.220	Probolinggo	65.420	Tuban	62.180	Probolinggo	57.400
Tulungagung	68.570	Pasuruan	65.470	Lamongan	63.680	Pasuruan	60.960
Blitar	67.570	Sidoarjo	51.650	Gresik	62.560	Mojokerto	56.530
Kediri	65.360	Mojokerto	64.510	Bangkalan	64.750	Madiun	55.560
Malang	65.450	Jombang	64.180	Sampang	59.370	Surabaya	54.580
Lumajang	61.890	Nganjuk	64.480	Pemekasan	70.050		
Jember	63.360	Madiun	58.400	Sumenep	69.990		
Banyuwangi	65.350	Magetan	65.730	Kediri	59.790		

Table 10. List of minimum aGCV in each l with $\mathbf{W}_i = (nt_i)^{-1}\mathbf{I}_{nt}$

l	Minimum aGCV	l	minimum mGCV	l	minimum aGCV
2	15.625	6	54.027	10	186.812
3	21.307	7	73.673		
4	29.054	8	100.463		
5	39.620	9	136.995		

Thirdly, we estimate aGCV used $\mathbf{W}_i = \mathbf{V}_i^{-1}$, where $\mathbf{V}_i = \text{Cov}(\mathbf{y}) = \text{diag}[\mathbf{V}_1, \mathbf{V}_2, \dots, \mathbf{V}_n]$. The result of optimal aGCV estimation can be seen in Table 11:

Table 11. List of minimum aGCV in each ρ with $\mathbf{W}_i = \mathbf{V}_i^{-1}$

l	Minimum aGCV	ρ	minimum aGCV	ρ	minimum aGCV
2	3053.112	6	12378.519	10	42801.892
3	4542.834	7	16879.798		
4	6656.8922	8	23017.906		
5	9077.5803	9	31388.054		

Table 11 shows that the minimum mGCV is located at $l = 2$. The minimum aGCV in this study with $\mathbf{W}_i = \mathbf{V}_i^{-1}$ is 3053.

Knots point in each district/city based on table 11 can be seen in table 6 because it has same knot points with Table 6.

After we know the minimum aGCV in each weighted matrix, then we compare the result based on R-Square and MSE in each weighted matrix. The results can be seen in Table 12. Table 12 shows that the biggest value of R-Square is 95.214 and the smallest value of MSE is 0.840. It is located in $\mathbf{W}_i = (nt)^{-1}\mathbf{I}_{nt}$ weighted matrix. So, the optimal knot points in aGCV method use weighted matrix with $\mathbf{W}_i = (nt)^{-1}\mathbf{I}_{nt}$.

Table 12. Comparing aGCV method based on weighted matrix

Pembobot	aGCV minimum	R-Square	MSE
$\mathbf{W}_i = (nt)^{-1}\mathbf{I}_{nt}$	1.042	95.214	0.840
$\mathbf{W}_i = (nt_i)^{-1}\mathbf{I}_{nt}$	15.625	95.214	0.840
$\mathbf{W}_i = \mathbf{V}_i^{-1}$	3053.112	92.979	1.346

4.3. Comparing mGCV and aGCV method

Comparison of mGCV and aGCV methods can be seen in Table 13.

Table 13. Comparing mGCV and aGCV method based on

Methods	Minimum Estimated value	R-Square	MSE
mGCV	0.570	95.907	0.719
aGCV	1.042	95.214	0.840

Table 13 shows that the R-Square value of the mGCV method is 95.907 and the MSE value is 0.719. The R-Square value of the aGCV method is 95.214 and the MSE value is 0.840. Based on this condition, the R-square value of the mGCV method is greater than the aGCV method and the MSE value of the mGCV method is smaller than the aGCV method. The results show that the mGCV method is better than the aGCV method for choosing the optimal knot point.

5. Conclusion

In this study we know that the difference in weighted matrix does not significantly affect the location of knot points in each district / city. This is because even though the weighted matrix is different but it will get the same knot points on some analysis. Based on overall analysis, we can conclude that the best method choosing the optimal knot point for semiparametric regression with spline truncated for longitudinal data life expectancy in East Java Province is the mGCV method.

References

- [1] Hardle W 1990 *Applied Nonparametric Regression* (New York: Cambrige University Press).
- [2] Cox D and O'Sullivan F 1996 *Journal of Multivariate Analysis*. **4**(56), 185-206.
- [3] Budiantara I N 2005 *Makalah Pembicara Umum pada Seminar Nasional Matematika FMIPA Universitas Diponegoro Semarang*.
- [4] Montoya E L, Ulloa N and Miller V 2014 *International Journal of Statistics and Probability*. **3**(3), 96-110,.
- [5] Wahba R 1990 *Spline Models for Observational Data* (Pensylvania: SIAM).
- [6] Lukas M A, Hoog F R. and Andersen R S 2012 *Scandinavian Journal of Statistics*. **39**, 97-115.
- [7] Han S W, Busch T M and Putt M E 2014 *Journal of Applied Statistic*. **60**, 26-45
- [8] Cummins D J, Fillon T G and Nychka D 2001 *Journal of Multivariate Analysis* **96**, 233-246.
- [9] Kim Y J and Gu C 2004 *Journal of The Royal Statistical Society Series B : Statistical Methodology*. **66**, 337-356
- [10] Lilliard L A and Weiss Y 1997 *Journal of Business & Economic Statistics*. **15**(2), 254-268.
- [11] Sugiantary A P 2013 *Analisis Faktor-faktor yang Mempengaruhi Angka Harapan Hidup di Jawa Timur Menggunakan Regresi Semiparametrik Spline* (Institut Teknologi Sepuluh Nopember : Surabaya)
- [12] Wu H and Zhang J T 2006 *Nonparametric Regression Methods for Longitudinal Data Analysis* (New Jersey: A Jhon Wiley and Son Inc).