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

Aviolla Terza Damaliana, I Nyoman Budiantara*, 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 RCGV 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 \((x_{ij}, z_{ij}, y_{ij})\), \(i = 1, 2, \ldots, n\), \(j = 1, 2, \ldots, t\) be paired data which is assumed to follow a semiparametric regression model for longitudinal data. The model can be expressed as below:

\[
y_{ij} = f(x_{ij}) + g(z_{ij}) + \epsilon_{ij} = f(x_{1ij}, x_{2ij}, \ldots, x_{p(i)}) + g(z_{1ij}, z_{2ij}, \ldots, z_{rij}) + \epsilon_{ij} \quad (1)
\]

Where, \(f(x_{1ij}, x_{2ij}, \ldots, x_{p(i)})\) is parametric component functions and \(g(z_{1ij}, z_{2ij}, \ldots, z_{rij})\) is nonparametric component functions.

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

\[
y_{ij} = f(x_{1ij}) + \cdots + f(x_{p(i)}) + g(z_{1ij}) + \cdots + g(z_{rij}) + \epsilon_{ij} = \sum_{u=1}^{p} f(x_{uij}) + \sum_{s=1}^{q} g(z_{sij}) \quad (2)
\]

Where, \(u = 1, 2, \ldots, p\), \(s = 1, 2, \ldots, q\).

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

\[
y_{ij} = \beta_{o(i)} + \sum_{u=1}^{p} \beta_{uij} x_{uij} + \sum_{s=1}^{q} \left( \alpha_{s1} z_{sij} + \sum_{k=1}^{r} \alpha_{s(k+1)} (z_{sij} - K_{ski})^{k} \right) \epsilon_{ij}
\]
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:
\[
y_1 = (y_{11} \ y_{12} \ ... \ y_{1t})^T \\
y_2 = (y_{21} \ y_{22} \ ... \ y_{2t})^T \\
\vdots \\
y_n = (y_{n1} \ y_{n2} \ ... \ y_{nt})^T
\]

The matrix of variables predictor:
\[
Z_1[K] = \begin{pmatrix} 1 & x_{111} & \cdots & x_{p11} & z_{111} & (z_{111} - K_{111})_+^1 & \cdots & z_{q11} & (z_{q11} - K_{q1r1})_+^1 \\
1 & x_{112} & \cdots & x_{p12} & z_{112} & (z_{112} - K_{11r1})_+^1 & \cdots & z_{q12} & (z_{q12} - K_{q1r1})_+^1 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
1 & x_{1lt} & \cdots & x_{plt} & z_{1lt} & (z_{1lt} - K_{11r1})_+^1 & \cdots & z_{qlt} & (z_{qlt} - K_{q1r1})_+^1 \\
\end{pmatrix}
\]

\[
Z_n[K] = \begin{pmatrix} 1 & x_{n11} & \cdots & x_{pn1} & z_{n11} & (z_{n11} - K_{11n})_+^1 & \cdots & z_{qn1} & (z_{qn1} - K_{q1rn})_+^1 \\
1 & x_{n12} & \cdots & x_{pn2} & z_{n12} & (z_{n12} - K_{11rn})_+^1 & \cdots & z_{qn2} & (z_{qn2} - K_{q1rn})_+^1 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
1 & x_{nlt} & \cdots & x_{pnt} & z_{nlt} & (z_{nlt} - K_{11rn})_+^1 & \cdots & z_{qnt} & (z_{qnt} - K_{q1rn})_+^1 \\
\end{pmatrix}
\]

The vector of parameters:
\[
\delta_1 = (\beta_{01} \ \beta_{11} \ \cdots \ \beta_{p11} \ \alpha_{11} \ \alpha_{1(1+1)} \ \cdots \ \alpha_{1(1+1+r-1)} \ \cdots \ \alpha_{q1} \ \alpha_{q(1+1)} \ \cdots \ \alpha_{q(1+r-1)})^T \\
\delta_n = (\beta_{0n} \ \beta_{1n} \ \cdots \ \beta_{pn1} \ \alpha_{1(1+1)} \ \cdots \ \alpha_{1(1+1+r-1)} \ \cdots \ \alpha_{qn} \ \alpha_{q(1+1)} \ \cdots \ \alpha_{q(1+r-1)})^T
\]

The vector of errors:
\[
\varepsilon_1 = (\varepsilon_{11} \ \varepsilon_{12} \ \cdots \ \varepsilon_{1t})^T \\
\varepsilon_2 = (\varepsilon_{21} \ \varepsilon_{22} \ \cdots \ \varepsilon_{2t})^T \\
\vdots \\
\varepsilon_n = (\varepsilon_{n1} \ \varepsilon_{n2} \ \cdots \ \varepsilon_{nt})^T
\]

The Model at \( i \)th subject can be expressed as below:
\[
y_i = Z_i[K] \delta + \varepsilon_i
\]

The model for overall data set can be written in matrix as:
\[
\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} Z_1[K] & 0 & 0 & 0 \\ 0 & Z_2[K] & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & Z_n[K] \end{pmatrix} \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}
\]
Based on equation (4) and (5), The model of semiparametric regression with spline truncated can be expressed as:

\[
\hat{y} = Z[K] \hat{\delta} + \varepsilon
\]

Where:
- \( y = Z[K] \delta + \varepsilon \)

The size of \( y \) is \( nt \times 1 \)
- The size of \( Z[K] \) is \( nt \times n(1 + p + q(1 + r)) \)
- The size of \( \delta \) is \( n(1 + p + q(1 + r)) \times 1 \)
- The size of \( \varepsilon \) is \( nt \times 1 \)

The result of partial equation (9), we get the equation as follows:

\[
\hat{\delta} = (Z[K]WZ[K])^{-1}Z[K]^TWy
\]

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

\[
\hat{y} = Z[K] \hat{\delta} = Z[K](Z[K]^TW^{-1}Z[K])^{-1}Z[K]^TWy
\]

Where:
- \( A[K] = Z[K](Z[K]^TW^{-1}Z[K])^{-1}Z[K]^TW \)

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

\[
\hat{y}_i = Z_i[K] \hat{\delta} = Z_i[K](Z[K]^T W^{-1}Z[K])^{-1}Z[K]^TWy
\]

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]:

\[
GCV(K) = \sum_{i=1}^{n} \left( \frac{y_i - \hat{y}_i}{\text{Trace}(A[K])} \right)^T W_i \left( \frac{y_i - \hat{y}_i}{\text{Trace}(A[K])} \right) = \sum_{i=1}^{n} \left( \frac{y_i - \hat{y}_i}{\text{Trace}(A[K])} \right)^T W_i \left( \frac{y_i - \hat{y}_i}{\text{Trace}(A[K])} \right) \left( 1 - \frac{\text{Trace}(A[K])}{nt} \right)^2
\]
Based on [8], [9], and [6], mGCV method is changing \( \left( 1 - \frac{\text{Trace}(A[K])}{nt} \right)^2 \) with \( \left( 1 - \frac{\rho \text{Trace}(A[K])}{nt} \right)^2 \). So, the function of mGCV is:

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

where \( 1 > \rho > 1 \)

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

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

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:

\[
MSE = \frac{||I - A(K)Y||^2}{nt}
\]

Where \( A(K) \) is obtained from equation (9)

\[
R^2 = \frac{SS \text{ Regression}}{SS \text{ Total}}
\]

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

| Variable | 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 \( W_i = (nt)^{-1}I_{nt} \) 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), \( \rho \) in this study are 0, 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:

| \( \rho \) | Minimum mGCV | \rho | minimum mGCV | \rho | 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 |
| 0.3 | 0.592 | 0.7 | 0.815 | 0.4 | 0.638 |
| 0.4 | 0.638 | 0.8 | 0.891 | 0.5 | 0.670 |

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.

| Districs/Cities | Knot Points | Districs/Cities | Knot Points | Districs/Cities | Knot Points | Districs/Cities | Knot Points |
|---|---|---|---|---|---|---|---|
| Pacitan | 68.625 | Bondowoso | 67.430 | Ngawi | 69.235 | Blitar | 64.085 |
| Ponorogo | 67.055 | Sutubondo | 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 \( W_i = (nt)^{-1}I_{nt} \) weighted matrix. The result of optimal mGCV estimation with \( \rho < 1 \) can be seen in Table 4:

| \( \rho \) | Minimum mGCV | \rho | minimum mGCV | \rho | 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 |
| 0.3 | 8.873 | 0.7 | 12.228 | 0.4 | 9.569 |
| 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 \( W_i = V_i^{-1} \), where \( V_i = Cov(\mathbf{y}) = \text{diag}(|V_1, V_2, ..., V_n|) \). 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 $W_i = V_i^{-1}$ 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 $\rho$ with $W_i = V_i^{-1}$

| $\rho$ | Minimum mGCV | $\rho$ | minimum mGCV | $\rho$ | 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 $W_i = V_i^{-1}$

| Districs/Cities | Knot Points | Districs/Cities | Knot Points | Districs/Cities | Knot Points | Districs/Cities | Knot Points |
|-----------------|-------------|-----------------|-------------|-----------------|-------------|-----------------|-------------|
| Pacitan          | 77.314      | Bondowoso       | 69.801      | Ngawi           | 71.484      | Blitar          | 70.311      |
| Ponorogo         | 71.995      | Situbondo       | 69.734      | Bojonegoro      | 70.104      | 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 $W_i = (nt)^{-1}I_{nt}$ weighted matrix. So, the optimal knot points in mGCV method use weighted matrix with $W_i = (nt)^{-1}I_{nt}$.

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

| Pembobot | mGCV minimum | R-Square | MSE     |
|----------|--------------|----------|---------|
| $W_i = (nt)^{-1}I_{nt}$ | 0.570 | 95.907 | 0.719 |
| $W_i = (nt)^{-1}I_{nt}$ | 7.692 | 95.907 | 0.719 |
| $W_i = V_i^{-1}$ | 1474.129 | 92.979 | 1.346 |

### 4.2. Estimation of aGCV method

Same with previously methods, we used $W_i = (nt)^{-1}I_{nt}$ 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 $W_i = V_i^{-1}$

| $l$ | 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 $W_i = (nt)^{-1}I_{nt}$. 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 seen in Table 9.

### Table 9. List of Knot Points in each district/city of East Java Province with $W_i = (nt)^{-1}I_{nt}$

| Districs/Cities | Knot Points | Districs/Cities | Knot Points | Districs/Cities | Knot Points | Districs/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    | 64.520      | 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      |                |             |                |             |

Table 10 shows that the minimum mGCV is located at $l = 2$. The minimum aGCV in this study with $W_i = (nt)^{-1}I_{nt}$ is 15.625. 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 $W_i = (nt)^{-1}I_{nt}$ weighted matrix. So, the optimal knot points in aGCV method use weighted matrix with $W_i = (nt)^{-1}I_{nt}$.

### Table 10. List of minimum aGCV in each $l$ with $W_i = (nt)^{-1}I_{nt}$

| $l$ | 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      |     |              |

### Table 11. List of minimum aGCV in each $\rho$ with $W_i = V_i^{-1}$

| $l$ | Minimum aGCV | $\rho$ | minimum aGCV | $\rho$ | 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    |       |              |

Thirdly, we estimate aGCV used $W_i = V_i^{-1}$, where $V_i = Cov(y) = diag[V_1, V_2, ..., V_n]$. The result of optimal aGCV estimation can be seen in Table 11.
Table 12. Comparing aGCV method based on weighted matrix

| Pembobot       | aGCV minimum | R-Square | MSE  |
|----------------|--------------|----------|------|
| $W_i = (nt)^{-1}_n$ | 1.042        | 95.214   | 0.840|
| $W_i = (nt)_i^{-1}$ | 15.625       | 95.214   | 0.840|
| $W_i = V_i^{-1}$ | 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: Cambridge 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).