Optimal knot selection in spline regression using unbiased risk and generalized cross validation methods

T W Utami¹, M A Haris¹, A Prahutama², E A Purnomo³
¹Department of Statistics, Universitas Muhammadiyah Semarang, Indonesia
²Statistics Department, Universitas Diponegoro, Indonesia
³Department of Mathematics Education, Universitas Muhammadiyah Semarang, Indonesia

*tiianiutami@unimus.ac.id

Abstract. Spline regression is a nonparametric regression method that estimates data patterns that do not form certain patterns with the help of knots. The best model is obtained from the optimal knot. There are several methods that can be used to select optimal knots, including Generalized Cross-Validation (GCV) and Unbiased Risk (UBR). The best model selection criteria used are based on the Mean Squared Error (MSE) and R-Square values. This study discusses the comparison of spline regression models using the UBR and GCV methods as a method for selecting optimal knots in data generation simulations. This research resulted in the best nonparametric spline regression model using the UBR method obtained by using three knots which produced an MSE value of 738.67 and R-Square of 85.65%. Whereas, the best nonparametric spline regression model of the GCV method was obtained using three knots which produced an MSE value of 121.43 and R-Square of 97.64%. It can be concluded that the more appropriate method used for the selection of optimal knot is the GCV method because it produces a smaller MSE value and a larger R-Square compared to the UBR method.

1. Introduction
Spline regression is able to overcome data patterns that show sharp up / down with the help of knots so that the resulting curve is relatively smooth [1]. Knots are joint fusion points that indicate changes in behaviour patterns of data [2]. With these knots, the spline is able to follow and adjust the data. Informing the spline regression model, important things to consider include determining the order of the model, the number of knots and the location of the knots [3]. Orde for the model can be known from the patterns formed in the data, while the number of knot points and knot point locations are determined based on changes in data patterns that occur at certain sub-intervals [4]. The accuracy of the spline regression model that is formed is influenced by the selection of optimal knot points. There are several methods that can be used in the selection of optimal knot points in spline regression, including Cross-Validation (CV) [5], Unlimited Risk (UBR) [6], Generalized Cross-Validation (GCV) [7], and Generalized Maximum Likelihood (GML) [8].

The Unbias Risk (UBR) method is the point of the optimal knot selection method that requires the estimated value of the variance of known errors [9]. The UBR method will produce good scores on non-Gaussian or data not patterned normal distribution GCV method will be better used in Gaussian or normal distribution data. The advantage of GCV is that it is efficient and simple in its calculations and does not require variants. GCV is asymptotic so if the sample used is small the results will not be maximal [10].
Research using UBR and GCV methods has been researched several times, with a study entitled Modeling Factors Affecting Morbidity Rates in East Java Using Spline Nonparametric Regression [11]. In this study using the GCV method in the selection of optimal knot points from three-knot points with R-Square of 89.72%. [12] conducted a study entitled Comparison of Multivariable Spline Nonparametric Regression Models Using GCV and UBR Methods in Optimal Knot Point Selection (Case Study of Maternal Mortality Rate Data in East Java). In this study, the GCV method is better used for the selection of optimal knot points with R-Square of 94.3%. A researcher has conducted a study entitled Unbias Risk (UBR) and Cross-Validation (CV) Methods for the Selection of Optimal Knot Points in Multivariable Spline Truncated Nonparametric Regression [13]. Based on the description above, this study focuses on the comparison of the UBR method and the GCV method in the selection of optimal knot points in spline nonparametric regression.

2. UBR and GCV Methods

The Unbias Risk (UBR) method is a method that can be used to select optimal knots. This method can be used when $\sigma^2$ is known. In this method, the most important thing is the estimated value $\sigma^2$. If the estimated value of $\sigma^2$ is good, the UBR method will be appropriate. The optimal knot point uses the smallest UBR value. According to [13], the formulation of the UBR method is as follows:

$$R(K) = E(L(K)) = E\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{f}_{Ki} - f_{i})^2\right)$$  \hspace{1cm} (1)

$$R(K) = \frac{1}{n}E\|A(k)(f + \xi) - f\|^2$$
$$= \frac{1}{n}E\|A(k)f + A(k)\xi - f\|^2$$
$$= \frac{1}{n}E\|(I - A(k))f + A(k)\xi\|^2$$
$$= \frac{1}{n}\| (I - A(k))f \|^2 + 0 + \frac{1}{n}E(\xi'A'(k)A(k)\xi)$$
$$= \frac{1}{n}\| (I - A(k))f \|^2 + \frac{\sigma^2}{n}tr(A'(k)A(k))$$  \hspace{1cm} (2)

The result of $R(K) = E(\hat{R}(K))$, where $\hat{R}(K)$ is the criterion of Unlock Risk (UBR). The $\hat{R}(K)$ is as follows:

$$\hat{R}(K) = \frac{1}{n}\| (I - A(k))y \|^2 - \frac{\sigma^2}{n}tr(I - A'(k))(I - A(k)) + \frac{\sigma^2}{n}tr(A'(k)A(k))$$

The selection of optimal knot points using the UBR method is obtained by finding the optimization value as below:

$$\hat{\sigma}^2 = \frac{\| (I - A(k))y \|^2}{tr[I - A(k)]}$$

$$Min \{ R(k_1, k_2, k_3, ..., k_j) \} = Min \left\{ \frac{1}{n}\| (I - A(k))y \|^2 - \frac{\sigma^2}{n}tr[I - A(k)]^2 + \frac{\sigma^2}{n}tr[A'^2](k) \right\}$$
\[ \text{MSE}(k) = \frac{[n^{-1}\text{trace}(I-A(k))]^{2}}{\sum_{i=1}^{n}(y_{i} - \hat{f}(x_{i}))^{2}} \]

\[ GCV(k) = \frac{\text{MSE}(k)}{[n^{-1}\text{trace}(I-A(k))]^{2}} \]

\[ = n^{-1} \frac{\sum_{i=1}^{n}(y_{i} - \hat{f}(x_{i}))^{2}}{[n^{-1}\text{trace}(I-A(k))]^{2}} \]

\[ = n^{-1} \frac{\text{trace}(I-A(k))^{2}(I-A(k)y)}{[n^{-1}\text{trace}(I-A(k))]^{2}} \]

The Generalized Cross-Validation (GCV) method is one of the methods often used in the selection of optimal knot points. This GCV method is the result of a modification of the CV method. According to [14], the optimal knot point is obtained from the smallest GCV value. Following are the functions of GCV [15]:

\[ \text{CV}(k) = n^{-1} \frac{\sum_{i=1}^{n}(y_{i} - \hat{f}(x_{i}))^{2}}{[1 - n^{-1}\text{trace}(A(k))]^{2}} \]  

\[ GCV(\hat{k}) \] is a vector that contains the GCV values from the knot points obtained from the division between the results of the sum of squared residuals from \( \hat{f}(x) \) with \( n\{1 - n^{-1}\text{trace}(A(k))\}^{2} \).

According to [12], the general equation of GCV is as follows

\[ GCV(k) = \frac{\text{MSE}(k)}{[n^{-1}\text{trace}(I-A(k))]^{2}} \]

\[ = n^{-1} \frac{\sum_{i=1}^{n}(y_{i} - \hat{f}(x_{i}))^{2}}{[n^{-1}\text{trace}(I-A(k))]^{2}} \]

\[ = n^{-1} \frac{\text{trace}(I-A(k))^{2}(I-A(k)y)}{[n^{-1}\text{trace}(I-A(k))]^{2}} \]

with:

\[ \hat{f}(x_{i}) = A(k)y = [X(X'X)^{-1}X']y \]

\[ A(k) = [X(X'X)^{-1}X'] \]

3. Method
The analysis steps in this research are as follows:

- generating data for response variables and predictor variables using simulation methods,
- generating data modelling with spline nonparametric regression using three-point knots with the UBR method,
- selecting the optimal knot point using the UBR method,
- generating data modelling with spline nonparametric regression using the point of the optimal knot from the UBR method
- generating modelling with spline nonparametric regression using three-point knots with the GCV method,
- selecting the optimal knot point using the GCV method,
- generating data modelling with spline nonparametric regression using the point of the optimal knot from the GCV method,
- generating a comparison of the spline nonparametric regression model with the optimal knot points obtained from the UBR method and the GCV method
- choosing the best model with criteria based on R-Square and MSE.

4. Results and Discussion
The first spline nonparametric regression modelling is modelling simulation data using the UBR method for the selection of optimal knot points.

The equation of the spline nonparametric regression model using three knots with five predictor variables, in general, is as follows:

\[ \hat{y} = \hat{\beta}_{0} + \hat{\beta}_{1}x_{1} + \hat{\beta}_{2}(x_{1} - k_{1})_{+} + \hat{\beta}_{3}(x_{1} - k_{2})_{+} + \hat{\beta}_{4}(x_{1} - k_{3})_{+} + \hat{\beta}_{5}x_{2} + \]
\[
\hat{\beta}_6(x_2 - k_4)_+ + \hat{\beta}_7(x_2 - k_5)_+ + \hat{\beta}_6(x_2 - k_6)_+ + \hat{\beta}_9 x_3 + \hat{\beta}_{10} (x_3 - k_7)_+ + \\
\hat{\beta}_{11} (x_3 - k_9)_+ + \hat{\beta}_{12}(x_3 - k_9)_+ + \hat{\beta}_{13} x_4 + \hat{\beta}_{14} (x_4 - k_{10})_+ + \hat{\beta}_{15}(x_4 - k_{11})_+ \\
+ \hat{\beta}_{16}(x_4 - k_{12})_+ + \hat{\beta}_{17} x_5 + \hat{\beta}_{18}(x_5 - k_{13})_+ + \hat{\beta}_{19}(x_5 - k_{14})_+ + \\
\hat{\beta}_{20}(x_5 - k_{15})_+ \\
\] (10)

The following are the results of the 5 smallest UBR values in the simulation data using three knots:

| Knot | X1 | X2 | X3 | X4 | X5 | UBR         |
|------|----|----|----|----|----|-------------|
| Knot 1 | 64.32 | 4.52 | 393.39 | 10.88 | 0.72 | 1.96 x 10^{-22} |
| Knot 2 | 82.15 | 7.34 | 1567.24 | 48.97 | 3.25 |               |
| Knot 3 | 99.98 | 10.17 | 2741.10 | 87.07 | 5.77 |               |
| Knot 1 | 92.34 | 8.96 | 2238.02 | 70.74 | 4.69 |               |
| Knot 2 | 96.58 | 9.63 | 2517.51 | 79.81 | 5.29 | 2.14 x 10^{-22} |
| Knot 3 | 98.28 | 9.90 | 2629.31 | 83.44 | 5.53 |               |
| Knot 1 | 76.20 | 6.40 | 1175.96 | 36.28 | 2.40 |               |
| Knot 2 | 92.34 | 8.96 | 2238.02 | 70.74 | 4.69 | 2.25 x 10^{-22} |
| Knot 3 | 96.58 | 9.63 | 2517.51 | 79.81 | 5.29 |               |
| Knot 1 | 75.35 | 6.27 | 1120.06 | 34.46 | 2.28 |               |
| Knot 2 | 93.19 | 9.09 | 2293.92 | 72.56 | 4.81 | 3.07 x 10^{-22} |
| Knot 3 | 98.28 | 9.90 | 2629.31 | 83.44 | 5.53 |               |
| Knot 1 | 77.05 | 6.53 | 1231.86 | 38.09 | 2.52 |               |
| Knot 2 | 92.34 | 8.96 | 2238.02 | 70.74 | 4.69 | 3.07 x 10^{-22} |
| Knot 3 | 95.73 | 9.49 | 2461.61 | 78.00 | 5.17 |               |

In Table 4.1 shows that the smallest UBR value of \(1.96 \times 10^{-22}\). The optimal knot points of each variable are as follows:

| Knot | X1 | X2 | X3 | X4 | X5 |
|------|----|----|----|----|----|
| Knot 1 | 64.32 | 4.52 | 393.39 | 10.88 | 0.72 |
| Knot 2 | 82.15 | 7.34 | 1567.24 | 48.97 | 3.25 |
| Knot 3 | 99.98 | 10.17 | 2741.10 | 87.07 | 5.77 |

The best spline nonparametric regression model is obtained from the optimal knot point using three knots. Based on the results of the parameter estimation, the spline nonparametric regression model with the optimal knot point using the UBR method formed is as follows:

\[
\hat{y} = 1.22 + 4.16 x_3 + 11.09(x_3 - 64.32)_+ - 33.47(x_1 - 82.15)_+ - 8.38(x_1 - 99.98)_+ \\
- 8.98x_2 - 35.61(x_2 - 4.52)_+ + 18.13(x_2 - 7.34)_+ + 2.25(x_2 - 10.17)_+ + \\
0.03x_3 - 0.04(x_3 - 393.39)_+ + 0.07(x_3 - 1567.24)_+ + 0.003(x_3 - 2741.10)_+ \\
+ 2.81x_4 - 2.71(x_4 - 10.88)_+ + 0.20(x_4 - 48.97)_+ - 51.30(x_4 - 87.07)_+ + \\
17.96x_5 - 18.18(x_5 - 0.72)_+ + 11.90(x_5 - 3.25)_+ - 448.22(x_5 - 5.77)_+ \\
\] (11)

The second spline nonparametric regression modelling is modelling simulation data with the GCV method for optimal knot point selection. The next step is to determine the smallest GCV value from the
three-knot points. The following are the results of the 5 smallest GCV values using three knots:

**Table 4.3 GCV values using 3 knot points**

| Knot  | $X_1$  | $X_2$  | $X_3$  | $X_4$  | $X_5$  | GCV    |
|-------|--------|--------|--------|--------|--------|--------|
| Knot 1| 67.71  | 5.05   | 616.98 | 18.14  | 1.20   | 514.70 |
| Knot 2| 74.51  | 6.13   | 1064.16| 32.65  | 2.16   | 521.52 |
| Knot 3| 93.19  | 9.09   | 2293.92| 72.56  | 4.81   | 520.01 |

In Table 4.3 shows that the smallest GCV value of 514.70. The optimal knot points of each variable is listed in Table 4.4.

**Table 4.4 The optimal knot points of each variable**

| Knot  | $X_1$  | $X_2$  | $X_3$  | $X_4$  | $X_5$  |
|-------|--------|--------|--------|--------|--------|
| Knot 1| 67.71  | 5.05   | 616.98 | 18.14  | 1.20   |
| Knot 2| 74.51  | 6.13   | 1064.16| 32.65  | 2.16   |
| Knot 3| 93.19  | 9.09   | 2293.92| 72.56  | 4.81   |

The best spline nonparametric regression model is obtained from the optimal knot point using three knots. Based on the results of the parameter estimation, the spline nonparametric regression model with optimal knots using the GCV method formed is as follows:

$$
\hat{y} = -3.10 - 7.23x_1 + 165.84(x_1 - 67.71) + 215.03(x_1 - 74.51) + 56.91(x_1 - 93.19) + 101.02x_2 - 17.21(x_2 - 5.05) - 120.60(x_2 - 6.13) - 5.34(x_2 - 9.09) + 0.02x_3 + 2.10(x_3 - 616.98) - 5.70(x_3 - 1064.16) + 10.62(x_3 - 2293.92) + 1.56x_4 - 5.49(x_4 - 18.84) + 6.99(x_4 - 32.65) - 13.56(x_4 - 72.56) + 45.19x_5 - 100.12(x_5 - 1.20) + 58.05(x_5 - 2.16) - 120.60
$$

**Table 4.5. Comparison between UBR and GCV methods**

| Method | MSE    | $R^2$   |
|--------|--------|---------|
| UBR    | 738.67 | 85.65   |
| GCV    | 121.43 | 97.64   |

The following is a comparison between UBR and GCV methods in spline nonparametric regression modelling on simulation data could be seen in Table 4.5. Table 4.5 shows that the GCV method produces a smaller MSE value of 121.43 and a greater $R^2$ value of 97.64% compared to the UBR method. Based on this, the GCV method is a better method used for the selection of optimal knot points compared to the UBR method in spline nonparametric regression modelling on simulation data applications.
5. Conclusions

Based on the results of the analysis and discussion that was conducted, it is possible to draw conclusions that the results of spline nonparametric regression modelling for optimal knot point selection using the UBR method produced the smallest UBR value of \(1.96 \times 10^{-22}\). Besides, the results of spline nonparametric regression modelling for optimal knot point selection using the GCV method produced the smallest GCV value of 738.67 with R-Square of 85.65%, while the best model with the GCV method produces an MSE value of 121.43 with R-Square of 97.64%. It can also be concluded that in this study the GCV method produces a better model compared to the UBR method because it produces a smaller MSE value and a larger R-Square value.

References

[1] Liu A, Tong T, and Wang Y 2014 *Journal of Computational and Graphical Statistics*, 16 p 312

[2] Budiantara I N 2009 Spline dalam regresi nonparametrik dan semiparametrik: sebuah pemodelan statistika masa kini dan masa mendatang In *Pidato Pengukuhan untuk Jabatan Guru Besar dalam Bidang Ilmu Matematika Statistika dan Probabilitas*. (Institut Teknologi Sepuluh Nopember: Departemen Statistika)

[3] Montoya E L, Ulloa N, and Miller V 2014 *International Journal of Statistics and Probability* 3 p 96

[4] Tibshirani R, and Wasserman L 2015 *Nonparametric regression statistical machine learning spring 2019* (https://doi.org/10.3109/9781439822463.147)

[5] Craven P, and Wahba G 1979a *Numerische Mathematik* 31 p 377

[6] Wang Y 2011 *Smoothing splines methods and applications*. (https://doi.org/10.1002/0470009675.ch6)

[7] Puspitasari E, and Sutanto H T 2017 *Journal Statistika UNIMUS* 6

[8] Utami T W 2018 *Journal Statistika UNIMUS* 6

[9] Wulandari K, Budiantara I N, and Ratna M 2017 *Jurnal Sains Dan Seni ITS* 6

[10] Devis S U R 2016 *Perbandingan model regresi nonparametrik spline multivariabel dengan menggunakan metode generalized cross validation (GCV) dan unbiased risk (UBR) dalam pemilihan titik knot optimal (studi kasus data angka kematian maternal di Jawa Timur)* (Surabaya: Institut Teknologi Sepuluh Nopember)

[11] Devi A R 2018 *Metode unbiased risk (UBR) dan cross validation (CV) untuk pemilihan titik knot optimal dalam regresi nonparametrik multivariabel spline truncated (studi kasus: data tingkat pengangguran terbuka di Provinsi Jawa Tengah tahun 2015)* (Surabaya: Institut Teknologi Sepuluh Nopember)