A Comparison of O’Sullivan Penalized Spline and Penalized Spline Based Truncated Power Basis Methods to Predict Ozone Concentrations

R Fitri¹, Y Widyaningsih² and D Lestari³

¹, ², ³ Department of Mathematics, Universitas Indonesia, Depok, 16424, Indonesia

Abstract. This paper aims to predict ozone concentrations based on the effects of ultraviolet light. Spline regression is piecewise polynomials that connect join points called knots. In spline regression, parameter estimation was fit by OLS (Ordinary Least Square) method. However, the OLS method leads to over parameterized and in the plot of estimated regression curve is fluctuated when using too much knots. Therefore, it needs an additional constraint which contains smoothing parameter, so that results an ideal fit. This parameter estimation method is known as PLS (Penalized Least Square) method. Spline regression using PLS method is called as penalized spline regression. O’Sullivan (1986) introduced a class of penalised splines based on B-spline basis functions. OPS (O’sullivan Penalized Spline) is a direct generalisation of smoothing splines that latter arises when the maximal number of B-spline basis functions included. One of the top performance measures the predicted regression curve that can be used is MSE (Mean Square Error). Results showed that the OPS method had a smaller MSE and GCV than the PSTP (Penalized Spline Based Truncated Power Basis) method, so the use of the OPS method for predicting ozone against ultraviolet light was suitable to use.

1. Introduction
Among others, spline is one included in nonparametric regression. There are two bases in spline regression, namely B spline and truncated power bases. The advantage of B spline base is that we can cope with high order spline, many knots, and the knot is too closed [8]. In spline regression, parameter estimation is fit by OLS (Ordinary Least Square) method. However, the OLS method leads to over parameterized and in the plot of estimated regression curve will be fluctuated when using too much knots. Therefore, it needs an additional constraint which contains smoothing parameter, so that results an ideal fit. This parameter estimation method is known as PLS (Penalized Least Square) method. Spline regression using PLS method is called as penalized spline regression. O’Sullivan uses a relatively large number of knots. To prevent over fitting, a penalty on the second derivative restricts the edibility of the A major problem of any smoothing technique is the choice of the optimal amount of smoothing, in our case the optimal weight of the penalty. Smoothing splines have a special place in nonparametric regression. O’Sullivan penalised splines become a direct generalisation and closer approximation of smoothing splines.

Ozone is one of many gases in stratosphere. Although the concentration of ozone is relatively small, it plays an important role for life on Earth due to its ability to absorb ultraviolet radiation (UV) from the sun. In the course of the last 20 years we have often heard about “the depletion of the ozone layer” and the so-called “ozone hole”, which have been considered as a threat to our health and environment [5]. Ozone concentrations, however, are not constant, and fluctuate quite a bit from day to day, depending on many factors. Therefore, this research aims to predict ozone concentrations
based on the effects of ultraviolet radiation. We use OPS and PSTPB methods to create a relationship curve between ozone and ultraviolet radiation. Then from both methods we compare which smaller MSE. Minimum MSE and GCV of both methods are used to predict ozone.

2. Smoothing Splines
The idea behind smoothing splines is to combine measures of the smoothness of a function and how well it fits the data. For the goodness of fit for the function, we use the residual sum of squares as a criterion [1]. Non-smoothness, like noise and rapid changes, can be suppressed by minimizing the penalty \( \int (f''(x))^2 \, dx \) the integrated second derivative. Together these two criteria formulate the smoothing spline technique as minimization problem

\[
\text{PLS}(\lambda) = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int (f''(x))^2 \, dx
\]  

(1)

where the estimated solution is \( \hat{f}(x) = \arg \min \text{PLS}(\lambda) \). As \( \lambda \to 0 \) we impose no penalty and end up with a very close fit, but the resulting curve could be very noisy as it follows every detail in the data. As \( \lambda \to \infty \) the penalty dominates and the solution converges to the OLS line, which is as smooth as you can get (the second derivative is always 0), but may be a very poor fit. Amazingly, it can be shown that minimizing \( \text{PLS}(\lambda) \) for a fixed \( \lambda \) over the space of all continuous differentiable functions leads to a unique solution, and this solution is a natural cubic spline with knots at the data points. A key point is the smoothing parameter \( \lambda \), which controls this trade-off between goodness-of-fit and roughness.

**Theorem 2.1** Given interpolation data \((t_i, y_i)\) \(n\). Among all functions \( f \in C^2[a, b] \) which interpolate \((t_i, y_i)\), the natural cubic spline is the smoothest, where smoothness is measure through \( \int (f''(x))^2 \, dx \).

Proof. We need to prove that \( \mu(f) \geq \mu(S) \) \( \forall f \in C^2[a, b] \) introduces \( g(x) = S(x) - f(x) \), \( g(x) \in C^2[a, b] \), \( g(t_i) = 0 \), \( i = 0, \ldots, n \). Inserting this yields

\[
\mu(f) = \int_{a}^{b} \left( S''(x) - g''(x) \right)^2 \, dx = \mu(S) + \mu(g) - 2 \int_{a}^{b} S''(x)g''(x) \, dx
\]

(2)

Since \( \mu(g) > 0 \), we have proved our result if we can show that

\[
\int_{a}^{b} S''(x)g''(x) \, dx = 0
\]

(3)

We have that,

\[
\int_{a}^{b} S''(x)g''(x) \, dx = g'(x)S''(x)|_{a}^{b} - \int_{a}^{b} g'(x)S'''(x) \, dx
\]

(4)

First part on the right hand size is zero since \( z_0 = z_n = 0 \). Second part we split an integral over each subdomain

\[
- \int_{a}^{b} g'(x)S'''(x) \, dx = - \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} g'(x)S'''(x) \, dx = - \sum_{i=0}^{n-1} 6a_i \int_{t_i}^{t_{i+1}} g'(x) \, dx
\]

\[
= - \sum_{i=0}^{n-1} 6a_i g(x)|_{t_i}^{t_{i+1}} = 0
\]

The role of the smoothing parameter in penalized smoothing is to control the smoothness of the fitted curve. In order to compute the optimal value of the smoothing parameter \( \lambda \) selection criteria are considered in this paper generalized cross validation (GCV) [6]. The GCV method is computationally simpler and very well used in the literature about smoothing splines. We consider two versions of GCV error, one for the smoothing splines and other for Penalized splines. The GCV method consists of selecting \( \lambda \) so that minimized

\[
\text{GCV}(\lambda) = \frac{1}{n} \frac{\sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2}{\left[1 - n^{-1} \text{tr}(I - H)\right]^2} = \frac{\text{MSE}(\lambda)}{\left[n^{-1} \text{tr}(I - H)\right]^2}
\]

(5)

where \( H \) is hat matrix in the case of the smoothing splines.
3. B-spline

The estimate $\hat{f}(x)$ is a cubic spline, meaning the function is a piecewise cubic polynomial. It is divided into intervals associated with a knot sequence

$$a = \tau_1 \leq \tau_2 \leq \ldots \leq \tau_{k-1} \leq \tau_k = b$$

(6)

The spline can be represented in terms of a basis, which give the term B-spline. The function $S(x)$ is specified by a coefficient or weight vector $w = (w_1, w_2, \ldots, w_n)^T$ and the basis matrix $B(x) = [B_{1,d}(x), \ldots, B_{n,d}(x)]$, where $B_{i,d}(x)$ are $i$ basis functions. The polynomial spline is given by

$$S(x) = \sum_{j=1}^{n} B_j(x)w_j$$

(7)

A B-spline on an interval $[a, b]$ has order $m = d + 1$, where $d$ is the degree of the polynomials, and number of internal knots $k$. An augmented knot sequence $\tau_1$ is defined by placing $m$ equal boundary knots on the end points and such a knot sequence can support a basis of order $l$ up to $l \leq m$

$$a = \tau_1 = \tau_2 = \ldots = \tau_m < \tau_{m+1} < \ldots < \tau_{k+m} < \tau_{K+m+1} = \ldots = \tau_{K+2m} = b$$

(8)

Each duplication of the boundary knots results in the loss of one continuous derivative. The number of basis function $B_l$ supported by the augmented sequence is $k + m$. This can be shown by counting the parameters needed to be specified. We will have $k + 1$ regions multiplied with $m$ function parameters per region, but we must subtract the parameters specified by the continuity constraints given by $k$ internal knots multiplied by $m - 1$ constraints on derivatives per knot. So the number of parameters and basis functions are

$$(k + 1), m - k, (m - 1) = k + m$$

(9)

The basis is found by the Cox-de Boor recursion formula for $j = 1, \ldots, k + 2m - 1$:

$$B_{j,1}(x) = \begin{cases} 1, & \text{if } \tau_j < x \leq \tau_{j+1} \\ 0, & \text{otherwise} \end{cases}$$

(10)

$$B_{j,m}(x) = \frac{x - \tau_j}{\tau_{j+m-1} - \tau_j}B_{j,m-1}(x) + \frac{\tau_{j+m} - x}{\tau_{j+m} - \tau_{j+1}}B_{j+1,m-1}(x)$$

(11)

If the knot sequence is uniform, the basic functions become shifted copies of another [3].

4. O’sullivan Penalized Spline (OPS)

The paper On semi parametric regression with O’Sullivan penalized splines (Wand and Ormerod, 2008) introduces penalized splines using quintile-based, non-uniform knots [7]. The penalty is based on the integral of B-spline basis functions, as introduced by O’Sullivan, and the spline estimate satisfies natural boundary conditions. The cubic B-spline basis functions $B_1, \ldots, B_{K+4}$ are defined by a knot sequence given as (10) and (11). The knots $\tau_i$ are chosen to be the $\frac{i}{K+1}$-th sample quintile of the unique data points $x_i$. The design matrix $B$ has entries $B_{ij} = B_j(x_i)$ and the $(K + 4) \times (K + 4)$ penalty matrix $\Omega$ is defined by

$$\Omega_{ij} = \int_a^b B_i''(x)B_j''(x)dx$$

(12)

The estimate $\hat{f}(x; \lambda)$ is the minimizer of

$$S(\omega) = (\mathbf{y} - B\omega)^T(\mathbf{y} - B\omega) + \lambda\omega^T\Omega\omega,$$

(13)

Giving

$$\hat{f}(x; \lambda) = B\hat{\omega}, \text{ where } \hat{\omega} = (B^TB + \lambda \Omega)^{-1}B^T\mathbf{y},$$

(14)

Where $B_x = [B_1, \ldots, B_{K+4}]$ dan $\lambda > 0$ is the smoothing parameter.

Wand and Ormerod (2008) comment that the cubic smoothing spline arises in the special case $k = n$ and $\tau_{i+4} = x_i$ for $1 \leq k \leq n$ provided that $x_i$ is distinct.

The estimate $\hat{f}(x; \lambda)$ satisfies the natural boundary condition, a constraint on the derivatives, meaning that,

$$\hat{f}''(a; \lambda) = \hat{f}'''(b; \lambda) = \hat{f}''(a; \lambda) = \hat{f}'''(b; \lambda) = 0$$

(15)
which implies that \( \hat{f}(x; \lambda) \) is approximately linear over the intervals \([a, k_5]\) and \([k_{K+4}, b]\). The linearity is exact if \( k_5 = \min(x_i) \) and \( k_{K+4} = \max(x_i) \).

The penalty matrix \( \Omega \) can be computed in R by the second derivative design matrix \( B'' \) and is given for a cubic basis
\[
\Omega = (\bar{B}'')^T \text{ diag } c (\bar{B}'')
\]  \hspace{1cm} (16)

Where \( \bar{B}'' \) is the \( \times (K + 4) \) matrix with entries \( \bar{B}''(\bar{x}_i) \), \( \bar{x}_i \) is from the vector
\[
\bar{x} = \left( k_1, \frac{k_1 + k_2}{2}, k_2, \frac{k_2 + k_3}{2}, k_3, \ldots, k_{K+7}, \frac{k_{K+7} + k_{K+8}}{2}, k_{K+8} \right)
\]  \hspace{1cm} (17)

and \( c \) is the \( 3(K + 7) \times 1 \) vector
\[
c = \left( \frac{4}{6} (\Delta k)_1, \frac{1}{6} (\Delta k)_1, \ldots, \frac{1}{6} (\Delta k)_K, \frac{4}{6} (\Delta k)_K, \frac{1}{6} (\Delta k)_{K+1}, \ldots, \frac{1}{6} (\Delta k)_{K+7}, \frac{4}{6} (\Delta k)_{K+7} \right)
\]  \hspace{1cm} (18)

where \( (\Delta k)_k = k_{k+1} - k_k, 1 \leq k \leq K + 7 \).

This result is given by applying Simpson’s rule over each of the inter-knot using the second derivative design matrix to calculate the integrals defining \( \Omega_{ij} \) since each function \( B''_i B''_j \) is piecewise-quadratic, Simpson’s rule will calculate the integral exactly. The greatest advantages with O’Sullivan splines is the direct use of the B-spline basis functions to calculate the penalty matrix, giving a better approximation to the smoothing spline penalty. The function \( f(x) \) is approximated by a weighted set of basis functions, as a parallel to finite element methods.

5. Penalized Spline Based Truncated Power Basis (PSTPB)

There are several options for the penalization criteria, but the easiest to implement is to choose a \( C \) such that \( \sum_{i=1}^K b_i^2 < C \), where \( b_i \) refers to the weight of each linear function. This is an excellent minimization criteria, because it reduces the overall effect of individual piecewise functions and avoids over-fitting the data. This minimization criteria is more formally stated as minimizing the equation \( \sum_{i=1}^n (y_i - \hat{y}_i)^2 \) subject to \( \beta^T D \beta \leq C \).

Using Lagrange multipliers, that this is equivalent to choosing \( \beta \) to minimize \( 4 \). Ruppert and Carroll (2003) used truncated polynomials as basic functions. In particular, with truncated polynomials of degree \( p \) based on \( K \) inner knots \( a < u_1 < \cdots < u_K < b \), the penalized spline estimator is defined as the solution to the penalized least squares criterion
\[
\sum_{i=1}^n (y_i - \hat{f}(x_i))^2 + \lambda^{2p} \beta^T D \beta
\]  \hspace{1cm} (20)

With \( \hat{f}(x) = \{a_0, a_1 x, \ldots, a_p x^p, \beta_1 (x - u_1)_+^p, \ldots, \beta_k (x - u_k)_+^p\} \). This has the solution
\[
\hat{\beta} = (X^T X + \lambda^{2p} D)^{-1} X^T y
\]  \hspace{1cm} (21)

The term is called a roughness penalty because it penalizes fits that are too rough, thus yielding a smoother result. The amount of smoothing is controlled by \( \lambda \), which is therefore usually referred to as a smoothing parameter.

6. Application Data

The data used in this research are UV index data and total data of layer columns downloaded from AURA satellite for Bandung area at position 6.9\(^{0}\) LS 107.5\(^{0}\) BT. OMI is an existing sensor on the AURA satellite launched in 2004. In the OMI sensor there is a total number of ozone layer and UV index. Then performed data processing by extracting data for each UV index data and total ozone columns obtained, then drawing a graph in the form to see patterns on the data. This pattern is made to look at the UV and ozone indices. The linkage of the UV index and the number of ozone columns in each season is expected to index the UV and ozone colour bias of Bandung in each season in 2013. The total ozone column data has Dobson Unit (DU), while the UV index has no units. Ozone concentration is very instrumental in the inhibition of solar UV radiation that reaches the surface of the
ider. The change of the concentration of ozone in every minute reduces the intensity of solar UV radiation. Using the OPS method and PSTPB method to create a curve of the relationship between ozone and ultraviolet radiation. Then from both methods we compare which minimum MSE and the results showed that,

**Table 1. Result of Comparison between OPS and PSTPB**

|       | GCV | \(\lambda\) | MSE   |
|-------|-----|--------------|-------|
| OPS   | 6.548 | 8.823        | 0.00358 |
| PSTPB | 7.098 | 12.755       | 0.00987 |

Based on Table 1, it can be seen that the result of comparison between OPS and PSTPB shows that OPS has a smaller MSE of 0.00358 than PSTPB with MSE of 0.00987, where the \(\lambda\) obtained from the OPS method is less than PSTPB, and so is the GCV.

**Figure 1.** The plot of estimated regression above is the curve resulting from the OPS method with \(\lambda=8.823\) and the curve which is obtained from the PSTPB method with \(\lambda=12.755\).

7. **Software**

All calculations in this research were performed on the 3.3.1 version of statistical software R project, making use of the spline package for the B-splines basis construction. The Penalized splines approach was implemented through our own routine considering Eilers and Marx’s definition and M.P. Wand and J.T. Ormerod. Smoothing parameters were chosen by our own routine based on the Generalized Cross Validation method. Smoothing Splines approximation was made by means of a function available in the spline library (smooth splines).

8. **Conclusion**

We have compared two methods namely OPS (O’sullivan Penalized Spline) and PSTPB (penalized spline truncated power base). Both methods have been compared in this paper to estimate a series of unobserved smooth curves from discrete noise observations. The comparison result between OPS and PSTPB shows that OPS has a smaller MSE that is 0.00358 compared to PSTPB with MSE 0.00987 where \(\lambda\) obtained from the OPS method is less than PSTPB, of which is similar to GCV. As the basis of this result, we can conclude that OPS and PSTPB spacing the splines lose their smooth control as the number of knots increases. Both inherited approaches can improve matches that give a mean squared error with respect to fine samples. On the other hand, PSTPB has numerical complexity that makes calculations easier and very insensitive to the choice of nodes, so it is enough to select a large number of vertices with equal distance. OPS methods are appropriate to avoid instability in numerical methods and at the same time have variations in the optimal smoothing rate to be found. Calculations using OPS are slightly more complicated than PSTPB methods. The OPS method has added flexibility due to the use of quartile based nodes, which provide more fundamental functions to areas with higher data density. This ensures that the computer power is not wasted in areas with small data and high uncertainty. Strategically placed knots, can reduce the number of basic functions. Another important feature of O’Sullivan splines is nature’s boundary properties divided by spline alignment. So the use of OPS methods to predict ozone for ultraviolet light is suitable for use.
9. References

[1] Eubank R L 1999 *Nonparametric Regression and Spline Smoothing* (New York: Marcel Dekker)

[2] Eilers P H C and Marx B D 2010 Splines, knots, and penalties *Wiley Interdisciplinary Reviews Computational Statistics* 2(6) pp 637–653

[3] De Boor C 2001 A Practical Guide to Splines Revised Edition *Applied Mathematical Sciences* chapter 27 (New York: Springer-Verlag)

[4] Ruppert D, Wand M P and Carroll R J 2009 Semiparametric Regression During 2003–2007 *Electronic Journal of Statistics* 3(2009) pp 1193–1256

[5] Heriksen 2009 (Online) Ozone and UV-Radiation http://www.uio.no/

[6] Hastie T J and Tibshirani R J 1990 *Generalized Additive Models* London: Chapman and Hall

[7] Wand M P and Ormerod J T 2008 On Semiparametric Regression with O’Sullivan Penalized Splines *Australian & New Zealand Journal of Statistics* 50(2) pp 179–198

[8] Khurotul L 2014 Election the Best B Spline Regression Model for Solar UV Radiation Estimation Based On Optimal Order and Point Knot *Thesis* University of Maulana Malik Ibrahim Malang

Acknowledgment

This research is funded by Universitas Indonesia via PITTA Research grant 2018 scheme.